universität wien

MASTERARBEIT

Titel der Masterarbeit

„Bacterial effector proteins in the evolution of pathogenic and symbiotic bacteria“

Verfasst von

Marcus Wieder B.rer.nat

angestrebter akademischer Grad

Master of Science (MSc)

Wien, 2013

Studienkennzahl lt. Studienblatt:	A 066 834
Studienrichtung lt. Studienblatt:	Molekulare Biologie
Betreuerin / Betreuer:	Prof. Dr. Thomas Rattei

Abstract

Introduction

Bacterial diseases and infections are one of the major causes of morbidity and mortality worldwide thus understanding key virulence factors is important to efficiently fight bacterial infections. It has been shown that effector proteins - as major virulence factors - contain very often a specific domain signature (so called eukaryotic-like domains - ELDs). In this thesis the hypothesis if these effectors are the result of horizontal gene transfer events with eukaryots as donor and pathogens as acceptor organism was tested. Furthermore the evolutionary and phylogenetic relationship between the donor species and the acceptor species was analyzed.

Methods

70 pathogens were chosen and their proteom was divided into ELD containing and non-ELD containing proteins based on the effector database classification. A computational method was implemented to use different available software solutions (e.g. phyloGenie, raxML, mtrees, blammer, phat, SIMAP) to automatize a pre-screening of the proteom based on similarity criteria (excluding e.g. housekeeping genes), calculating a multiple alignment, constructing a phylogenetic tree and analyzing the relationship of taxa in the tree (e.g. finding a monophyletic prokaryotic group with eukaryotic proteins) to identify horizontal gene transfer. This was done using full-length proteins as well as only domain sequences. The parameters used in these programs were tested and validated exhaustingly. Subsequently to the constructions of the trees was the analysis of the phylogenetic trees and the classification in HGT event with pathogen as gene acceptor and eukaryote as donor.

Results

The results indicated that pre-filtering with alien index was applicable - reducing the proteom by 40-50% yet conserving proteins with an ELD score and gaining a speed up factor of about 2 for the subsequent tree construction. Tree reconstruction parameters were tested by comparing positive signals for HGTs in Legionella pneumophila str. Paris with known positive signals in the literature - showing that the results are identical with some exceptions resulting from proteins with little homology. It was shown that the differences in the trees obtained from maximum-likelihood in comparison to neighbor-joining methods were negligible - thus the much faster neighbor-joining method could be used. Using full length proteins tree analysis showed for 14 bacteria a significant ($p \leq 0.05$) connection between ELDs and HGTs. Most of these bacteria are highly pathogenic and virulent - and most protein donors are not from the phyla these bacteria infects primarily. Utilizing domain sequences (domain structures are evolutionary more stable than proteins) for multiple alignment and tree construction resulted in 33 positive p-values out of 60 analyzed proteoms - with an overlap between positive signals from protein sequences an domain sequences of 8 bacteria.

Summary

The results of this thesis shows that for a certain set of organisms a significant connection between horizontal gene transfer events and eukaryotic-like domains can be stated and this link becomes even more established when using domain sequences. The 8 pathogens which showed a significant connection between ELDs and HGTs on protein and on domain level were either pathogens which were known for their numerous HGT events (Legionella pneumophila str. Lens, Legionella pneumophila str. Paris, Legionella pneumophila subsp. pneumophila str. Philadelphia 1), their endosymbiotic life-style (Wolbachia endosymbiont of Culex quinquefasciatus Pel) or their high virulence and pathogenicity (Coxiella burnetii RSA 493, Francisella tularensis subsp. tularensis NE061598, Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130) with the exception of Acidaminococcus intestini RyC-MR95 which colonize the human gut flora as commensal.

Abstract

Einleitung

Bakterielle Erkrankungen und Infektionen sind eine der häufigsten Todesursachen weltweit - ein tieferes Verständnis von Virulenzfaktoren ist unumgänglich für eine effektive Bekämpfung von bakterielle Infektionen. In früheren Arbeiten wurde bereits gezeigt, dass Effektor-Proteine - welche als Haupt-Virulenzfaktoren anzusehen sind - sehr oft eine spezifische Domainsignatur besitzen (so genannte eukaryotic-like domains - ELDs). In dieser Arbeit wurde die Hypothese, dass Proteine mit eukaryotic-like domains das Resultat von horizontalen Gentransfer (HGT) mit Eukaryonten als Gen-Donatoren und Prokaryonten als Gen-Akzeptoren sind, untersucht. Weiters wurde das evolutionäre und phylogenetische Verhältnis zwischen Donor und Akzeptor des HGT analysiert.

Methoden

70 Pathogene wurden randomisiert aus der effector.org Datenbank ausgewählt und ihr Proteom in Proteine mit ELDs (verwendete Definition von ELDs entspricht der effector Datenbank Definition) und Proteine ohne ELDs unterteilt. Eine Softwarelösung wurde implementiert, welche folgende Schritte für die Menge aller Proteine der Proteome mit Hilfe vorhandener Software (verwendet wurden unter anderem phyloGenie, raxML, mtrees, blammer, phat und SIMAP) automatisierte: (1) Vorselektion des Proteoms anhand von Sequenzähnlichkeit, (2) Erstellen eines multiplen alignments, (3) Konstruktion eines phylogenetischen Baumes und die (4) Analyse des Baumes (z.B. das Auffinden einer monophyletischen, prokaryontischen Gruppe mit eukaryontischen Proteinen im Ast) zum Auffinden eines HGT. Dieser Ansatz wurde sowohl auf Proteine als auch für Proteindomains angewandt. Anschließend an die Parametersuche für die einzelnen Programme wurde der methodische Ansatz angewendet, die phylogenetischen Bäume analysiert und nach positiven/negativen Signalen für HGTs mit Pathogene als Gen-Akzeptor und Eukaryonten als Gen-Donor klassifiziert.

Resultate

Die Selektion der Proteine mittels Ähnlichkeitskriterien stellte sich als sehr effektiv heraus - es reduzierte die zu analysierenden Proteine um 40-50% und führte zu vernachlässigbarem Verlust an potentiellen HGT Kandidaten. Die Suche nach optimalen Parametern zur Konstruktion von phylogenetischen Bäume wurde mittels bekannten Beispielen aus der Literatur vollzogen - die Ergebnisse waren bis auf erklärbare Ausnahmen ident. Der Vergleich zwischen Maximum-likelihood und Neighbor-joining Methoden zur Konstruktion von phylogenetischen Bäumen zeigte am Beispiel eines Chlamydia Stammes keinen Unterschied im Ergebnis - was zum Verwenden einer Neighbor-joining Methode für die weitere Analyse führte. Angewendet auf Proteine zeigt die entwickelte Methode für14 Pathogene einen signifikanten ($p \leq 0.05$) Zusammenhang zwischen ELDs und HGTs. Die meisten dieser Pathogene sind hoch virulent und pathogen - und nach Analyse des Baums ist ersichtlich das nur wenige Gene-Donoren dem Phylum angehören welches das Pathogen infiziert. Wendet man die entwickelte Methode auf Proteindomains an (Domains sind evolutionär stabiler als komplette Proteinsequenzen) ergeben sich in 33 von 60 Fällen signifikante p-values und damit eine signifikante Korrelation zwischen ELDs und HGTs. Die Ergebnisse für Proteine und Proteindomains überschneiden sich in acht Pathogenen.

Zusammenfassung

Die Ergebnisse dieser Arbeit zeigen, dass man für eine bestimmte Gruppe an Bakterien einen signifikanten Zusammenhang zwischen HGTs und ELDs zeigen kann und dieser Zusammenhang bei Verwendung von Domaininsequenzen tendenziell stärker wird. Die acht Pathogene, welche sowohl auf Proteinsequenz- als auch auf Domainsequenzebene einen signifikanten Zusammenhang zwischen ELDs und HGTs zeigten, rekrutieren sich aus Bakterien, welche für ihre hohe Anzahl an HGTs aus der Literatur bekannt sind (Legionella pneumophila str. Lens, Legionella pneumophila str. Paris und Legionella pneumophila subsp. pneumophila str. Philadelphia 1), eine endosymbiontische Lebensweise besitzen (Wolbachia endosymbiont of Culex quinquefasciatus Pel) oder hoch virulent und pathogen sind (Coxiella burnetii RSA 493, Francisella tularensis subsp. tularensis NE061598, Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130) mit der Ausnahme von Acidaminococcus intestini RyC-MR95.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bacterial diseases and infections are one of the major causes of morbidity and mortality worldwide (e.g. chronic lower respirator diseases ranking as number three on the list of leading causes of death in the US in 2008/9[1]) - thus it should come as no surprise that much research has been done to elucidate molecular mechanisms by which pathogenic bacteria have evolved and developed pathogenicity. One of the crucial keystones for pathogens to function efficiently is their ability to impair the host enough to promote their survival, yet not killing their host prior to dissemination. This is primarily done by protein secretion - which is one of the main mechanism bacteria utilise to modulate their interactions with the biotic environment. This is especially noteworthy for bacteria which have a pathogenic associations with host eukaryotic organisms - secreted bacterial proteins are in this cases a key virulence mechanism. A stable relationship between bacteria and eukaryotes is often initiated and maintained by the secretion of a broad arsenal of secreted virulence factors (which are not expressed in related non-pathogenic strains) which are able to utilize the host cellular machinery - so called effectors [Tseng *et al.*, 2009, Shames *et al.*, 2009].

Not only pathogens but also symbionts interact with their host protein network by utilizing effector proteins - since bacteria make up about 90 % of the cellular population of our bodies it is important to understand how these bacteria influence and effect our lives. This is especially true since it became clear that populations of human cells have

[1] data provided by the U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

a constant relationship with the human microbiome making use of protein-protein interactions and under certain conditions these interactions can become pathogenic. The human microbial flora has been associated with obesity, metabolic disease, infant growth, colorectal cancer and only recently it was shown that genes acquired by horizontal gene transfer are enriched in cancer samples [Arthur *et al.*, 2012, Riley *et al.*, 2013]. Bacterial effectors which are successfully transported from the bacteria to the host cell are able to fulfill diverse functions - ranging from manipulation the immune response of their host, rearranging host cytoskeleton to hijacking the MAPK or ubiquitination pathway to name just a few [Mattoo *et al.*, 2007]. For some organisms it has been shown that their effector proteins resemble not only the function of eukaryotic proteins but have also similarity to eukaryotic proteins on the sequence and structure level (these domains/proteins are called eukaryotic-like domains/proteins) - leading to an ongoing investigation of the evolution of effector proteins [Gomez-Valero *et al.*, 2011a, Al-Khodor *et al.*, 2008].

Also in the light of the finding that more and more bacterial strains are resistant to common antibiotics there is a rising need for new targets in fighting bacterial infections - understanding the evolution and behaviour of effectors may lead to a better understanding of pathogen-host interaction, virulence, host adaptions, resistance and infection on a molecular level which in turn provides new targets for drug development. The aim of this master thesis is to investigate the evolutionary dynamic of effectors and to test the hypothesis that many eukaryotic-like domain containing proteins are typically acquired by horizontal gene transfer comprehensively - utilizing the fact that horizontal gene transfer can be detected by tree reconziliation - and analyzing their evolutionary dynamic.

1.1 Bacterial effector proteins

Bacterial effectors are proteins/small molecules that fulfill their function outside the prokaryotic cell and interact with proteins in the host cell to selectively bind and regulate the biological activity of proteins [Crow *et al.*, 2012]. To meet this requirement both Gram-negative and Gram-positive bacteria have evolved specialized transport mecha-

nisms to translocate proteins from the bacterial interior to the exterior or host cell cytoplasm - a process named secretion [Beeckman & Vanrompay, 2010].
Seven different bacterial secretion systems and the Sec pathway have been described so far as possible transport systems for effectors. There is little homology between the systems, each of them is unique in terms of molecular structure and way of translocation.
A single bacterium can deliver up to 100 different effector proteins into individual host cells. These effectors are often multifunctional proteins with overlapping properties and interactions for the main purpose of triggering specific host responses [Dean, 2011].
These effectors can be organized in classes using their different sub-cellular targets (e.g. Mitochondrial-targeting sequences, Membrane targeting motifs, nuclear localization sequences and transcription.related domains, effectors that exploit E3 ligases or mimic host E3 ubiquitin ligases). They can also be classified utilizing their biochemical function (e.g. Phosphatases, PiPases, Kinases, Lipases, Proteases, Cyclases, Lyases). A prominent example for proteases is e.g. the YopJ effector which is present in Yersinia and acts as an anti-inflammatory effector by binding the mitogen-activated protein kinase (MAPK) and preventing phosphorylation [Dean, 2011]. Another example for a well studied and important class of effectors are the TAL (Transcription activator-like) effectors which are injected by Xanthomonas bacteria via type 3 secretion system to infect plant cells. The TAL effectors are of great interest in various fields of molecular biology for their ability to bind DNA with high sequence specificity and their relative ease of retargeting them to different DNA sequences [Mak *et al.*, 2012, Boch *et al.*, 2009].

1.1.1 Bacterial Secretion Systems

Overview

In Gram-negative bacteria there are six known secretion pathways (called type I to type VI secretion systems and abbreviated T1SS to T6SS) - each of them displaying considerable diversity. In Gram-positive bacteria there are some of the same secretion systems present as in gram-negative bacteria and one specific to this group: the type VII secretion system. A summary is shown in Figure 1.1.

For Gram-negative bacteria the barrier to be crossed for proteins consists of an inner and an outer membrane, in between lies the periplasmic space. Gram-positive bacteria only have one membrane and therefor no periplasmic space.

Gram-negative bacteria transport proteins across the inner and outer membranes in a single step using the Type I, Type III, Type IV or Type VI secretion pathway. Sec or Tat (for Type II and type IV secretion) systems often transport unfolded or folded proteins via a stop at the periplasmic space - where their final three-dimensional structure is obtained - before transporting them across the outer membrane via the Type II, Type V or less commonly the Type I or Type IV secretion pathway.
A characteristic of proteins being secreted using sec-dependent pathways (Type II secretion system and type V secretion system) is a short terminal signal sequence ($\sim$ 30 amino acids) which is cleaved by peptidases in the periplasmic space, in contrast to type I secretion system, type III secretion system and usually also type IV secretion systems which are able to secrete proteins without any such signal [Tseng *et al.*, 2009].

In Gram-positive bacteria proteins are normally translocated across the membrane by the two-arginine (Tat) or Sec pathway - in the case of Gram-positive bacteria with a hydrophobic, nearly impermeable cell wall the Type VII secretion system translocates proteins [Shames *et al.*, 2009, Hacker & Carniel, 2001, Tseng *et al.*, 2009].
Type III, type IV and VI penetrate the host cell membranes and inject proteins into the cytosol in contrast to the other systems which can be used if effector entry into the cytosol is not required [Arnold *et al.*, 2012]. Type III and especially type IV systems have been shown to be of great importance in interactions between bacterial and eukaryotic cells and in the translocation of bacterial effector molecules directly into the eukaryotic cell. Commonly encoded in large clusters of contiguous genes this two secretion systems are often considered as pathogenicity islands in bacteria [Bingle *et al.*, 2008].

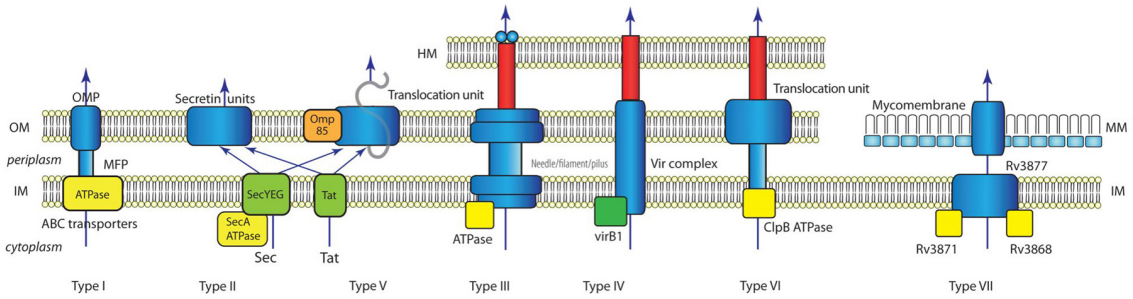

Figure 1.1: Summary of known bacterial secretion systems. The secretion systems are simplified sketched - only the basics of each mechanism are shown. HM: Host membrane; OM: outer membrane, IM: inner membrane; MM: mycomembrane; OMP: outer membrane protein; MFP: membrane fusion protein. ATPases and chaperones are shown in yellow [Tseng *et al.*, 2009].

Sec pathway and twin-arginine pathways

The general secretion pathway (Sec) and the two-arginine (Tat) translocation pathway are universal to all three domains of life - even though there are differences in number of translocase components and energy sources. The sec machinery is used to insert proteins into the endoplasmatic reticulum of eukaryotes and into the inner membrane (cytoplasmic membrane) of bacteria and archaea - but it cannot transfer folded proteins across the membrane. Whereas the Tat system is able to translocate folded proteins across the cytoplasmic membrane in bacteria and archaea (also across the chloroplast thylakoid membrane in eukaryotes). The sec and tat pathways both function to insert proteins into the membrane as well as transport proteins across the inner and - using the secretion and translocation unit - outer membrane [Yuan *et al.*, 2010]. The Sec pathway uses a hydrophobic N-terminal sequence for recognizing proteins destined for secretion, the tat secretion pathway utilizes a motif rich in basic amino acid residues (S-R-R-x-F-L-K) in the N-terminal region of large co-factor containing proteins [Tseng *et al.*, 2009].

The Sec translocon core compound consists of a heterotrimeric protein complex designated as SecYEG in Bacteria and Sec61$\alpha\beta\gamma$ in eukaryotes. The translocon can facilitate the transport across or integration of proteins in the membrane in either a co-translational

or post-translational way. In bacteria the majority of inner membrane proteins use the co-translational, the majority of secreted proteins the post-translational pathway. Noteworthy is that the selection step for both pathways lie at an early stage of translation once the nascent chain emerges from the ribosomal exit tunnel [du Plessis *et al.*, 2011].

Type I secretion system

The type I secretion system (T1SS) is a sec-independent transport system and bypasses the periplasm. It recognizes a specific secretion signal at the C-terminus of the secreted protein which is not cleaved during secretion. Many proteins secreted by the T1SS have distinctive glycine rich repeats (GGXGXDXXX) that specifically bind calcium ions. This repeats are in various numbers (up to 50) present in most secreted proteins, forming distinctive beta-sandwiches or beta-roll structures with calcium ions in the turns whereas there are also cases without any repeats (HasA) or other kind of repeats.
This glycin rich repeats have been found up to now only in cyanobacteria and in all classes of proteobacteria. Proteins with this repeat are widely diverse in size (varying from 90 up to 9000 residues), with only little exceptions very acidic (pI around 4) and have very few or no cystein residues. In addition to glycin rich repeats many proteins have repeats which are homologue to regions in adhesion molecules. The translocator consists of three proteins that span the cell envelope - one outer membrane protein and two cytoplasmic membrane proteins (an ATP-binding cassette (ABC) and a membrane fusion protein (MFP)) [Delepelaire, 2004].

Type II secretion system

Type II secretion systems (T2SS) are conserved and common (but not universal) among Gram-negative bacteria. Bacterial pathogens of plants (Pseudomonas fluorescens, Erwinia spp., Xanthomonas spp.), animals (Aeromonas hydrophila) and humans (Klebsiella oxytoca, Pseudomonas aeruginosa, Legionella pneumophila) utilize predominantly the T2SS to secrete extracellular toxins, surface-associated virulence factors and hydrolytic enzymes (e.g., cellulases, elastases, amylases, proteases, phoPHATases, lipases) [Douzi

et al., 2012]. A detailed description of its component is given in Figure 1.2.
The secretion into the extracellular medium is a two-step process in which proteins are released into the periplasm before transported across the outer membrane. The Tat export pathway is responsible for the transport across the inner membrane of proteins that require cytoplasmic folding. Proteins that are translocated as unfolded precursors are transported via the Sec pathway in the periplasmic space.

Type III secretion system

Unlike the two secretion systems already discussed (Type I and Type II) Type III secretion systems (T3SS) predominantly translocate proteins in the cytoplasm of its eukaryotic host cells. T3SS secretion system is a key feature of the virulence of many bacterial pathogens (including species of Chlamydia, Xanthomonas, Pseudomonas, Ralstonia, Shigella, Salmonella, Escherichia and Yersinia), causing several of the world's most important disease. It allows injection of effector proteins into the cytosol of target animal or plant cells. Also bacteria which have a symbiotic relationships with plant or animal hosts often possess a T3SS [Beeckman & Vanrompay, 2010].
The T3S system - called the injectisome - is build up from two pairs of rings spanning across the IM and OM, connected by a rod structure and an about 60 nm long needle protruding outside the bacteria. It is by far the most complex protein secretion system know so far - composed of approximately 25 proteins. The injectisome is shown in Figure 1.3. The injectisome has a common evolutionary origin with the bacterial flagellum, as suggested by the fact that the basal bodies are similar in both and include many proteins with a significant similarity and homology [Troisfontaines & Cornelis, 2005, Dean, 2011].

Type IV secretion system

In comparison to other secretion systems the type IV secretion system's (T4SS) key feature is it's unique ability to transport nucleic acids in addition to proteins into plant and animal cells, as well as into yeast and other bacteria. The T4SS complex is present in

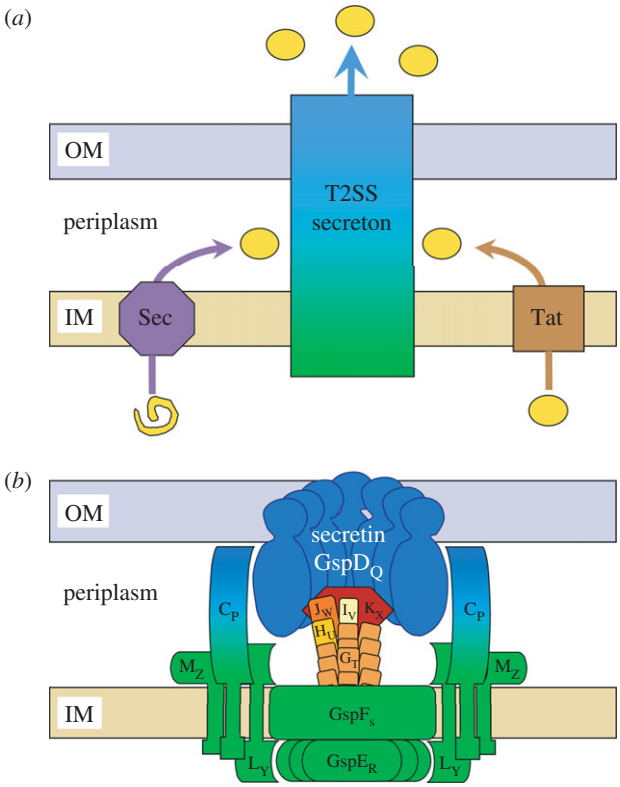

Figure 1.2: Sketch of the type II secretion pathway in Gram-negative bacteria. (a) Proteins which are secreted via the T2SS are shown as yellow circles. They are first transported across the IM utilizing either the Sec (purple) or Tat (brown) mechanism. The secreton (shown in blue/green) further recognizes and transports this exoproteins subsequently across the OM. (b) A more detailed but still schematic view of the secreton. It is divided into three sub complexes: Shown in blue is the Secretin $GspD_Q$, it C-terminus forms a dodecameric pore in the OM while its N-terminus protrudes in the periplasm. Colored in green are the IM surface proteins (F_S, L_Y, M_Z) and the ATPase $GspE_R$. The blue/green part are the connection between the secretin and the IM surface ($GspC_P$). In orange and red the pseudopilus are shown. Gsp proteins are indicated by their corresponding letters [Douzi *et al.*, 2012].

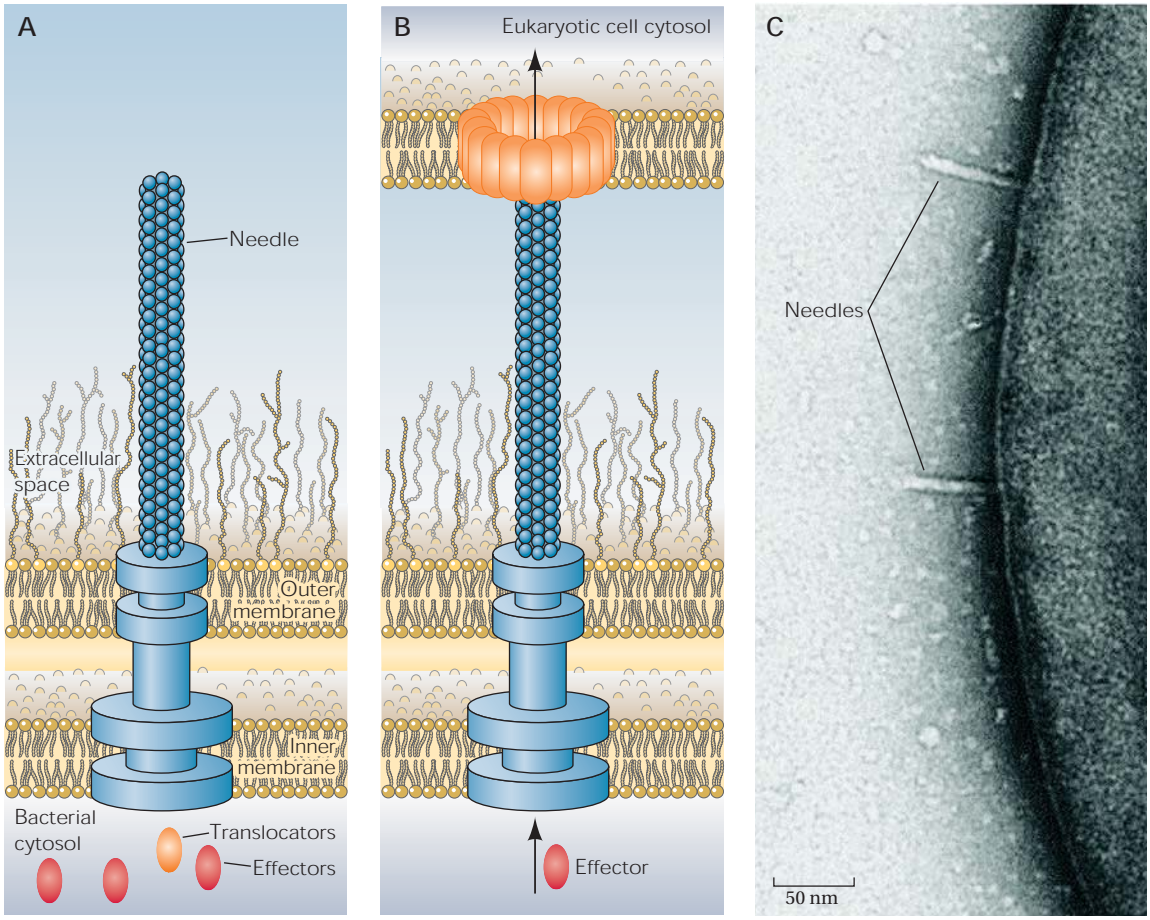

Figure 1.3: Sketch and electron microscope image of T3S Injectosome. (A): a resting T3S injectisome is shown - spanning from IM across the periplasmatic space and OM. The needle protrudes beyond the extracellular space outside the bacterium. Effectors as well as translocators are stored in the cytosol.
(B): when activated the translocators form a pore into the target cell membrane and the effectors are transported into the cyotosol of the target cell.
(C): electron micrograph of the surface of Yersinia enterocolitica with protruding needles [Troisfontaines & Cornelis, 2005].

Gram-positive as well as in Gram-negative bacteria.
Type IV secretion systems is homologous in many bacteria (but not present in all), including the pathogens Agrobacterium tumefaciens C58 (VirB), Helicobacter pylori (CAG; ComB), Pseudomonas aeruginosa (TraS/TraB), Bordetella pertussis (Ptl), E. coli (Tra), Legionella pneumophila (Dot) and the nitrogen-fixing plant mutualist Mesorhizobium loti [Tseng *et al.*, 2009]. Although functional similar, they are not all derived from the same sets of genes - which the only known exception of VirB10 (TrbI) which is present in all characterized systems [Cao & Saier, 2003].
The T4SS family can be divided into three groups based on its functions - schematically shown in Figure 1.4.
The first group is able to mediate the conjugative transfer of plasmid DNA or transposons into a wide range of bacterial species. For at least two Gram-negative species (Eschericha coli and Agrobacterium tumefaciens) it has been shown that they are able to deliver DNA substrates into fungal, plant or human cells. Conjugation provides a possibility to exchange genetic material therefor promoting genomic plasticity and adaptive responses of bacteria to changes in the environment.
The second group is able to mediate DNA uptake from and release into the extracellular milieu - promoting further genetic exchange. This has been shown for some Gram-negative bacteria including Helicobacter pylori and Neisseria gonorrhoeae.
The third group delivers virulence proteins into eukaryotic (and mammalian) host cells. This has been shown for pathogenic Gram-negative bacteria like H.pylori, Brucella suis and Legionella pneumophila [Fronzes *et al.*, 2009].

Type V secretion system

A very large number of proteins are secreted utilizing the type V secretion pathway (T5SS), even more than via the T2SS. Most of this proteins contribute to the virulence of animal or human pathogens [Tseng *et al.*, 2009].
The T5SS is perhaps the most simple secretion mechanism and commonly present in the outer membrane of Gram-negative bacteria. It is subdivided into the auto-transporter

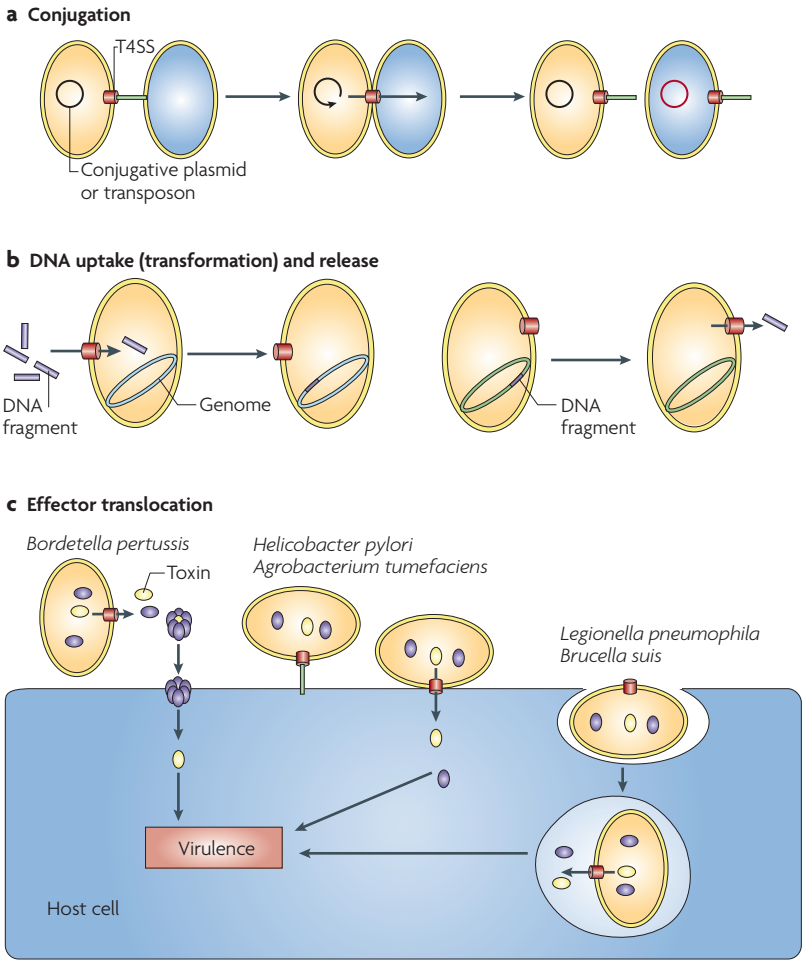

Figure 1.4: Schematic view of the role of type IV secretion in bacteria. (a): Conjugative type IV secretion systems - transport of plasmids or transposoms between a donor and recipient cell, possible in Gram-negative and Gram-positive bacteria.
(B): Transformative type IV secretion systems - DNA uptake and release mediated by the T4SS system. (C): Effector translocation via the T4SS - transport of DNA or protein virulence factors inside eukaryotic cells [Fronzes *et al.*, 2009].

system (type Va or AT-1), the two-partner secretion pathway (type Vb), and the type Vc system (also termed AT-2) - all using the Sec machinery for transport in the periplasm. [Henderson *et al.*, 2004]. Proteins which are secreted using the T5SS all have a β-strand which is inserted in the OM and forms the pore in the bacterial OM for translocation of the folded passenger domain. In some cases (like adhesins) the passenger domain stays attached to the β-strand and the protein is anchored in the OM - while in other cases the passenger domain is cleaved and forms a soluble enzyme or toxin [Tseng *et al.*, 2009]. More shared features include a N-terminal signal sequence, a functional passenger domain, a linker region necessary for translocation and a C-terminal region associated with the formation of the transmembrane pore [Beeckman & Vanrompay, 2010]. A schematic overview of the three different classes of T5SS is shown in Figure 1.5.
The similarities in the primary structures as well as the biogenesis of the secreted proteins are noteworthy, as well as the N-terminal signal peptide for transport via the Sec machinery in the periplasm.

Type VI secretion system

The type VI secretion system (T6SS) is a newly described (first named in 2006) machinery for protein transport across the cell envelope of Gram-negative bacteria, widespread in nature and not confined to known pathogens although T6SS gene clusters are usually found within pathogenicity islands - for example: Pseudomonas aeruginosa HSi (Hcp-secretion island), EaEc (enteroaggregative Escherichia coli) pheU, Salmonella typhimurium Sci (Salmonella centrisome island), Francisella tularensis Fpi (Francisella pathogenicity island) or on chromosomes that show a bias towards virulence or survival in the eukaryotic host cell. Core components conserved among all T6SSs include the T4SS IcmF- and IcmH-like proteins, a putative lipoprotein, the ClvP AAA+ ATPase (a potential energy source) and the Hcp (Haemolysing co-regulated protein) and VgrG (valine-glycine repeats) proteins. The needle complex shares similarities with the T4 bacteriophage tail spike [Miyata *et al.*, 2013]. The secretion is done in one step and Sec independent [Cascales, 2008]. An overview of the type VI secretion systems assembly and function is

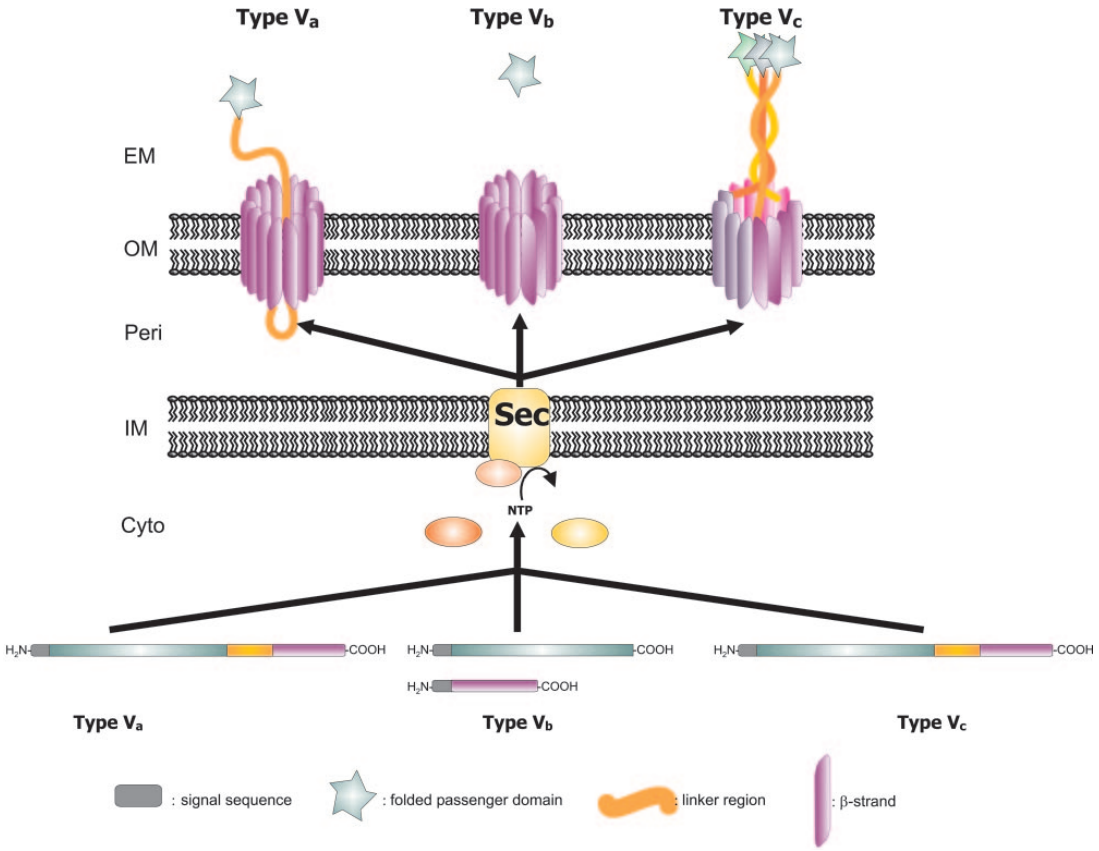

Figure 1.5: Schematic view of the role of type V secretion systems in bacteria. From left to right the auto-transporter secretion pathway (type Va), the two-partner system (type Vb) and the type Vc or AT-2 family is depicted. For all three different classes the transport across the IM is performed by the Sec machinery. The signal sequence is cleaved in the periplasmatic space and the β-domain inserts into the OM forming a pore. The passenger domain is translocated through the pose to the bacterial cell surface [Henderson *et al.*, 2004].

shown in Figure 1.6.

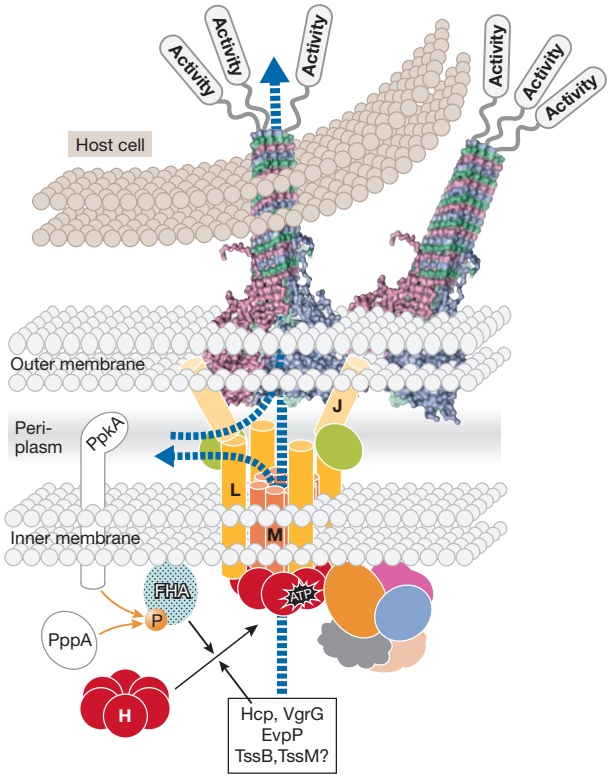

Figure 1.6: Schematic overview of the type VI secretion systems assembly and function. Starting at the inner membrane, a channel is formed by the IcmF-like and IcmH-like proteins associated at the cytoplasmic side of the IM with the ClpVAAA^{+}ATPase and the type VI secretion (T6S) subunits. The ClpV multimer (shown in red) is recruited by a pathway starting with PpkA and PpkA (in the presence of Hcp), which regulate the forkhead-associated (FHA) phosphorylation. A multimer of associated lipoproteins in combination of the periplasmic subunits is shown at the outer membrane. The blue arrow indicate the route for substrate translocation through the cell envelope and the host-cell membrane - a one-step mechanism through an inner channel and a two-step mechanism with accumulation in the periplasm is shown. J, L, M and H (TssJ, TssL, TssM and TssH) represent the T6S core components [Cascales, 2008].

Recent studies showed that some effectors transported by the type VI secretion system (type VI lipase effectors) target phosphatidylethanolamine, the major component of bacterial membranes, thus functioning as antibacterial effectors [Russell *et al.*, 2013, Leroux *et al.*, 2012].

Type VII secretion system

The type VII secretion system is a recently suggested name for the former known ESAT-6 secretion system (ESX-1-ESX-5). This system, intensely studied in Mycobacterium tuberculsis but also present in other actionobacteria and Gram-positive bacteria, transports small, highly immunogenic proteins that lack classical signal sequences. These proteins are characterized by the amino acid sequence Trp-Xaa-Gly (WXG) and a size of about 100 amino acids (WXG-100 proteins). In mycobacteria the secreted proteins typically contain additional N-terminal motifs (PE - proline-glutamic acid and PEE - prolin-proline-glutamic acid) [Simeone *et al.*, 2009].

1.1.2 Computational methods for the prediction of secreted virulence factors

Identification of secreted proteins for all seven secretion systems still remains a challenge. The detection of secreted proteins with high sensitivity and specificity is important to identify virulence factors and interaction partners in host-pathogen interaction networks. Three different methods for identification of secreted proteins are discussed below.

Homology-based methods

Since there is experimental evidence for the characterization of some proteins as effectors homology searches can be applied to detect effector candidates. This has all the advantages and disadvantages of homology search - i.e., this approach is restricted to known effector protein families, susceptible to sequence without functional change but on the other hand HGTs are detectable [Arnold *et al.*, 2012]. This can be done using a simple BLAST search against known effectors.

This has been done e.g. to add type III secreted effectors to the repertoire of the Sakai strain of enterohemorrhagic E.coli (EHEC). In the case of EHEC this approach identified 65 candidates of which 60% were experimentally verified to be secreted [Tobe

et al., 2006].

Signal based methods

Signal based methods utilize secretion signals (peptide sequences which are recognized by the respective secretion system) to predict secreted proteins. This method is only usable for secretion systems which have a defined and known secretion system. Type II and V secretion systems (sec-dependent pathways) have a defined cleavage signal and hydrophobic N-terminal signal sequence - proteins utilizing one of these two systems can be characterized with high accuracy (sensitivity and selectivity $> 88\%$) as implemented in the SignalP detection software [Bendtsen *et al.*, 2004].

For type III secretion system the exact signal leading to secretion is not known - but it has been shown that the first 100 amino acid at the N-teriminus may contain the signal peptides and chaperone-binding sequences needed to translocate proteins with the T3S. Due to low sequence similarities and lack of common features among different T3S signal sequences all signal based approaches use amino-acid frequencies of k-meres from the N-terminus (which is responsible for substrate recognition in T3SS) of proteins and binary classification algorithms (neural networks, Naive Bayes classifiers and support vector machines) [Arnold *et al.*, 2012]. An example for a classifier that uses N-terminal position-specific amino acid composition (AAc) features is BPBAac. It is trained using Support Vector Machine based on the AAc feature extracted utilizing a Bi-profile Bayes model [Wang *et al.*, 2011]. Cross-validation experiments showed a sensitivity up to 90 % and a specificity ranging between 88-97 percentage. For the other secretion systems signal sequences are not well defined [Arnold *et al.*, 2012]. Type IV secreted proteins are hard to identify - a number of bioinformatic screening strategies were adopted to identify candidates but these methods are not exhaustive and are not generally applicable for secretion signal detection. Thus a SVM model using features calculated from residue composition and position specific scoring matrix profiles was implemented - this was tested on experimentally validated TIVSS effectors resulting in the identification of TIVSS effectors with an accuracy of 93% [Zou *et al.*, 2013].

Genomic and functional based methods

Genomic and functional based methods enable the detection of effectors for which the signal of substrate recognition for the transport system is unknown. To do this more general features of effectors must be exploit like their genomic organization (pathogenicity islands or CG content) or phylogenetic distribution of orthologs (since some of them have been acquired by horizontal gene transfer from eukaryotes). A summary of this features is given in Figure 1.8. Also proteins with eukaryotic like domains (ELDs - eukaryotic like domains are being introduced in section subsection 1.2.3) serve as a basis for detection and indicate furthermore their function as effectors in the host [Arnold *et al.*, 2012, Burstein *et al.*, 2009].

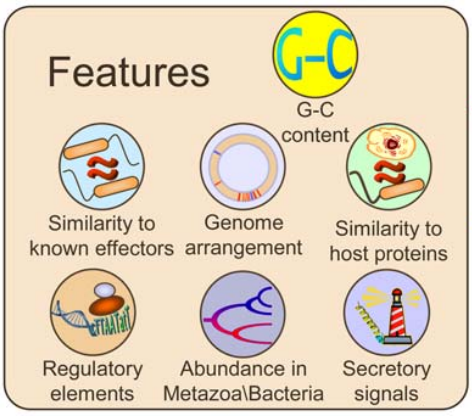

Figure 1.7: Genomic and functional based methods to detect effectors [Burstein *et al.*, 2009].

1.1.3 Host-Pathogen interaction networks

It is evident that the interaction between a host and a microbe is an important contribution to both health and disease of the host. There are three forms of species-species interaction:

1. Neutral interaction - two organisms do not depend on each other for proliferation.

2. Commensal interaction - one organism depends on the other for growth. If the host

pays a fitness price this is called parasitism (or host-pathogen interaction).

3. Symbiosis - species-species interaction in which both partners gain fitness.

Focusing on pathogens/symbionts, interaction networks can be studied on the level of genome-scale metabolic networks. This is usually combined with constraint-based modeling (e.g., steady-state conditions) to investigate the (metabolic) interaction between microbes and their host. Modeling of metabolic networks enables prediction of the phenotypic characteristics of single organisms to communities [Thiele *et al.*, 2013].
Host-pathogen interaction networks can also be studied on the interactome level - mainly the interaction of secreted proteins with host proteins [Arnold *et al.*, 2012]. To give a rough estimation of the huge and complex network different PPIs give rise to: it has been estimated that the size of the binary human interactome comprise of about 130,000 PPIs (only about 8% are yet identified) [Thiel *et al.*, 2012]. Focusing on interactoms the interaction with the host can be viewed as interference of pathogenic proteins with the interaction network of the host. For pathogens this happens after bacterial attachment to host cell surface receptors since bacterial pathogenic proteins are - in contrast to viral proteins - either secreted or injected into host cells. There are some typically key signaling events bacterial effectors target - usually to induce rearrangements of the cytoskeleton and promote bacterial-host association or internalization (depending on their lifestyle) subsequently followed by manipulation of host cell signaling pathways involved in attenuation or evasion of the innate defenses of the host. One example for effectors that target the host cytoskeleton of eukaryotes are SopE, SopE2 and SptP (present in Salmonella) which act as guanidine exchange factors and activate either CdC42 alone or Cdc42 and Rac-1 - exploiting the fact that many eukaryotic signal pathways are controlled by members of the Rho family of GTPases. Many pathogens mimic eukaryotic protein-protein interaction domains to remodel the host's PPI network. A well studied example is the influenza A non-structural protein 1 which contains regions to bind the SH2 and SH3 domain of PI3K. These proteins are mainly responsible for the virulence of a bacteria and disturbing their interaction with the host may often lead to the death of the pathogen

- therefor being of great interest for new antibacterial therapeutics [de Chassey *et al.*, 2012, Mattoo *et al.*, 2007]. But it has also been recognized that the modulation of PPIs is one of the most challenging tasks in drug discovery - therefor deeper understanding of the network is a crucial step [Thiel *et al.*, 2012].
To understand PPI networks it is necessary to solve a number of sub-problems: (I) Prediction of possible pathogen effectors; (2) Identifying the host binding partners and (III) Understanding the impact of this interaction on the host system. There are few experimental data for host-pathogen interaction available but their relative paucity emphasizes the need for computational predictions.

The identification of host binding partners for proteins can be done on different level: predicting links among conserved protein domains or predicting interaction between protein structures. An example for a tool which predicts links among conserved protein domains is the Domain interaction MAp (DIMA) - which combines contact evidence provided by iPFAM and phylogenetic domain profiling [Pagel *et al.*, 2006].

1.2 Eukaryotic-like proteins

It has been shown in a number of single-organism studies that many effector proteins contain a specific domain signature which is usually found only in eukaryotes - their detection is a computational prediction method for effector proteins. These proteins are called eukaryotic-like proteins and the domain signature is called eukaryotic-like domain [Price *et al.*, 2010, Voth & Heinzen, 2009, Cazalet *et al.*, 2004].
It has been known for some time that these structures exist - to the author's knowledge one of the first appearances of the term ELD was in 1992 in the course of investigating Bordetella pertussis and one of its multiple adhesin proteins: pertussis toxin. This toxin can bind to glycoconjugates on human cilia and macrophages and sequence comparison of the carbohydrate recognition domain to plant and animal lectins showed that these sequences are strongly related to a subset of eukaryotic carbohydrate recognition

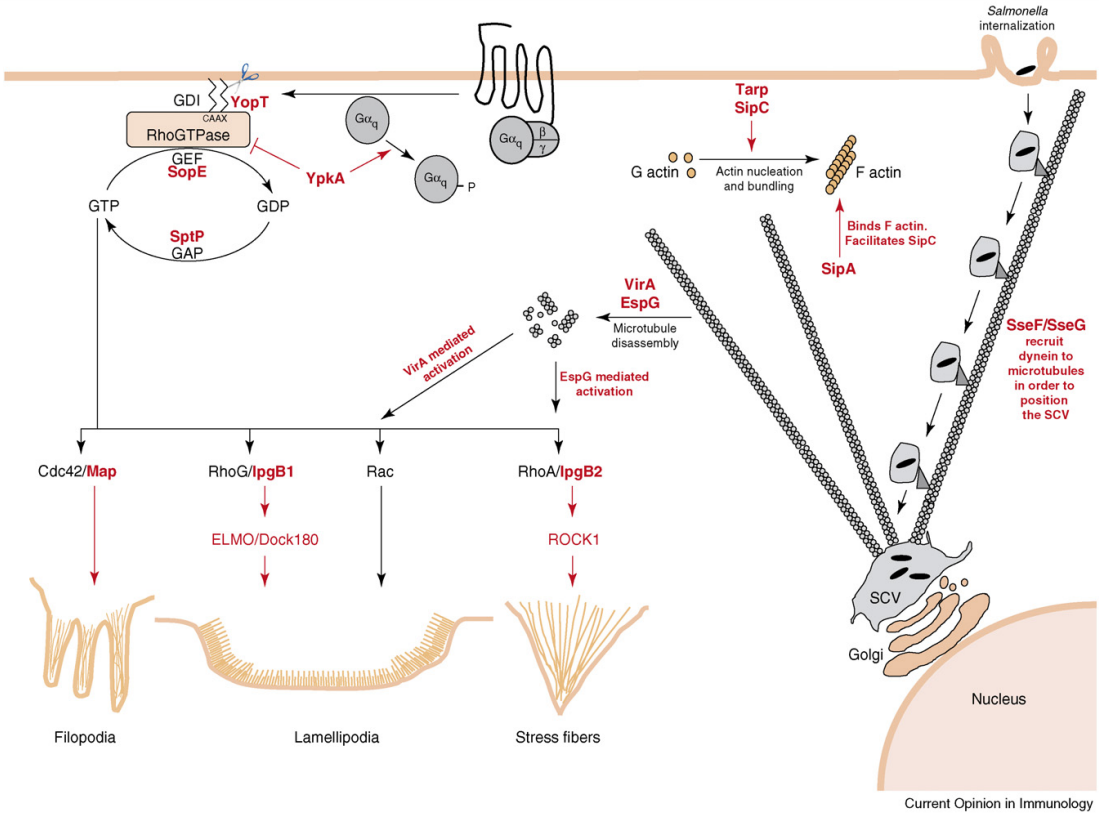

Figure 1.8: Effectors which target the host cytoskeleton. Cytoskeletal rearrangement is mediated by the Rho family of GTPases which are controlled by activating via G proteins. YpkA (present in *Yersinia* phosphorylates and inactivates Gαq-mediated pathways. YopT cleaves and inactivates RhoGTPases while SopE and SptP (both present in *Salmonella*) modulate and control membrane invagination by mimicking guanidine exchange factors (GEF) and GTPases activating proteins (GAP). The WxxxE effectors (Map, IpgB1 and IpgB2) elude endogenous RhoGTPases to directly induce formation of filopodia, lamellipodia and stress fibers. Actin polymerization is done by Tarp (present in *Chlamydia*) and SipC (*Salmonella*). SipA facilitates and spatially controls SipC-mediated actin polymerization. *Salmonella* utilizes SseF/G to position the SCVs (Salmonella containing vesicle - a vesicular compartment within Salmonella replicates) in close proximity to the the golgi network - this is done by recruiting dynein (dark gray triangles) to microtubule networks. VirA (present in *Shigella*) - and its EPEC homolog EspG - disassemble microtubules to induce RhoGTPase-mediated cytokeletal rearrangements. This enables bacteria to attache themselves to the host, to enter or move within the cytoplasm [Mattoo *et al.*, 2007].

domains of the C type [Saukkonen *et al.*, 1992]. In the following years eukaryotic-like proteins/domains were found in a variety of species (e.g. Sensory rhodopsin I in the archeon Halbacterium salinarium [Munoz-Dorado *et al.*, 1993], Serin/Threonine kinases in Myxococcus xanthus [Spudich, 1993] or src homology 3 domain in Staphylococcus aureus [Whisstock & Lesk, 1999]) - but since sequenced organisms were rare it was hard to distinguish between orthology or horizontal gene transfer. This changed around 1999 when genome sequences of organisms from each of the three divisions of life came available - making it possible to detect horizontal gene transfer (as shown in Figure 1.9) [Ponting *et al.*, 1999].

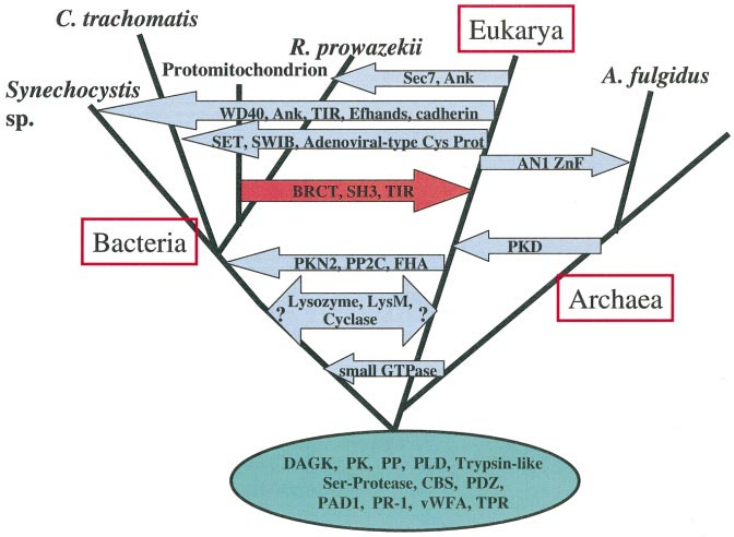

Figure 1.9: Schematic representation of suggested horizontal gene transfer events in selected signaling domains. Arrows in blue indicate proposed HGT events and the red arrow represents gene acquisition from mitochondrial endosymbiosis. Domains represented within the green oval are suggested to have been present in the last common ancestor of the three domains of live. For lysozyme, LysM, LRR and cyclase domains the direction of the transfer between eukaryotes and bacteria was not clear from this study [Ponting *et al.*, 1999].

In the following years it became more and more clear that many of these eukaryotic-like domains/proteins are part of the effector repertoire of pathogens and play a key role in their virulence [Milani *et al.*, 2012].

1.2.1 Protein domains

Protein structures can be analyzed on different structural levels: the primary structure refers to the amino acid composition; secondary structure refers to different motifs like α-helix or β-sheet, the tertiary structure refers to the sum of structural units of a polypeptide chain and quaternary structure refers to the 3D structure of a multi-polypeptide complex.

Domains are the building blocks of the tertiary structure with distinct function, conserved structure and sequence and can often function and retain its correct three-dimensional structure even when separated from the rest of the polypeptide chain. Domains also function as evolutionary building blocks - retaining its function even when integrated in new proteins [Nelson & Cox, 2004]. Domains vary in sequence length from between 25 up to 500 amino acids - shorter domains such as zinc fingers are stabilized by metal ions or disulfide bridges whereas longer domains are usually stabilized by hydrophobic and electrostatic interactions.

Protein domains are clustered into protein families whose members have diverged from a common ancestor and have similar structure, sequence and usually also similar function. A strong evolutionary relationship is normally evident within a protein family [Chothia, 1992, Hegyi & Gerstein, 1999]. To determine the function of a protein it is often more suitable to start with known domains rather than trying to identify protein function based on sequence similarity with proteins of known function. The relationship between sequence similarity and functional similarity is not straightforward - whereas the common elements in distantly related protein sequences with similar function are often their underlying structures or folds [Hegyi & Gerstein, 1999].

1.2.2 PFAM

Related domains are summarized into protein families which are collected in the Pfam database in which each family is represented by a multiple sequence alignment and an inherent hidden Markov model (HMM).

Families are multitudes of proteins that have a significant degree of sequence similarity suggesting that members of the same PFAM family share a common evolutionary history and some functional aspect.

PFAM contains two types of families: manually curated high quality Pfam-A families and automatically generated Pfam-B families. The Pfam-A families are built following a four-step procedure:

(1) first, a high-quality multiple sequence alignment (seed alignment) is built,

(2) followed by the construction of a profile hidden Markov model (HMM) from the seed alignment,

(3) this HMM model is then used to search against the UniProtKB sequence database;

(4) all hits above a certain threshold are then included in the full alignment for the family.

In addition to the UniProtKB also the NCBI non-redundant database and a collection of metagenomic samples are included in the analysis [Punta *et al.*, 2012a].

1.2.3 Eukaryotic-like domains/proteins

For Legionella pneumophila strands it has been shown that many eukaryotic-like proteins (the definition varies but the minimum agreement is that the normalized blast result of the protein of interest against the bacterial NCBI nr-db must be lower than against the eukaryotic NCBI nr-db to be characterized as ELP) are involved in exploitation and modulation of host cellular functions. Also, it hast been shown that many ELPs not only have high sequence similarity to eukaryotic proteins but also cluster in phylogenetic trees with eukaryots therefor suggesting that these proteins are obtained by horizontal gene transfer rather than convergent evolution [Lurie-Weinberger *et al.*, 2010].

Eukaryotic-like domain (ELD) is a term used to describe domain signatures in effectors that are typically found in eukaryots. Therefor it can be conjectured that ELDs may function as possible identification of yet unknown effectors. This has been shown in particular for bacterial proteins that contain ankyrin repeats which are involved in protein-protein interaction in various eukaryotic systems. In Legionella pneumophila all ankyrin containing proteins are translocated into the host cell via the Dot/Icm type IV

secretion system and are required for intracellular proliferation in human macrophages as well as in protozoa (therefor being an important virulence factor) [Cazalet *et al.*, 2004]. Other domains - like the F-box eukaryotic domain that is involved in ubiquitination of eukaryotic proteins - have also been seen in bacterial proteins which are translocated into the host cell [Price *et al.*, 2010, Al-Khodor *et al.*, 2008].

One of the first studies suggesting horizontal gene transfer as a possible evolutionary scenario for acquiring ELDs/ELPs was done about an effector protein called RalF that functions as an exchange factor for the ADP ribosylation factor (ARF) family of guanosine triphosphatases (GTPases) in 2002. The high degree of sequence similarity between the Sec7-like domains in RalF and ARF-GEFs in eukaryots combined with the fact that Sec7-homology domains have only been found in two prokaryots gave rise to the hypothesis that this gene was acquired by a horizontal gene transfer event from eukaryots to prokaryots [Nagai *et al.*, 2002]. This finding led to studies identifying a variety of ELDs/ELPs in L. pneumophila and discussing their possible evolutionary history - this is further discussed in section 1.3 [Cazalet *et al.*, 2010, Gomez-Valero *et al.*, 2009, Gomez-Valero *et al.*, 2011b].

1.2.4 Computational methods to identify ELDs - Effectors Database

Effective is a database of predicted secreted bacterial proteins. It provides pre-calculated results for all publicly available pathogenic and symbiotic bacteria [Jehl *et al.*, 2011]. Proteome data and features from a variety of different sources are integrated into the effectors database and structured in a genome repository. The annotated proteins of all publicly available completely sequenced genomes listed in the RefSeq database as well as all domain signatures detected by Pfam (Version 26) were also included [Rattei *et al.*, 2010, Punta *et al.*, 2012b]. The effectors database has implemented two different strategies to determine the likelihood for protein secretion: (1) the identification of eukaryotic-like protein domains (ELD) and (2) the recognition of amino acid sequences corresponding to signal peptides.

The systematic, large-scale approach to identify eukaryotic-like protein domains is based

on the comparison of domain frequencies in all genomes available in the genome repository. The aim is to use a numeric score to distinguish between a uniform distribution of domains over different classes of organisms and an enrichment of eukaryotic-like domains in the proteomes of pathogenic bacteria. The possibility of bacterial contamination in eukaryotic genomes was precluded by restriction to protein domains that are present both in pathogenic and at least in 3 eukaryotic genomes. For every remaining domain a background model is estimated, using the average and standard deviation of its frequencies in all non-pathogenic genomes.

The following example is taken from the Effectors database homepage (`www.effectors.org`) to show the actual calculation of the ELD-score:

The MACPF (membrane-attack complex/perforin domain - OF01823) domain is present in the proteom of 28 pathogens/symbionts, 3 non-pathogens and 33 eukaryotes. The average background frequency for non-pathogens can be calculated with knowledge of the number of non-pathogens the domain is present in (3) and the number of non-pathogenic bacteria listed in the genome repository (292):

$$\bar{X}_{non-pathogens} = 3/292 \tag{1.1}$$

The result for the average background frequency in this example is 0.01. The standard deviation in non-pathogens can be estimated as:

$$\sigma_{non-pathogens} = \sqrt{\bar{X}_{non-pathogens} * 2^{3-0.01}} \tag{1.2}$$

The result is 0.1. The calculation of the score for a particular pathogen with a proteome that contains 1 protein with this domain (therefor c=1) is as following:

$$SCORE_{Pathogen} = \frac{c - \bar{X}_{non-pathogens}}{\sigma_{non-pathogens}} \tag{1.3}$$

Although the distribution of domains has varying shapes, it has been shown that the domain enrichment scores follow the characteristics of Z-scores and therefor scores higher

than 3-5 can be considered as significant. Domains that occur only in proteome of pathogens and eukaryotes are listed with a score of 10000, domains which occur in just one pathogenic proteom have a score of 9000 [Jehl *et al.*, 2011].

In this thesis the pre-calculated predictions for whole bacterial genomes and the calculated ELD scores were used to divide the genomes into proteins containing ELDs and proteins with no such domains.

All pre-calculated bacterial genomes were downloaded (version: november 2012, 1058 bacteria) and 70 bacteria were randomly chosen for further analysis.

1.3 Proposed evolution of eukaryotic like domains/proteins

The detection of sequences that have close proximity to eukaryotic proteins follows the question of the evolutionary origin and function of such sequences. The analysis of L. pneumophila genomes has revealed that many of the predicted or experimentally verified Dot/Icm secreted substrates are proteins similar to eukaryotic proteins or contain motifs mainly or only found in eukaryotic proteins. The hypothesis is that this is not only true for L. pneumophila - specifically that most proteins which are characterized as ELPs or contain ELDs are effector proteins and there is an evolutionary benefit in mimicking/obtaining eukaryotic-like proteins/domains. Analysis of ELPs/ELDs in Legionella (Legionella pneumophila str. Paris, str. Philadelphia 1 and Lens) revealed that many such events are likely to be involved in interfering with eukaryotic cellular functions and therefore constitute virulence factors but also that some are important for the life cycle of L. pneumophila. Ankyrin repeats, SEL1 (TPR), Set domain, Sec7, serine threonine kinase domains (STPK), U-box, and F-box motifs are all eukaryotic domains which were previously described in the proteom of L. pneumophila and characterized as ELDs - sphingosine-1-phosphate lyase and sphingosine kinase, eukaryotic like glycoamylase, cytokinin oxidase, zinc metalloprotease and an RNA binding precursor are examples for ELPs in L. pneumophila. Functional prediction and functional studies have confirmed their involvement in eukaryotic cell modulation [Gomez-Valero *et al.*, 2011b,Brüggemann *et al.*, 2006,Cazalet *et al.*, 2010]. There are two possible scenarios to explain the evolution of effector proteins: (a) convergent evolution or (b) horizontal gene transfer. As discussed in subsection 1.4.1 there is an ongoing discussion about the importance of HGTs in respect of genetic viability and the evolution of bacteria [Gomez-Valero *et al.*, 2009].
A closer look at Figure 1.8 shows how specific these effectors act in the host cell and how they modulate the interaction between proteins.
The evolutionary origin is normally determined by phylogenetic tree reconstruction and parsimony - e.g., in Legionella a species tree was constructed and a protein tree was mapped on the species tree topology. This method showed the possibility that at least

some ELPs and also ELDs have been acquired by horizontal gene transfer from eukaryots. This has been suggested by recent studies for L. pneumophila [Nagai *et al.*, 2002]. Given the lifestyle of L. pneumophila the acquisition of some of the eukaryotic like genes by HGT events from protozoa seems plausible - the ralF was the first gene for which a HGT from eukaryotes has been suggested, another one is the eukaryote-like sphingosine-1-phosphate lyase of L. pneumophila (the phylogenetic tree is shown in Figure 1.10) that is encoded by lpp212 [García *et al.*, 2004, Gomez-Valero *et al.*, 2011b, Cazalet *et al.*, 2010, Cazalet *et al.*, 2004].

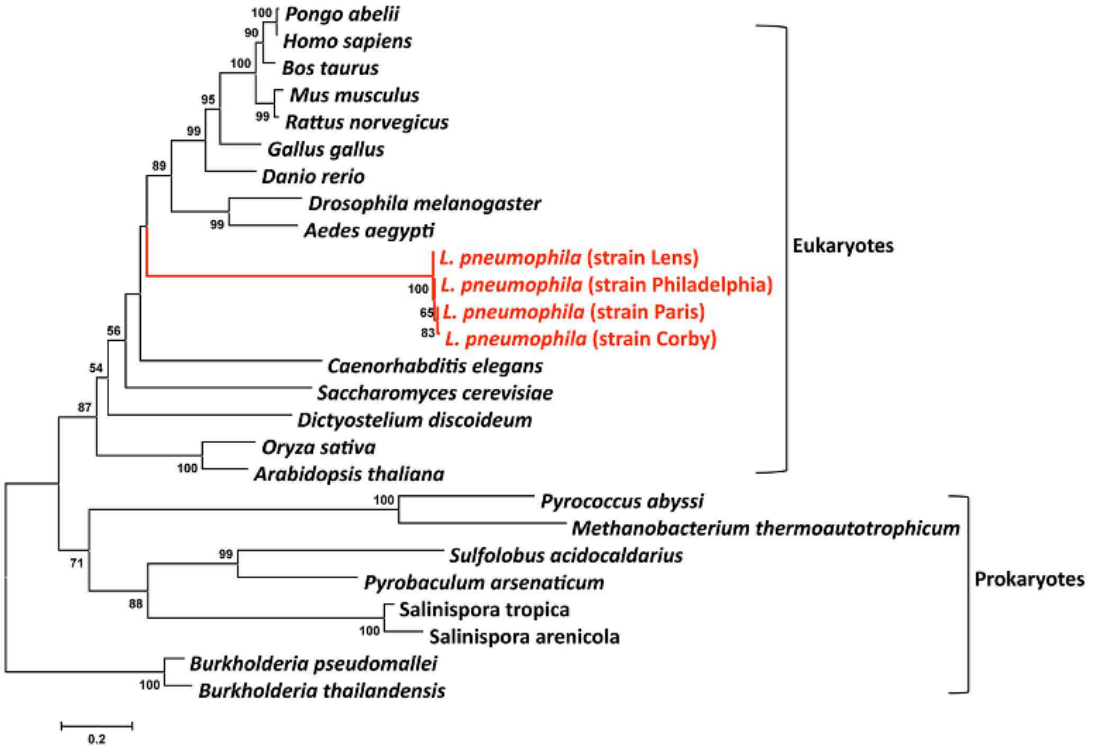

Figure 1.10: Phylogenetic tree built with a Neighbor-Joining method from a multiple sequence alignment for sphingosine-phosphate lyase proteins. In red are the L. pneumophila sequences shown that are embedded in the eukaryotic clad thus suggesting a HGT [Gomez-Valero *et al.*, 2011b].

Given the fact that HGT events are independent evolutionary processes it is surprising that many ELPs/ELDs are conserved in all six L. pneumophila strains. This might indicating an ancient gene transfers - the favored explanation for this is that these early

HGT events might have allowed a common Legionella ancestor to colonize an intracellular niche or to adapt better to the intercellular environment of a specific protozoan species leading to the evolution of the species L. pneumophila [Gomez-Valero *et al.*, 2011b]. ELPs/ELDs that are not conserved and of eukaryotic origin are acquired at different time points throughout the evolution of Legionella which can be roughly estimated based on their phylogenetic distribution in the protein tree [de Felipe *et al.*, 2005, Price *et al.*, 2010, Gomez-Valero *et al.*, 2011b, Lurie-Weinberger *et al.*, 2010].

1.4 Horizontal gene transfer

1.4.1 Overview

Genetic variability - and organismic variability mostly boils down to genetic variability (this is not without exceptions [Diard *et al.*, 2013]) - is one of the keystones for the survival of life. In eukaryotes, variability in the genome is primarily due to sexual reproduction which involves chromosomal recombination - shuffling of pre-existing sequences - during meiosis. In prokaryotes other factors determine the rate of evolution, including frequent occurrence of point mutations, high levels of recombinations, gene silencing and the transfer of genetic material between different species - named horizontal or lateral gene transfer [Hacker & Carniel, 2001].

Horizontal gene transfer (HGT) - described as 'the non-genealogical transmission of genetic material from one organism to another' is of considerable importance for the understanding of the evolution of Bacteria and Archaea [Goldenfeld & Woese, 2007]. The phylogenetic incongruence - emerging in the 1990s while studying phylogenetic relationships in microorganisms - between molecular, morphological and physiological markers and certain genes revealed a pressing need to understand the evolution of organism and its genes to a greater extend. The identity of and relation between the three domains of life (Bacteria, Eukaryots and Archaea) is based on the sequences of small subunit ribosomal RNAs (SSU rRNAs) - rooting has been done independently, using certain key proteins [Woese *et al.*, 1990, Gupta, 1998]. This tree is widely accepted as the most

important single guide to understanding genealogical relationships and is supported by many molecular data sets. However, there are also molecular data sets which disagree - rooting and/or branching differently than then the rRNA tree. On this background the concept of horizontal gene transfer between organisms emerged and offered a way to unify these opposing data sets [Hilario & Gogarten, 1993].

It has been shown that transfer of genetic material cannot only occur among but also between domains of life in all possible directions: from Bacteria to Archaea, from Archaea to Bacteria, from Archaea to Eukarya, from Bacteria to Eukarya, from Eukarya to Bacteria and even within Eukarya [Boto, 2010, Doolittle, 1999b]. The importance of horizontal gene transfer for evolution is widely debated - there are voices pointing out that HGT is a mechanism that permits the acquisition of evolutionary novelties and therefor questions the conception of a gradualist process driving the appearance of novel traits and functions. This calls - more drastically speaking - the tree-based view of the evolution of life into question. [Boto, 2010, Doolittle, 1999a, Doolittle, 1999b].
For a long time the prevailing opinion seemed to be that HGTs are rare and therefore insignificant for the general understanding of evolution - this changed quickly with the emerging of sequenced-based genomics [Ochman *et al.*, 2000]. Nowadays, it is estimated that the genome of each microbial organism consist of between 1.6 and 32.6 per cent genes that were acquired by horizontal gene transfer [Boto, 2010]. This leads, among other reasons, to huge differences in gene repertoires even among closely related bacteria such as E.coli and Haemophilus influenzae - indicating that vertical transfer of genetic material is not enough to describe genome evolution conclusively [Koonin *et al.*, 2001]. Besides E.coli and Haemophilus influenzae, another example is the pathogenic sister species of E.coli: Salmonella enterica. The phenotypical diversification between the two species are mostly due to new genetic material acquired by horizontally transferred genes or loss of genes but not derived from the stepwise mutational changes resulting from point mutations. Stepwise mutational changes only rarely lead to novel functions, whereas traits encoded by acquired DNA will occasionally enable the ability to explore

new environments [Lawrence & Ochman, 1998].

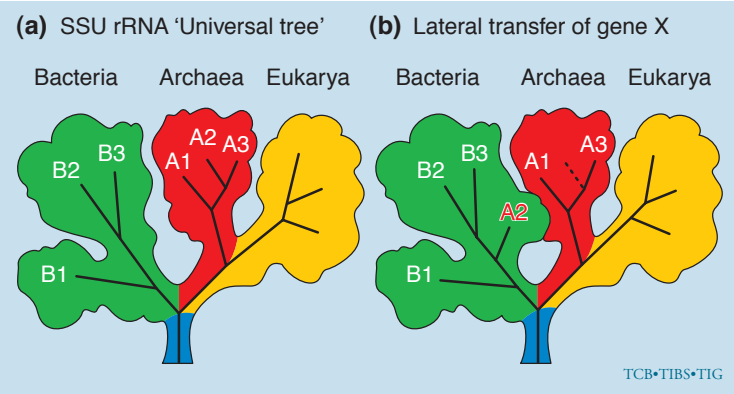

Figure 1.11: Horizontal Gene Transfer. (a) shows a simplified version of a universal phylogenetic tree derived from SSU rRNA. B1-B3 are three bacterial and A1-A3 are three archaeal taxa. (b) shows a genomic tree for the case that gene X has been transferred (HGT) from an ancestor of B2 and B3 to an ancestor of A2. It is also possible to explain this example by paralogy and differential gene loss, but for such a mechanism many steps are purely speculative and constructed rendering such a scenario much more unlikely [Koonin *et al.*, 2001].

In bacteria, gene transfer takes place by using three different mechanisms: transformation - responsible for the uptake and incorporation of naked DNA; transduction - encapsidated host DNA is transferred into a bacteriophage which functions as a vector for injecting DNA into a recipient cell; and conjugation - cell contact depending DNA transfer mechanism nearly ubiquitous in bacteria [Davison, 1999].

1.4.2 Biological mechanisms of HGT in bacteria

Transformation

The term transformation describes a mechanism which is responsible for acquisition of naked DNA from the extracellular environment, integration and functional expression of this DNA. The ability of a bacteria to undergo transformation is called genetic competence - it is a transient physiological state, which is tightly controlled by approximately 20 to 50 proteins and a reaction to environmental variables like nutritional signals, altered growth condition, cell density or starvation. Up till now it is not possible to determine if

a given bacterial isolate can develop competence during its life cycle - at the momentary point of research it seems that only about 1 % of the described bacteria are transformable. Many human pathogenic bacteria are naturally transformable - this includes bacteria of the genera Campylobacter, Haemophilus, Helicobacter, Neisseria, Pseudomonas, Staphylococcus and Streptococcus [Thomas & Nielsen, 2005].
A requirement for the uptake of free DNA is it's abundance in the environment - and free DNA is indeed present in great abundance since it is released from dead organisms [Lorenz & Wackernagel, 1994].
The uptake of genetic material - by definition achieved when exogenous, DNase-sensitive DNA is converted into DNase-protected DNA - means for Gram-negative and Gram-positive bacteria slightly different things in regard to where the DNA is located after the uptake. In Gram-negative bacteria uptake of DNA can be achieved by translocation of DNA across the outer membrane into the periplasmic space, whereas for Gram-positive bacteria it means that the DNA is located in the cytosol.
Gram-positive and Gram-negative microorganisms use related proteins to import chromosomal DNA and Plasmids which are partly homolog to proteins involved in the assembly of type IV pili and T2SS [Chen & Dubnau, 2004]. The schematic model of the mechanism of natural transformation is shown in Figure 1.12.

Conjugation

Conjugative transfer is specifically linked to plasmid acquisition or mobile transposable elements, event though it is also possible but much more unlikely to acquire whole chromosomes from different bacteria. It is mediated by cell-to-cell junctions and the formation of a pore connecting the cytoplasm of both bacteria thus allowing DNA to pass through this structure.
Conjugative transfer is considered to be a diverse collection of processes - ranging from large self-transmissible plasmids in Gram-negative bacteria using the type IV secretion apparatus for unspecific plasmid trading to a very tightly regulated transfer system only allowing transfer to a limited group of bacteria species in some Gram-positive bacte-

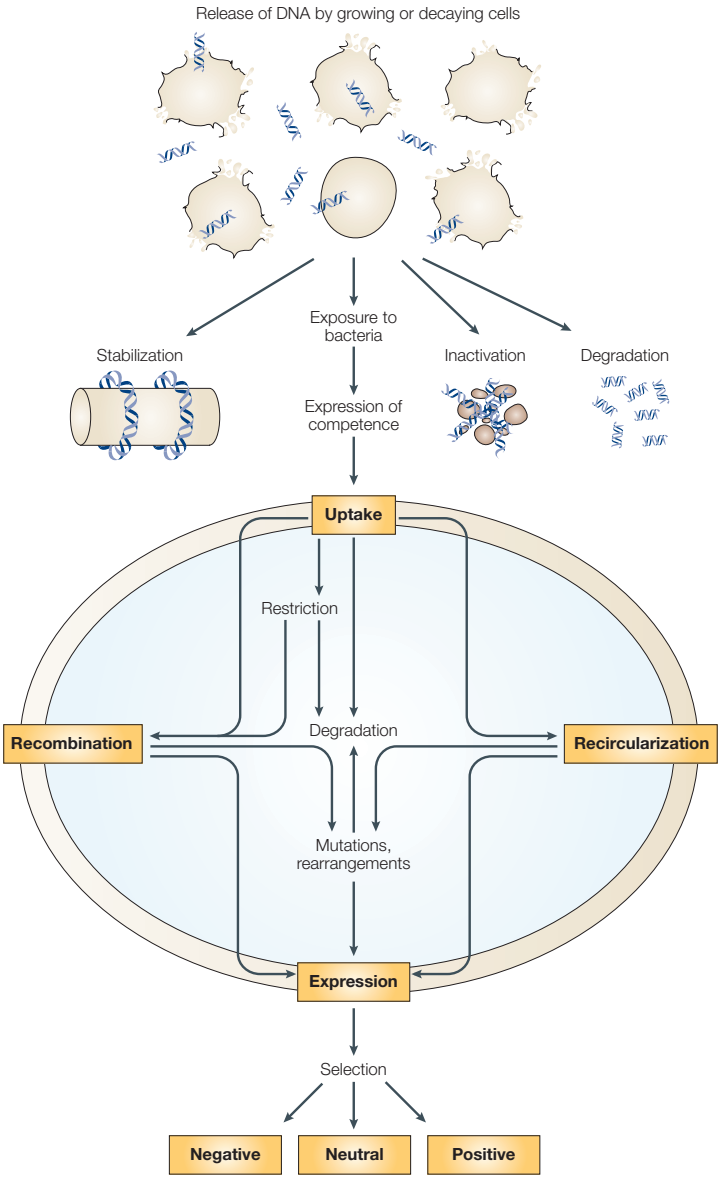

Figure 1.12: Schematic view of the natural transformation of recipient bacteria and selection of transformants. The first step of transformation involves the binding of exogenous DNA to the surface of competent cells. The mechanism for the translocation varies among bacteria and is yet not fully characterized. The DNA is transported single-stranded into the cytoplasm - in B. subtilis it has been shown that the present of surface endonucleases (such as NucA) which introduces double-stranded cleavages increase the rate of DNA intake. Finally, following uptake, the transferred DNA has to be integrated into the bacterial genome for its persistence. Integration can be achieved by homologous recombination or by sequence-independent, illegitimate recombination. Plasmids that managed to reconstitute a replication-proficient form do not need to integrate into the host genome [Thomas & Nielsen, 2005].

ria [Thomas & Nielsen, 2005]. A schematic overview of the mechanism of conjugation is discussed and shown in Figure 1.13.

Transduction and gene transfer agents

For a fair share of horizontal gene transfer between different bacterial species phages are responsible - it was shown that two-thirds of all gamma-proteobacteria and low G+C gram-positive bacteria harbor prophages (including Salmonella spp. and E. coli). Transduction only occurs after phage infection. The transfer of genetic material can occur in two cases: (a) imprecise excision of the phage genome from bacterial chromosome - termed specialized transduction - or (b) faulty packaging of host DNA instead of phage DNA - called generalized transduction [Brüssow *et al.*, 2004]. A process similar to generalized transduction but with differences is HGT via gene transfer agents (GTAs). This are prophages which are defective and incapable of self-replication and tend to package bacterial genome fragments [Stanton, 2007].

1.4.3 Biological mechanisms of HGT in eukaryotes

When the need arise to predict whether a prokaryotic gene was transferred via HGT from eukaryotes the contrary case should also be considered: that probably the gene is of prokaryotic origin and was transferred to an eukaryotic ancestor.
There are two different ways of gene transfer in eukaryotes: (a) endosymbiotic gene transfer: the transfer of genes from organelles (mitochondria or plastids) with an endosymbiotic origin to the nucleus of the eukaryotic cell and (b) HGT between unrelated species. While the first scenario is widely accepted, the later is very controversial.
Claims that over 100 human genes are of prokaryotic origin and only recently transferred were soon rejected by more in-depth reanalysis using phylogenetic methods [Lander *et al.*, 2001, Stanhope *et al.*, 2001, Andersson, 2005]. While it appears for now as if this is an extremely rare scenario, it cannot be formally excluded [Genereux & Logsdorn, 2003].

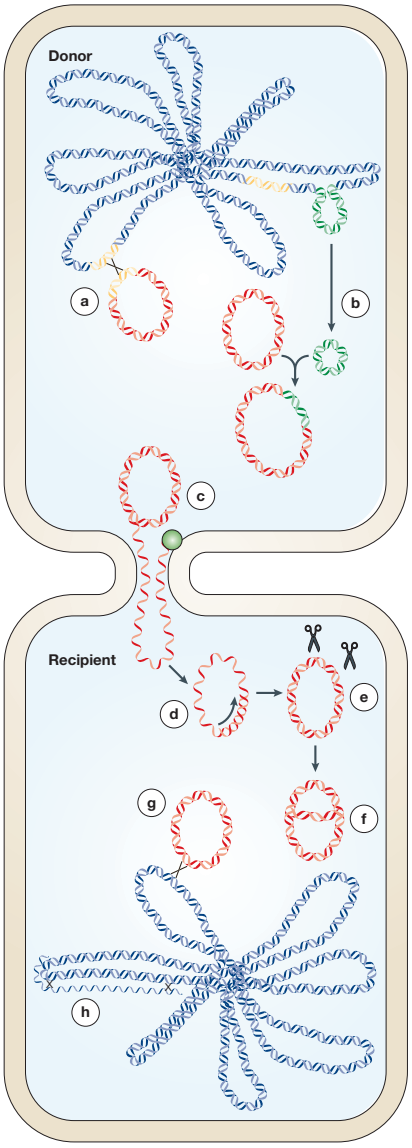

Figure 1.13: Schematic overview of plasmids and conjugative transfer in the horizontal/lateral transfer of genes. Starting in the donor cell, the events for the successful transfer of plasmids are: (a) the plasmid is integrated into the chromosome by recombination between insertion sequence elements; (b) a transposable element is transferred through a circular intermediate from the chromosome to the plasmid; (c) at the mating-pare apparatus a rolling-circle replication is initiated. The events in the recipient cell are as following: (d) the genetic material is re-circularized (e) and attacked by restriction endonucleases (shown as scissors); (f) replication and (g) integration into the chromosome by illegitimate Campbell recombination. Since it is also possible that chromosomal DNA is transferred (yet not as numerous as plasmids) (h) shows the transfer between chromosomal DNA and resident chromosome [Thomas & Nielsen, 2005].

1.4.4 Computational methods to identify horizontal gene transfer

The task to identify horizontal gene transfer inevitably rely on some unusual feature of a subset of genes that distinguishes them from other genes in the genome. Since for the vast majority of HGTs it is not possible to directly observe the event (but sometimes it is [Babic *et al.*, 2008]) the proof of such an event remains probabilistic. Thus several different criteria and methods based on them have been developed in order to identify their presents in the genome of a host organism [Koonin *et al.*, 2001].

Sequence composition

This method is based on the 'genome hypothesis' according to which codon usage, oligonucleotide frequency and GC content are distinct signatures of each genome. The assumption of this method is that directional mutation pressures within bacterial genomes impart a specific bias to the composition of long-term members of the genome. This happens in a way such that recently acquired genes will appear aberrant by comparison if they have evolved in a genome with a different mutational bias [Lawrence & Ochman, 2002].

Therefor every gene that is - compared to the mean of the genome in this aspects - different may be a good candidate for HGTs. This approach identifies many recently transferred genes - however, there are also other causes for compositional bias and it is also important to keep in mind that an untypical nucleotide composition is no proof of HGT since selective effects or mutation bias could also account for this composition [Lawrence & Ochman, 1998, Koonin *et al.*, 2001, Gogarten & Townsend, 2005]. Furthermore not all recently acquired genes show compositional bias, also ancient HGTs - due to sequence adoptions - are not detectable rendering this method insensitive to some events and scenarios. In principle, examination of the phylogenetic conflict among loci is the most direct approach to screen for horizontally transferred genes [Koonin *et al.*, 2001].

An advantage of this method is that it does not rely on comparison with other genomes - therefore providing an independent mean of assessing the impact of HGT [Lawrence & Ochman, 2002].

Phyletic Patterns

The availability of many completely sequenced genomes enables the building of clusters of orthologous (direct evolutionary counterparts related by vertical descent) genes (COGs). The presence or absence of a species in a cluster of orthologous genes is called a phyletic pattern. If this pattern is for a given species not in accordance with the species tree, this is a strong hint for a HGT (e.g., presence of a gene in all eukaryotes and one bacterial lineage). The advantages of this method - besides its independency of other genomic data - are the possibility to identify potential donors, ancient transfers and it is possible to analyze large sets of genomes in reasonable time [Koonin *et al.*, 2001]. One of the disadvantage is that gene loss or lineage-specific mutation rates can also generate unusual species distributions in COGs [Lawrence & Ochman, 2002].

Sequence similarity based methods

This and the following method (Phylogenic tree reconstruction) both use the concept of homologs. Homologs are genes that share a common origin therefore displaying a greater similarity to each other than to most other genes [Koonin, 2005]. Probably the most intuitive and simplest way to identify possible HGTs is to conduct a similarity search against a sequence database and evaluate the taxonomy of the best hits - if they belong to homologs from distant taxa the chances for a HGT are not bad. A standard way to identify similarities of a given gene to its homologs is to use a BLAST search [Altschul *et al.*, 1990] against the NCBI genome database. This has been done in this thesis and will be discussed in the Methods section in more detail but for now the basic workflow should held as example: genes from bacteria were compared to their best hits - whenever the gene had significantly better 'similarity' to an eukaryotic gene than to a prokaryotic one this gene was noted as possible HGT. This method is fast and automatable - these advantages come at the expense of precision due to some reasons including e.g., varying mutation rates in different taxa as well as the similarity score threshold chosen for a particular search [Koonin, 2005]. Also, it was previously shown that the best BLAST hit does not necessarily represent the closest sequence relatives - as shown by the predicted

113 lateral gene transfers from bacteria to vertebrates published by the human genome consortium based on BLAST results - none of which were supported by phylogenetic analysis [Frickey & Lupas, 2004, Lander *et al.*, 2001].
To describe quantitatively with a numeric value for a given gene (e.g., bacterial gene) the relationship between the best in-group (bacteria) against best out-group (eukaryotes or archaea) a so-called alien index was introduced in 2008 [Gladyshev *et al.*, 2008]. First, E-values were used to calculate the difference, resulting in the following equation:

$$\text{AI} = ln((\text{In-group E-value}) + e^{-200}) - ln((\text{Out-group E-value}) + e^{-200}) \tag{1.4}$$

This quickly changed and E-values were replaced by Bit-scores [Boschetti *et al.*, 2012] and not the difference but the ratio was calculated [Williams *et al.*, 2012] resulting in the following relationship for the AI:

$$\text{AI} = \frac{\text{Bit-score of In-group}}{\text{Bit-score of Out-group}} \tag{1.5}$$

Phylogenic tree reconstruction and reconcialiation

Phylogenic tree reconstruction can be separated in two different steps:
(1) distances between character sets are estimated (either using Bayesian inference, maximum likelihood or maximum parsimony), (2) and subsequently a tree is inferred from the estimated distances. The implicit assumption - that a set of sequences has evolved from a common ancestor by mutation and selection without recombination events - should be kept in mind. This excludes several evolutionary mechanisms like hybridization, horizontal gene transfer, recombination or gene duplication followed by gene loss. To identify gene duplications, losses and transfers the gene tree is mapped on the species tree. This can be done manually or programs can be used. Examples for the implementation are: the GeneTree program (using a modification of Eulenstein's linear time algorithm to perform the tree-mapping) or the T-rex web server [Page, 1998, Koonin *et al.*, 2001, Boc *et al.*, 2012]. In this thesis the T-rex web server was tested as an alternative for the phat

java application which does not map the gene tree on the phylogenetic tree but analyzes the relationship of species in the gene tree. One of the reasons why this method was not used in the final investigation was its inability to handle homologs in the gene tree - resulting from its use of bipartition dissimilarity for tree inferring [Boc *et al.*, 2010].
In this thesis two different tree reconstructing methods were used: (a) the Neighbor Joining (NJ) algorithm and (b) the Maximum-likelihood (ML) approach. Both methods produce with the same data and same evolutionary model very similar trees. They differ greatly in their time consumption (also due to the fact that the NJ method generates only one tree in contrast to the ML method) and NJ is regarded as being not as sensitive to subtle sequence dissimilarities as the ML mehtod [Felsenstein, 1981, Saitou & Nei, 1987, Diard *et al.*, 2013].
This approach is regarded as the most reliable prediction method for detecting HGT events, allowing in addition the detection of ancestral transfer and determination the direction of the transfer. On the other hand the problem of phylogenetic tree construction is considered to be an NP-complete problem therefor very computationally intensive and almost non-applicable for larger datasets [Foulds & Graham, 1982]. Even the polynomial heuristic approaches (UPGMA and Neighbor-Joining) are computational intensive and the probabilistic methods like Maximum-likelihood are so even more [Koonin *et al.*, 2001]. The NJ tree reconstruction method is
Since the amount of sequences being generated exceeds our ability to manually analyze them it became necessary to investigate these sequences with automated methods in a reasonable time. If this is done for one or more organism this method is called genome-wide phylome calculations - incorporating all steps from the seed sequence to the multiple alignment and the final phylogenic tree. An example for such a pipeline is PhyloGenie which was used as major method in this thesis [Frickey & Lupas, 2004].

1.5 Aims of the Master Thesis

It has been shown that many ELD containing proteins and ELPs are effector proteins and modulate host protein functions. There are two possible ways to obtain such proteins: by (a) convergent evolution or (b) horizontal gene transfer. The aim of this thesis was to investigate the possibility of a significant link between ELD containing proteins/domains and horizontal gene transfer from eukaryotes to prokaryotes - in other words to test the hypothesis that proteins containing ELDs are the result of an horizontal gene transfer at some point in the past. This was investigated using a dataset of 70 proteoms of known pathogens/symbionts, which were chosen randomly from 648 bacteria present in the effector database the distribution of the taxonomic lineages are shown in section 3.1). To analyze these proteoms a large-scale phylogenetic tree reconstruction was performed and subsequently the relationship of the prokaryotic seed protein with its homologs in prokaryotes and/or eukaryotes was analyzed. A first dataset was generated using proteins as input for tree construction, a second dataset utilized domains as input. If there is a direct relationship between ELDs and horizontal gene transfer significantly more proteins containing an ELD should cluster with eukaryotic species in the phylogenetic tree in comparison to proteins without ELDs. This has not been shown for many organism till now and hast not been not performed - to the author's knowledge - in such a large scale approach (full proteom of 70 pathogens/symbionts). Phylogenetic trees which show HGTs for proteins with ELDs would also enable to investigate the origin of the horizontal gene transfer event. It would be interesting to see if there is a taxonomic connection between the donor of the HGT and the host organism of the pathogen. which These findings would show the evolutionary history of these effector proteins.

Chapter 2

Methods

2.1 Large-scale construction of multiple-alignment and phylogenetic trees for known bacterial symbionts/pathogens

2.1.1 Overview

A workflow (a graphic representation is shown in 2.1) was implemented to analyze the phylogenetic relationship between a protein of interest (or seed protein) and its homologs in other organism. First the similarity between the seed protein and its best hit in eukaryotes and prokaryotes was compared - failing to reach a certain threshold value resulted in its exclusion from further analysis (these excluded proteins are e.g., housekeeping proteins unique to the domain prokaryotes) - thus functioning as a first filter. The remaining proteins were aligned with its homologs and phylogenetic trees were constructed. This was done utilizing a grid engine to calculate multiple sequence alignments and phylogenetic trees in parallel since this calculations had to be done for every protein of the proteom of 70 bacteria that fulfilled the AI criteria rendering cpu power the rate determing part. This workflow can be sectioned in the following parts:

1. Retrieving a complete list of proteins for all bacteria classified as pathogens/symbionts and present in the effectors database[1]. This list is further divided into two sub-

[1] www.effectors.org

groups based on the score used in the effectors database for the presence of eukaryotic-like domains (ELD) in prokaryotic proteins. If a protein contains a eukaryotic-like domain the 'uniqueness' of this domain in prokaryotes is scored from 4 to 10000 - the higher the more significant is the finding. The proteome of a prokaryote was divided into two groups: non-ELD containing proteins (nELDP) and ELD containing proteins (ELDP) - a protein with a ELD-score of four or higher is regarded as member of the ELD group whereas proteins with a score of four or lower are regarded as members of the non-ELD group [Jehl *et al.*, 2011]. This leads to two lists for each organism.

2. For each bacteria a similarity search of its proteins against an eukaryotic and prokaryotic (excluding the family of the bacteria) database is performed. For example for Legionella pneumophila str. Paris a similarity search against all eukaryotes and all prokaryotes except bacteria that belong to the family Legionellaceae is done. Based on the ratio between the best bit-score for a eukaryotic and the best bit-score for a prokaryotic protein the proteome is further filtered. Between 40 to 60 % of the proteins are excluded based on either no match in eukaryotes or a low ratio of eukaryotic vs. prokaryotic bit-scores.

3. A FASTA file with the protein sequences is built based on the protein's acc number making use of Bio Perl's Database object interface to GenBank [Stajich *et al.*, 2002]. This FASTA file is the starting point for using the PhyloGenie pipeline with some minor modifications. PhyloGenie is an established work flow for phylogenetic tree construction and analysis - for every protein a multiple alignment based on BLASTP results is built using blammer and a profile-Hidden-Markov Model (HMM) search against the full-length sequences [Frickey & Lupas, 2004].

4. Using the multiple alignments tree construction can be done either by utilizing the PhyloGenie specific tree construction package mtrees.jar (utilizing a neighbor-joining clustering method for creating phenograms) or raxML (utilizing a maximum-likelihood clustering method for creating phenograms [Stamatakis, 2006]), both

producing trees in newick format (New Hampshire bracket format).

5. Trees were analyzed for differences between the species and gene tree. This was done either manually (using the improved ATV software named 'archaeopteryx') or automatically (using the PHAT program which is part of the phyloGenie package) for the taxonomic relationship between the seed protein (the original protein of interest), eukaryotic and other prokaryotic proteins [Han & Zmasek, 2009, Frickey & Lupas, 2004]. The obtained trees are unrooted - in the course of analyzing the relationship between the different nodes of the tree rooting is performed for the seed sequence and the tree is analyzed in relation to this tip note.. Using an iterative approach the requirements for a positive signal increased - starting from the question: Are there any monophyletic nodes with the seed organism and eukaryotes? The final search gave only a positive signal if there were monophyletic nodes with more than three eukaryotes and less than 20 prokaryotes which are not in the taxonomic family than the seed organism. This was quantitatively analyzed and compared to their affiliation to the ELD or non-ELD classification from the effectors database - based on this knowledge a chi-square test was performed to analyze if this two binary properties (classified as ELD/non-ELD and signal/no-signal with PHAT) are in/dependent from each other.

There was no pipeline available to combine pre-filtering of a proteom with a sequence similarity criteria with phylogenetic tree reconstruction, tree analysis and automatic result processing for a variety of organisms. It was crucial to implement a pipeline that can deal with the large data set and the grid system, making the calculations feasible in a reasonable timeframe and performing all steps automatically - the pipeline implemented in this thesis was able to do all of the mentioned steps.

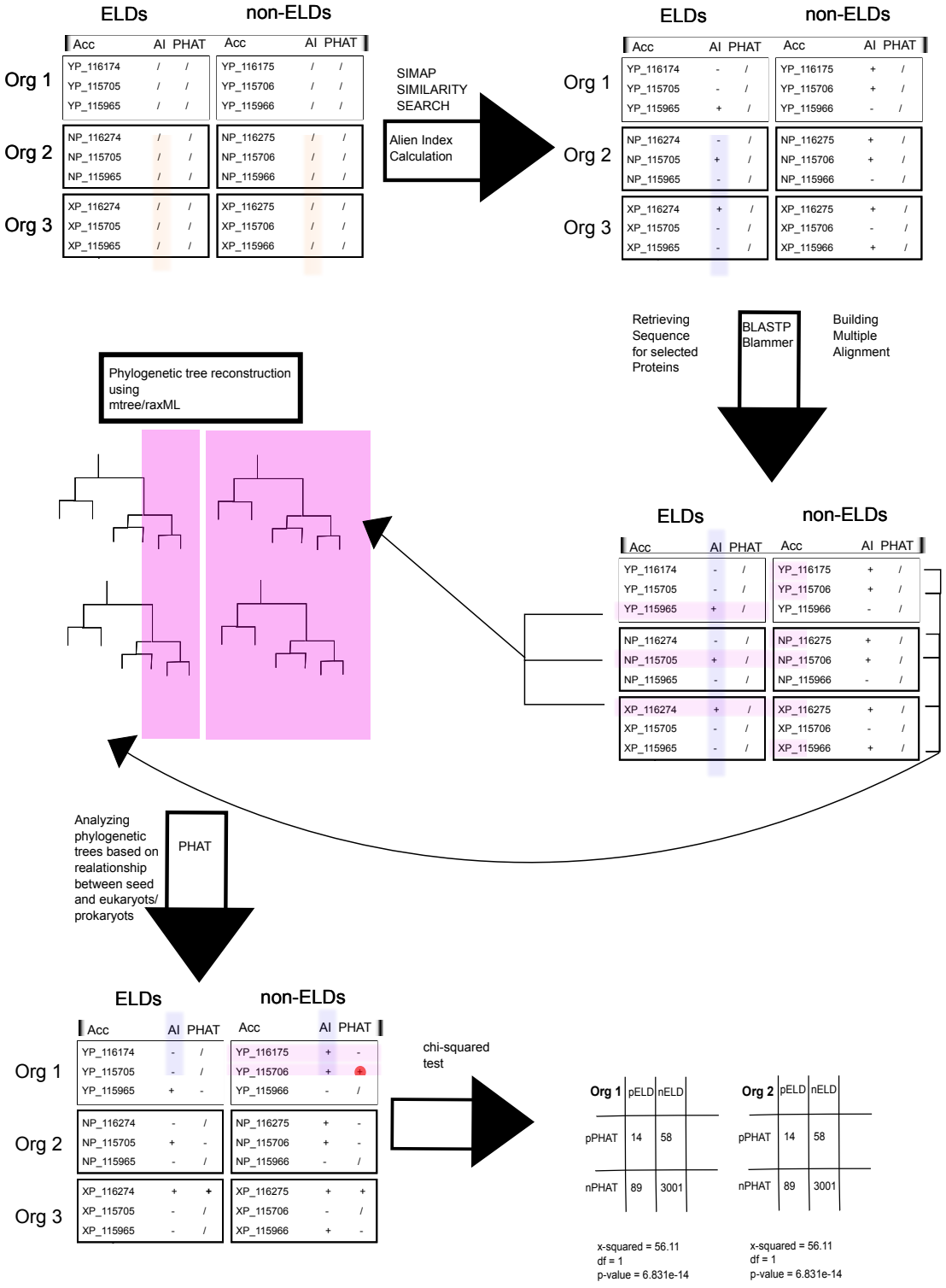

Figure 2.1: Overview of the workflow

2.1.2 Methods and parameters

Method and parameters used for calculating the alien index

A similarity search for the protein of interest against a database of all known eukaryotic proteins (excluding Pinus, Populus trichocarpa and Oryza sativa Indica Group for their known prokaryotic contamination) and all known prokaryotic proteins (excluding the all prokaryotes belonging to the same taxonomic family as the bacteria of interest) was performed using the pre-calculated Similarity Matrix of Proteins (SIMAP) [Rattei *et al.*, 2010]. This saved much computational time in comparison to a BLASTP search.
The E-value cut-off was set to $1*10^{-4}$ and the results were limited to 20000 entries.

Method and parameters used for multiple alignment and tree reconstruction

For phylogenetic tree reconstruction and the preceding multiple alignment the PhyloGenie package was used [Frickey & Lupas, 2004] - automating the steps from seed sequence to phylogenetic tree reconstruction. PhyloGenie used for sequence comparison the BLASTP v2.26 and for the multiple sequence alignments the Java program Blammer.
Certain characters ('(',')',',',';') had to be substituted by underscores in the input Fasta sequence, the NCBI non-redundant sequence database and the NCBI taxonomy files to work with the PhyloGenie pipeline. BLASTP was evoked with the following parameter set: -e 10 - b 2000 -v 2000. The parameter -e describes the expectation value threshold; parameter -b sets the upper limit of database sequences with HSPs to the query (database sequences to show alignments for); -v controls the upper limit of one line descriptions to show in the result. The BLASTP search was performed against the NCBI non-redundant (nr) sequence database obtained from the NCBI website[2]. It was - together with the NCBI taxonomy files - stored locally at the 28. December 2012.
Blammer performs five post-processing steps for BLAST results converting sets of high-scoring segment pairs (HSPs) to multiple alignments. In the first step full-length sequences for HSPs below an expectation values of 10 are extracted, enabling the search of

[2] www.ncbi.nlm.nih.gov

the sequence database with a profile hidden Markov model (HMM) in a later step. The HSPs of the query sequence with an E-value better than 10^{-5} and a coverage greater than 60 % are extracted and a multiple alignment is built based on the most dissimilar 600 sequences using static 'alignment pillars' of minimal 100 consecutive ungapped columns and realigning all residues between such pillars using ClustalW [Larkin *et al.*, 2007,Frickey & Lupas, 2004]. The HMM generated from the HSPs of the most dissimilar sequences is then used to search the database of full-length sequences generated in the first step - the alignment for phylogenetic tree reconstruction consists of a maximum of 2000 HMM-HSPs with E-values better than 10^{-6}. Phylogenetic trees were calculated using the neighbor-joining (NJ) tree reconstruction method in combination with the Poisson distance correction scheme and bootstrapping with 100 replicates.

Method and parameters used for scoring of phylogenetic tree reconstruction

The JAVA program PHAT was used to search the phylogenetic trees the previously described method yielded for both the ELD containing and non-ELD containing group. Three iterative query strings were used to increase the selectivity of every analysis.
The first query searched for every node that contained the taxonomic name of the bacteria the tree was built from and furthermore an undefined number of eukaryotes. To obtain this nodes the query string 'Taxonomic name of seed prokaryote & Eukaryots' was used.
The second query searched for every node that contained the taxonomic name of the bacteria the tree was build from and at least five eukaryotes. The query string used for this analysis is 'Taxonomic name of seed prokaryote & {Eukaryots>5}'.
The third query searched for every node that contained the taxonomic name of the bacteria the tree was build from, not more than 20 other prokaryotic species and at least three eukaryotes. The query string used is '(Taxonomic name of seed prokaryote & {Eukaryots>3} & {*Bacteria<20})'.
These search strings were the result of an extensive search and validation for strings that give - with reasonable sensitivity and specificity - similar results than manually investigation of the trees. Sensitivity and specificity of the used phat string was evaluated

for three organisms (for further information see 3.5) resulting in a sensitivity of 95 % and specificity of 93 %.

2.2 Construction of domain trees

Construction of domain trees for Legionella pneumophila str. Paris

Tree construction was done using the hidden Markov models and multiple sequence alignment for all unique protein domains present in Legionella pneumophila str. Paris provided by the Pfam protein families database [Fukami-Kobayashi *et al.*, 2007, Punta *et al.*, 2012a, Yang & Bourne, 2009]. This and other pre-calculated files are in the 27th PFAM release available at `ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam27.0/` - Pfam-A.full.gz was used to extract the relevant multiple alignments for the domain trees. Also a list with all domains present in the proteome of Legionella pneumophila str. Paris is provided by the Pfam Database.
For every domain a multiple sequence alignment was obtained - many of them exceeding 15,000 sequences and making it impossible to generate phylogenetic trees. therefore similar sequences were clustered based on sequence identity and only one representative was used in order to decrease the total amount of sequences. This was done using Cd-Hit - starting with clustering all sequences with a sequence similarity of above 90% [Li & Godzik, 2006]. All multiple sequence alignments that were still not processable by mtree after this clustering step the sequence similarity was gradually reduced from 80% to 75%, 70%, 60% and finally 50 % [Li & Godzik, 2006].
The phylogenetic trees were constructed using mtrees and analyzed utilizing the PHAT package.

Construction of domain trees for all other analyzed prokaryotes

The starting point for the analysis was the same as above: the pre-calculated full hmm models provided by the Pfam protein families database. The protein alignments were converted from Stockholm to Multiple Fasta Format and for every domain a file with

its multiple sequence alignment was created resulting in 14831 files. These multiple sequence alignments were searched for the occurrence of one of the 70 organisms of interest - if none was present they were excluded from the analysis pool resulting in 6755 multiple sequence alignments for domains of interest. Phylogenetic domain trees were obtained for these multiple sequence alignments - if the number of sequences were below 1500 without further pre-processing, if the number of sequences was above 1500 a Cd-Hit run was performed with 0.8 sequence identity and a word length of 5 and if the number of sequences after this first Cd-Hit run was still above 2000 a second Cd-Hit run was performed using 0.6 sequence identity and a word length of 3. If - after the second Cd-Hit run - the multiple sequence alignment contained below 6000 sequences tree reconstruction was performed, if it contained above 6000 sequences the multiple sequence alignment was skipped. Tree reconstruction was not always successful - mostly due to memory limitations for multiple sequence alignments above 2500 species. This was regarded as tolerable limitation since it can be assumed that the more species are present in a multiple alignment the lesser are the chances for the detection of a HGT in this alignment. Normally, domains that fall in the region of above 6000 species in the multiple alignments are present in a wide range of proteins and house keeping genes - often present in more than one domain of life (e.g. PF00004 - ATPases Associated with diverse cellular Activities).

Chapter 3

Results and Discussion

3.1 Distribution of phyla in the effectors database

The distribution of phyla in the effectors database is not evenly; this is contingent on their cultivability, on the difference in members of bacterial phyla and the fact that different bacteria are of different interest for biology and research in general. It is not possible to analyze all 648 bacteria present in the effectors database (in average a bacteria has between 2000-3000 proteins - analyzing all of them would lead to an unmanageable amount of data but over all the requirements for computational power and time would be unfeasible for the used grid system) - so it was important to take a representative sample from the phyla present in the database.

70 bacteria were randomly choose from a total of 648 bacteria present in the effectors database (as shown in Table 3.1 and in Figure 3.1).

It is clear from the numbers and the figure that mostly Proteobacteria (34), Firmicutes (13), Actinobacteria (10) were analyzed. Fusobacteria (0), Fibrobacteres/Acidobacteria group (0), Synergistetes (0), Cyanobacteria (0) and Elusimicrobia (0) represent only a small part of the pathogens (in sum 9 taxa from these five phyla are present in the effector database) and were excluded from the analysis.

Phylum	Amount of organisms present in effectors database	Amount of organisms analyzed in this thesis
Proteobacteria	352	34
Firmicutes	137	13
Actinobacteria	56	10
Tenericutes	31	4
Chlamydiae/Verrucomicrobia group	25	3
Spirochaetes	21	3
Bacteroidetes/Chlorobi group	17	3
Fusobacteria	4	0
Fibrobacteres/Acidobacteria group	2	0
Synergistetes	1	0
Cyanobacteria	1	0
Elusimicrobia	1	0

Table 3.1: Distribution of phyla in the effector database and analyzed in this thesis

3.2 Parameter screening for the alien index

Phylogenetic tree reconstruction is both time and CPU consuming. Thus a filter criteria to diminish the amount of proteins for which tree reconstruction was necessary - excluding e.g., housekeeping genes and other genes only present in prokaryotes - was applied. The filter criteria was a simple test which determines for every protein of interest the ratio between the best prokaryotic and best eukaryotic homology hit.

$$AI = \frac{\text{bit score of best eukaryotic hit}}{\text{bit score of best prokaryotic hit}} \tag{3.1}$$

The higher the ratio (normally ranging between 0.1 to 1.5) the more likely the protein will cluster with eukaryotic proteins in a phylogenetic tree reconstruction.
This method's advantage are the simplicity of the test, automatability and the small time requirement (which was mostly due to the fact that the SIMAP database was used instead of a BLASTP search) - the disadvantage are the poor sensitivity and specificity making it only usable as a pre-filter. Older HGTs are almost impossible to determine by blast results - therefore using only a bit score ration to determine phylogenetic relation-

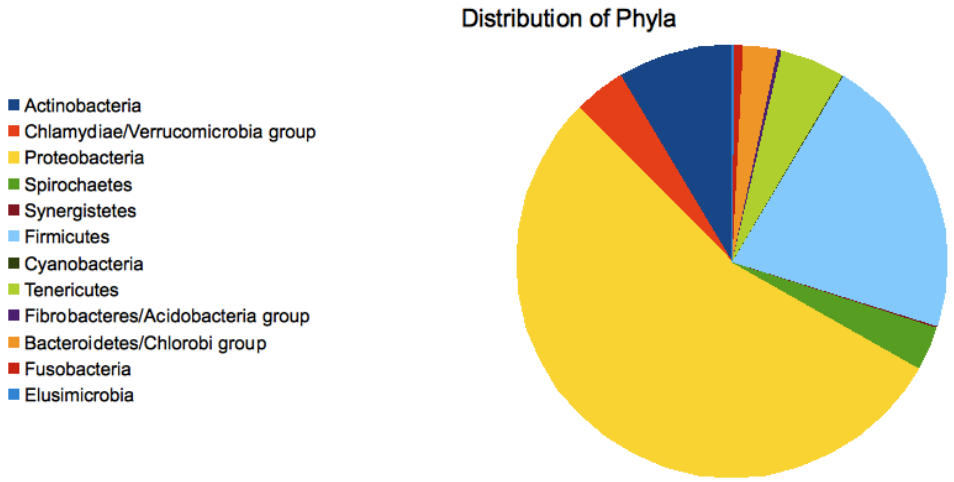

Figure 3.1: Distribution of phyla in the effector database

ships is not enough.

It was important to determine a specific value for the AI that allowed to diminish the amount of proteins for which a phylogenetic tree reconstruction had to be calculated while preventing the loss of potential proteins that underwent horizontal gene transfer. Therefore the workflow as described was performed with changing bit score ratio values for three bacteria. The impact of the AI cut-off was monitored and is shown in the following section.

For three organisms (Legionella pneumophila st. Paris, Coxiella burnetii RSA 493, Chlamydia trachomatis D/UW-3/CX - the first two had shown promising candidates for horizontal gene transfer in former analysis) phylogenetic tree reconstruction was performed for all proteins exceeding a bit score ratio of 0.1. To determine which threshold value should be used the amount of trees which were both positive for having a ELD (the characterization of the effectors database was used) and PHAT (PHAT is a software solution for analyzing phylogenetic relationships in a tree - the program is further discussed and its parameters validated in section 3.4) selected as positive match (e.g., for Legionella pneumophila str. Paris a positive match was defined as finding a node in the tree that had Legionella pneumophila str. Paris and at least five eukaryotes but not

more than 20 other bacterias) were monitored. The biggest loss of proteins for the analysis resulted from the initial, lowest AI-score of 0.1. This is a finding that is consistent through all bacteria that were analyzed and important to note for the interpretation of Table 3.7.

Legionella pneumophila str. Paris

The genome of Legionella pneumophila st. Paris translates to 3166 proteins - after setting the cut-off to 0.1 2428 proteins remain in the analysis (76%). 96 ELDs were characterized for Legionella pneumophila str. Paris in 89 proteins - 82 of them remained in the analysis pool after a bit-score ratio cut-off of 0.1. Details for all proteins with an ELD score greater or equal to 4 and remaining in the analysis pool after applying an AI cut-off of 0.1 are shown in Table 3.2. This analysis was only done for Legionella pneumophila str. Paris. It should be noted that the proteins for which either manual classification (investigation of trees using a tree viewer and manually interpreting the tree) or classification using PHAT (automatized analysis of the phylogenetic relationship of taxa in a tree using the PHAT package) resulted in a positive signal for eukaryotic origin are all present at a bit-score ratio cut-off of 0.5. This finding indicates that there is no bias introduced with the large-scale usage of automatized tree analysis (using the PHAT package) towards a certain bit-score ratio.

Many proteins get eliminated just by applying an AI of 0.1 - this also includes all genes that have no homologs in eukaryotes.

Increasing the bit score ratio had only minor influence on the proteins which were classified as ELDs by PHAT until exceeding an AI score of 0.5 but reducing the quantity of proteins to about 57% of the initial amount. Figure 3.2 and Figure 3.3 show how the figures change when gradually increasing the AI. The category named 'pELD+pPHAT' was the most important - this showed that applying the cut-off did not lead to a loss of positive hits from PHAT in the ELD category. This must be kept in mind as it remains as a potential bias for the χ^2-test. In Table 3.2 this is further described by showing that the manually assigned positive hits in the ELD category do not change until an AI of 0.5

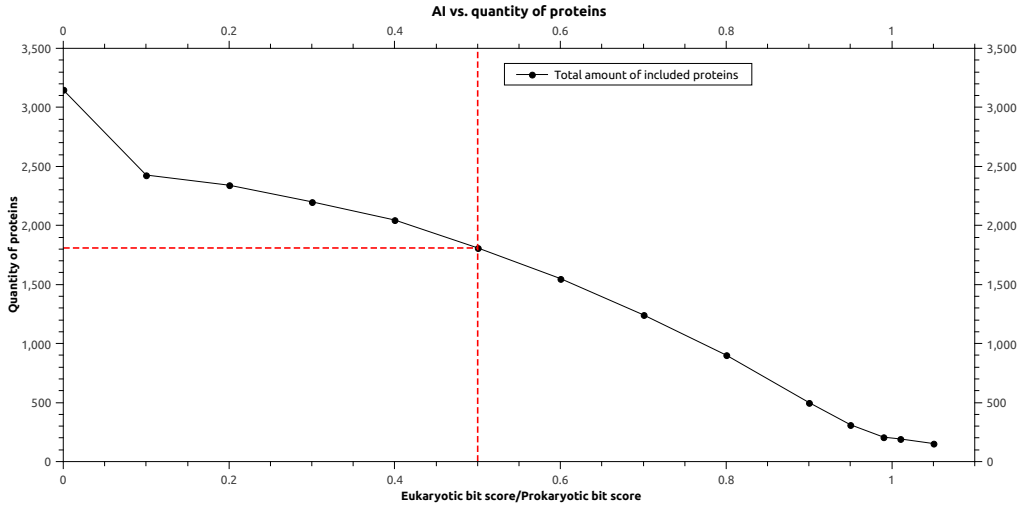

Figure 3.2: Legionella pneumophila: Impact of changing AI filter criteria on total quantity of proteins. The vertical red line indicates x = 0.5, whereas the horizontal red line is a projection on the y-axis.

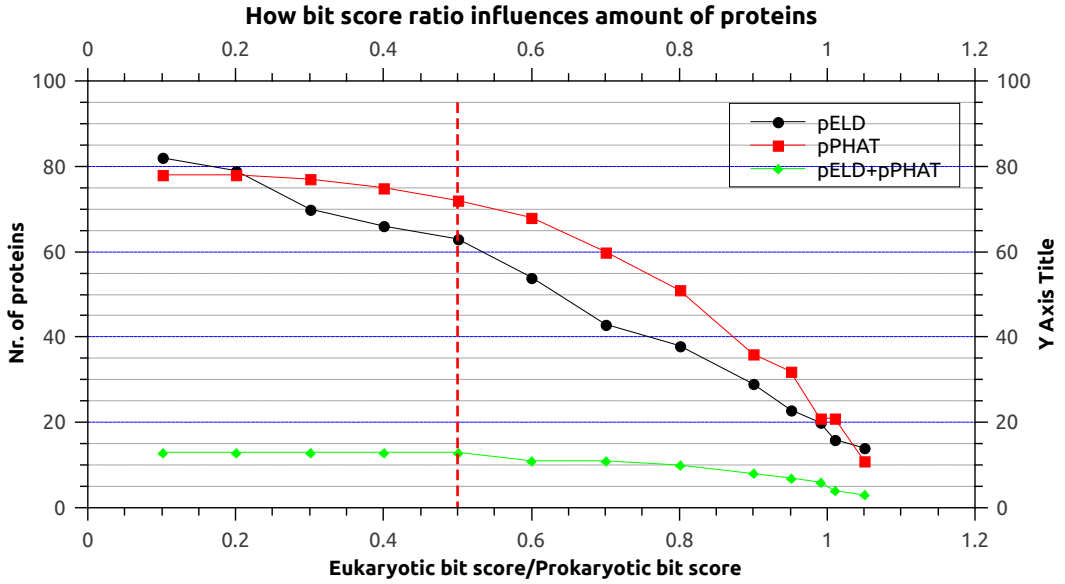

Figure 3.3: Legionella pneumophila: Impact of changing AI filter criteria on subsets of proteins.

is superseded.

ACC-number	ELD-score	AI-score	>0.2	>0.5	>1.0	Manually assigned	L1PHAT	L2PHAT
YP_122247	9	0.63	P	P	-	-	-	-
YP_125351	4	0.58	P	P	-	-	-	-
YP_125158	4	0.89	P	P	-	-	-	-
YP_122671	8	0.69	P	P	-	-	-	-
YP_124367	8	0.69	P	P	-	-	-	-
YP_124197	48	0.82	P	P	-	P	P	P
YP_124239	4	0.27	P	-	-	-	-	-
YP_124636	5	0.5	P	P	-	-	-	-
YP_123283	7	0.5	P	P	-	-	-	-
YP_124764	7	1.18	P	P	P	-	-	-
YP_124830	5	0.62	P	P	-	-	-	-
YP_123814	6	0.19	n	-	-	-	-	-
YP_124394	10000	0.98	P	P	-	-	-	
YP_124559	5	0.19	n	n	-	-	-	-
YP_123088	5	0.98	P	P	-	P	P	P
YP_122224	6	0.28	P	n	-	-	-	-
YP_123444	4	0.68	P	P	-	P	P	P
YP_122870	4	0.6	P	P	-	-	-	-
YP_123607	7	1.38	P	P	P	-	-	-
YP_123468	8	0.99	P	P	-	-	-	-
YP_124726	10	0.95	P	P	-	P	P	P
YP_123778	4	0.95	P	P	-	-	-	-
YP_122547	5	0.92	P	P	-	-	-	-
YP_124246	10000	0.51	P	P	-	P	P	P
YP_123903	6	0.28	P	-	-	-	-	-
YP_124079	10000	1.17	P	P	P	-	-	-
YP_122807	5	0.99	P	P	-	P	P	P
YP_124228	7	0.99	P	P	-	-	-	-
YP_124377	5	1.13	P	P	P	-	-	-
YP_122389	5	1.11	P	P	P	-	-	-
YP_122460	5	0.57	P	P	-	P	P	P
YP_125327	14	1.27	P	P	P	P	P	P

ACC-number	ELD-score	AI-score	>0.2	>0.5	>1.0	Manually assigned	L1PHAT	L2PHAT
YP_125022	5	0.36	P	-	-	-	-	-
YP_124275	10	0.85	P	P	-	-	-	-
YP_125189	10000	1.16	P	P	P	-	-	-
YP_122366	4	0.2	P	-	-	-	-	-
YP_122841	5	1.12	P	P	P	-	-	-
YP_123514	4	0.87	P	P	-	-	-	-
YP_123956	4	0.67	P	P	-	-	-	-
YP_122371	4	0.58	P	P	-	-	-	-
YP_123944	4	0.58	P	P	-	-	-	-
YP_123539	5	0.38	P	-	-	-	-	-
YP_124370	5	1.04	P	P	P	P	P	P
YP_124506	5	0.88	P	P	-	-	-	-
YP_123342	4	0.49	P	-	-	-	-	-
YP_125156	6	0.31	P	-	-	-	-	-
YP_123335	7	0.81	P	P	-	P	P	P
YP_122161	6	0.25	P	-	-	-	-	-
YP_124576	4	0.2	P	-	-	-	-	-
YP_123705	4	0.84	P	P	-	-	-	-
YP_124373	5	1.06	P	P	P	P	P	P
YP_122690	10000	1.09	P	P	P	-	-	-
YP_123361	48	0.74	P	P	-	P	P	P
YP_125198	8	0.69	P	P	-	-	-	-
YP_124257	10	0.83	P	P	-	-	-	-
YP_123657	6	0.41	P	-	-	-	-	-
YP_124254	7	1.15	P	P	P	P	P	P
YP_124602	4	0.6	P	P	-	-	-	-
YP_122708	4	0.16	-	-	-	-	-	-
YP_124507	4	0.88	P	P	-	-	-	-
YP_124945	4	0.73	P	P	-	-	-	-
YP_124784	4	0.78	P	P	-	-	-	-
YP_125190	4	0.75	P	P	-	-	-	P
YP_124175	7	1.07	P	P	P	-	-	-
YP_122223	9	0.27	P	-	-	-	-	-
YP_124001	5	1.11	P	P	P	-	-	-

ACC-number	ELD-score	AI-score	>0.2	>0.5	>1.0	Manually assigned	L1PHAT	L2PHAT
YP_124240	8	0.43	P	-	-	-	-	-
YP_124751	4	0.69	P	P	-	-	-	-
YP_125196	4	0.91	P	P	-	-	-	-
YP_124822	5	0.96	P	P	-	-	-	-
YP_125160	4	0.92	P	P	-	-	-	-
YP_123559	4	0.72	P	P	-	-	-	-
YP_122579	8	0.31	P	-	-	-	-	-
YP_123416	9	0.61	P	P	-	-	-	-
YP_124460	6	0.21	P	-	-	-	-	-
YP_124388	5	1.12	P	P	P	-	-	-
YP_125003	8	0.68	P	P	-	-	-	-
YP_122719	10000	1	P	P	P	P	P	P
YP_123064	4	0.26	P	-	-	-	-	-
YP_124222	5	1	P	P	-	P	P	P
YP_123922	8	0.55	P	P	-	-	-	-
YP_124376	4	0.9	P	P	-	-	-	-
Sum			79	63	16	16	16	16

Table 3.2: The ACC numbers for all proteins which have an ELD score greater or equal to 4 and remained in the analysis pool after the cut-off of 0.1 for Legionella pneumophila str. Paris are shown. The minus sign and P correspond to either a negative or positive signal. Bit score ratios were tested starting with an AI of 0.1 and finishing with 1.01 - not all steps are shown. The numbers in the last row are the sums of the positive signals for the corresponding column. In red are the acc numbers of proteins which were of eukaryotic origin based on manually investigation shown. L2PHAT and L1PHAT (Level 2 and Level 1 PHAT constraints) means the signals for HGTs detected by PHAT - Level 2 stands for the constraints used for the final results and Level 1 stands for a loose set of the rules normally resulting in more signals but less precision of the results (for a discussion of these constrains see section 3.4).

Coxiella burnetii RSA 493

Coxiella burnetii RSA 493 has 1847 proteins in its proteom - an applied bit-score ratio cut-off of 0.1 excludes 571 proteins (69% of proteins remained in the analysis). The overall impression of the influence of the cut-off on the protein quantity is similar to what was observed with the Legionella pneumophila proteome. Increasing the bit score ratio had only minor influence on the proteins which were classified being of eukaryotic origin by PHAT until exceeding an AI score of 0.5 but about 46% of all proteins were excluded thus considerably diminishing the amount of necessary tree reconstructions and significantly reducing the computational time for the analysis. Figure 3.4 and Figure 3.5 show how the figures change when gradually increasing the AI. What should be noted is that the an AI of 0.1 lead from 26 proteins containing ELDs to 22 ELDs (84 % of proteins with ELDs remained in the analysis pool).

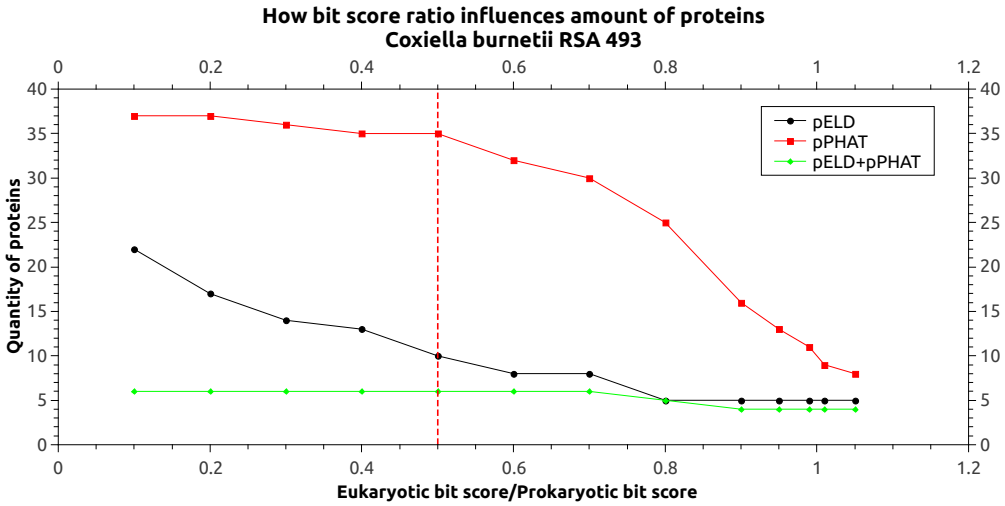

Figure 3.4: Coxiella burnetii: Impact of changing AI filter criteria on different subsets of proteins.

Chlamydia trachomatis D/UW-3/CX

The Chlamydia trachomatis D/UW-3/CX proteome contains 895 proteins. A bit-score ratio cut-off of 0.1 led to the exclusion of 205 proteins (76 % of proteins remained in the

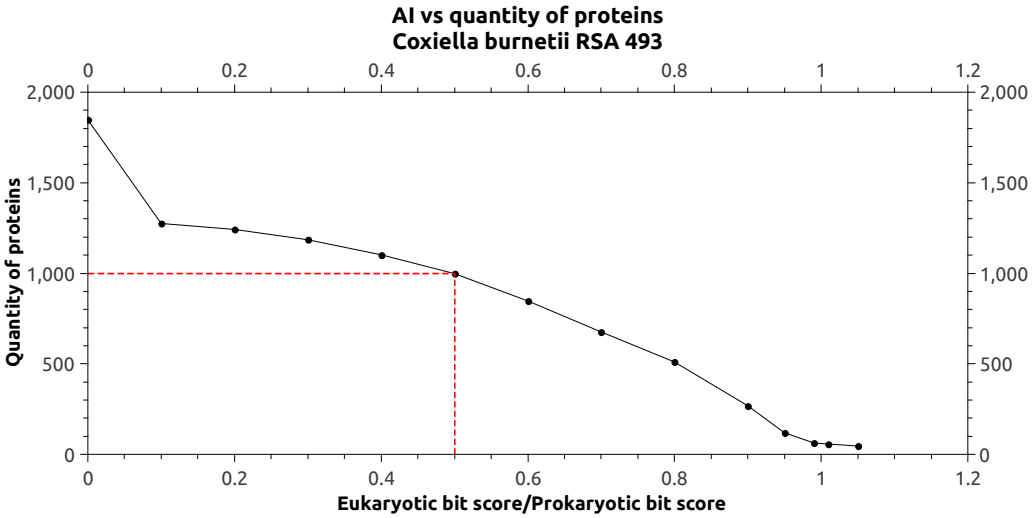

Figure 3.5: Coxiella burnetii: Impact of changing AI filter criteria on total amount of proteins.

analysis).

Increasing the bit score ratio had only minor influence on the proteins which were classified as of eukaryotic origin by PHAT until exceeding an AI score of 0.5 - but the total number of proteins decreased by about 43%. The number of proteins which were categorized as containing an ELD and having a positive signal for eukaryotic origin based on the PHAT analysis did not change until an AI score ratio cut-off of 0.7 was superseded. Figure 3.6 and Figure 3.7 show how the figures change when gradually increasing the AI. This reduction was essential since for Chlamydia trachomatis D/UW-3/CX phylogenetic tree reconstruction was not only done using a NJ method but also a maximum-likelihood tree construction method - the time consumption of more trees in the analysis pool would have rendered this unfeasible with the available means.

In conclusion the assumption that an alien index of above 0.5 diminishes the amount of proteins significantly and on the other hand keeps potential HGTs in the pool for further analysis seems valid. As noted before the decrease of proteins which were classified as having a ELD in their sequence was mainly due to the introduction of a cut-off of 0.1.

The step from no cut-off to a cut-off for proteins below an alien index of 0.1 is responsible for the major loss in ELD containing proteins.

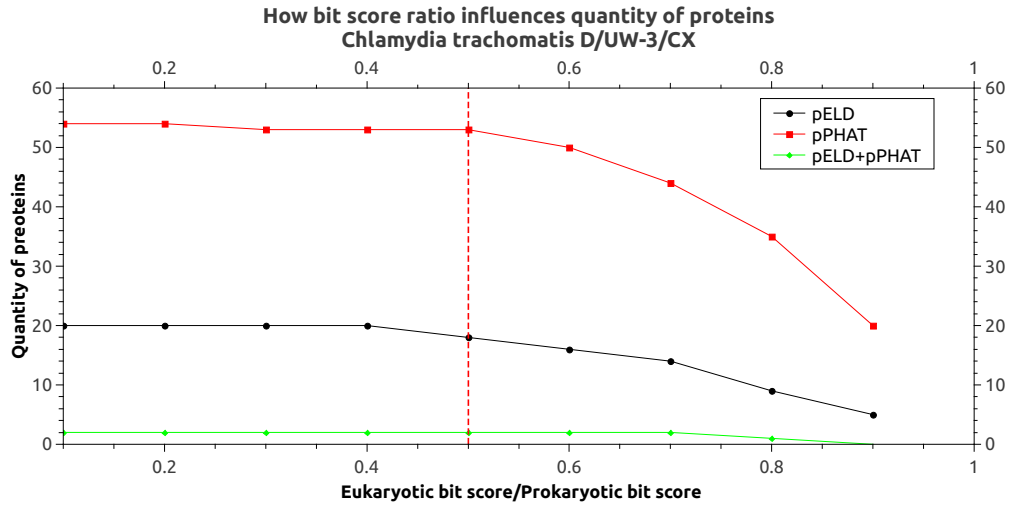

Figure 3.6: Chlamydia trachomatis D/UW-3/CX: Impact of changing AI filter criteria on different subsets of proteins.

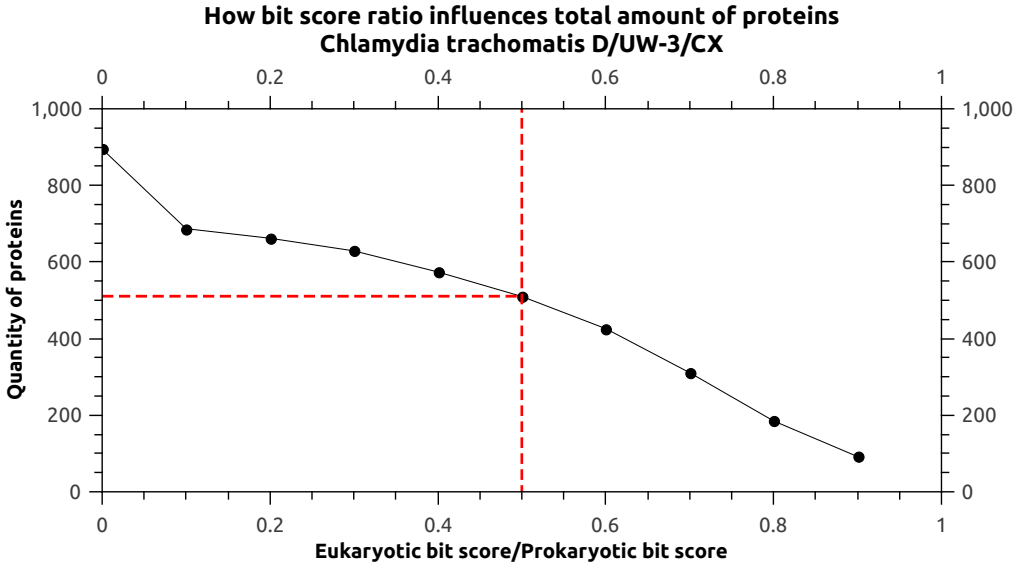

Figure 3.7: Chlamydia trachomatis D/UW-3/CX: Impact of changing AI filter criteria on total amount of proteins.

3.3 Validation of the tree reconstruction and tree analysis parameters

In previous studies the analysis of L. pneumophila genomes showed the presence of an unexpected high number and variety of eukaryotic-like proteins (ELPs) and proteins containing eukaryotic-like domains - possibly reflecting its co-evolution with fresh-water amoebae [Cazalet *et al.*, 2010]. To validate the quality of the tree reconstruction the obtained results were compared to previously conducted analysis of the Legionella pneumophila str. Paris genome done by [Lurie-Weinberger *et al.*, 2010] and [Gomez-Valero *et al.*, 2011b]. In these studies proteins with ELDs and eukaryotic-like proteins were identified and their phylogenetic origin based on manual phylogenetic tree analysis affirmed.

In the study done by Lurie-Weinberger et al. only eukaryotic-like proteins (ELPs) were studied. Using the normalized BLASTP bit score proteins were compared against bacteria and eukarya protein datasets - if the BLASTP bit score of the protein of interest against the best eukaryotic hit had a greater bit score than the BLASTP bit score of the protein of interest against the best bacterial hit and the difference superseded 0.05 the protein is regarded as eukaryotic-like protein and phylogenetic tree analysis was performed. The results of the study by Lurie-Weinberger et al. were compared to the results obtained in this thesis - a summary is shown in Table 3.3.

The classification as ELP is similar to a positive AI signal - although the constraints for a positive signal in this study for the AI are milder (AI score ratio of above 0.5 was regarded as positive signal whereas for ELPs the AI score ratio need to be above 1.00). As a result every ELP should have a positive signal in the AI - with the exception of lpp2146 (YP_124458) and lpp2248 (YP_124559) this is true. These two proteins were further analyzed and instead of the SIMAP database a BLASTP search against the nr NCBI database was performed (even though normally these two methods are equivalent) which resulted also in an AI of below 0.5. What should be noted is that proteins like lpp0037, which was previously described as having only eukaryotic homologs, had, besides a high

enough AI score to pass through the ratio threshold, only Legionella specific hits in the multiple alignment, resulting in a Legionella-only phylogenetic tree. Using another approach to evaluate this protein a BLASTP search was performed on the NCBI homepage and the COBALT (Constraint-based Multiple Alignment Tool) multiple alignment tool was used to reconstruct a phylogenetic tree using a neighbor joining tree reconstruction method [Papadopoulos & Agarwala, 2007]. The tree is shown in Figure 3.12.

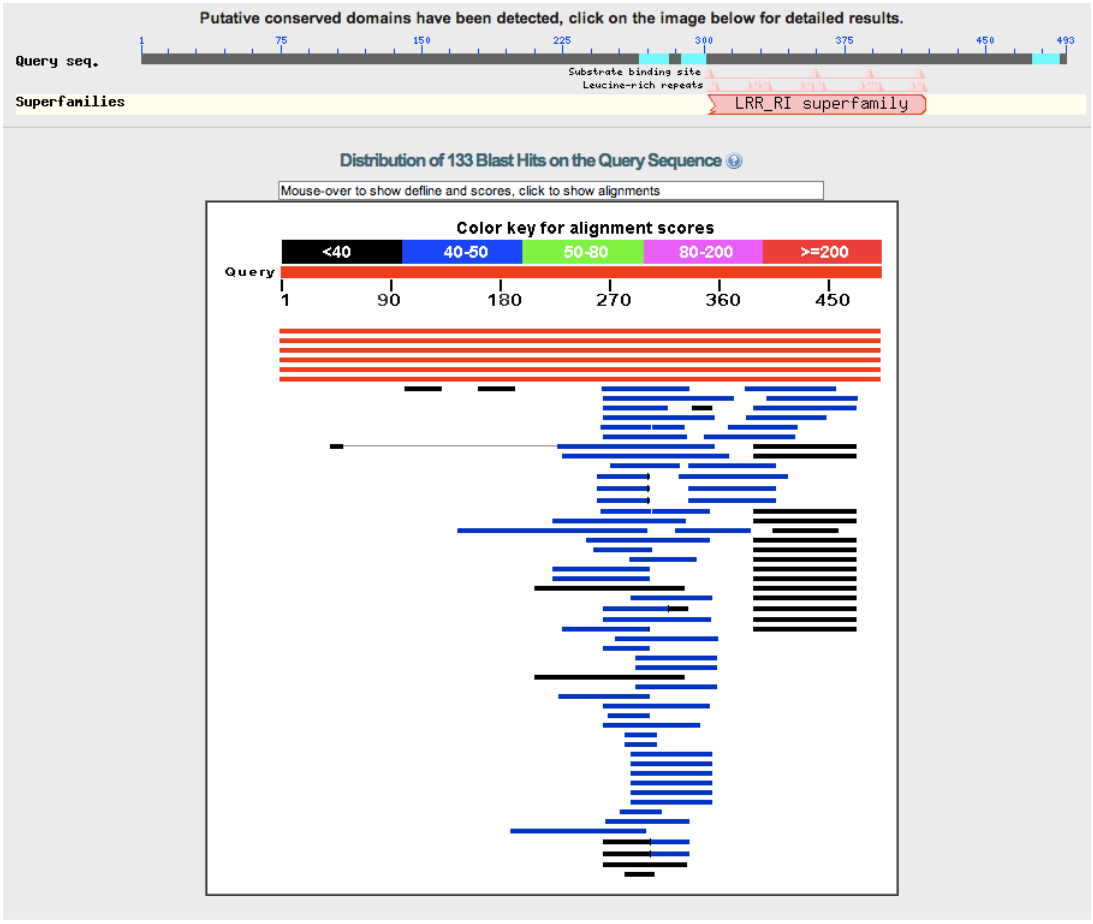

Figure 3.8: Multiple sequence alignment of YP_122389 done by COBALT.

A closer look at the blast result (shown in Figure 3.11) showed that there are only Legionella pneumophila specific hits with high alignment scores. This protein produces no significant alignments for non-legionella proteins and therefore no signal for HGTs in the workflow used in this thesis but were characterized as of eukaryotic origin by

Lurie-Weinberger et al. in 2010. The differences in results may be a consequence of different methods and parameters for multiple sequence alignments. In this thesis multiple sequence alignment was achieved using the provided PhyloGenie routine which uses the BLASTP results and Blammer to generate a hidden Markov model (HMM) of the most dissimilar high-scoring segment pairs (HSPs) to search the sequence database and generating a multiple alignment. In contrast COBALT and also MUSCLE (as used by Lurie-Weinberger et al.) utilizing protein domains for better performance. Very often (for example YP_124001) showed only for Legionella homologs a coverage above 30 %. All other bacteria or eukaryotes had coverage below 30 %, yet the used parameters allowed blammer only to accept proteins for the HMM calculation with at least 60 % sequence coverage.

In the shown example (YP_122389) the blammer result agrees with the COBALT result - but e.g., for YP_123949 they disagree. The sequence coverage is still not enough for blammer to construct a multiple sequence alignment for proteins other than Legionella - yet COBALT can. The multiple sequence alignment and the calculated tree are shown in Figure 3.9 and Figure 3.10.

28 proteins were described as of eukaryotic origin, the approach used in this thesis could confirm the claim for 15 of these proteins. Of the remaining 13 four were - after manually investigation - re-evaluated and considered as of eukaryotic origin. Of the 9 remaining proteins 2 were confirmed as potential HGTs using exclusively the COBALT workflow on the NCBI homepage but the multiple alignment and tree reconstruction provided by the PhyloGenie package did not yield enough homologs for tree reconstruction due to the sequence coverage restriction. The origin of the other six proteins could either not be confirmed or were of bacterial descent.

Effective - the bacterial effector database - showed 96 ELD containing proteins, with a minimal ELD score of 4. Gomez-Valero et al. identified 42 proteins with ELDs, 37 of them highly conserved in other L. pneumophila strains [Gomez-Valero *et al.*, 2011b]. Of this 42 proteins only 18 were also present in the selection of the proposed ELD containing

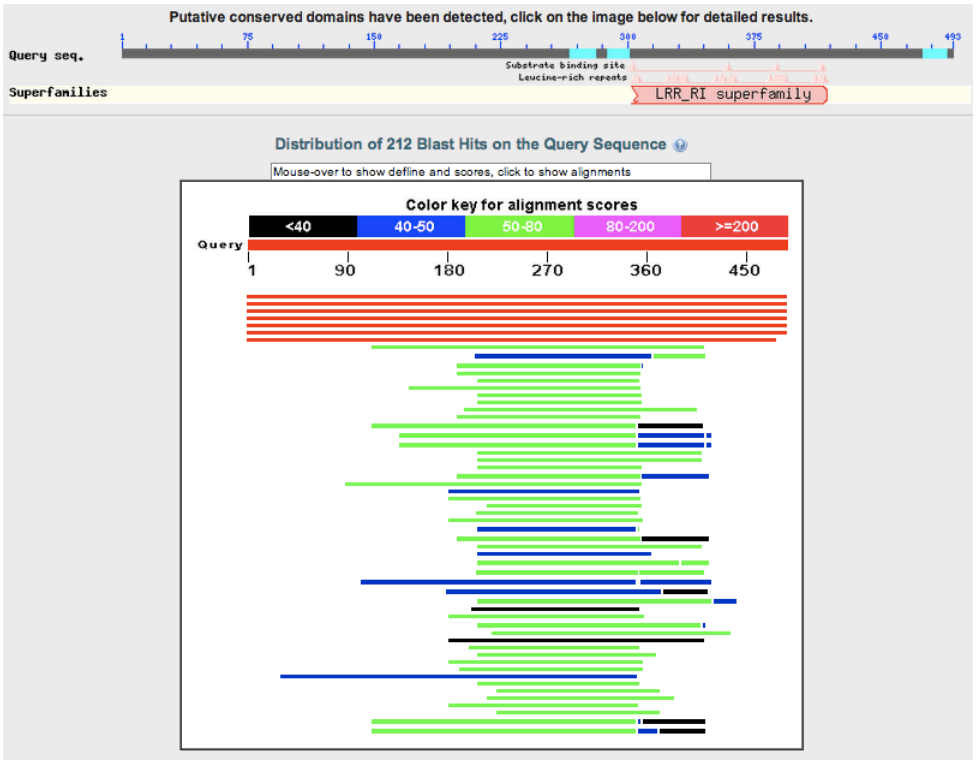

Figure 3.9: Blast results of YP_123949.

proteins provided by the effectors database and of this 18 proteins 6 resulted in a positive signal after testing for HGT with phylogenetic tree reconstruction. All 6 proteins were also analyzed by Lurie-Weinberger et al., 4 of them were regarded as of eukaryotic origin. It seems like there is little agreement on the method to evaluate eukaryotic domains.

In conclusion it seems that the phylogenetic tree reconstruction methods lead to similar results for the exception of homologs with little coverage. Representative for the multiple alignment made with COBALT of as Legionella-only categorized trees the multiple alignment of YP_122389 is shown in Figure 3.11. Coverage was very often below 30 % and E-values above 0.05. The alignment was built around the putative conserved domain but with little alignment in other regions.

To sum up it seems that of the 28 proteins with ELDs with proposed eukaryotic origin 15 were found in with the normal workflow applied in this thesis. Of the 13 proteins

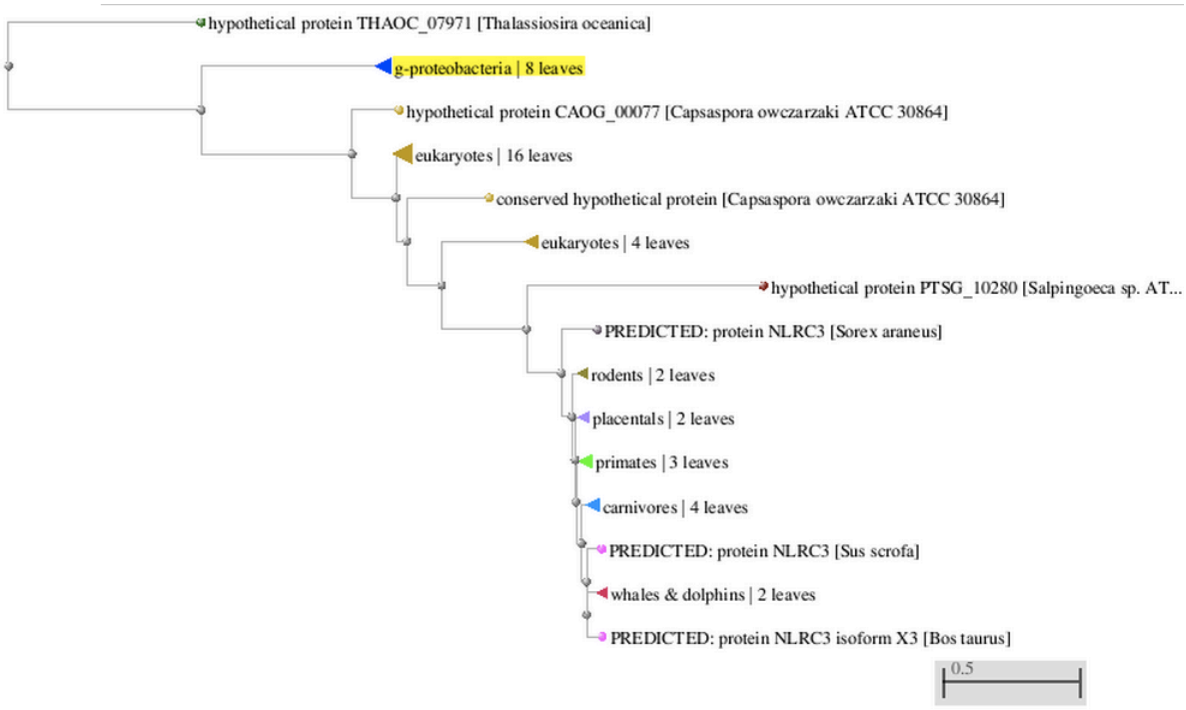

Figure 3.10: Neighbor joining tree calculated based on the multiple sequence alignment for YP_123949 done by COBALT.

that were not categorized as potential HGTs from eukaryotes four were - after manual re-evaluation - categorized as positive for HGTs from eukaryotes - the analysis of the other nine remained inconclusive and are therefore not regarded as positive signals.

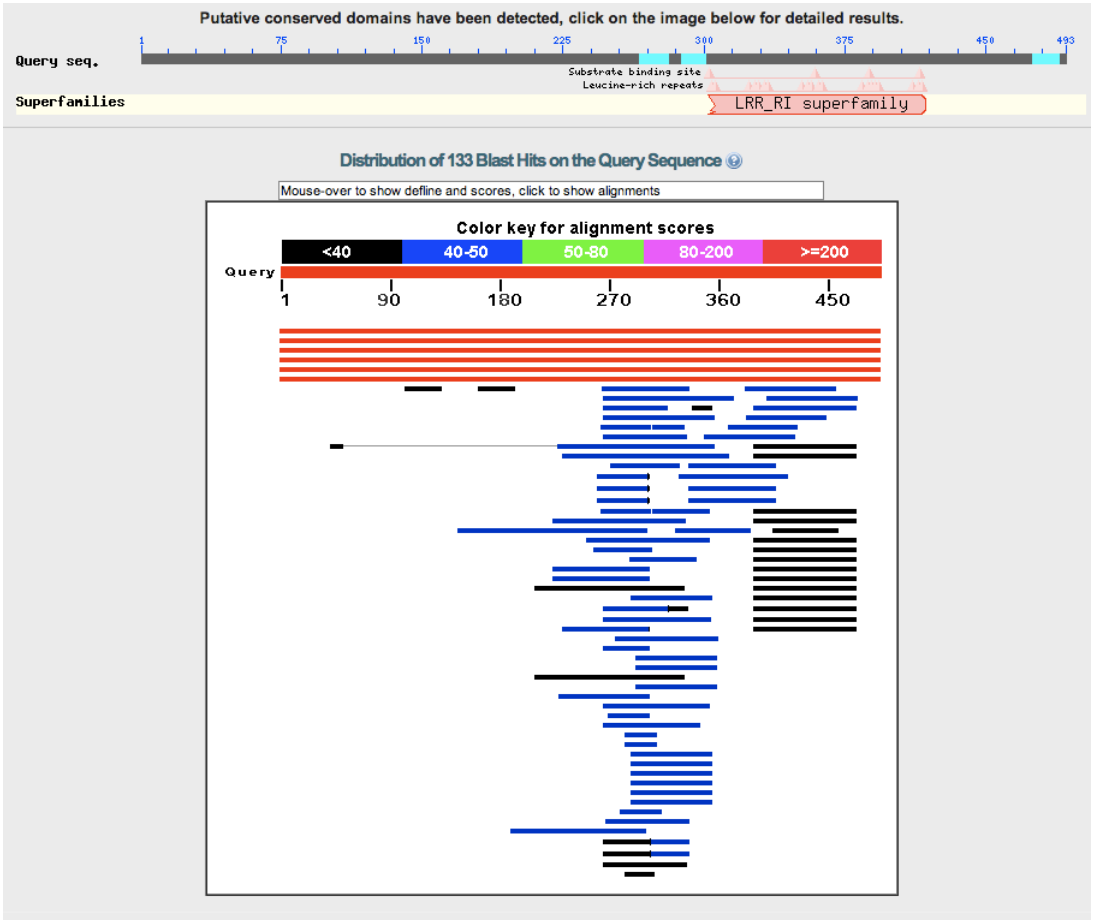

Figure 3.11: Multiple sequence alignment of YP_122389 done by COBALT.

ACC Number	Gene name	AI	L2PHAT	L1PHAT	M	C	Bootstrap support	
							this thesis	Lurie-Weinberger
YP_124254	lpp1940	+	+	+	+	/	Low	Low
YP_122719	lpp0379	+	+	+	+	/	LE	LE
YP_124222	lpp1905	+	+	+	+	/	Low	Low
YP_124246	lpp1932	+	+	+	+	/	LE	LE
YP_124370	lpp2058	+	+	+	+	/	Low	Low
YP_122205	lpp0050	+	+	+	+	/	Low	Low
YP_123763	lpp1439	+	+	+	+	/	LE	Medium
YP_123846	lpp1522	+	+	+	+	/	High (100)	High (100)
YP_123965	lpp1647	+	+	+	+	/	Low	High (82)
YP_124273	lpp1959	+	+	+	+	/	Low10	Low
YP_124440	lpp2128	+	+	+	+	/	Low	Low
YP_124446	lpp2134	+	+	+	+	/	Low	Low
YP_124773	lpp2468	+	+	+	+	/	Low	Low
YP_125324	lpp3022	+	+	+	+	/	Low	High (100)
YP_122661	lpp0321	+	+	+	+	/	Low	Low
YP_122389	lpp0037	+	-	-	-	-	LO	LE
YP_122807	lpp0469	+	-	-	+	+	High (83)	Low
YP_123481	lpp1157	+	-	+	+	+	Medium (63)	High (100)
YP_123949	lpp1631	+	-	-	-	+	LO	High (82)
YP_124001	lpp1683	+	-	-	-	+	LO	LE
YP_124173	lpp1855	+	-	-	-	-	LO	LE
YP_124377	lpp2065	+	-	-	-	-	LO	LE
YP_124458	lpp2146	-	/	/	/	-	/	LE
YP_124559	lpp2248	-	/	/	/	-	/	Low
YP_124764	lpp2459	+	-	-	-	-	Low	LE
YP_124931	lpp2626	+	-	+	+	+	Low	Low
YP_125053	lpp2748	+	-	+	?	?	?	Low
YP_125082	lpp2777	+	-	-	+	+	LE	LE

Table 3.3: Evaluation of the quality of the phylogenetic tree reconstruction. Shown are all proteins which were characterized as ELP and of eukaryotic origin in the study by Lurie-Weinberger et al. in 2010. The approach used in this thesis showed for the 28 ELP only 15 of eukaryotic origin. Explanation of abbreviations: p=positive; n=negative; LE represents genes with only eukaryotic homologs; NEH stands for not enough homologs for tree analysis; low bootstrap support is below 50; LO is used to mark genes which had only homologs in Legionella; AI=Alien Index; L2PHAT and L1PHAT (Level 2 and Level 1 PHAT constraints) means the signals for HGTs detected by PHAT - Level 2 stands for the constraints used for the final results and Level 1 stands for a loose set of the rules normally resulting in more signals but less precision of the results, M=manually checked to determine phylogenetic origin, C=workflow consisting of the NCBI BLASTP and COBALT multiple alignment and tree reconstruction. +/- means either positive or negative signal; / stands for not done; ? means that it was not possible to unambiguous interpret a tree reconstruction.

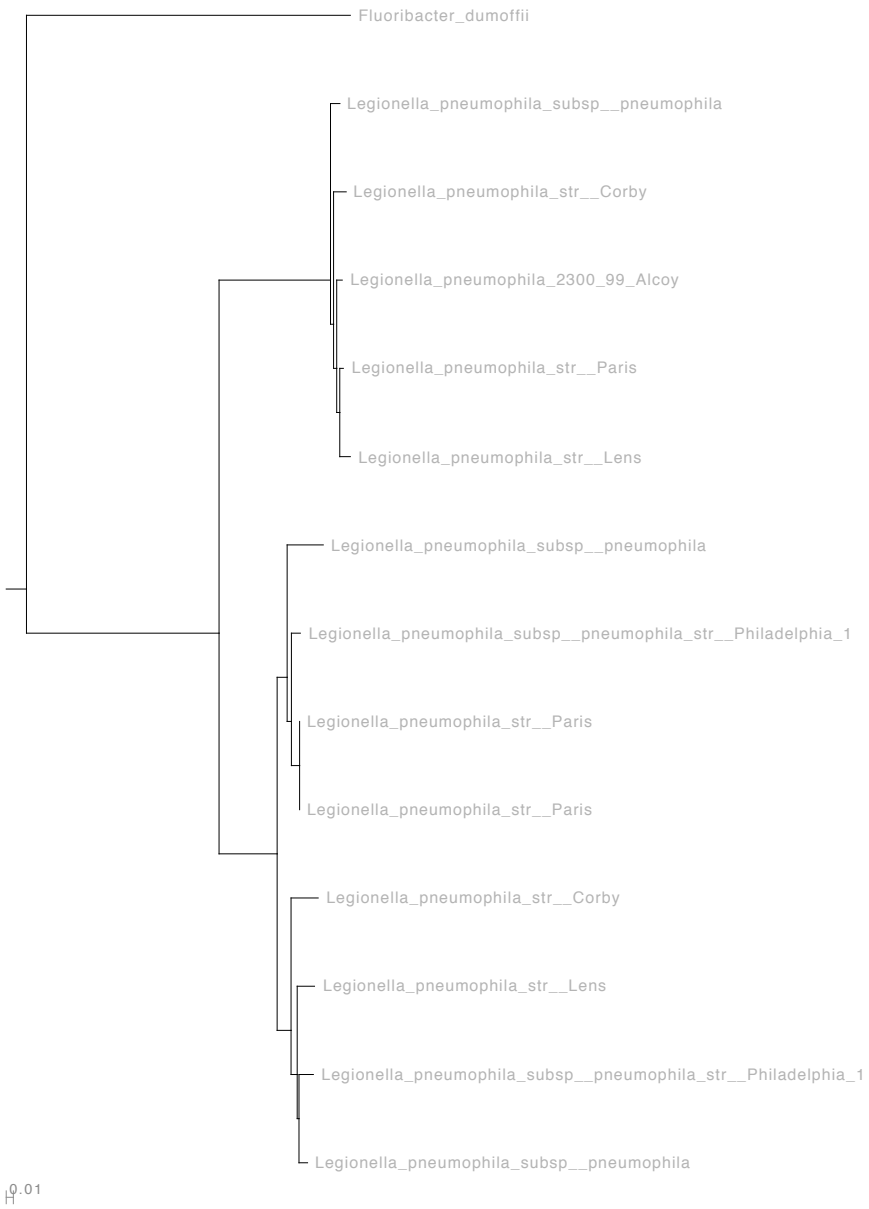

Figure 3.12: Phylogenetic tree reconstruction using COBALT for a protein of Legionella pneumophila Str. Paris - YP_122389.1. The parameters used for the multiple alignment with the built-in routine of the PhyloGenie package did result in a Legionella-only alignment and tree.

3.4 Discussion of the used PHAT search strings

Three different search strategies were applied to analyze the relationship between eukaryotes and the seed organism in the phylogenetic trees:

1. '(Taxonomic name of seed prokaryote & Eukaryota)' - for the taxonomic name of seed prokaryote the species name of the prokaryote was inserted. This string was positive if there was a eukaryote present in the tree reconstruction. It was necessary to investigate this in addition to the AI score because of the differences in multiple alignment and pairwise alignment.

2. '(Taxonomic name of seed prokaryote & {Eukaryota>3})' - taxonomic name of seed prokaryote corresponds to the species name of the prokaryote. This string was positive if there was a node with more than 3 different species (neither paralogs nor homologs in the same species but different strain were counted) of eukaryotes and the seed prokaryote. Since there is no root the tree is analyzed in relation to the seed species name - this means that this search string was positive if there were more than five different eukaryotic species present.

3. '(Taxonomic name of seed prokaryote & {Eukaryota>3} {*Bacteria<20})' taxonomic name of seed prokaryote corresponds to the species name of the prokaryote. This string was positive if there was a node with more than 3 different species of eukaryotes, the seed prokaryote and less than 20 other prokaryotic species.

The search strings results in a positive signal if any node in the tree is found that correspond to the statement.
The syntax of the PHAT tree for the used elements is as following: brackets are used to describe a node, curved brackets are used to describe a quantitative statement, ampersand is used as AND-junction, the exclamation mark is used as a NOT-statement and the asterisk before bacteria means count only bacteria OHTER than the seed prokaryote. The search term in (3) was developed using a trial and error strategy. For Legionella pneumophila str. Paris, Wolbachia endosymbiont of Culex quinquefasciatus Pel and Francisella tularensis subsp. tularensis NE061598 all proteins having an AI greater or

equal than 0.5 and a ELD score greater or equal than 4 were manually categorized in two groups: proteins containing ELDs and having an eukaryotic origin or proteins containing ELDs and are not of eukaryotic descent. Afterwards different search terms were applied with the aim to reproduce the manual categorization using PHAT. Also different phylogenetic levels (e.g., instead of using the seed species using the seed genus or family) were sampled with the same aim. For Legionella pneumophila str. Paris the following search terms were sampled:

1. (Legionella pneumophila str. Paris & {Eukaryota>3} & {*Bacteria<20})
2. (Legionella pneumophila str. Paris & Eukaryota{>3} & {*Bacteria<20})
3. (Legionella & Eukaryota{>3} & {*Bacteria<20})
4. (Legionella pneumophila str. Paris & Eukaryota{>3} & !(*Bacteria))
5. (Legionella & Eukaryota{>3} & !(*Bacteria))
6. (Legionella & Eukaryota{>3} & (!(!Legionella)))
7. (Legionella & {Eukaryota>3} & (!(!Legionellaceae)))
8. (Legionella pneumophila str. Paris & {Eukaryota>3} & !(*cellular organism))
9. (Legionella pneumophila str. Paris & {Eukaryota>3} & !(*Legionellaceae))
10. (Legionella pneumophila & {Eukaryota>3} !(*Legionellaceae))
11. (Legionella & {Eukaryota>3} & !(*Legionellaceae))
12. (Legionella pneumophila str. Paris & {Eukaryota>3} & {*Legionellaceae<2})

In the following table Table 3.4 all proteins with ELD scores greater or equal than 4 and a bit-score ratio greater or equal than 0.5 are shown. 12 different search strategies were applied - out of shortage of space in the table the different sampling terms were called P1-P12 - the number corresponds to the different PHAT strings shown above.

ACC number	M	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
YP_122247	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_122371	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_122389	x	-	-	-	-	-	-	-	-	-	-	-	-

ACC number	M	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
YP_122460	P	P	P	P	-	-	P	P	-	-	-	-	-
YP_122547	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_122671	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_122690	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_122719	P	P	P	P	P	P	P	P	-	P	P	P	P
YP_122807	P	-	P	P	-	-	P	-	-	-	-	-	-
YP_122841	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_122870	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_123088	P	P	P	P	-	-	P	P	-	-	-	-	P
YP_123283	?	-	-	-	-	-	P	P	-	-	-	-	-
YP_123335	P	P	P	P	-	-	P	P	-	P	P	P	P
YP_123361	P	P	P	P	-	-	P	P	-	-	-	-	-
YP_123416	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_123444	P	P	P	P	-	-	P	P	-	-	-	-	-
YP_123468	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_123514	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_123559	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_123607	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_123705	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_123778	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_123922	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_123944	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_123956	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124001	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124079	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124175	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124197	P	P	P	P	-	-	P	P	-	-	-	-	-
YP_124222	P	P	P	P	-	-	P	P	-	P	P	P	P
YP_124228	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124246	P	P	P	P	-	-	P	P	-	-	P	P	-
YP_124254	P	P	P	P	-	P	P	P	-	-	P	P	-
YP_124257	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124275	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124367	x	-	-	-	-	-	-	-	-	-	-	-	-

ACC number	M	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
YP_124370	P	P	P	P	-	P	P	P	-	P	P	P	P
YP_124373	P	P	P	P	-	P	P	P	-	P	P	P	-
YP_124376	x(?)	-	-	-	-	-	P	P	-	-	-	-	-
YP_124377	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124388	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124394	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124506	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_124507	x(?)	-	-	-	-	-	P	P	-	-	-	-	-
YP_124559	x(?)	-	-	-	-	-	-	-	-	-	-	-	-
YP_124602	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_124636	x	-	-	-	-	-	P	P	-	P	P	P	P
YP_124726	P	P	P	P	-	-	P	P	-	P	P	P	P
YP_124751	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_124764	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124784	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124822	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124830	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_124945	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_125003	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_125158	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_125160	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_125189	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_125190	x	-	P	P	-	-	P	-	-	-	-	-	-
YP_125196	x	-	-	-	-	-	P	P	-	-	-	-	-
YP_125198	x	-	-	-	-	-	-	-	-	-	-	-	-
YP_125327	P	P	P	P	-	-	P	P	-	-	-	-	-
YP_125351	x	-	-	-	-	-	P	P	-	-	-	-	-
Sum	15	14	16	16	1	4	32	30	0	7	9	9	7

Table 3.4: Tree analysis of all proteins of Legionella pneumophila str. Paris with ELD score equal to or greater than 4 and a bit-score ratio of equal to or greater than 0.5 using different PHAT search strings. The ACC numbers are given; M stands for manually evaluated; P stands for positive for eukaryotic origin whereas '-' indicates a negative signal for eukaryotic origin.

For Wolbachia endosymbiont of Culex quinquefasciatus Pel and for Francisella tularensis subsp. tularensis NE061598 the following search terms were used in the PHAT analysis (the search strings are only shown for Wolbachia as seed species):

1. (Wolbachia endosymbiont of Culex quinquefasciatus Pel & {Eukaryota>3} & {*Bacteria<20})
2. (Wolbachia endosymbiont of Culex quinquefasciatus Pel & Eukaryota{>3} & {*Bacteria<20})
3. (Wolfbachia & Eukaryota{>3} & {*Bacteria<20})
4. (Anaplasmataceae & Eukaryota{>3} & {*Bacteria<20})
5. Wolbachia endosymbiont of Culex quinquefasciatus Pel & Eukaryota{>3} & !(*Bacteria)
6. Wolbachia & Eukaryota{>3} & !(*Bacteria)
7. Wolbachia & {Eukaryota>3} & !(*Bacteria)
8. Anaplasmataceae & Eukaryota{>3} & !(*Bacteria)
9. (Anaplasmataceae & {Eukaryota>3} & !(*Bacteria))
10. (Wolbachia & Eukaryota{>3} & !(!Wolbachia))
11. (Wolbachia & {Eukaryota>3} & !(!Wolbachia))
12. (Anaplasmataceae & Eukaryota{>3} & !(!Anaplasmataceae))
13. (Wolbachia endosymbiont of Culex quinquefasciatus Pel & {Eukaryota>3} & !(*cellular organism))
14. (Wolbachia endosymbiont of Culex quinquefasciatus Pel & Eukaryota{>3} & !(*cellular organism))
15. (Wolbachia endosymbiont of Culex quinquefasciatus Pel & {Eukaryota>3} & !(*Anaplasmataceae))
16. (Wolbachia & {Eukaryota>3} & !(*Anaplasmataceae))
17. (Anaplasmataceae & {Eukaryota>3} & !(*Anaplasmataceae))
18. (Wolbachia endosymbiont of Culex quinquefasciatus Pel & {Eukaryota>3} & {*Anaplasmataceae<2})

In table Table A.1 all proteins from Wolbachia and Francisella with ELD scores equal to or greater than 4 and a bit-score ratio of equal to or greater to 0.5 are shown. The 18 different search strategies were applied - out of shortage of space in the table the different sampling terms were called P1-P18 - the number corresponds to the different PHAT strings shown above.

In the final results presented in Table 3.8 to Table 3.10 and in Table D.1 protein trees in the category pELD+pPHAT were pre-filtered with PHAT and the positive PHAT signals were manually verified.

What should be noted is that the first three search strings for Legionella pneumophila as well as for Wolbachia and Francisella tularensis were - for the most part - in accordance to the manual classification of the tree reconstruction. Even though generic terms like 'under 20 bacteria' seem to be random and hard to argue for this particular search string yielded the best result of all approaches (shown in Table 3.5). The false positive and false negative rate was very low compared to all other queries resulting in a sensitivity of 95 % and specificity of 93 % for this three organisms.

	Legionella pneumophila		Wolbachia		Francisella tularensis	
	Positive	Negative	Positive	Negative	Positive	Negative
True	14	49	24	29	3	20
False	-	1	-	4	2	2

Table 3.5: Evaluated results for the used PHAT query string.

Protein trees which resulted in a positive signal using this PHAT string and having an ELD score equal to or greater than 4 were manually verified. Often trees like the one shown in Figure 3.13 resulted in positive signals - in the case of YP_005661992 the phylogenetic tree can be explained by a HGT (a possible donor phylum for the genetic material would be Stramenopiles) from eukaryotes to Proteobacteria or Chlamydiae and further horizontal gene transfer in prokaryotes. Another possibility - and probably more likely - is that after changing the early branching of the tree (which has very low bootstrap support) prokaryotes and eukaryotes cluster in different branches - making this a protein

that was present in the last common ancestor of prokaryotes and eukaryotes and was only preserved in the shown species. This theory is also supported by the different deep branches of the seed family. In conclusion this tree's topology is far from a clear case - selective gene loss is next to HGT a valid interpretation. A clear, qualitative interpretation of this phylogenetic tree is not possible. This tree should stand as an example for many trees that showed such a topology - it was rarely possible to determine the origin unambiguous and a HGT scenario with just one HGT event was rarely enough to explain the topology of the tree. This also explains partly why the PHAT search string '(Taxonomic name of seed prokaryote & {Eukaryota>3} {*Bacteria<20})' works better than most other strings that only allow one HGT. .

The used PHAT string also detected trees like the one shown in Figure 3.14 which were - after manual evaluation - removed from the pPHAT+pELD group. The observed topology of the protein tree in Figure 3.14 can be best interpreted by rearranging the early branching of the tree, clustering eukaryotes and prokaryotes in two different branches. The phylogenetic distribution of only fungi containing this protein makes selective gene loss in eukaryotes a possible mechanism to construct such a tree topology. But nevertheless the possibility of convergent evolution and the unlikely but still possible scenario of HGT events from prokaryotes to eukaryotes should be kept in mind.

Tree topologies like the one shown in Figure 3.15 form by far the majority. Very often even less prokaryotes are present. The topology of the tree shows similarities to the tree shown in Figure 3.13 - the topology of the tree can be explained by two possible HGT events. The first HGT had most likely as gene donor a species belonging to the phylum Viridiplantae. In the second HGT event genetic material was presumably exchanged between Gammaproteobacteria and the delta/epsilon subdivisions of Proteobacteria which is one way to explain the protein tree topology.

What should be kept in mind while analyzing the phylogenetic relationship of organisms in neighbor-joining trees is that the deeper the branching of the tree the more reliable is the topology - the topology near the root - and especially the position of the root - are very doubtable and should at best be used as hints rather than facts.

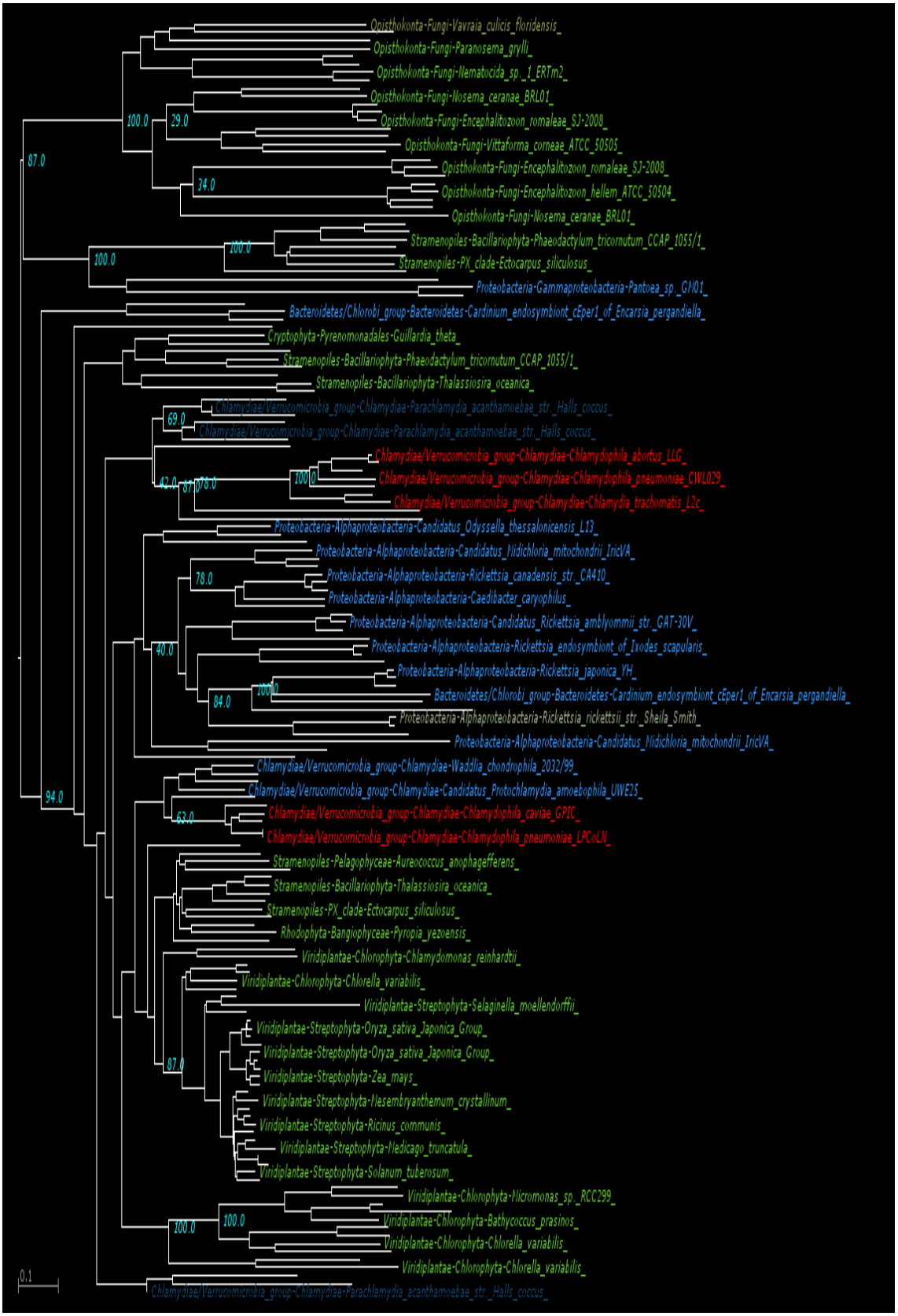

Figure 3.13: Protein tree for YP_005661992 in Chlamydophila pneumoniae LPCoLN. Eukaryots are shown in green, Prokaryotes are shown in light blue, Parachlamydia in dark blue and Chlamydia/Chlamydophila in red. The seed organism is the lowest red entry.

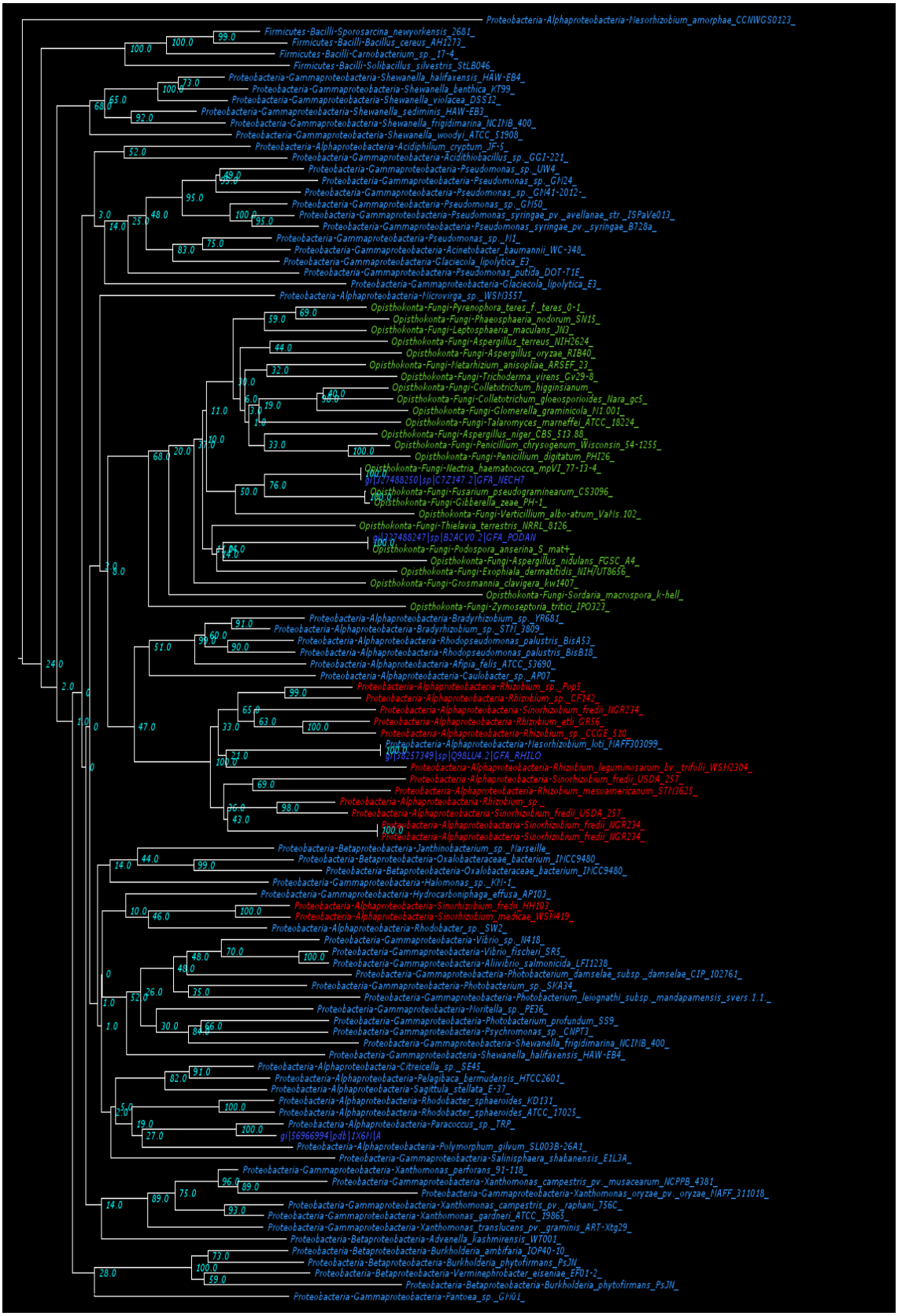

Figure 3.14: Protein tree for YP_002823667 in Sinorhizobium fredii NGR234. Eukaryots are shown in green, Prokaryotes are shown in light blue and the family of the family of the seed organism (Rhizobium/Sinorhizobium) in red.

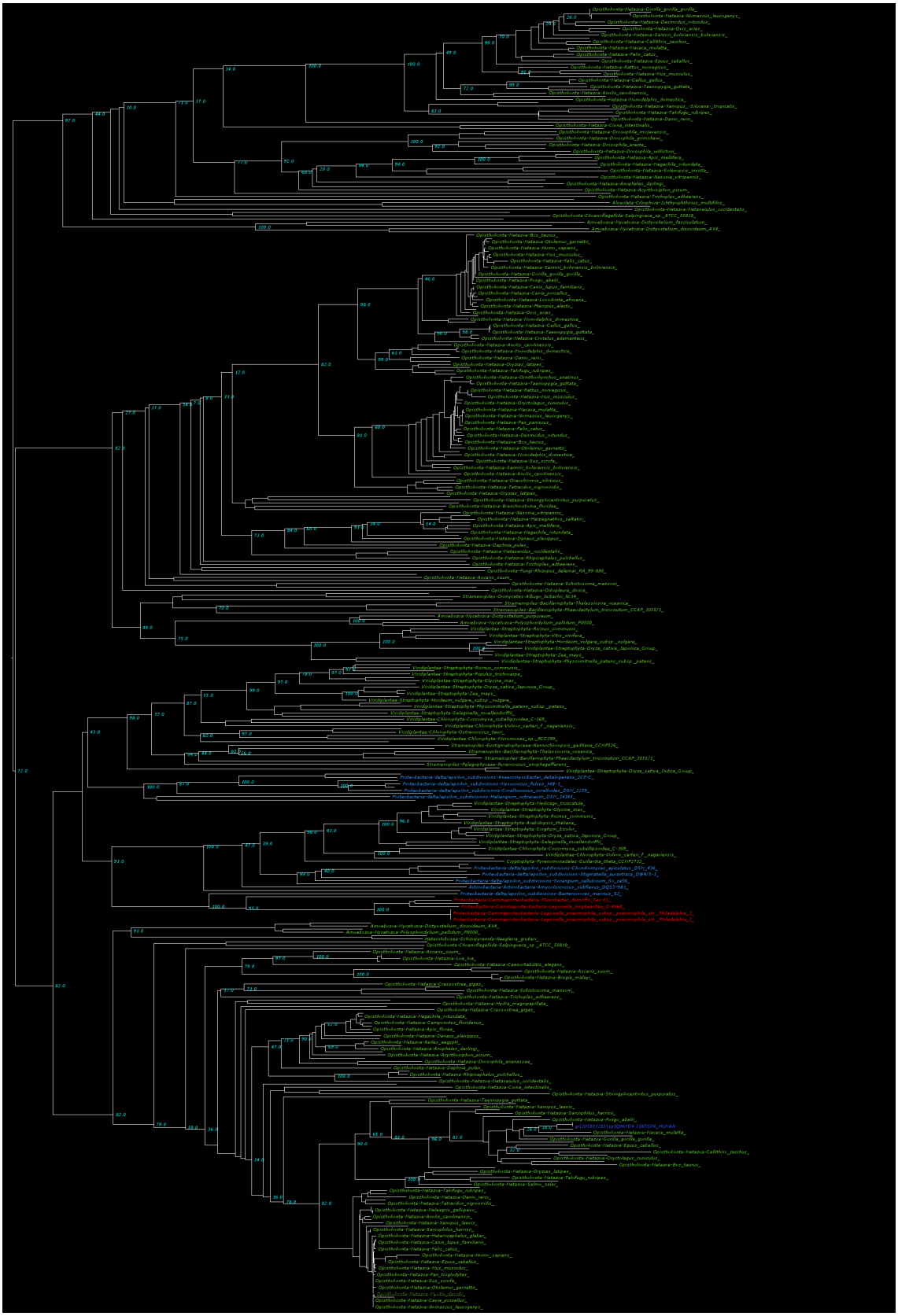

Figure 3.15: Protein tree for YP_094149 in Legionella pneumophila subsp. pneumophila str. Philadelphia 1. Eukaryots are shown in green, Prokaryotes are shown in light blue and and the family of the seed organism (Legionellacaea) in red.

3.5 Comparison of neighbor-joining and maximum-likelihood tree reconstruction

It was important to determine the quality and robustness of the neighbor-joining (NJ) trees. To evaluate this for one organism (Chlamydia trachomatis D/UW-3/CX) phylogenetic trees were calculated using RAxML - a maximum-likelihood (ML) tree reconstruction method. This was only done for one organism - after the AI cut-off of 0.5 473 proteins remained in the analysis and the same number of multiple alignments were calculated, functioning as starting points for the phylogenetic tree construction. This took - with an average of 10x(8 CPUs) working on this problem - about three months.
Chlamydia trachomatis D/UW-3/CX was used mostly because of its small proteome (895 proteins - divided into 22 proteins containing ELDs and 873 proteins without ELDs). Table 3.6 shows the differences in the result. For proteins which were categorized as containing ELDs and the PHAT analysis resulted in a positive signal there is no difference in the two different tree reconstruction methods.

Closer investigation of the two trees (NP_219537 and NP_219568) which are positive for eukaryotic origin and have an ELD-score ≥ 4 constructed using RAxML show that NP_219537 is distinctively of eukaryotic origin while for NP_219568 this question can not be answered unambiguous but seems very likely.
The same trees were also positive in the analysis with the neighbor-joining trees. The topology of the trees did not differ in crucial nodes.
The category nPHAT+pELD also resulted in the same amount of positive signals for the different PHAT queries. pPHAT+nELD and nPHAT+nELD showed different results - the neighbor-joining tree construction method seems to tend to give more positive signals in contrast to maximum-likelihood method. This might result from the fact that 3 trees build with the NJ method in contrast to 37 trees build with the ML method did not yield a tree (this happened out of different reasons - one was that a time cut-off of 1 month on 8 CPUs was set and every analysis that took longer was discarded) and were

therefore regarded as belonging to the nPHAT and nELD group. This was only done in this analysis, otherwise they were excluded.

Chlamydia trachomatis D/UW-3/CX	ELDP:22	nELDP:873		
	pAI+pELD	pAI+nELD	nAI+pELD	nAI+nELD
	19	495	3	378
NJ	pPHAT+pELD	pPHAT+nELD	nPHAT+pELD	nPHAT+nElD
PHAT1	4	384	18	489
PHAT2	4	300	18	573
PHAT3	2	58	20	815
ML	pPHAT+pELD	pPHAT+nELD	nPHAT+pELD	nPHAT+nElD
PHAT1	4	356	18	517
PHAT2	4	267	18	606
PHAT3	2	59	20	814

Table 3.6: Comparison of the results obtained with different PHAT search strings for trees from Chlamydia trachomatis D/UW-3/CX using either a neighbor-joining or a maximum-likelihood tree reconstruction method. PHAT1 used (Chlamydia trachomatis D/UW-3/CX & Eukaryota), PHAT2 (taxname & {Eukaryota>3}) and PHAT3 (Chlamydia trachomatis D/UW-3/CX & {Eukaryota>3} & {*Bacteria<20}) as search string.

3.6 Results obtained for analysis of proteins using the AI

The list Table 3.7 shows the number of total proteins, number of proteins containing an ELD, proteins which had an AI above 0.5 and the intersection of both groups - the number of proteins which had an AI above 0.5 and contained an ELD - for every analyzed organism. The final results add to this list the phylogenetic tree analysis - but with the alien index alone some results can be obtained.

Name	Porteins	ELDP	pAI	pAI+pELD	pELP	pELP+pELD	χ^2
Pseudomonas aeruginosa DK2	5863	360	2711	205	162	13	0.3968
Staphylococcus aureus subsp. aureus 71193	2620	40	1129	18	64	1	0.6224
Chlamydophila pneumoniae CWL029	1051	23	567	17	75	3	0.4819
Mycoplasma gallisepticum NC06_2006.080-5-2P	737	4	576	4	84	0	NA
Helicobacter pylori PeCan18	1491	16	666	8	37	0	NA
Mycoplasma wenyonii str. Massachusetts	651	9	323	9	68	0	NA
Francisella tularensis subsp. tularensis SCHU S4	1529	36	1043	22	33	2	0.4014
Streptococcus pyogenes MGAS315	1858	36	638	7	31	0	NA
Escherichia coli CFT073	5324	155	2017	80	303	3	0.06115
Streptococcus mutans UA159	1952	29	677	9	12	0	NA
Yersinia pestis CO92	3882	64	1634	32	135	2	0.8503
Coxiella burnetii RSA 493	1810	26	999	12	61	6	4.162e-07
Rickettsia typhi str. Wilmington	836	10	647	9	159	0	NA
Haemophilus influenzae R2846	1631	23	846	8	15	0	NA
Onion yellows phytoplasma OY-M	721	89	500	83	80	4	0.05267
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	3662	157	1905	113	173	10	0.4232
Chlamydia trachomatis D/UW-3/CX	895	22	514	19	40	2	0.5893
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	2931	62	1769	49	202	11	0.001602

Name	Porteins	ELDP	pAI	pAI+pELD	pELP	pELP+pELD	χ^2
Staphylococcus aureus RF122	2509	40	1111	16	68	1	0.6831
Bartonella quintana str. Toulouse	1139	1	695	1	36	0	NA
Mycoplasma hyopneumoniae 232	691	7	504	7	93	0	NA
Legionella pneumophila str. Lens	2920	67	1735	51	186	15	2.237e-07
Legionella pneumophila str. Paris	3147	89	1812	63	200	16	1.431e-05
Campylobacter lari RM2100	1544	20	636	9	18	0	NA
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1	2074	28	1274	20	32	1	0.9164
Xanthomonas campestris pv. vesicatoria str. 85-10	4628	104	2198	48	122	3	0.8811
Actinobacillus succinogenes 130Z	2079	22	1016	11	22	1	0.5757
Campylobacter fetus subsp. fetus 82-40	1715	2	667	1	6	0	NA
Sinorhizobium fredii NGR234	6362	867	2583	451	107	8	0.08393
Shewanella putrefaciens 200	4179	41	1492	12	16	0	NA
Chlamydophila pneumoniae LPCoLN	1105	25	562	17	75	2	0.8742
Borrelia recurrentis A1	977	3	715	3	104	0	NA
Lactococcus garvieae Lg2	1938	29	702	13	12	0	NA
Bacteroides vulgatus ATCC 8482	3983	302	736	59	27	0	NA
Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	5497	177	1445	41	49	0	NA
Polynucleobacter necessarius subsp. necessarius STIR1	1507	2	920	2	16	0	NA
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476	4758	105	1798	52	223	4	0.8441
Streptococcus gordonii str. Challis substr. CH1	2050	51	665	14	20	0	NA
Cryptobacterium curtum DSM 15641	1355	22	721	6	9	0	NA
Marinomonas posidonica IVIA-Po-181	3489	40	1765	17	22	1	0.6186

Name	Porteins	ELDP	pAI	pAI+pELD	pELP	pELP+pELD	χ^2
Francisella tularensis subsp. tularensis NE061598	1709	42	1105	27	34	3	0.06257
Streptococcus mutans NN2025	1893	24	671	10	13	0	NA
Candidatus Liberibacter asiaticus str. psy62	1102	18	643	12	22	0	NA
Staphylococcus aureus subsp. aureus TCH60	2693	45	1154	21	68	1	0.7274
Listeria monocytogenes HCC23	2972	114	1042	56	23	1	0.677
Bifidobacterium animalis subsp. lactis DSM 10140	1565	14	714	6	15	0	NA
Pseudomonas aeruginosa LESB58	5902	374	2752	214	164	12	0.7188
Pectobacterium wasabiae WPP163	4431	102	1779	37	151	5	0.5718
Bifidobacterium longum subsp. infantis 157F	1975	47	784	19	15	0	NA
Acidaminococcus intestini RyC-MR95	2383	70	866	31	6	0	NA
Wolbachia endosymbiont of Culex quinquefasciatus Pel	1161	73	794	59	145	34	= undef;
Mycobacterium africanum GM041182	3814	142	928	48	60	5	0.1193
Bifidobacterium animalis subsp. lactis V9	1569	14	718	6	15	0	NA
Arcobacter nitrofigilis DSM 7299	3125	15			35	50	= undef;
Escherichia coli IAI1	4326	110	1840	59	271	13	0.02538
Helicobacter pylori B38	1367	14	634	6	23	0	NA
Streptococcus dysgalactiae subsp. equisimilis ATCC 12394	2050	44	686	7	26	1	0.937
Treponema pallidum subsp. pallidum str. Chicago 1002	980	4	550	2	26	0	NA
Corynebacterium pseudotuberculosis 1002	2090	10	677	3	12	0	NA
Helicobacter mustelae 12198	1398	3	637	2	18	0	NA
Rothia mucilaginosa DY-18	1903	9	887	3	92	0	NA
Corynebacterium diphtheriae 31A	2352	19	736	5	11	0	NA

Name	Porteins	ELDP	pAI	pAI+pELD	pELP	pELP+pELD	χ^2
Novosphingobium sp. PP1Y	4613	225	1875	87	38	6	0.005821
Bacillus thuringiensis BMB171	5327	323	1736	116	70	5	0.8975
Erwinia amylovora ATCC 49946	3565	20	1449	10	135	1	0.7624
Herbaspirillum seropedicae SmR1	4732	106	2396	79	122	1	0.4448
Rothia dentocariosa ATCC 17931	2217	24	802	7	40	0	NA

Table 3.7: Overview of the number of proteins, number of positive ELD signals, number of all proteins with an AI score above 0.5, number of proteins with an AI score above 0.5 and positive ELD signal for all analyzed prokaryotes, number of proteins with an AI score above 1.00 and number of proteins with an AI score above 1.00 and positive ELD signal.

The concept of ELPs (an ELP is defined as a protein that has an AI score of above 1.05) is also of interest in this analysis. Proteins that are categorized as ELPs should - in theory - also contain ELDs. The two concepts are closely related - the one is a description of eukaryotic origin on domain the other on protein level. The data generated in this thesis does not support this - the intersection between this two groups (proteins with an alien index above 1.00 and proteins containing ELDs) is very little.

3.7 Large-scale search for HGTs from eukaryotes to bacteria utilizing AI and phylogenetic tree reconstruction

3.7.1 Variation of ELD-scores

The used threshold for the classification of an eukaryotic like domain was a ELD-score of 4 or above. The ELD-score allows to distinguish between protein domains that are uniformly distributed and eukaryotic like domains that are enriched in pathogenic bacteria. It was shown that this score has all the characteristics of a Z-scores - hence constituting the amount of standard deviations above a gaussian bell-shaped curve (e.g., a Z-score of +4 means that the difference of a given value to the mean is four times the standard deviation) and is significant for values of 4 or higher. If this score is gradually increased the

impact on the amount of members of the different groups and on the p-value is shown in Table B.1. It was not possible to detect a clear trend in the χ^2 p-values - the significance of the relationship between containing a ELD and showing a horizontal gene transfer event from eukaryotes to prokaryotes in the protein-tree does not increase or decrease consistently with a change in the ELD score threshold.

3.7.2 Analysis with ELDs-score ≥ 4

70 bacteria were analyzed with the implemented workflow using blammer for multiple sequence alignment and mtrees (neighbor-joining) for phylogenetic tree calculation. An overview of the results are given in Table 3.8, Table 3.9 and Table 3.10 and a comprehensive table of all results is given in Table D.1. Proteins of all bacteria were classified in four groups:

1. pPHAT+pELD - protein was characterized as containing a ELD (ELD = ELD-score of ≥ 4) and the investigation of the phylogenetic tree reconstruction with PHAT resulted in a positive signal for an eukaryotic origin.

2. nPHAT+pELD - even though the protein was characterized as containing an ELD either the AI was too low or the analysis of the phylogenetic tree did not result in a positive signal for an eukaryotic origin.

3. pPHAT+nELD - this group contains proteins which had no predicted ELD in their sequence and phylogenetic tree reconstruction and investigation with PHAT resulted in a positive signal for eukaryotic origin.

4. nPHAT+nELD - protein had no predicted ELD in its sequence and either had a too little AI or PHAT returned a negative signal for eukaryotic origin.

A Pearson's χ^2 test was performed using this four groups to establish a possible significant dependency between predicted ELDs and a positive signal for eukaryotic origin. It was not possible to obtain χ^2 results for all bacteria investigated - there were cases with zero proteins in one of the four groups (usually the pELD+pPHAT group) described

above which lead to an inconclusive test result. For this bacteria only NA (Not Available) is listed in the χ^2 and p-value column. The results are shown in Table 3.8, Table 3.9 and Table 3.10. There is a significant dependency between ELDs and HGTs if the p-value is below 0.5 - otherwise this result can not be regarded as significant and could also be produced by chance.

It must be noted that very often the amount of ELDs does not correspond to the amount of proteins containing ELDs. An extreme example is Francisella tularensis subsp. tularensis NE061598 which has 90 ELDs but only 42 proteins with ELDs - in average two ELDs on every protein. This becomes obvious when the sum of the two groups pPHAT+pELD and nPHAT+nELD (in this case 42) does not correspond to the amount of ELDs. For the calculation of the χ^2-test the amount of proteins containing an ELD was consulted.

Taxonomic name	Excluded family (TaxID)	ELD	nELD	pPHAT	nPHAT	pPHAT+ pELD	pPHAT+ nELD	nPHAT+ pELD	nPHAT+ nELD	χ^2	p-value
Pseudomonas aeruginosa DK2	135621	363	5521	17	5846	4	13	356	5490	6.1754	0.01295
Staphylococcus aureus subsp. aureus 71193	90964	40	2583	3	2618	0	3	40	2577	NA	NA
Chlamydophila pneumoniae CWL029	809	23	1029	62	989	2	60	21	968	0.0164	0.898
Mycoplasma gallisepticum NC06_2006.080-5-2P	2092	4	740	2	723	0	14	4	719	NA	NA
Helicobacter pylori PeCan18	72293	16	1479	2	1489	0	2	16	1473	NA	NA
Mycoplasma wenyonii str. Massachusetts	2092	9	643	14	637	0	14	9	627	NA	NA
Francisella tularensis subsp. tularensis SCHU S4	34064	82	1522	40	1487	2	38	34	1455	0.3479	0.5553
Streptococcus pyogenes MGAS315	1300	36	1829	0	1858	0	0	36	1822	NA	NA
Escherichia coli CFT073	543	155	5214	2	5322	0	2	155	5167	NA	NA
Streptococcus mutans UA159	1300	29	1931	2	1950	0	2	29	1921	NA	NA
Yersinia pestis CO92	543	159	3907	6	3818	0	6	64	3812	NA	NA
Coxiella burnetii RSA 493	118968	26	1821	35	1784	7	28	19	1756	74.0123	$<$ 2.2E-16
Rickettsia typhi str. Wilmington	775	10	827	36	800	3	33	7	784	7.6476	0.005685
Haemophilus influenzae R2846	712	23	1613	2	1630	1	2	22	1606	5.0318	0.002489
Onion yellows phytoplasma OY-M	2146	99	650	12	709	7	5	82	627	19.7269	8.93E-006
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	170	157	3510	83	3579	14	83	143	3422	22.5193	0.00000208
Chlamydia trachomatis D/UW-3/CX	809	22	873	48	847	2	58	20	815	0.0005	0.9827
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	444	62	2881	59	2872	11	63	51	2806	53.449	2.654E-013
Staphylococcus aureus RF122	90964	40	2469	2	2501	0	2	40	2459	NA	NA
Bartonella quintana str. Toulouse	772	1	1141	1	1138	0	1	1	1137	NA	NA
Mycoplasma hyopneumoniae 232	2092	7	684	6	685	0	6	7	678	NA	NA
Legionella pneumophila str. Lens	444	67	2867	64	2856	12	52	55	2801	71.7078	$<$ 2.2e-16
Legionella pneumophila str. Paris	444	96	3070	72	3142	14	58	75	3000	67.9736	6.83E-014
Campylobacter lari RM2100	72294	20	1525	2	1542	0	2	20	1522	NA	NA
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1	119060	28	2049	9	2065	0	9	28	2037	NaN	NaN
Xanthomonas campestris pv. vesicatoria str. 85-10	32033	143	4583	33	4616	2	31	102	4493	0.7992	0.3713
Campylobacter fetus subsp. fetus 82-40	72294	2	1717	6	1709	0	6	2	1707	NA	NA

Table 3.8: Overview of the results - Part 1/3. In red cells are the names of bacteria for which the Pearson's χ^2 test returns a p-value below 0.05. The column named 'Excluded' shows the TaxID of the family that was excluded from the AI calculation in the all-prokaryotes database.

Taxonomic name	Excluded family (TaxID)	ELD	nELD	pPHAT	nPHAT	pPHAT+ pELD	pPHAT+ nELD	nPHAT+ pELD	nPHAT+ nELD	χ^2	p-value
Chlamydophila pneumoniae LPCoLN	809	25	1080	58	1047	2	56	23	1024	0.029	0.8647
Borrelia recurrentis A1	137	3	987	20	857	0	20	3	954	NA	NA
Lactococcus garvieae Lg2	1300	29	1939	1	1937	0	1	29	1908	NA	NA
Bacteroides vulgatus ATCC 8482	815	307	3759	8	3975	0	8	302	3673	NA	NA
Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	83676	177	5320	13	5484	0	13	177	5307	NaN	NaN
Polynucleobacter necessarius subsp. necessarius STIR1	119060	2	1506	5	1502	0	5	2	1500	NA	NA
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476	543	105	4675	6	4752	1	5	104	4648	1.0449	0.3067
Streptococcus gordonii str. Challis substr. CH1	1300	51	2000	3	2047	0	3	51	1996	NA	NA
Cryptobacterium curtum DSM 15641	84107	22	1335	6	1349	0	6	22	1327	NA	NA
Marinomonas posidonica IVIA-Po-181	135620	40	3451	20	3469	0	20	40	3429	NA	NA
Francisella tularensis subsp. tularensis NE061598	34064	90	1746	40	1786	5	35	37	1632	13.2088	0.0002786
Streptococcus mutans NN2025	511691	24	1871	0	1883	0	0	24	1869	NA	NA
Candidatus Liberibacter asiaticus str. psy62	82115	20	1089	18	1084	1	17	17	1067	0.1492	0.6993
Staphylococcus aureus subsp. aureus TCH60	90964	45	2648	5	2688	0	5	45	2643	NA	NA
Listeria monocytogenes HCC23	186820	114	2860	7	2965	1	6	113	2852	0.208	0.6483
Bifidobacterium animalis subsp. lactis DSM 10140	31953	14	1552	5	1560	0	5	14	1546	NA	NA
Pseudomonas aeruginosa LESB58	135621	379	5546	16	5883	3	16	371	5512	1.4942	0.2216
Pectobacterium wasabiae WPP163	543	102	4335	10	4421	5	5	97	4324	81.248	$<$ 2.2e-16
Bifidobacterium longum subsp. infantis 157F	31953	47	1952	7	1968	0	7	38	1930	NA	NA
Acidaminococcus intestini RyC-MR95	909930	70	2331	7	2376	2	5	68	2308	8.4193	0.003713
Wolbachia endosymbiont of Culex quinquefasciatus Pel	942	73	1202	80	1081	24	56	42	1039	89.9378	$<$ 2.2e-16
Mycobacterium africanum GM041182	1762	142	3688	16	3798	1	15	141	3657	0.016	0.8992
Bifidobacterium animalis subsp. lactis V9	31953	14	1555	5	1564	0	5	14	1550	NA	NA
Arcobacter nitrofigilis DSM 7299	72294	15	3111	34	3091	0	34	15	3076	NA	NA
Escherichia coli IAI1	543	110	4239	2	4324	0	2	110	4214	NA	NA
Helicobacter pylori B38	72293	14	1368	2	1365	0	2	14	1351	NA	NA

Table 3.9: Overview of the results - Part 2/3. In red cells are the names of bacteria for which the Pearson's χ^2 test returns a p-value below 0.05.

Taxonomic name	Excluded family (TaxID)	ELD	nELD	pPHAT	nPHAT	pPHAT+ pELD	pPHAT+ nELD	nPHAT+ pELD	nPHAT+ nELD	χ^2	p-value
Actinobacillus succinogenes 130Z	712	22	2057	8	2049	0	8	22	2047	NA	NA
Sinorhizobium fredii NGR234	82115	867	5495	18	5513	3	15	843	5399	0.0022	0.9629
Shewanella putrefaciens 200	267890	41	4138	9	4055	0	9	34	4021	NA	NA
Streptococcus dysgalactiae subsp. equisimilis ATCC 12394	1300	45	2011	0	2050	0	0	44	2006	NA	NA
Treponema pallidum subsp. pallidum str. Chicago	137	4	977	9	971	0	9	4	967	NA	NA
Corynebacterium pseudotuberculosis 1002	1653	10	2080	1	2089	0	1	10	2079	NA	NA
Helicobacter mustelae 12198	72293	3	1395	4	1394	0	4	3	1391	NA	NA
Rothia mucilaginosa DY-18	1268	9	1895	0	1904	0	0	9	1894	NA	NA
Corynebacterium diphtheriae 31A	1653	20	2361	3	2349	0	3	19	2330	NA	NA
Novosphingobium sp. PP1Y	41297	239	4425	31	4582	6	25	219	4363	11.1324	0.0008483
Bacillus thuringiensis BMB171	186817	323	5029	5	5322	1	4	322	5000	0.1362	0.7121
Erwinia amylovora ATCC 49946	543	20	3545	4	3559	0	4	20	3539	NA	NA
Herbaspirillum seropedicae SmR1	75682	106	4629	7	4725	0	7	106	4619	NA	NA
Rothia dentocariosa ATCC 17931	1268	24	2193	4	2208	0	4	24	2184	NA	NA

Table 3.10: Overview of the results - Part 3/3. In red cells are the names of bacteria for which the Pearson's χ^2 test returns a p-value below 0.05.

In Table D.1 a detailed list for all bacteria is shown. It contains the name of the bacteria, if available the KEGG name and number of the disease caused by the pathogen, the amount of ELPs found in the genome (ELP = AI above or equal to 1), a short and longer, more detailed description of the associated pathogenicity and a number ranging from 0 (symbiont) to 4 (extraordinary high virulence) roughly mirroring the virulence of the pathogen. Also all proteins which were classified as ELD containing proteins and PHAT analysis detected potential eukaryotic origin are included with its accession number, alien index, domains and - if available - a short description of the protein function.
In the following prokaryotes the χ^2-text showed a dependency between ELD containing proteins and proteins which were characterized as of eukaryotic origin:

1. Pseudomonas aeruginosa DK2 - Pseudomonas aeruginosa is the predominant microorganism in chronic lung infection of cystic fibrosis patients - Virulence: 2 - Human pathogen (secondary infection) - risk group[1] 2

2. Coxiella burnetii RSA 493 - The etiological agent of Q-fever - Virulence: 4 - Human pathogen (primary infection) - risk group 3

3. Rickettsia typhi str. Wilmington - Rickettsia typhi is the etiological agent of typhus feve - Virulence: 4 - Human pathogen (primary infection) - risk group 3

4. Haemophilus influenzae R2846 - Serotype b is the most frequent cause of bacterial meningitis - Virulence: 4 - Human pathogen (primary infection) - risk group 2

5. Onion yellows phytoplasma OY-M - Phytoplasmal infections can cause extensive phloem necrosis - Virulence: 3 - Plant pathogen (primary infection)

6. Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 - Mostly asymptotic but small numbers of the patients may develop febrile illness with icterus, splenomegaly and nephritis, febrile illness with pulmonary haemorrhages or ocular manifestations - Virulence: 3 - Human pathogen, animal pathogen (primary infection) - risk group 2

[1]This category is due to the list of pathogens and their required biosafety level published by the Bundesamt für Verbraucherschutz und Lebensmittelsicherheit, Germany

7. Legionella pneumophila subsp. pneumophila str. Philadelphia 1 - Etiological agent of Legionaires disease and Pontiac fever - Virulence: 3 - Human pathogen (primary/secondary infection) - risk group 2

8. Legionella pneumophila str. Lens - Etiological agent of Legionaires disease and Pontiac fever - Virulence: 3 - Human pathogen (primary/secondary infection) - risk group 2

9. Legionella pneumophila str. Paris - Etiological agent of Legionaires disease and Pontiac fever - Virulence: 3 - Human pathogen (primary/secondary infection) - risk group 2

10. Francisella tularensis subsp. tularensis NE061598 - the etiological agent of tularemia - Virulence:4 - Human pathogen (primary infection) - risk group 3

11. Pectobacterium wasabiae WPP163 - causing soft rotting in potatoes - Virulence: 3 - plant pathogen (primary infection) - risk group 1

12. Acidaminococcus intestini RyC-MR95 - occasionally associated to infective processes but always as part of polymicrobial infections - Virulence: 1 - Human pathogen/symbiont (secondary infection) - risk group ?

13. Wolbachia endosymbiont of Culex quinquefas- ciatus Pel - Wolbachia bacteria colonize Onchocerca volvulus as mutualistic symbionts - Virulence: 3 - Endosymbiont of athropoda - risk group 1

14. Novosphingobium sp. PP1Y - discussed as a potential human pathogen - Virulence: ? - Possible human pathogen - risk group ?

3.7.3 Phylogenetic origin and dynamics

All protein trees belonging to the pPHAT+pELD group were manually verified. Also, if possible, the origin of the HGT was determined. It was rarely easy to determine the donor of the genetic material - very often it was not even possible to give a clear answer

regarding the phylum of the donor. It must be kept in mind that many of the phylogenetic trees are very large and paralogs from the seed organism (not always coming up on the same leaf in the tree) and from all other organisms are present in the phylogenetic tree. Furthermore the root is determined by midpoint rooting - this means the longest edge of the graph is split at its half-point and the root is assigned to this point. It is normal that this rooting method is applied since there is no out-group which could be used instead but nevertheless it must be kept in mind that the root is doubtable. Furthermore very often bootstrap values are low - resulting in an uncertain topology at early branches. And - as with all tree constructing methods - a phylogenetic tree is the solution of fitting a n∗m distance matrix with n∗m dimensions in a two dimensional representation - a task that is by definition only an approximation. The proposed donor-acceptor relationship for the genetic material (always from eukaryotes to prokaryotes) is shown in Table 3.11 for the manually verified pELD+pPHAT group.

Revealing the phylogenetic relationship between the HGT donor and the acceptor of effector proteins is an important step in understanding the evolution of pathogens (and its effectors and pathogenicity). Since the NJ trees have bad bootstrap support in early branches the following donor-acceptor relationships should only be used as rough guidance. What can be said is that the donors are very often not from the phyla the bacteria infects primarily - e.g., Xanthomonas campestris pv. Vesicatoria is a plant pathogen yet its two ELD containing effectors come from fungi - the same situation applies to Sinorhizobium fredii and its three effectors. Pseudomonas aeruginosa DK2 is a human pathogen - its effector proteins are derived from Fungi. Different is the situation e.g., in Chlamydophila pneumoniae where the two effector proteins have their origin very likely in Viridiplantaea - reflecting an ancestral relationship between Chlamydiaceae, Cyanobacteria, and the Chloroplast [Brinkman *et al.*, 2002].

TaxID: 1093787 TaxName: Pseudomonas aeruginosa DK2

Protein Acc	Domains	Origin
YP_006483810	PF03088	unclear relationship
YP_006480122	PF06355	Fungi

YP_006483846	PF01989, PF04412	Fungi
YP_006485602	PF12710	Fungi
TaxID: 115713 TaxName: Chlamydophila pneumoniae CWL029		
Protein Acc	Domains	Origin
NP_224551	PF032119	Viridiplantaea
NP_224810	PF032119	Viridiplantaea
TaxID: 177416 TaxName: Francisella tularensis subsp. tularensis SCHU S4		
Protein Acc	Domains	Origin
YP_170294	PF02931	Opisthokonta-Metazoa-Caenorhabditis elegans
YP_170045	PF00328	Opisthokonta-Metazoa-Crassostrea gigas
TaxID: 227377 TaxName: Coxiella burnetii RSA 493		
Protein Acc	Domains	Origin
NP_820201	PF01222	Amoebozoa-Centramoebida
NP_819913	PF00086	Opistokonta:Metazoa
NP_820468	PF08613	Euglenozoa
NP_819085	PF00368	Opisthokonta-Fungi
NP_820155	PF01222	Opisthokonta-Metazoa
NP_819228	PF00961	Viridiplantaea-Chlorophyta
YP_002332981	PF07671	Viridiplantaea
TaxID: 257363 TaxName: Rickettsia typhi str. Wilmington		
Protein Acc	Domains	Origin
YP_067421	PF03219	unclear relationship
YP_067327	PF03219	unclear relationship
YP_067440	PF03219	unclear relationship
TaxID: 262727 TaxName: Haemophilus influenzae R2846		
Protein Acc	Domains	Origin
YP_005829448	PF00312	Viridiplantaea
TaxID: 262768 TaxName: Onion yellows phytoplasma OY-M		
Protein Acc	Domains	Origin
NP_950553	PF09419	unclear relationship

NP_950369	PF00004	unclear relationship
NP_950784	PF00004	unclear relationship
NP_950947	PF00004	unclear relationship
NP_950929	PF00004	unclear relationship
NP_950296	PF00004	Opisthokonta
NP_950813	PF00004	bacterial origin

TaxID: 267671 TaxName: Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130

Protein Acc	Domains	Origin
YP_002595	PF10520	Opisthokonta
YP_001985	PF01344	Opisthokonta
YP_003198	PF12796	Opisthokonta
YP_000337	PF01731	Opisthokonta
YP_002923	PF12796	unclear relationship
YP_001274	PF04768	Opisthokonta
YP_003211	PF02493	Alveolata
YP_002479	PF08713	unclear relationship
YP_001020	PF12796	Opisthokonta
YP_000239	PF12796	unclear relationship
YP_000473	PF02493	Opisthokonta
YP_003271	PF03088	Opisthokonta
YP_002325	PF12796	Opisthokonta
YP_003053	PF01210, PF07479	unclear relationship

TaxID: 272561 TaxName: Chlamydia trachomatis D/UW-3/CX

Protein Acc	Domains	Origin
NP_219568	PF03219	Viridiplantaea
NP_219973	PF02201	unclear relationship

TaxID: 272624 TaxName: Legionella pneumophila subsp. pneumophila str. Philadelphia 1

Protein Acc	Domains	Origin
YP_096400	PF01150	unclear relationship
YP_095005	PF13516	Parabasalia:Trichomonidida
YP_095964	PF00328	unclear relationship
YP_095152	PF01369	Parabasalia:Trichomonidida
YP_095966	PF11901	Euglenozoa:Kinetoplastida

YP_094355	PF13516	Opisthokonta:Fungi
YP_095974	PF08123	Opisthokonta:Metazoa
YP_096947	PF01150	Alveolata:Ciliophora
YP_095922	PF05761	Parabasalia:Trichomonidida
YP_094149	PF05761	Opisthokonta
YP_094979	PF13855	unclear relationship

TaxID: 297245 TaxName: Legionella pneumophila str. Lens

Protein Acc	Domains	Origin
YP_125472	PF05761	Viridiplantaea
YP_127384	PF12796	Opisthokonta:Fungi
YP_125721	PF11901	Opisthokonta:Fungi
YP_126094	PF12796	Opisthokonta:Fungi
YP_127648	PF13516	unclear relationship
YP_127387	PF12796	unclear relationship
YP_127630	PF09445	Stramenopiles:Pelagophyceae
YP_127207	PF01150	Parabasalia:Trichomonidida
YP_127255	PF01369	Opisthokonta
YP_126475	PF00328	unclear relationship
YP_126358	PF01150	Opisthokonta:Fungi
YP_128208	PF08123	unclear relationship

TaxID: 297246 TaxName: Legionella pneumophila str. Paris

Protein Acc	Domains	Origin
YP_124197	PF01150	Alveolata
YP_123088	PF12796	Opisthokonta
YP_123444	PF00328	Alveolata
YP_124726	PF09445	Alveolata
YP_124246	PF01369	Opisthokonta
YP_122460	PF05761	Viridiplantae
YP_125327	PF08123	unclear relationship
YP_124370	PF12796	Opisthokonta
YP_123335	PF13516	Viridiplantaea
YP_124373	PF12796	Opisthokonta
YP_123361	PF01150	Alveolata
YP_124254	PF13516	unclear relationship

YP_122719	PF11901	unclear relationship
YP_124222	PF12796	unclear relationship
TaxID: 316273 TaxName: Xanthomonas campestris pv. vesicatoria str. 85-10		
Protein Acc	Domains	Origin
YP_364278	PF04768	Opisthokonta:Fungi
YP_365163	PF03080	Opisthokonta:Fungi
TaxID: 394 TaxName: Sinorhizobium fredii NGR234		
Protein Acc	Domains	Origin
YP_002824454	PF03328	Opisthokonta:Fungi
YP_002823667	PF04828	Opisthokonta:Fungi
YP_002826446	PF04828	unclear relationship
TaxID: 406984 TaxName: Chlamydophila pneumoniae LPCoLN		
Protein Acc	Domains	Origin
YP_005661727	PF03219	Viridiplantaea
YP_005661992	PF03219	Viridiplantaea
TaxID: 454169 TaxName: Salmonella enterica subsp. enterica serovar Heidelberg str. SL476		
Protein Acc	Domains	Origin
YP_002047662	PF14269	
TaxID: 510831 TaxName: Francisella tularensis subsp. tularensis NE061598		
Protein Acc	Domains	Origin
YP_005831663	PF02274	Alveolata:Ciliophora
YP_005831889	PF02931	Opisthokonta:Metazoa
YP_005831982	PF11940	Viridiplantaea:Chlorophyta
YP_005830479	PF02274	unclear relationship
YP_005831228	PF00328	Alveolata:Ciliophora
TaxID: 537021 TaxName: Candidatus Liberibacter asiaticus str. psy62		
Protein Acc	Domains	Origin
YP_003064736	PF03219	unclear relationship
TaxID: 552536 TaxName: Listeria monocytogenes HCC23		

Protein Acc	Domains	Origin
YP_002350681	PF08719	unclear relationship
TaxID: 561231 TaxName: Pectobacterium wasabiae WPP163		
Protein Acc	Domains	Origin
YP_003260359	PF13343	Opisthokonta:Fungi
YP_003259508	PF12796	Opisthokonta:Metazoa
YP_003259880	PF13343	Opisthokonta:Fungi
YP_003258518	PF12796	Opisthokonta:Metazoa
YP_003258335	PF06045	Opisthokonta:Fungi
TaxID: 568816 TaxName: Acidaminococcus intestini RyC-MR95		
Protein Acc	Domains	Origin
YP_004897753	PF05049	Opisthokonta:Metazoa
YP_004897754	PF05049	Opisthokonta:Metazoa
TaxID: 570417 TaxName: Wolbachia endosymbiont of Culex quinquefasciatus Pel		
Protein Acc	Domains	Origin
YP_001976067	PF12796	Opisthokonta:Metazoa
YP_001975172	PF12796	Opisthokonta:Metazoa
YP_001975493	PF12796	Opisthokonta:Metazoa
YP_001975547	PF12796	Opisthokonta:Metazoa
YP_001975413	PF12796	Opisthokonta:Metazoa
YP_001975955	PF12796	Opisthokonta:Metazoa
YP_001975179	PF12796	Opisthokonta:Metazoa
YP_001976103	PF12796	Opisthokonta:Metazoa
YP_001975536	PF12796	Opisthokonta:Metazoa
YP_001975139	PF12796	Opisthokonta:Metazoa
YP_001975025	PF12796	Opisthokonta:Metazoa
YP_001975234	PF12796	Opisthokonta:Metazoa
YP_001976012	PF12796	Opisthokonta:Metazoa
YP_001975451	PF12796	Opisthokonta:Metazoa
YP_001975284	PF12796	Opisthokonta:Metazoa
YP_001975372	PF12796	Opisthokonta:Metazoa
YP_001975491	PF12796	Opisthokonta:Metazoa

YP_001975140	PF12796,PF00023	Opisthokonta:Metazoa
YP_001975417	PF12796,PF00023	Opisthokonta:Metazoa
YP_001975197	PF12796,PF00023	Opisthokonta:Metazoa
YP_001975764	PF12796	Opisthokonta:Metazoa
YP_001975619	PF12796	Opisthokonta:Metazoa
YP_001975403	PF12796,PF13637	Opisthokonta:Metazoa
YP_001975193	PF12796,PF00023,PF13637	Opisthokonta:Metazoa
TaxID: 702113 TaxName: Novosphingobium sp. PP1Y		
Protein Acc	Domains	Origin
YP_004532680	PF03080,PF14365	Opisthokonta:Fungi
YP_004533567	PF03332	Opisthokonta:Fungi
YP_004533187	PF02359,PF02933	unclear relationship
YP_004533918	PF13577	Opisthokonta:Fungi
YP_004534061	PF07110	Opisthokonta:Fungi
YP_004533910	PF13577	Opisthokonta:Fungi

Table 3.11: Possible phylogenetic donor-acceptor relationship for HGTs.

3.8 Tree construction based on protein domains

Utilizing protein domains for tree construction instead of the full length of a protein has some immediate advantages: (1) protein domains are evolutionary units and their occurrence and descent can be detected over long evolutionary distances, (2) using only the most important part of the protein - structural as well as functional - for the multiple alignment minimizes the bias introduced by different mutation rates for different segments of the protein.

3.8.1 Protein domains for Legionella pneumophila str. Paris

Tree construction was done using the hidden Markov models and multiple sequence alignments for all unique protein domains present in Legionella pneumophila str. Paris provided by the Pfam protein families database [Fukami-Kobayashi *et al.*, 2007, Punta

et al., 2012a, Yang & Bourne, 2009]. The total number of proteins in the proteome of Legionella pneumophila str. Paris is 3145, the total number of domains (domain instances) is 3643 and 1809 unique protein domains (domain occurrences) are present.
For 1809 domain occurrences multiple sequence alignments were obtained. It was possible to calculate starting from 1809 multiple sequence alignments 1704 protein domain trees - for the other 105 domain occurrences the number of species who possessed this domain were - even after clustering with CD-HIT based on a sequence identity of 50% - too high to be handled by mtrees (normally above 3000-4000).

Legionella pneumophila str. Paris Domain Trees			
	pPHAT	nPHAT	Sum
pELD	21	61	82
nELD	92	1635	1727
Sum	113	1696	

Table 3.12: Results for the analyzed domain trees of Legionella pneumophila str. Paris.

The pPHAT+pELD results for domain trees were compared to the pPHAT+pELD results for protein trees: 14 phylogenetic trees constructed from proteins were both positive for the pPHAT search string and had an ELD-score ≥ 4. Seven of these proteins contain domains which are positive for eukaryotic origin in phylogenetic trees based on domains (Q5X0S1 - YP_125327, Q5X874 - YP_122719, Q5X2H2 - YP_124726, Q5X8Y3 - YP_122460, Q5X3V2 - YP_124246, Q5X650 - YP_123444, Q5X6D3 - YP_123361).
Investigation of the remaining 1551 phylogenetic trees based on domains showed in further 113 trees positive hits for eukaryotic origin. All details for the positive PHAT hits are shown in Table C.1 - the ELD-Scores in this analysis are obtained from the updated Effective database (2012-12-08).
Of the 113 proteins that are characterized as containing ELDs 21 are positive for eukaryotic origin. The χ^2-test for the domain trees shows a p-value of $6.8*10^{-13}$. It appears as if the ELD classifications correspond to a greater degree to the results gained from

domain trees than to protein trees. It is known that protein domains reflect their origin much better but it must also be kept in mind that their is much phylogenetic information lost when discarding the rest of the protein sequence and focusing only on the preserved structures.

3.8.2 Analysis of domain trees for 60 bacteria

The results obtained from the analysis of the domain occurrence in Legionella pneumophila str. Paris showed that it is essential to analyze - besides proteins - domain trees to be able to understand the evolution of effector proteins.
Domain trees for 60 bacteria - for all of which proteins were previously analyzed - were obtained and the results are summarized in Table 3.14. It should be noted beforehand that from 6755 multiple sequence alignments for domains only 6630 trees were obtained (loss of 125 trees or 1.8%) due to memory limitations and calculation errors for multiple sequence alignments superseding 6000 species. For individual organisms this lead to a loss of about 5% of domain trees - these domains are automatically added to the nPHAT+nELD category (these domain trees were never ELDs).

Significant p-values were obtained for 33 out of 60 bacteria :

1. Acidaminococcus intestini RyC-MR95
2. Arcobacter nitrofigilis DSM 7299
3. Bacillus thuringiensis BMB171
4. Bacteroides vulgatus ATCC 8482
5. Bartonella quintana str. Toulouse
6. Bifidobacterium animalis subsp. lactis DSM 10140
7. Bifidobacterium animalis subsp. lactis V9
8. Campylobacter fetus subsp. fetus 82-40

9. Campylobacter lari RM2100
10. Clostridiales genomosp. BVAB3 str. UPII9-5
11. Coxiella burnetii RSA 493
12. Cryptobacterium curtum DSM 15641
13. Erwinia amylovora ATCC 49946
14. Francisella tularensis subsp. tularensis NE061598
15. Francisella tularensis subsp. tularensis SCHU
16. Herbaspirillum seropedicae SmR1
17. Legionella pneumophila str. Lens
18. Legionella pneumophila str. Paris
19. Legionella pneumophila subsp. pneumophila str. Philadelphia 1
20. Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130
21. Listeria monocytogenes HCC23
22. Marinomonas posidonica IVIA-Po-181
23. Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111
24. Pectobacterium wasabiae WPP163
25. Polynucleobacter necessarius subsp. asymbioti- cus QLW-P1DMWA-1
26. Polynucleobacter necessarius subsp. necessarius STIR1
27. Pseudomonas aeruginosa LESB58
28. Rothia dentocariosa ATCC 17931
29. Rothia mucilaginosa DY-18

30. Shewanella putrefaciens 200

31. Sinorhizobium fredii NGR234

32. Wolbachia endosymbiont of Culex quinquefasciatus Pel

33. Xanthomonas campestris pv. vesicatoria str. 85- 10

These results can be compared to the significant p-values obtained for the same categories in the χ^2-test with protein trees (summarized in Table 3.8, Table 3.9 and Table 3.10) which resulted in 14 significant p-values. Merging both lists result in the table shown in Table 3.13. There are 8 prokaryotes for which a positive p-value for the chi-square test for proteins as well as for domains is present. It seems like the relationship between ELDs and eukaryotic origin is much stronger on domain than on protein level.

Taxonomic name	Domain trees	Protein trees
Acidaminococcus intestini RyC-MR95	positive	positive
Wolbachia endosymbiont of Culex quinquefasciatus Pel	positive	positive
Coxiella burnetii RSA 493	positive	positive
Francisella tularensis subsp. tularensis NE061598	positive	positive
Legionella pneumophila str. Lens	positive	positive
Legionella pneumophila str. Paris	positive	positive
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	positive	positive
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	positive	positive
Pectobacterium wasabiae WPP163	positive	positive
Arcobacter nitrofigilis DSM 7299	positive	negative
Bacillus thuringiensis BMB171	positive	negative
Bacteroides vulgatus ATCC 8482	positive	negative
Bartonella quintana str. Toulouse	positive	negative
Bifidobacterium animalis subsp. lactis DSM 10140	positive	negative
Bifidobacterium animalis subsp. lactis V9	positive	negative
Campylobacter fetus subsp. fetus 82-40	positive	negative
Campylobacter lari RM2100	positive	negative
Chlamydophila pneumoniae LPCoLN	positive	negative
Clostridiales genomosp. BVAB3 str. UPII9-5	positive	negative
Cryptobacterium curtum DSM 15641	positive	negative
Erwinia amylovora ATCC 49946	positive	negative
Francisella tularensis subsp. tularensis SCHU	positive	negative
Haemophilus influenzae R2846	np	positive
Herbaspirillum seropedicae SmR1	positive	negative
Listeria monocytogenes HCC23	positive	negative
Marinomonas posidonica IVIA-Po-181	positive	negative
Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	positive	negative
Novosphingobium sp. PP1Y	negative	positive
Onion yellows phytoplasma OY-M	negative	positive
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1	positive	negative
Polynucleobacter necessarius subsp. necessarius STIR1	positive	negative
Pseudomonas aeruginosa DK2	negative	positive
Pseudomonas aeruginosa LESB58	positive	negative
Rickettsia typhi str. Wilmington	negative	positive
Rothia dentocariosa ATCC 17931	positive	negative
Rothia mucilaginosa DY-18	positive	negative
Shewanella putrefaciens 200	positive	negative
Sinorhizobium fredii NGR234	positive	negative

Table 3.13: Comparison of the significant results obtained from protein and domain tree analysis. The table shows a merged list of the significant chi-square results obtained for protein and domain tree analysis. The 'np' for Haemophilus influenzae should indicate that this particular prokaryote was not among the 60 analyzed bacteria.

Taxonomic Name	pELDs	nELDs	pPHAT+ pELD	pPHAT+ nELD	nPHAT+ pELD	nPHAT+ nELD	χ^2	p-value
Shewanella putrefaciens 200	16	2386	2	38	14	2348	5.8466	0.01561
Streptococcus mutans NN2025	9	1357	1	20	8	1337	0.9664	0.3256
Borrelia recurrentis A1	4	745	1	49	3	696	0.219	0.6398
Campylobacter fetus subsp. fetus 82-40	2	1329	1	30	1	1299	4.5256	0.03339
Wolbachia endosymbiont of Culex quinquefasciatus Pel	11	872	4	79	7	793	6.5735	0.01035
Legionella pneumophila str. Lens	34	1753	19	59	15	1694	207.9642	2.2e-16
Bifidobacterium animalis subsp. lactis DSM 10140	13	1135	2	25	11	1110	4.8318	0.02794
Rothia mucilaginosa DY-18	8	1213	3	17	5	1196	43.8267	3.588e-11
Clostridiales genomosp. BVAB3 str. UPII9-5	12	1044	3	38	9	1006	9.3455	0.002235
Chlamydophila pneumoniae LPCoLN	9	787	3	81	6	706	2.8613	0.09073
Candidatus Liberibacter asiaticus str. psy62	9	918	0	36	9	882	0	1
Erwinia amylovora ATCC 49946	24	2206	4	33	20	2173	24.8361	6.242e-07
Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	51	2007	8	54	43	1953	24.4724	7.538e-07
Streptococcus gordonii str. Challis substr. CH1	20	1375	0	0	20	1375	NaN	NA
Francisella tularensis subsp. tularensis NE061598	19	1242	5	50	14	1192	17.2663	3.249e-05
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1	13	1575	4	34	9	1541	33.7675	6.211e-09
Legionella pneumophila str. Paris	43	1766	21	61	22	1705	189.4429	2.2e-16
Listeria monocytogenes HCC23	33	1743	5	19	28	1724	38.0672	6.835e-10
Bacteroides vulgatus ATCC 8482	88	1730	7	40	81	1690	8.4642	0.003622

Taxonomic Name	pELDs	nELDs	pPHAT+ pELD	pPHAT+ nELD	nPHAT+ pELD	nPHAT+ nELD	χ^2	p-value
Escherichia coli CFT073	56	2618	1	28	55	2590	0	1
Campylobacter lari RM2100	9	1233	2	26	7	1207	8.5454	0.003464
Marinomonas posidonica IVIA-Po-181	23	1985	8	42	15	1943	86.9254	2.2e-16
Haemophilus influenzae R2846	15	1443	0	9	15	1434	0	1
Rickettsia typhi str. Wilmington	7	834	2	64	5	770	1.8002	0.1797
Corynebacterium diphtheriae 31A	13	1449	0	0	13	1449	NaN	NA
Helicobacter pylori B38	6	1105	0	20	6	1085	0	1
Prevotella denticola F0289	19	1389	1	25	18	1364	0.0655	0.798
Bifidobacterium longum subsp. infantis 157F	13	1263	0	24	13	1239	0	1
Mycobacterium africanum GM041182	34	1772	2	37	32	1735	0.832	0.3617
Francisella tularensis subsp. tularensis SCHU S4	20	1203	6	48	14	1155	25.6733	4.044e-07
Chlamydia trachomatis D/UW-3/CX	11	755	2	80	9	675	0.1003	0.7514
Herbaspirillum seropedicae SmR1	16	2221	2	31	14	2190	6.9195	0.008526
Bifidobacterium animalis subsp. lactis V9	10	1141	2	24	8	1117	7.417	0.006461
Actinobacillus succinogenes 130Z	17	1689	0	10	17	1679	0	1
Rothia dentocariosa ATCC 17931	12	1324	6	19	6	1305	127.446	2.2e-16
Coxiella burnetii RSA 493	9	1242	6	32	3	1210	103.8018	2.2e-16
Lactococcus garvieae Lg2	11	1289	0	14	11	1275	0	1
Onion yellows phytoplasma OY-M	12	413	2	16	10	397	2.0796	0.1493
Polynucleobacter necessarius subsp. necessarius STIR1	2	1251	1	19	1	1232	6.9856	0.008217
Xanthomonas campestris pv. vesicatoria str. 85-10	54	2328	8	68	46	2260	20.4735	6.046e-06

Taxonomic Name	pELDs	nELDs	pPHAT+ pELD	pPHAT+ nELD	nPHAT+ pELD	nPHAT+ nELD	χ^2	p-value
Mycoplasma hyopneumoniae 232	4	480	1	23	3	457	0.4867	0.4854
Staphylococcus aureus subsp. aureus TCH60	18	1688	1	30	17	1658	0.0941	0.759
Staphylococcus aureus RF122	17	1636	0	23	17	1613	0	1
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	45	1609	12	118	33	1491	20.0021	7.736e-06
Acidaminococcus intestini RyC-MR95	18	1411	2	25	16	1386	4.0835	0.0433
Escherichia coli IAI1	51	2585	2	29	49	2556	1.3943	0.2377
Streptococcus dysgalactiae subsp. equisimilis ATCC 12394	11	1338	0	19	11	1319	0	1
Pseudomonas aeruginosa LESB58	63	2617	4	40	59	2577	6.1197	0.01337
Treponema pallidum subsp. pallidum str. Chicago	4	771	1	41	3	730	0.3933	0.5306
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476	50	2594	0	24	50	2570	0	1
Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	4	840	0	23	4	817	0	1
Pectobacterium wasabiae WPP163	50	2415	6	31	44	2384	31.1464	2.393e-08
Bacillus thuringiensis BMB171	81	2332	7	33	74	2299	20.8421	4.988e-06
Cryptobacterium curtum DSM 15641	9	1109	2	26	7	1083	7.4527	0.006334
Sinorhizobium fredii NGR234	98	2285	8	39	90	2246	17.059	3.624e-05
Arcobacter nitrofigilis DSM 7299	8	1727	2	56	6	1671	5.9044	0.0151
Corynebacterium pseudotuberculosis 1002	7	1393	1	26	6	1367	1.0113	0.3146
Helicobacter mustelae 12198	3	1154	0	26	3	1128	0	1

Taxonomic Name	pELDs	nELDs	pPHAT+ pELD	pPHAT+ nELD	nPHAT+ pELD	nPHAT+ nELD	χ^2	p-value
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	38	1731	21	55	17	1676	232.8403	2.2e-16
Bartonella quintana str. Toulouse	1	1072	1	30	0	1042	7.918	0.004894

Table 3.14: Details for domain trees obtained for 60 prokaryotes. Taxonomic names are shown in red if the p-value returned by the χ^2 test was significant.

Chapter 4

Conclusion and outlook

The foundation of this thesis is based on three observations:

1. prokaryotes interact with their biotic environment using proteins (so called effectors) - in the case of pathogens these effector proteins are virulence factors and play a key role in the pathogenicity of the prokaryote [Mattoo *et al.*, 2007];

2. effector proteins often exhibit a certain domain signature - so called eukaryotic-like domains. Eukaryotic-like domains are defined through their abundance in pathogens and eukaryots and their absence in non-pathogenic prokaryotes. Proteins with eukaryotic-like domains act in eukaryotic cells as binding partners to eukaryotic proteins - showing great homology to the eukaryotic binding domain [Jehl *et al.*, 2011];

3. horizontal gene transfer between prokaryotes plays an important role in explaining genetic viability and adaption to new niches [Gogarten & Townsend, 2005] - but there are also many examples of horizontal gene transfer events between prokaryotes and eukaryotes [Hotopp, 2011, Sanchez, 2011].

Combining these three observations it is possible to formulate the main hypothesis the research done in this thesis wanted to investigate: are proteins with eukaryotic-like domains the result of horizontal gene transfer? Are there significantly more horizontal-gene

transfer events when looking at proteins with eukaryotic-like domains in comparison to proteins without such a domain - and can this be shown at the whole protein level as well as on the domain level? The null hypothesis in this research question was that eukaryotic-like proteins and their similarity to the eukaryotic binding domain are the result of convergent evolution or vertical ancestry resulting in evenly distributed horizontal gene transfer events in proteins with eukaryotic-like domains and proteins without ELDs. To decide which hypothesis is applicable the proteom of 70 randomly chosen pathogens was analyzed. The proteom of each organism was divided in ELD containing proteins and non-ELD containing proteins based on the ELD-score introduced in the effector database [Jehl *et al.*, 2011].

To the author's knowledge there are no previous studies using phylogenetic tree analysis and reconciliation in such a large scale approach and no studies using this methodology to detect HGTs in multiple pathogens.
Preceding the experiment to decide the research question was the search for the optimal parameters of the used software.

Parameter search for the alien index

For three organisms phylogenetic trees were calculated for the whole proteom and the impact of different alien indices on these sets was analyzed retrospectively. It was shown that an alien index of 0.5 reliably removes prokaryotic proteins that have no homologs in eukaryotes (e.g. household proteins) but leaves the set of proteins with ELDs intact. It removes between 40-50% of the initial amount of proteins.

Parameter search for tree reconstruction and multiple-alignment

For parameter validation of the tree reconstruction method the results obtained for Legionella pneumophila str. Paris were compared to the results from previous studies with this organism. The used parameters for the phylogenetic tree reconstruction led to similar results except for homologs with little coverage. But it should be kept in mind that

there is no standard or truth against which phylogenetic trees can be compared - so even when the overlap with previous studies is reassuring the accuracy of the phylogenetic trees should be handled with caution.

Parameters for tree analysis

For automatic tree analysis the JAVA package PHAT was used. This program analyses the relationship of different taxa in a phylogenetic tree based on a search string provided by the user. The aim was to find tree topologies that are the result of a HGT event with eukaryotes as gene donors and prokaryotes as gene acceptors - this was done by searching for monohpyletic branches of the tree which contain eukaryotes as well as prokaryotes. Rooting was done for the node of the seed species.
But there are also other possible HGT events that shape the genome of pathogens - e.g., HGT events from bacteria to bacteria which are not detectable with this method and combined with HGT events from eukaryotes can lead to wrongly assigned donor phyla or undetected HGT events with eukaryotes as gene donors.
Different search strings were used for three bacteria and the results compared to manually positively classified trees for these prokaryotes - this approach resulted in a search string with a sensitivity of 95 % and specificity of 93 % for this three organisms. The findings in this thesis are based on the sensitivity and specificity of the used phat search string and the assumption that the conclusive results for the final phat string in the three tested organisms can be generalized to other pathogens. Before further discussing the results it should be stated that manually classification of HGT vs non-HGT was rarely trivial - a certain error rate should and must be considered.
The final phat string searched for nodes with more than three eukaryotes, the seed species and less than 20 bacteria which are not from the taxonomic family of the seed species. This search string excludes some evolutionary scenarios: horizontal gene transfer from genes that are only present in one eukaryotic species or early HGT events to a large monophyletic family of pathogens. But despite these constrains the benefits prevailed - contamination of prokaryotic genomes with eukaryotic genes was avoided by requiring a

minimum of three eukaryotes and the constrain of a maximum of 20 bacteria excluded positive nodes near the root in trees with prokaryotes and eukaryotes.

Comparison of different tree reconstruction methods

Due to the high requirements maximum-likelihood methods like raxML have in regards of time and CPU consumption the decision reached in this thesis to use a neighbor-joining method like mtrees while facing the proteom of 70 bacteria was easy. Still, the fact remains that maximum-likelihood methods better reflect the evolutionary reality and their resolution of sequence similarities near the root is a given - therefore it was important to estimate the differences in the outcome of the experiment resulting from the decision against a maximum-likelihood method. This was done using phylogenetic trees from the proteom of Chlamydia trachomatis D/UW-3/CX obtained with a maximum-likelihood method and a neighbor joining method. The two different sets of trees were analyzed using the previously discussed PHAT string and the results were compared. It was shown that these two different methods were equivalent in their results. For developing a methodology which is suitable to investigate the relationship between ELDs and HGTs in a large scale approach this step was crucial - if the results between this two methods had differ to a larger extend the tree reconstruction would have become either unreliable with neighbor-joining methods or very time and CPU consuming with maximum likelihood methods.

Subsequent to the parameter search was the main experiment - analyzing the amount of horizontal-gene transfer events in proteins with eukaryotic-like domains in comparison with protein without eukaryotic-like domains.
The results - using full length proteins - showed for 14 bacteria significant p-values ($p \leq 0.5$) - the others were tested negative for a correlation between ELDs and HGTs. Most of the positively tested organisms are highly pathogenic and virulent (e.g. Haemophilus inuenzae R2846, Coxiella burnetii RSA 493, Francisella tularensis subsp. tularensis

NE061598), some are known for their numerous HGTs (e.g. Legionella pneumophila str. Lens and Paris) and others are tightly associated with either a specific host organism or an obligate intracellular lifestyle (e.g. Rickettsia typhi str. Wilmington, Wolbachia endosymbiont of Culex quinquefasciatus Pel, Onion yellows phytoplasma OY-M). It appears as if the conclusion from these findings may very well indicate different types of pathogens - e.g. one of the types is evolutionary adapted to a certain host via HGTs (among other factors) obtained from eukaryotes and another type is either not competent for HGTs or their ability to infect eukaryotes is not based on genetic material obtained from an eukaryotic donor. For the last group it is hard to eliminate the possibility of ancient horizontal gene transfer which is not detectable with the implemented methodology (rendering the distinction non-existing, even though it may exist at certain points in time) - but this possibility is also true for the first group.

Variation in the ELD score did not show a clear trend (higher ELD score did not show an increased significants).
Analyzing the phylogenetic relationship between the HGT donor and the acceptor for proteins which were classified as being of eukaryotic origin and having an ELD score above 4 showed that donors are very often not from the phyla the bacteria infects primarily - e.g., Xanthomonas campestris pv. Vesicatoria is a plant pathogen yet its two ELD containing effectors come from fungi, the same situation applies to Sinorhizobium fredii and its three effectors and Pseudomonas aeruginosa DK2 (human pathogen yet its effector proteins are derived from Fungi). But it must be said that it was often not easy to interpret the tree topology and it was seldomly possible to determine the origin unambiguous. Furthermore a HGT scenario with just a single HGT event was rarely enough to explain the topology of the tree.

The result changed when using domain sequences instead of the full length protein sequences - it was possible to obtain significant p-values in 33 out of 60 analyzed proteoms. These set of pathogens - the overlap is discussed below - contains bacteria which are

pathogenic (e.g. Bacillus thuringiensis BMB171, Bartonella quintana str. Toulouse, Campylobacter fetus subsp. fetus 82-40, Chlamydophila pneumoniae LP-CoLN, Erwinia amylovora ATCC 49946), ubiquitous present in a wide variety of habitats (e.g. Shewanella putrefaciens 200, Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111) and Symbionts (e.g. Arcobacter nitroř gilis, Bacteroides vulgatus ATCC 8482, Biřdobacterium longum subsp. incantis 157F, Bifidobacterium animals subs. lactis DSM 10140). It appears that the majority of these group are also highly pathogenic prokaryotes - but there are also symbionts and pathogens that appear mostly as parasites of immunodeficient organisms.

Merging the resulting organism lists gave an overlap of 8 prokaryotes. These prokaryotes are either known for their numerous HGT events (Legionella pneumophila str. Lens, Legionella pneumophila str. Paris, Legionella pneumophila subsp. pneumophila str. Philadelphia 1), their endosymbiotic life-style (Wolbachia endosymbiont of Culex quinquefasciatus Pel) or their high virulence and pathogenicity (Coxiella burnetii RSA 493, Francisella tularensis subsp. tularensis NE061598, Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130). It seems like the relationship between ELDs and eukaryotic origin is much stronger on domain sequence than on protein sequence level. No clear trend between number of ELDs and HGTs in these ELDs was visible.

For further work the question remain if effectors are a closed group of proteins that were transferred at some point in the evolution of a bacteria which marked its transformation to a pathogen - therefore setting the focus on the function of this set of proteins since they may be the key to understand the evolution of pathogens in general. Another possibility is that proteins are continuously obtained during the evolution of prokaryotes - leading to new pathogens and/or more virulent bacteria. The finding of the thesis point in both directions to some extend. On protein sequence level only very virulent pathogens show a significant link between ELDs and HGTs, pointing in the direction that these proteins are a finite set of effectors - and only pathogens have obtained them.

On domain level the second scenario can be favoured - highly virulent pathogens have a certain set of domains but also commensal bacteria have obtained some in the past. This leaves the question if there is a general functional difference between sets of pathogenic effectors and effectors in symbiotic bacteria, and also if a symbiotic prokaryote can become pathogenic by obtaining further effectors.

Another topic for further work is the in-depth analysis of the obtained phylogenetic trees which were only analyzed to answer the question of HGT events between eukaryotes and prokaryotes. It would be very interesting to see if prokaryotes with a greater number of such HGT events also have a tendency to obtain genetic material from other prokaryotes. It would also be very interesting to try and assign time frames to HGT events relative to each other in a pathogen - this might give insight in the evolution of pathogens. This could be done by analyzing HGTs with eukaryotes as donors on different taxonomic levels - in this thesis the HGTs were only examined on family level - other levels could be used as well and would provide this kind of temporal information.

Bibliography

[Al-Khodor *et al.*, 2008] Al-Khodor, Souhaila, Price, Christopher T, Habyarimana, Fabien, Kalia, Awdhesh, & Kwaik, Yousef Abu. 2008. A Dot/Icm-translocated ankyrin protein of Legionella pneumophilais required for intracellular proliferation within human macrophages and protozoa. *Molecular Microbiology*, Nov, 1–16.

[Alsmark *et al.*, 2004] Alsmark, Cecilia M, Frank, A Carolin, Karlberg, E Olof, Legault, Boris-Antoine, Ardell, David H, Canbäck, Björn, Eriksson, Ann-Sofie, Näslund, A Kristina, Handley, Scott A, Huvet, Maxime, Scola, Bernard La, Holmberg, Martin, & Andersson, Siv G E. 2004. The louse-borne human pathogen Bartonella quintana is a genomic derivative of the zoonotic agent Bartonella henselae. *Proc Natl Acad Sci USA*, **101**(26), 9716–21.

[Altekruse *et al.*, 1999] Altekruse, S F, Stern, N J, Fields, P I, & Swerdlow, D L. 1999. Campylobacter jejuni–an emerging foodborne pathogen. *Emerging Infect Dis*, **5**(1), 28–35.

[Altschul *et al.*, 1990] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–10.

[Alvarez-Ordóñez *et al.*, 2011] Alvarez-Ordóñez, Avelino, Begley, Máire, Prieto, Miguel, Messens, Winy, López, Mercedes, Bernardo, Ana, & Hill, Colin. 2011. Salmonella spp. survival strategies within the host gastrointestinal tract. *Microbiology*, **157**(Pt 12), 3268–81.

[Andersson, 2005] Andersson, J. O. 2005. Lateral gene transfer in eukaryotes. *CMLS, Cell. Mol. Life Sci.*, **62**(11), 1182–1197.

[Arnold *et al.*, 2012] Arnold, Roland, Boonen, Kurt, Sun, Mark G F, & Kim, Philip M. 2012. Computational analysis of interactomes: current and future perspectives for bioinformatics approaches to model the host-pathogen interaction space. *Methods*, **57**(4), 508–18.

[Arthur *et al.*, 2012] Arthur, J. C, Perez-Chanona, E, Muhlbauer, M, Tomkovich, S, Uronis, J. M, Fan, T.-J, Campbell, B. J, Abujamel, T, Dogan, B, Rogers, A. B, Rhodes, J. M, Stintzi, A, Simpson, K. W, Hansen, J. J, Keku, T. O, Fodor, A. A, & Jobin, C. 2012. Intestinal Inflammation Targets Cancer-Inducing Activity of the Microbiota. *Science*, **338**(6103), 120–123.

[Babic *et al.*, 2008] Babic, Ana, Lindner, Ariel B, Vulic, Marin, Stewart, Eric J, & Radman, Miroslav. 2008. Direct visualization of horizontal gene transfer. *Science*, **319**(5869), 1533–6.

[Baldani & Baldani, 2005] Baldani, José I, & Baldani, Vera L D. 2005. History on the biological nitrogen fixation research in graminaceous plants: special emphasis on the Brazilian experience. *An Acad Bras Cienc*, **77**(3), 549–79.

[Barrangou *et al.*, 2009] Barrangou, Rodolphe, Briczinski, Elizabeth P, Traeger, Lindsay L, Loquasto, Joseph R, Richards, Melissa, Horvath, Philippe, Coûté-Monvoisin, Anne-Claire, Leyer, Gregory, Rendulic, Snjezana, Steele, James L, Broadbent, Jeffery R, Oberg, Taylor, Dudley, Edward G, Schuster, Stephan, Romero, Dennis A, & Roberts, Robert F. 2009. Comparison of the complete genome sequences of Bifidobacterium animalis subsp. lactis DSM 10140 and Bl-04. *Journal of Bacteriology*, **191**(13), 4144–51.

[Beeckman & Vanrompay, 2010] Beeckman, Delphine Sylvie Anne, & Vanrompay, Daisy C G. 2010. Bacterial secretion systems with an emphasis on the chlamydial Type III secretion system. *Curr Issues Mol Biol*, **12**(1), 17–41.

[Bendtsen *et al.*, 2004] Bendtsen, Jannick Dyrløv, Nielsen, Henrik, von Heijne, Gunnar, & Brunak, Søren. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**(4), 783–95.

[Bingle *et al.*, 2008] Bingle, Lewis Eh, Bailey, Christopher M, & Pallen, Mark J. 2008. Type VI secretion: a beginner's guide. *Curr Opin Microbiol*, **11**(1), 3–8.

[Bjarnsholt *et al.*, 2010] Bjarnsholt, Thomas, Jensen, Peter Østrup, Jakobsen, Tim Holm, Phipps, Richard, Nielsen, Anne Kirstine, Rybtke, Morten Theil, Tolker-Nielsen, Tim, Givskov, Michael, Høiby, Niels, Ciofu, Oana, & Consortium, Scandinavian Cystic Fibrosis Study. 2010. Quorum sensing and virulence of Pseudomonas aeruginosa during lung infection of cystic fibrosis patients. *PLoS ONE*, **5**(4), e10115.

[Boc *et al.*, 2010] Boc, Alix, Philippe, Hervé, & Makarenkov, Vladimir. 2010. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Syst Biol*, **59**(2), 195–211.

[Boc *et al.*, 2012] Boc, Alix, Diallo, Alpha Boubacar, & Makarenkov, Vladimir. 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res*, **40**(Web Server issue), W573–9.

[Boch *et al.*, 2009] Boch, Jens, Scholze, Heidi, Schornack, Sebastian, Landgraf, Angelika, Hahn, Simone, Kay, Sabine, Lahaye, Thomas, Nickstadt, Anja, & Bonas, Ulla. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**(5959), 1509–12.

[Bogdanos & Vergani, 2009] Bogdanos, Dimitrios P, & Vergani, Diego. 2009. Bacteria and primary biliary cirrhosis. *Clinic Rev Allerg Immunol*, **36**(1), 30–9.

[Boschetti *et al.*, 2012] Boschetti, Chiara, Carr, Adrian, Crisp, Alastair, Eyres, Isobel, Wang-Koh, Yuan, Lubzens, Esther, Barraclough, Timothy G, Micklem, Gos, & Tunnacliffe, Alan. 2012. Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers. *PLoS Genet*, **8**(11), e1003035.

[Boto, 2010] Boto, Luis. 2010. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences*, **277**(1683), 819–27.

[Boudewijns *et al.*, 2003] Boudewijns, M., Magerman, K., Verhaegen, J., Debrock, G., Peetermans, W. E., Donkersloot, P., Mewis, A., Peeters, V., Rummens, J. L., & Cartuyvels, R. 2003. Rothia dentocariosa, endocarditis and mycotic aneurysms: case report and review of the literature. *Clinical Microbiology and Infection*, **9**(3), 222–229.

[Brinkman *et al.*, 2002] Brinkman, Fiona S L, Blanchard, Jeffrey L, Cherkasov, Artem, Av-Gay, Yossef, Brunham, Robert C, Fernandez, Rachel C, Finlay, B Brett, Otto, Sarah P, Ouellette, B F Francis, Keeling, Patrick J, Rose, Ann M, Hancock, Robert E W, Jones, Steven J M, & Greberg, Hans. 2002. Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. *Genome Research*, **12**(8), 1159–67.

[Brüggemann *et al.*, 2006] Brüggemann, Holger, Cazalet, Christel, & Buchrieser, Carmen. 2006. Adaptation of Legionella pneumophila to the host environment: role of protein secretion, effectors and eukaryotic-like proteins. *Curr Opin Microbiol*, **9**(1), 86–94.

[Brüssow *et al.*, 2004] Brüssow, Harald, Canchaya, Carlos, & Hardt, Wolf-Dietrich. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev*, **68**(3), 560–602, table of contents.

[Burstein *et al.*, 2009] Burstein, David, Zusman, Tal, Degtyar, Elena, Viner, Ram, Segal, Gil, & Pupko, Tal. 2009. Genome-Scale Identification of Legionella pneumophila Effectors Using a Machine Learning Approach. *PLoS Pathog*, **5**(7), e1000508.

[Cao & Saier, 2003] Cao, Thien B, & Saier, Milton H. 2003. The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. *Biochim Biophys Acta*, **1609**(1), 115–25.

[Carratala & Garcia-Vidal, 2010] Carratala, Jordi, & Garcia-Vidal, Carolina. 2010. An update on Legionella. *Current Opinion in infectious diseases*, **23**(2), 152–157.

[Cascales, 2008] Cascales, Eric. 2008. The type VI secretion toolkit. *EMBO Rep*, **9**(8), 735–41.

[Cazalet *et al.*, 2004] Cazalet, Christel, Rusniok, Christophe, Brüggemann, Holger, Zidane, Nora, Magnier, Arnaud, Ma, Laurence, Tichit, Magalie, Jarraud, Sophie, Bouchier, Christiane, Vandenesch, François, Kunst, Frank, Etienne, Jérôme, Glaser, Philippe, & Buchrieser, Carmen. 2004. Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity. *Nat Genet*, **36**(11), 1165–73.

[Cazalet *et al.*, 2010] Cazalet, Christel, Gomez-Valero, Laura, Rusniok, Christophe, Lomma, Mariella, Dervins-Ravault, Delphine, Newton, Hayley J, Sansom, Fiona M, Jarraud, Sophie, Zidane, Nora, Ma, Laurence, Bouchier, Christiane, Etienne, Jerôme, Hartland, Elizabeth L, & Buchrieser, Carmen. 2010. Analysis of the Legionella longbeachae genome and transcriptome uncovers unique strategies to cause Legionnaires' disease. *PLoS Genet*, **6**(2), e1000851.

[Cejková *et al.*, 2012] Cejková, Darina, Zobaníková, Marie, Chen, Lei, Pospíšilová, Petra, Strouhal, Michal, Qin, Xiang, Mikalová, Lenka, Norris, Steven J, Muzny, Donna M, Gibbs, Richard A, Fulton, Lucinda L, Sodergren, Erica, Weinstock, George M, & Smajs, David. 2012. Whole genome sequences of three Treponema pallidum ssp. pertenue strains: yaws and syphilis treponemes differ in less than 0.2sequence. *PLoS Negl Trop Dis*, **6**(1), e1471.

[Chen & Dubnau, 2004] Chen, Inês, & Dubnau, David. 2004. DNA uptake during bacterial transformation. *Nat Rev Micro*, **2**(3), 241–9.

[Chothia, 1992] Chothia, C. 1992. Proteins. One thousand families for the molecular biologist. *Nature*, **357**(6379), 543–4.

[Crow *et al.*, 2012] Crow, Allister, Hughes, Richard K, Taieb, Frédéric, Oswald, Eric, & Banfield, Mark J. 2012. The molecular basis of ubiquitin-like protein NEDD8 deamidation by the bacterial effector protein Cif. *Proceedings of the National Academy of Sciences*, **109**(27), E1830–8.

[D'Auria *et al.*, 2011] D'Auria, Giuseppe, Galán, Juan-Carlos, Rodríguez-Alcayna, Manuel, Moya, Andrés, Baquero, Fernando, & Latorre, Amparo. 2011. Complete genome sequence of Acidaminococcus intestini RYC-MR95, a Gram-negative bacterium from the phylum Firmicutes. *Journal of Bacteriology*, **193**(24), 7008–9.

[Davison, 1999] Davison, J. 1999. Genetic exchange between bacteria in the environment. *Plasmid*, **42**(2), 73–91.

[de Chassey *et al.*, 2012] de Chassey, Benoît, Meyniel-Schicklin, Laurène, Aublin-Gex, Anne, André, Patrice, & Lotteau, Vincent. 2012. New horizons for antiviral drug discovery from virus-host protein interaction networks. *Current Opinion in Virology*, **2**(5), 606–13.

[de Felipe *et al.*, 2005] de Felipe, Karim Suwwan, Pampou, Sergey, Jovanovic, Oliver S, Pericone, Christopher D, Ye, Senna F, Kalachikov, Sergey, & Shuman, Howard A. 2005. Evidence for acquisition of Legionella type IV secretion substrates via interdomain horizontal gene transfer. *J Bacteriol*, **187**(22), 7716–26.

[Dean, 2011] Dean, Paul. 2011. Functional domains and motifs of bacterial type III effector proteins and their roles in infection. *FEMS Microbiology Reviews*, **35**(6), 1100–1125.

[Deleo *et al.*, 2009] Deleo, Frank R, Diep, Binh An, & Otto, Michael. 2009. Host defense and pathogenesis in Staphylococcus aureus infections. *Infectious Disease Clinics of North America*, **23**(1), 17–34.

[Delepelaire, 2004] Delepelaire, P. 2004. Type I secretion in gram-negative bacteria. *Biochim Biophys Acta*, **1694**(1-3), 149–61.

[Dhakal *et al.*, 2008] Dhakal, B K, Kulesus, R R, & Mulvey, M A. 2008. Mechanisms and consequences of bladder cell invasion by uropathogenic Escherichia coli. *European Journal of Clinical Investigation*, **38 Suppl 2**(Oct), 2–11.

[Diard *et al.*, 2013] Diard, Médéric, Garcia, Victor, Maier, Lisa, Remus-Emsermann, Mitja N P, Regoes, Roland R, Ackermann, Martin, & Hardt, Wolf-Dietrich. 2013. Stabilization of cooperative virulence by the expression of an avirulent phenotype. *Nature*, **494**(7437), 353–6.

[Doolittle, 1999a] Doolittle, W F. 1999a. Lateral genomics. *Trends Cell Biol*, **9**(12), M5–8.

[Doolittle, 1999b] Doolittle, W. F. 1999b. Phylogenetic Classification and the Universal Tree. *Science*, **284**(5423), 2124–2128.

[dos Santos *et al.*, 2012] dos Santos, Andrea P, Guimaraes, Ana M S, do Nascimento, Naíla C, SanMiguel, Phillip J, & Messick, Joanne B. 2012. Complete genome sequence of Mycoplasma wenyonii strain Massachusetts. *Journal of Bacteriology*, **194**(19), 5458–9.

[Douzi *et al.*, 2012] Douzi, Badreddine, Filloux, Alain, & Voulhoux, Romé. 2012. On the path to uncover the bacterial type II secretion system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**(1592), 1059–72.

[du Plessis *et al.*, 2011] du Plessis, David J F, Nouwen, Nico, & Driessen, Arnold J M. 2011. The Sec translocase. *Biochim Biophys Acta*, **1808**(3), 851–65.

[Duan *et al.*, 2009] Duan, Yongping, Zhou, Lijuan, Hall, David G, Li, Wenbin, Doddapaneni, Harshavardhan, Lin, Hong, Liu, Li, Vahling, Cheryl M, Gabriel, Dean W, Williams, Kelly P, Dickerman, Allan, Sun, Yijun, & Gottwald, Tim. 2009. Complete genome sequence of citrus huanglongbing bacterium, 'Candidatus Liberibacter asiaticus' obtained through metagenomics. *Mol Plant Microbe Interact*, **22**(8), 1011–20.

[et al., 1995] et al., Fleischmann. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **169**(5223), 496–512.

[Felsenstein, 1981] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, **17**(6), 368–76.

[Foulds & Graham, 1982] Foulds, L R, & Graham, R L. 1982. The Steiner Problem in Phylogeny is NP-Complete. *Advances in Applied Mathematics*, **3**(May), 43–49.

[Fricke *et al.*, 2011] Fricke, W Florian, Mammel, Mark K, McDermott, Patrick F, Tartera, Carmen, White, David G, Leclerc, J Eugene, Ravel, Jacques, & Cebula, Thomas A. 2011. Comparative genomics of 28 Salmonella enterica isolates: evidence for CRISPR-mediated adaptive sublineage evolution. *Journal of Bacteriology*, **193**(14), 3556–68.

[Frickey & Lupas, 2004] Frickey, Tancred, & Lupas, Andrei N. 2004. PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Research*, **32**(17), 5231–8.

[Fronzes *et al.*, 2009] Fronzes, Rémi, Christie, Peter J, & Waksman, Gabriel. 2009. The structural biology of type IV secretion systems. *Nat Rev Micro*, **7**(10), 703–14.

[Fukami-Kobayashi *et al.*, 2007] Fukami-Kobayashi, Kaoru, Minezaki, Yoshiaki, Tateno, Yoshio, & Nishikawa, Ken. 2007. A tree of life based on protein domain organizations. *Molecular Biology and Evolution*, **24**(5), 1181–9.

[Furukawa & Dale, 2013] Furukawa, Yoko, & Dale, Jason R. 2013. The surface properties of Shewanella putrefaciens 200 and S. oneidensis MR-1: the effect of pH and terminal electron acceptors. *Geochemical Transactions*, **14**(1), 3.

[García *et al.*, 2004] García, C, Ugalde, E, Campo, A B, Miñambres, E, & Kovács, N. 2004. Fatal case of community-acquired pneumonia caused by Legionella longbeachae in a patient with systemic lupus erythematosus. *Eur J Clin Microbiol Infect Dis*, **23**(2), 116–8.

[Genereux & Logsdorn, 2003] Genereux, D P, & Logsdorn, J M. 2003. Much ado about bacteria-to-vertebrate lateral gene transfer. *TRENDS in Genetics*, **19**(4), 191–195.

[Genova *et al.*, 2011] Genova, Suzanne G, Streeter, Robert N, Velguth, Karen E, Snider, Timothy A, Kocan, Katherine M, & Simpson, Katharine M. 2011. Severe anemia associated with Mycoplasma wenyonii infection in a mature cow. *Can Vet J*, **52**(9), 1018–21.

[Gilmour *et al.*, 2010] Gilmour, Matthew W, Graham, Morag, Domselaar, Gary Van, Tyler, Shaun, Kent, Heather, Trout-Yakel, Keri M, Larios, Oscar, Allen, Vanessa, Lee, Barbara, & Nadon, Celine. 2010. High-throughput genome sequencing of two Listeria monocytogenes clinical isolates during a large foodborne outbreak. *BMC Genomics*, **11**(Jan), 120.

[Gladyshev *et al.*, 2008] Gladyshev, Eugene A, Meselson, Matthew, & Arkhipova, Irina R. 2008. Massive horizontal gene transfer in bdelloid rotifers. *Science*, **320**(5880), 1210–3.

[Gogarten & Townsend, 2005] Gogarten, J. Peter, & Townsend, Jeffrey P. 2005. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Micro*, **3**(9), 679–687.

[Goldenfeld & Woese, 2007] Goldenfeld, Nigel, & Woese, Carl. 2007. Biology's next revolution. *Nature*, **445**(7126), 369.

[Gomez-Valero *et al.*, 2009] Gomez-Valero, L, Rusniok, C, & Buchrieser, C. 2009. Legionella pneumophila: population genetics, phylogeny and genomics. *Infect Genet Evol*, **9**(5), 727–39.

[Gomez-Valero *et al.*, 2011a] Gomez-Valero, Laura, Rusniok, Christophe, Cazalet, Christel, & Buchrieser, Carmen. 2011a. Comparative and functional genomics of legionella identified eukaryotic like proteins as key players in host-pathogen interactions. *Front. Microbio.*, **2**(Jan), 208.

[Gomez-Valero *et al.*, 2011b] Gomez-Valero, Laura, Rusniok, Christophe, Jarraud, Sophie, Vacherie, Benoit, Rouy, Zoé, Barbe, Valerie, Medigue, Claudine, Etienne, Jerome, & Buchrieser, Carmen. 2011b. Extensive recombination events and horizontal gene transfer shaped the Legionella pneumophila genomes. *BMC Genomics*, **12**(Jan), 536.

[Guettler *et al.*, 1999] Guettler, M V, Rumler, D, & Jain, M K. 1999. Actinobacillus succinogenes sp. nov., a novel succinic-acid-producing strain from the bovine rumen. *Int J Syst Bacteriol*, **49 Pt 1**(Jan), 207–16.

[Gupta, 1998] Gupta, R S. 1998. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev*, **62**(4), 1435–91.

[Gürlebeck *et al.*, 2006] Gürlebeck, Doreen, Thieme, Frank, & Bonas, Ulla. 2006. Type III effector proteins from the plant pathogen Xanthomonas and their role in the interaction with the host plant. *J Plant Physiol*, **163**(3), 233–55.

[Hacker & Carniel, 2001] Hacker, J, & Carniel, E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep*, **2**(5), 376–81.

[Hahn *et al.*, 2009] Hahn, Martin W, Lang, Elke, Brandt, Ulrike, Wu, Qinglong L, & Scheuerl, Thomas. 2009. Emended description of the genus Polynucleobacter and the species Polynucleobacter necessarius and proposal of two subspecies, P. necessarius subsp. necessarius subsp. nov. and P. necessarius subsp. asymbioticus subsp. nov. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, **59**(Pt 8), 2002–9.

[Han & Zmasek, 2009] Han, Mira V, & Zmasek, Christian M. 2009. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**(Jan), 356.

[He *et al.*, 2010] He, Jin, Shao, Xiaohu, Zheng, Huajun, Li, Mingshun, Wang, Jieping, Zhang, Qingye, Li, Lin, Liu, Ziduo, Sun, Ming, Wang, Shengyue, & Yu, Ziniu. 2010. Complete genome sequence of Bacillus thuringiensis mutant strain BMB171. *Journal of Bacteriology*, **192**(15), 4074–5.

[Hegyi & Gerstein, 1999] Hegyi, H, & Gerstein, M. 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol*, **288**(1), 147–64.

[Henderson *et al.*, 2004] Henderson, Ian R, Navarro-Garcia, Fernando, Desvaux, Mickaël, Fernandez, Rachel C, & Ala'Aldeen, Dlawer. 2004. Type V protein secretion pathway: the autotransporter story. *Microbiol Mol Biol Rev*, **68**(4), 692–744.

[Herron-Olson *et al.*, 2007] Herron-Olson, Lisa, Fitzgerald, J Ross, Musser, James M, & Kapur, Vivek. 2007. Molecular correlates of host specialization in Staphylococcus aureus. *PLoS ONE*, **2**(10), e1120.

[Highlander *et al.*, 2007] Highlander, Sarah K, Hultén, Kristina G, Qin, Xiang, Jiang, Huaiyang, Yerrapragada, Shailaja, Mason, Edward O, Shang, Yue, Williams, Tiffany M, Fortunov, Régine M, Liu, Yamei, Igboeli, Okezie, Petrosino, Joseph, Tirumalai, Madhan, Uzman, Akif, Fox, George E, Cardenas, Ana Maria, Muzny, Donna M, Hemphill, Lisa, Ding, Yan, Dugan, Shannon, Blyth, Peter R, Buhay, Christian J, Dinh, Huyen H, Hawes, Alicia C, Holder, Michael, Kovar, Christie L, Lee, Sandra L, Liu, Wen, Nazareth, Lynne V, Wang, Qiaoyan, Zhou, Jianling, Kaplan, Sheldon L, & Weinstock, George M. 2007. Subtle genetic changes enhance virulence of methicillin resistant and sensitive Staphylococcus aureus. *BMC Microbiol*, **7**(Jan), 99.

[Hilario & Gogarten, 1993] Hilario, Elena, & Gogarten, Johann Peter. 1993. Horizontal transfer of {ATPase} genes — the tree of life becomes a net of life. *Biosystems*, **31**(2–3), 111 – 119.

[Hotopp, 2011] Hotopp, Julie C Dunning. 2011. Horizontal gene transfer between bacteria and animals. *Trends Genet*, **27**(4), 157–63.

[Jehl *et al.*, 2011] Jehl, Marc-André, Arnold, Roland, & Rattei, Thomas. 2011. Effective–a database of predicted secreted bacterial proteins. *Nucleic Acids Research*, **39**(Database issue), D591–5.

[Jong *et al.*, 2010] Jong, Bouke C De, Antonio, Martin, & Gagneux, Sebastien. 2010. Mycobacterium africanum–review of an important cause of human tuberculosis in West Africa. *PLoS Negl Trop Dis*, **4**(9), e744.

[Jousimies-Somer, 1997] Jousimies-Somer, H. 1997. Recently described clinically important anaerobic bacteria: taxonomic aspects and update. *Clin Infect Dis*, **25 Suppl 2**(Sep), S78–87.

[Kabir, 2011] Kabir, Shahjahan. 2011. The role of interleukin-17 in the Helicobacter pylori induced infection and immunity. *Helicobacter*, **16**(1), 1–8.

[Kalman *et al.*, 1999] Kalman, S, Mitchell, W, Marathe, R, Lammel, C, Fan, J, Hyman, R W, Olinger, L, Grimwood, J, Davis, R W, & Stephens, R S. 1999. Comparative genomes of Chlamydia pneumoniae and C. trachomatis. *Nat Genet*, **21**(4), 385–9.

[Klasson *et al.*, 2008] Klasson, Lisa, Walker, Thomas, Sebaihia, Mohammed, Sanders, Mandy J, Quail, Michael A, Lord, Angela, Sanders, Susanne, Earl, Julie, O'Neill, Scott L, Thomson, Nicholas, Sinkins, Steven P, & Parkhill, Julian. 2008. Genome evolution of Wolbachia strain wPip from the Culex pipiens group. *Molecular Biology and Evolution*, **25**(9), 1877–87.

[Koonin, 2005] Koonin, Eugene V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**(Jan), 309–38.

[Koonin *et al.*, 2001] Koonin, Eugene V, Makarova, Kira S, & Aravind, L. 2001. HORIZONTAL GENE TRANSFER IN PROKARYOTES: Quantification and Classification. *Annu. Rev. Microbiol.*, **55**(Aug), 709–42.

[Kumar *et al.*, 2003] Kumar, P S, Griffen, A L, Barton, J A, Paster, B J, Moeschberger, M L, & Leys, E J. 2003. New bacterial species associated with chronic periodontitis. *J Dent Res*, **82**(5), 338–44.

[Ladhani, 2012] Ladhani, Shamez N. 2012. Two decades of experience with the Haemophilus influenzae serotype b conjugate vaccine in the United Kingdom. *Clin Ther*, **34**(2), 385–99.

[Lander *et al.*, 2001] Lander, E S, Linton, L M, Birren, B, & et al, C Nusbaum. 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

[Larkin *et al.*, 2007] Larkin, M A, Blackshields, G, Brown, N P, Chenna, R, McGettigan, P A, Mcwilliam, H, Valentin, F, Wallace, I M, Wilm, A, Lopez, R, Thompson, J D, Gibson, T J, & Higgins, D G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**(21), 2947–8.

[Larsson *et al.*, 2005] Larsson, Pär, Oyston, Petra C F, Chain, Patrick, Chu, May C, Duffield, Melanie, Fuxelius, Hans-Henrik, Garcia, Emilio, Hälltorp, Greger, Johansson, Daniel, Isherwood, Karen E, Karp, Peter D, Larsson, Eva, Liu, Ying, Michell, Stephen, Prior, Joann, Prior, Richard, Malfatti, Stephanie, Sjöstedt, Anders, Svensson, Kerstin, Thompson, Nick, Vergez, Lisa, Wagg, Jonathan K, Wren, Brendan W, Lindler, Luther E, Andersson, Siv G E, Forsman, Mats, & Titball, Richard W. 2005. The complete genome sequence of Francisella tularensis, the causative agent of tularemia. *Nat Genet*, **37**(2), 153–9.

[Lawrence & Ochman, 1998] Lawrence, J G, & Ochman, H. 1998. Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci USA*, **95**(16), 9413–7.

[Lawrence & Ochman, 2002] Lawrence, Jeffrey G, & Ochman, Howard. 2002. Reconciling the many faces of lateral gene transfer. *Trends Microbiol*, **10**(1), 1–4.

[Lee *et al.*, 2000] Lee, I M, Davis, R E, & Gundersen-Rindal, D E. 2000. Phytoplasma: phytopathogenic mollicutes. *Annu Rev Microbiol*, **54**(Jan), 221–55.

[Leroux *et al.*, 2012] Leroux, Michele, Leon, Justin A De, Kuwada, Nathan J, Russell, Alistair B, Pinto-Santini, Delia, Hood, Rachel D, Agnello, Danielle M, Robertson, Stephen M, Wiggins, Paul A, & Mougous, Joseph D. 2012. Quantitative single-cell characterization of bacterial interactions reveals type VI secretion is a double-edged sword. *Proceedings of the National Academy of Sciences*, **109**(48), 19804–9.

[Levisohn & Kleven, 2000] Levisohn, S., & Kleven, S.H. 2000. Avian mycoplasmosis (Mycoplasma gallisepticum). *Rev Sci Tech*, **19**(2), 425–42.

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

[Lorenz & Wackernagel, 1994] Lorenz, M G, & Wackernagel, W. 1994. Bacterial Gene Transfer by Natural Genetic Transformation in theEnvironment. *Microbiological Reviews*, **58**(3), 563–602.

[Lucas-Elío *et al.*, 2012] Lucas-Elío, Patricia, Goodwin, Lynne, Woyke, Tanja, Pitluck, Sam, Nolan, Matt, Kyrpides, Nikos C, Detter, Janine C, Copeland, Alex, Lu, Megan, Bruce, David, Detter, Chris, Tapia, Roxanne, Han, Shunsheng, Land, Miriam L, Ivanova, Natalia, Mikhailova, Natalia, Johnston, Andrew W B, & Sanchez-Amat, Antonio. 2012. Complete genome sequence of Marinomonas posidonica type strain (IVIA-Po-181(T)). *Stand. Genomic Sci.*, **7**(1), 31–43.

[Lurie-Weinberger *et al.*, 2010] Lurie-Weinberger, Mor N, Gomez-Valero, Laura, Merault, Nathalie, Glöckner, Gernot, Buchrieser, Carmen, & Gophna, Uri. 2010. The origins of eukaryotic-like proteins in Legionella pneumophila. *Int J Med Microbiol*, **300**(7), 470–81.

[Mak *et al.*, 2012] Mak, Amanda Nga-Sze, Bradley, Philip, Cernadas, Raul A, Bogdanove, Adam J, & Stoddard, Barry L. 2012. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science*, **335**(6069), 716–9.

[Margaret *et al.*, 2011] Margaret, Isabel, Becker, Anke, Blom, Jochen, Bonilla, Ildefonso, Goesmann, Alexander, Göttfert, Michael, Lloret, Javier, Mittard-Runte, Virginie, Rückert, Christian, Ruiz-Sainz, José E, Vinardell, José María, & Weidner, Stefan. 2011. Symbiotic properties and first analyses of the genomic sequence of the fast growing model strain Sinorhizobium fredii HH103 nodulating soybean. *Journal of Biotechnology*, **155**(1), 11–9.

[Maruyama *et al.*, 2009] Maruyama, Fumito, Kobata, Mitsuhiko, Kurokawa, Ken, Nishida, Keishin, Sakurai, Atsuo, Nakano, Kazuhiko, Nomura, Ryota, Kawabata, Shigetada, Ooshima, Takashi, Nakai, Kenta, Hattori, Masahira, Hamada, Shigeyuki, & Nakagawa,

Ichiro. 2009. Comparative genomic analyses of Streptococcus mutans provide insights into chromosomal shuffling and species-specific content. *BMC Genomics*, **10**(Jan), 358.

[Mattoo *et al.*, 2007] Mattoo, Seema, Lee, Yvonne M, & Dixon, Jack E. 2007. Interactions of bacterial effector proteins with host proteins. *Curr Opin Immunol*, **19**(4), 392–401.

[Mavrommatis *et al.*, 2009] Mavrommatis, Konstantinos, Pukall, Rüdiger, Rohde, Christine, Chen, Feng, Sims, David, Brettin, Thomas, Kuske, Cheryl, Detter, John C, Han, Cliff, Lapidus, Alla, Copeland, Alex, Rio, Tijana Glavina Del, Nolan, Matt, Lucas, Susan, Tice, Hope, Cheng, Jan-Fang, Bruce, David, Goodwin, Lynne, Pitluck, Sam, Ovchinnikova, Galina, Pati, Amrita, Ivanova, Natalia, Chen, Amy, Palaniappan, Krishna, Chain, Patrick, D'haeseleer, Patrik, Göker, Markus, Bristow, Jim, Eisen, Jonathan A, Markowitz, Victor, Hugenholtz, Philip, Rohde, Manfred, Klenk, Hans-Peter, & Kyrpides, Nikos C. 2009. Complete genome sequence of Cryptobacterium curtum type strain (12-3). *Stand. Genomic Sci.*, **1**(2), 93–100.

[Mckinlay *et al.*, 2010] Mckinlay, James B, Laivenieks, Maris, Schindler, Bryan D, Mckinlay, Anastasia A, Siddaramappa, Shivakumara, Challacombe, Jean F, Lowry, Stephen R, Clum, Alicia, Lapidus, Alla L, Burkhart, Kirk B, Harkins, Victoria, & Vieille, Claire. 2010. A genomic perspective on the potential of Actinobacillus succinogenes for industrial succinate production. *BMC Genomics*, **11**(Jan), 680.

[Meincke *et al.*, 2012] Meincke, Linda, Copeland, Alex, Lapidus, Alla, Lucas, Susan, Berry, Kerrie W, Rio, Tijana Glavina Del, Hammon, Nancy, Dalin, Eileen, Tice, Hope, Pitluck, Sam, Richardson, Paul, Bruce, David, Goodwin, Lynne, Han, Cliff, Tapia, Roxanne, Detter, John C, Schmutz, Jeremy, Brettin, Thomas, Larimer, Frank, Land, Miriam, Hauser, Loren, Kyrpides, Nikos C, Ivanova, Natalia, Göker, Markus, Woyke, Tanja, Wu, Qinglong L, Pöckl, Matthias, Hahn, Martin W, & Klenk, Hans-Peter. 2012. Complete genome sequence of Polynucleobacter necessarius subsp. asymbioticus type strain (QLW-P1DMWA-1(T)). *Stand. Genomic Sci.*, **6**(1), 74–83.

[Milani *et al.*, 2012] Milani, Andrea, Vecchietti, Davide, Rusmini, Ruggero, & Bertoni, Giovanni. 2012. TgpA, a protein with a eukaryotic-like transglutaminase domain, plays a critical role in the viability of Pseudomonas aeruginosa. *PLoS ONE*, **7**(11), e50323.

[Miller *et al.*, 2008] Miller, William G, Wang, Guilin, Binnewies, Tim T, & Parker, Craig T. 2008. The complete genome sequence and analysis of the human pathogen Campylobacter lari. *Foodborne Pathogens and Disease*, **5**(4), 371–86.

[Minion *et al.*, 2004] Minion, F Chris, Lefkowitz, Elliot J, Madsen, Melissa L, Cleary, Barbara J, Swartzell, Steven M, & Mahairas, Gregory G. 2004. The genome sequence of Mycoplasma hyopneumoniae strain 232, the agent of swine mycoplasmosis. *Journal of Bacteriology*, **186**(21), 7123–33.

[Mitchell *et al.*, 2010] Mitchell, Candice M, Hovis, Kelley M, Bavoil, Patrik M, Myers, Garry S A, Carrasco, Jose A, & Timms, Peter. 2010. Comparison of koala LPCoLN and human strains of Chlamydia pneumoniae highlights extended genetic diversity in the species. *BMC Genomics*, **11**(Jan), 442.

[Miyata *et al.*, 2013] Miyata, Sarah T, Bachmann, Verena, & Pukatzki, Stefan. 2013. Type VI secretion system regulation as a consequence of evolutionary pressure. *Journal of Medical Microbiology*, **62**(Pt 5), 663–76.

[Molins *et al.*, 2010] Molins, Claudia R, Delorey, Mark J, Yockey, Brook M, Young, John W, Sheldon, Sarah W, Reese, Sara M, Schriefer, Martin E, & Petersen, Jeannine M. 2010. Virulence differences among Francisella tularensis subsp. tularensis clades in mice. *PLoS ONE*, **5**(4), e10205.

[Morita *et al.*, 2011] Morita, Hidetoshi, Toh, Hidehiro, Oshima, Kenshiro, Yoshizaki, Mariko, Kawanishi, Michiko, Nakaya, Kohei, Suzuki, Takehito, Miyauchi, Eiji, Ishii, Yasuo, Tanabe, Soichi, Murakami, Masaru, & Hattori, Masahira. 2011. Complete genome

sequence and comparative analysis of the fish pathogen Lactococcus garvieae. *PLoS ONE*, **6**(8), e23184.

[Munoz-Dorado *et al.*, 1993] Munoz-Dorado, J, Inouye, S, & Inouye, M. 1993. Eukaryotic-like protein serine/threonine kinases in Myxococcus xanthus, a developmental bacterium exhibiting social behavior. *J Cell Biochem*, **51**(1), 29–33.

[Myers *et al.*, 2009] Myers, G S A, Mathews, S A, Eppinger, M, Mitchell, C, O'Brien, K K, White, O R, Benahmed, F, Brunham, R C, Read, T D, Ravel, J, Bavoil, P M, & Timms, P. 2009. Evidence that human Chlamydia pneumoniae was zoonotically acquired. *Journal of Bacteriology*, **191**(23), 7225–33.

[Nagai *et al.*, 2002] Nagai, Hiroki, Kagan, Jonathan C, Zhu, Xinjun, Kahn, Richard A, & Roy, Craig R. 2002. A bacterial guanine nucleotide exchange factor activates ARF on Legionella phagosomes. *Science*, **295**(5555), 679–82.

[Nakazawa *et al.*, 1999] Nakazawa, F, Poco, S E, Ikeda, T, Sato, M, Kalfas, S, Sundqvist, G, & Hoshino, E. 1999. Cryptobacterium curtum gen. nov., sp. nov., a new genus of gram-positive anaerobic rod isolated from human oral cavities. *Int J Syst Bacteriol*, **49 Pt 3**(Jul), 1193–200.

[Nalbantoglu *et al.*, 2010] Nalbantoglu, Ufuk, Sayood, Khalid, Dempsey, Michael P, Iwen, Peter C, Francesconi, Stephen C, Barabote, Ravi D, Xie, Gary, Brettin, Thomas S, Hinrichs, Steven H, & Fey, Paul D. 2010. Large direct repeats flank genomic rearrangements between a new clinical isolate of Francisella tularensis subsp. tularensis A1 and Schu S4. *PLoS ONE*, **5**(2), e9007.

[Nambu *et al.*, 2013] Nambu, Takayuki, Yamane, Kazuyoshi, Yamanaka, Takeshi, Mashimo, Chiho, Maruyama, Hugo, Yoshida, Masahiro, Hayashi, Hiroyuki, Leung, Kai-Poon, & Fukushima, Hisanori. 2013. Identification of disulphide stress-responsive extracytoplasmic function sigma factors in Rothia mucilaginosa. *Archives of Oral Biology*, **58**(6), 681–9.

[Nelson & Cox, 2004] Nelson, David L, & Cox, Michael M. 2004. *Lehninger Principles of Biochemistry*. Fourth edition edn. W. H. Freeman.

[Noto & Peek, 2012a] Noto, Jennifer M, & Peek, Richard M. 2012a. The Helicobacter pylori cag Pathogenicity Island. *Methods Mol Biol*, **921**(Jan), 41–50.

[Noto & Peek, 2012b] Noto, JenniferM., & Peek, RichardM., Jr. 2012b. Helicobacter pylori: An Overview. **921**, 7–10.

[Ochman *et al.*, 2000] Ochman, H, Lawrence, J G, & Groisman, E A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**(6784), 299–304.

[Opriessnig *et al.*, 2011] Opriessnig, T, Giménez-Lirola, L G, & Halbur, P G. 2011. Polymicrobial respiratory disease in pigs. *Anim. Health. Res. Rev.*, **12**(2), 133–48.

[O'Toole *et al.*, 2010] O'Toole, Paul W, Snelling, William J, Canchaya, Carlos, Forde, Brian M, Hardie, Kim R, Josenhans, Christine, Graham, Robert LJ, Mcmullan, Geoff, Parkhill, Julian, Belda, Eugenio, & Bentley, Stephen D. 2010. Comparative genomics and proteomics of Helicobacter mustelae, an ulcerogenic and carcinogenic gastric pathogen. *BMC Genomics*, **11**(Jan), 164.

[Pagani *et al.*, 2003] Pagani, Leonardo, Lang, Adolf, Vedovelli, Claudio, Moling, Oswald, Rimenti, Giovanni, Pristerà, Raffaele, & Mian, Peter. 2003. Soft tissue infection and bacteremia caused by Shewanella putrefaciens. *Journal of Clinical Microbiology*, **41**(5), 2240–1.

[Page, 1998] Page, R D. 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, **14**(9), 819–20.

[Pagel *et al.*, 2006] Pagel, Philipp, Oesterheld, Matthias, Stümpflen, Volker, & Frishman, Dmitrij. 2006. The DIMA web resource–exploring the protein domain network. *Bioinformatics*, **22**(8), 997–8.

[Papadopoulos & Agarwala, 2007] Papadopoulos, J. S, & Agarwala, R. 2007. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**(9), 1073–1079.

[Parkhill *et al.*, 2001] Parkhill, J, Wren, B W, Thomson, N R, Titball, R W, Holden, M T, Prentice, M B, Sebaihia, M, James, K D, Churcher, C, Mungall, K L, Baker, S, Basham, D, Bentley, S D, Brooks, K, Cerdeño-Tárraga, A M, Chillingworth, T, Cronin, A, Davies, R M, Davis, P, Dougan, G, Feltwell, T, Hamlin, N, Holroyd, S, Jagels, K, Karlyshev, A V, Leather, S, Moule, S, Oyston, P C, Quail, M, Rutherford, K, Simmonds, M, Skelton, J, Stevens, K, Whitehead, S, & Barrell, B G. 2001. Genome sequence of Yersinia pestis, the causative agent of plague. *Nature*, **413**(6855), 523–7.

[Parola *et al.*, 2005] Parola, Philippe, Davoust, Bernard, & Raoult, Didier. 2005. Tick- and flea-borne rickettsial emerging zoonoses. *Vet. Res.*, **36**(3), 469–92.

[Pati *et al.*, 2010] Pati, Amrita, Gronow, Sabine, Lapidus, Alla, Copeland, Alex, Rio, Tijana Glavina Del, Nolan, Matt, Lucas, Susan, Tice, Hope, Cheng, Jan-Fang, Han, Cliff, Chertkov, Olga, Bruce, David, Tapia, Roxanne, Goodwin, Lynne, Pitluck, Sam, Liolios, Konstantinos, Ivanova, Natalia, Mavromatis, Konstantinos, Chen, Amy, Palaniappan, Krishna, Land, Miriam, Hauser, Loren, Chang, Yun-Juan, Jeffries, Cynthia D, Detter, John C, Rohde, Manfred, Göker, Markus, Bristow, James, Eisen, Jonathan A, Markowitz, Victor, Hugenholtz, Philip, Klenk, Hans-Peter, & Kyrpides, Nikos C. 2010. Complete genome sequence of Arcobacter nitrofigilis type strain (CI). *Stand. Genomic Sci.*, **2**(3), 300–8.

[Pitman & Harrow, 2010] Pitman, Andrew R, & Harrow, Sally A. 2010. Genetic characterisation of Pectobacterium wasabiae causing soft rot disease of potato in New Zealand. *Eur J Plant Pathol*, **126**, 423–435.

[Ponting *et al.*, 1999] Ponting, C P, Aravind, L, Schultz, J, Bork, P, & Koonin, E V. 1999. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J Mol Biol*, **289**(4), 729–45.

[Price *et al.*, 2010] Price, Christopher T D, Al-Khodor, Souhaila, Al-Quadan, Tasneem, & Kwaik, Yousef Abu. 2010. Indispensable role for the eukaryotic-like ankyrin domains of the ankyrin B effector of Legionella pneumophila within macrophages and amoebae. *Infection and Immunity*, **78**(5), 2079–88.

[Punta *et al.*, 2012a] Punta, Marco, Coggill, Penny C, Eberhardt, Ruth Y, Mistry, Jaina, Tate, John, Boursnell, Chris, Pang, Ningze, Forslund, Kristoffer, Ceric, Goran, Clements, Jody, Heger, Andreas, Holm, Liisa, Sonnhammer, Erik L L, Eddy, Sean R, Bateman, Alex, & Finn, Robert D. 2012a. The Pfam protein families database. *Nucleic Acids Research*, **40**(Database issue), D290–301.

[Punta *et al.*, 2012b] Punta, Marco, Coggill, Penny C, Eberhardt, Ruth Y, Mistry, Jaina, Tate, John, Boursnell, Chris, Pang, Ningze, Forslund, Kristoffer, Ceric, Goran, Clements, Jody, Heger, Andreas, Holm, Liisa, Sonnhammer, Erik L L, Eddy, Sean R, Bateman, Alex, & Finn, Robert D. 2012b. The Pfam protein families database. *Nucleic Acids Research*, **40**(Database issue), D290–301.

[Rattei *et al.*, 2010] Rattei, Thomas, Tischler, Patrick, Götz, Stefan, Jehl, Marc-André, Hoser, Jonathan, Arnold, Roland, Conesa, Ana, & Mewes, Hans-Werner. 2010. SIMAP–a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Research*, **38**(Database issue), D223–6.

[Ren *et al.*, 2003] Ren, Shuang-Xi, Fu, Gang, Jiang, Xiu-Gao, Zeng, Rong, Miao, You-Gang, Xu, Hai, Zhang, Yi-Xuan, Xiong, Hui, Lu, Gang, Lu, Ling-Feng, Jiang, Hong-Quan, Jia, Jia, Tu, Yue-Feng, Jiang, Ju-Xing, Gu, Wen-Yi, Zhang, Yue-Qing, Cai, Zhen, Sheng, Hai-Hui, Yin, Hai-Feng, Zhang, Yi, Zhu, Gen-Feng, Wan, Ma, Huang, Hong-Lei, Qian, Zhen, Wang, Sheng-Yue, Ma, Wei, Yao, Zhi-Jian, Shen, Yan, Qiang, Bo-Qin, Xia, Qi-Chang, Guo, Xiao-Kui, Danchin, Antoine, Girons, Isabelle Saint, Somerville, Ronald L, Wen, Yu-Mei, Shi, Man-Hua, Chen, Zhu, Xu, Jian-Guo, & Zhao, Guo-Ping. 2003. Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature*, **422**(6934), 888–93.

[Riley *et al.*, 2013] Riley, David R, Sieber, Karsten B, Robinson, Kelly M, White, James Robert, Ganesan, Ashwinkumar, Nourbakhsh, Syrus, & Hotopp, Julie C Dunning. 2013. Bacteria-Human Somatic Cell Lateral Gene Transfer Is Enriched in Cancer Samples. *PLoS Comput Biol*, **9**(6), e1003107.

[Ruiz *et al.*, 2011] Ruiz, Jerônimo C, D'Afonseca, Vívian, Silva, Artur, Ali, Amjad, Pinto, Anne C, Santos, Anderson R, Rocha, Aryanne A M C, Lopes, Débora O, Dorella, Fernanda A, Pacheco, Luis G C, Costa, Marcília P, Turk, Meritxell Z, Seyffert, Núbia, Moraes, Pablo M R O, Soares, Siomar C, Almeida, Sintia S, Castro, Thiago L P, Abreu, Vinicius A C, Trost, Eva, Baumbach, Jan, Tauch, Andreas, Schneider, Maria Paula C, McCulloch, John, Cerdeira, Louise T, Ramos, Rommel T J, Zerlotini, Adhemar, Dominitini, Anderson, Resende, Daniela M, Coser, Elisângela M, Oliveira, Luciana M, Pedrosa, André L, Vieira, Carlos U, Guimarães, Cláudia T, Bartholomeu, Daniela C, Oliveira, Diana M, Santos, Fabrício R, Rabelo, Élida Mara, Lobo, Francisco P, Franco, Glória R, Costa, Ana Flávia, Castro, Ieso M, Dias, Sílvia Regina Costa, Ferro, Jesus A, Ortega, José Miguel, Paiva, Luciano V, Goulart, Luiz R, Almeida, Juliana Franco, Ferro, Maria Inês T, Carneiro, Newton P, Falcão, Paula R K, Grynberg, Priscila, Teixeira, Santuza M R, Brommonschenkel, Sérgio, Oliveira, Sérgio C, Meyer, Roberto, Moore, Robert J, Miyoshi, Anderson, Oliveira, Guilherme C, & Azevedo, Vasco. 2011. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two Corynebacterium pseudotuberculosis strains. *PLoS ONE*, **6**(4), e18551.

[Russell *et al.*, 2013] Russell, Alistair B, Leroux, Michele, Hathazi, Krisztina, Agnello, Danielle M, Ishikawa, Takahiko, Wiggins, Paul A, Wai, Sun Nyunt, & Mougous, Joseph D. 2013. Diverse type VI secretion phospholipases are functionally plastic antibacterial effectors. *Nature*, **496**(7446), 508–12.

[Saitou & Nei, 1987] Saitou, N, & Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**(4), 406–25.

[Sanchez, 2011] Sanchez, Cesar. 2011. Horizontal gene transfer: Eukaryotes under a new light. *Nature Publishing Group*, **9**(4), 228–228.

[Saukkonen *et al.*, 1992] Saukkonen, K, Burnette, W.N, Mar, V.L, & Masure, H.R. 1992. Pertussis toxin has eukaryotic-like carbohydrate recognition domains. *Proc. Natl. Acad. Scie*, **89**, 119–122.

[Schmeisser *et al.*, 2009] Schmeisser, Christel, Liesegang, Heiko, Krysciak, Dagmar, Bakkou, Nadia, Quéré, Antoine Le, Wollherr, Antje, Heinemeyer, Isabelle, Morgenstern, Burkhard, Pommerening-Röser, Andreas, Flores, Margarita, Palacios, Rafael, Brenner, Sydney, Gottschalk, Gerhard, Schmitz, Ruth A, Broughton, William J, Perret, Xavier, Strittmatter, Axel W, & Streit, Wolfgang R. 2009. Rhizobium sp. strain NGR234 possesses a remarkable number of secretion systems. *Applied and Environmental Microbiology*, **75**(12), 4035–45.

[Schmidt *et al.*, 1980] Schmidt, U., Chmel, H., Kaminski, Z., & Sen, P. 1980. The clinical spectrum of Campylobacter fetus infections: report of five cases and review of the literature. *O J Med*, **49**, 431–442.

[Sebaihia *et al.*, 2010] Sebaihia, M, Bocsanczy, A. M, Biehl, B. S, Quail, M. A, Perna, N. T, Glasner, J. D, Declerck, G. A, Cartinhour, S, Schneider, D. J, Bentley, S. D, Parkhill, J, & Beer, S. V. 2010. Complete Genome Sequence of the Plant Pathogen Erwinia amylovora Strain ATCC 49946. *Journal of Bacteriology*, **192**(7), 2020–2021.

[Sela *et al.*, 2008] Sela, D A, Chapman, J, Adeuya, A, Kim, J H, Chen, F, Whitehead, T R, Lapidus, A, Rokhsar, D S, Lebrilla, C B, German, J B, Price, N P, Richardson, P M, & Mills, D A. 2008. The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proc Natl Acad Sci USA*, **105**(48), 18964–9.

[Seshadri *et al.*, 2003] Seshadri, Rekha, Paulsen, Ian T, Eisen, Jonathan A, Read, Timothy D, Nelson, Karen E, Nelson, William C, Ward, Naomi L, Tettelin, Hervé, Davidsen,

Tanja M, Beanan, Maureen J, Deboy, Robert T, Daugherty, Sean C, Brinkac, Lauren M, Madupu, Ramana, Dodson, Robert J, Khouri, Hoda M, Lee, Kathy H, Carty, Heather A, Scanlan, David, Heinzen, Robert A, Thompson, Herbert A, Samuel, James E, Fraser, Claire M, & Heidelberg, John F. 2003. Complete genome sequence of the Q-fever pathogen Coxiella burnetii. *Proc Natl Acad Sci USA*, **100**(9), 5455–60.

[Shames *et al.*, 2009] Shames, Stephanie R, Auweter, Sigrid D, & Finlay, B Brett. 2009. Co-evolution and exploitation of host cell signaling pathways by bacterial pathogens. *Int J Biochem Cell Biol*, **41**(2), 380–9.

[Simeone *et al.*, 2009] Simeone, Roxane, Bottai, Daria, & Brosch, Roland. 2009. ESX/type VII secretion systems and their role in host-pathogen interaction. *Curr Opin Microbiol*, **12**(1), 4–10.

[Spudich, 1993] Spudich, J L. 1993. Color sensing in the Archaea: a eukaryotic-like receptor coupled to a prokaryotic transducer. *J Bacteriol*, **175**(24), 7755–61.

[Stajich *et al.*, 2002] Stajich, Jason E, Block, David, Boulez, Kris, Brenner, Steven E, Chervitz, Stephen A, Dagdigian, Chris, Fuellen, Georg, Gilbert, James G R, Korf, Ian, Lapp, Hilmar, Lehväslaiho, Heikki, Matsalla, Chad, Mungall, Chris J, Osborne, Brian I, Pocock, Matthew R, Schattner, Peter, Senger, Martin, Stein, Lincoln D, Stupka, Elia, Wilkinson, Mark D, & Birney, Ewan. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, **12**(10), 1611–8.

[Stamatakis, 2006] Stamatakis, Alexandros. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**(21), 2688–90.

[Stanhope *et al.*, 2001] Stanhope, Michael J., Lupas, Andrei, Italia, Michael J., Koretke, Kristin K., Volker, Craig, & Brown, James R. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, **411**(6840), 940–944.

[Stanton, 2007] Stanton, Thad B. 2007. Prophage-like gene transfer agents—Novel mechanisms of gene exchange for Methanococcus, Desulfovibrio, Brachyspira, and Rhodobacter species. *Anaerobe*, **13**(2), 43 – 49.

[Stephens *et al.*, 1998] Stephens, R S, Kalman, S, Lammel, C, Fan, J, Marathe, R, Aravind, L, Mitchell, W, Olinger, L, Tatusov, R L, Zhao, Q, Koonin, E V, & Davis, R W. 1998. Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science*, **282**(5389), 754–9.

[Suzuki *et al.*, 2011] Suzuki, Haruo, Lefébure, Tristan, Hubisz, Melissa Jane, Bitar, Paulina Pavinski, Lang, Ping, Siepel, Adam, & Stanhope, Michael J. 2011. Comparative genomic analysis of the Streptococcus dysgalactiae species group: gene content, molecular adaptation, and promoter evolution. *Genome Biology and Evolution*, **3**(Jan), 168–85.

[Thiberge *et al.*, 2010] Thiberge, Jean-Michel, Boursaux-Eude, Caroline, Lehours, Philippe, Dillies, Marie-Agnès, Creno, Sophie, Coppée, Jean-Yves, Rouy, Zoé, Lajus, Aurélie, Ma, Laurence, Burucoa, Christophe, Ruskoné-Foumestraux, Anne, Courillon-Mallet, Anne, Reuse, Hilde De, Boneca, Ivo Gomperts, Lamarque, Dominique, Mégraud, Francis, Delchier, Jean-Charles, Médigue, Claudine, Bouchier, Christiane, Labigne, Agnès, & Raymond, Josette. 2010. From array-based hybridization of Helicobacter pylori isolates to the complete genome sequence of an isolate associated with MALT lymphoma. *BMC Genomics*, **11**(Jan), 368.

[Thiel *et al.*, 2012] Thiel, Philipp, Kaiser, Markus, & Ottmann, Christian. 2012. Small-molecule stabilization of protein-protein interactions: an underestimated concept in drug discovery? *Angew Chem Int Ed Engl*, **51**(9), 2012–8.

[Thiele *et al.*, 2013] Thiele, Ines, Heinken, Almut, & Fleming, Ronan M T. 2013. A systems biology approach to studying the role of microbes in human health. *Current Opinion in Biotechnology*, **24**(1), 4–12.

[Thieme *et al.*, 2005] Thieme, Frank, Koebnik, Ralf, Bekel, Thomas, Berger, Carolin, Boch, Jens, Büttner, Daniela, Caldana, Camila, Gaigalat, Lars, Goesmann, Alexander, Kay, Sabine, Kirchner, Oliver, Lanz, Christa, Linke, Burkhard, McHardy, Alice C, Meyer, Folker, Mittenhuber, Gerhard, Nies, Dietrich H, Niesbach-Klösgen, Ulla, Patschkowski, Thomas, Rückert, Christian, Rupp, Oliver, Schneiker, Susanne, Schuster, Stephan C, Vorhölter, Frank-Jörg, Weber, Ernst, Pühler, Alfred, Bonas, Ulla, Bartels, Daniela, & Kaiser, Olaf. 2005. Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium Xanthomonas campestris pv. vesicatoria revealed by the complete genome sequence. *J Bacteriol*, **187**(21), 7254–66.

[Thomas & Nielsen, 2005] Thomas, Christopher M, & Nielsen, Kaare M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Micro*, **3**(9), 711–21.

[Tobe *et al.*, 2006] Tobe, Toru, Beatson, Scott A, Taniguchi, Hisaaki, Abe, Hiroyuki, Bailey, Christopher M, Fivian, Amanda, Younis, Rasha, Matthews, Sophie, Marches, Olivier, Frankel, Gad, Hayashi, Tetsuya, & Pallen, Mark J. 2006. An extensive repertoire of type III secretion effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci USA*, **103**(40), 14941–6.

[Touchon *et al.*, 2009] Touchon, Marie, Hoede, Claire, Tenaillon, Olivier, Barbe, Valérie, Baeriswyl, Simon, Bidet, Philippe, Bingen, Edouard, Bonacorsi, Stéphane, Bouchier, Christiane, Bouvet, Odile, Calteau, Alexandra, Chiapello, Hélène, Clermont, Olivier, Cruveiller, Stéphane, Danchin, Antoine, Diard, Médéric, Dossat, Carole, Karoui, Meriem El, Frapy, Eric, Garry, Louis, Ghigo, Jean Marc, Gilles, Anne Marie, Johnson, James, Bouguénec, Chantal Le, Lescat, Mathilde, Mangenot, Sophie, Martinez-Jéhanne, Vanessa, Matic, Ivan, Nassif, Xavier, Oztas, Sophie, Petit, Marie Agnès, Pichon, Christophe, Rouy, Zoé, Ruf, Claude Saint, Schneider, Dominique, Tourret, Jérôme, Vacherie, Benoit, Vallenet, David, Médigue, Claudine, Rocha, Eduardo P C, & Denamur, Erick. 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet*, **5**(1), e1000344.

[Troisfontaines & Cornelis, 2005] Troisfontaines, Paul, & Cornelis, Guy R. 2005. Type III secretion: more systems than you think. *Physiology (Bethesda)*, **20**(Oct), 326–39.

[Trost *et al.*, 2012] Trost, Eva, Blom, Jochen, de Castro Soares, Siomar, Huang, I-Hsiu, Al-Dilaimi, Arwa, Schröder, Jasmin, Jaenicke, Sebastian, Dorella, Fernanda A, Rocha, Flavia S, Miyoshi, Anderson, Azevedo, Vasco, Schneider, Maria P, Silva, Artur, Camello, Thereza C, Sabbadini, Priscila S, Santos, Cíntia S, Santos, Louisy S, Hirata, Raphael, Mattos-Guaraldi, Ana L, Efstratiou, Androulla, Schmitt, Michael P, Ton-That, Hung, & Tauch, Andreas. 2012. Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. *Journal of Bacteriology*, **194**(12), 3199–215.

[Tseng *et al.*, 2009] Tseng, Tsai-Tien, Tyler, Brett M, & Setubal, João C. 2009. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol*, **9 Suppl 1**(Jan), S2.

[Velho *et al.*, 2003] Velho, Paulo Eduardo Neves Ferreira, Cintra, Maria Letícia, Uthida-Tanaka, Ana Maria, de Moraes, Aparecida Machado, & Mariotto, Andréia. 2003. What do we (not) know about the human bartonelloses? *Braz J Infect Dis*, **7**(1), 1–6.

[Vickerman *et al.*, 2007] Vickerman, M M, Iobst, S, Jesionowski, A M, & Gill, S R. 2007. Genome-wide transcriptional changes in Streptococcus gordonii in response to competence signaling peptide. *Journal of Bacteriology*, **189**(21), 7799–807.

[Voth & Heinzen, 2009] Voth, Daniel E, & Heinzen, Robert A. 2009. Coxiella type IV secretion and cellular microbiology. *Curr Opin Microbiol*, **12**(1), 74–80.

[Wang *et al.*, 2011] Wang, Yejun, Zhang, Qing, Sun, Ming-An, & Guo, Dianjing. 2011. High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics*, **27**(6), 777–84.

[Whisstock & Lesk, 1999] Whisstock, J C, & Lesk, A M. 1999. SH3 domains in prokaryotes. *Trends Biochem Sci*, **24**(4), 132–3.

[Williams *et al.*, 2012] Williams, D, Gogarten, J. P, & Papke, R. T. 2012. Quantifying Homologous Replacement of Loci between Haloarchaeal Species. *Genome Biology and Evolution*, **4**(12), 1223–1244.

[Woese *et al.*, 1990] Woese, C R, Kandler, O, & Wheelis, M L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*, **87**(12), 4576–9.

[Xu *et al.*, 2007] Xu, Jian, Mahowald, Michael A, Ley, Ruth E, Lozupone, Catherine A, Hamady, Micah, Martens, Eric C, Henrissat, Bernard, Coutinho, Pedro M, Minx, Patrick, Latreille, Philippe, Cordum, Holland, Brunt, Andrew Van, Kim, Kyung, Fulton, Robert S, Fulton, Lucinda A, Clifton, Sandra W, Wilson, Richard K, Knight, Robin D, & Gordon, Jeffrey I. 2007. Evolution of symbiotic bacteria in the distal human intestine. *Plos Biol*, **5**(7), e156.

[Xue *et al.*, 2011] Xue, Huping, Lu, Hong, & Zhao, Xin. 2011. Sequence diversities of serine-aspartate repeat genes among Staphylococcus aureus isolates from different hosts presumably by horizontal gene transfer. *PLoS ONE*, **6**(5), e20332.

[Yang & Bourne, 2009] Yang, Song, & Bourne, Philip E. 2009. The evolutionary history of protein domains viewed by species phylogeny. *PLoS ONE*, **4**(12), e8378.

[Yuan *et al.*, 2010] Yuan, Jijun, Zweers, Jessica C, van Dijl, Jan Maarten, & Dalbey, Ross E. 2010. Protein transport across and into cell membranes in bacteria and archaea. *Cell. Mol. Life Sci.*, **67**(2), 179–99.

[Zou *et al.*, 2013] Zou, Linyun, Nan, Chonghan, & Hu, Fuquan. 2013. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics*, Sep.

Appendices

Appendix A

Tree analysis of all proteins of Wolbachia endosymbiont of Culex quinquefasciatus Pel and Francisella tularensis subsp. tularensis NE061598 with ELD score equal to or greater than 4 and a bit-score ratio of equal to or greater than 0.5 using different PHAT search strings

ACC number	M	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18
Wolbachia endosymbiont of Culex quinquefasciatus Pel																			
YP_001974936	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001974998	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975006	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975025	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_001975026	P	-	-	-	-	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975058	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975087	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975093	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975094	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975097	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975139	P	P	P	P	P	P	-	P	-	-	P	P	P	P	P	-	P	P	-
YP_001975140	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975144	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975145	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975172	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	-	P	P	-
YP_001975177	P	-	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975179	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	-	P	P	-
YP_001975190	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975193	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975196	P	-	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975197	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975209	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975234	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975284	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975372	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	-	P	P	-
YP_001975403	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975413	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975417	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975440	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975443	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975451	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	-	P	P	P
YP_001975452	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975489	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975490	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

ACC number	M	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18
YP_001975491	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975493	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_001975523	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975527	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975528	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975529	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975534	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975536	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
YP_001975547	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	P
YP_001975619	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975745	x	-	-	-	-	-	-	-	-	-	P	P	P	P	P	-	-	P	-
YP_001975764	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001975847	P	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001975955	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	-	P	P	-
YP_001975998	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001976008	?	-	P	P	P	P	-	P	-	-	P	P	P	-	P	-	-	-	-
YP_001976012	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
YP_001976023	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001976025	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001976034	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001976039	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001976067	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
YP_001976080	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_001976103	P	P	P	P	P	-	-	-	-	-	P	P	P	P	P	-	P	P	-
Francisella tularensis subsp. tularensis NE061598																			
YP_005830323	P(?)	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005830479	P(?)	P	P	P	P	P	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005830518	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_005830523	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_005830564	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_005830612	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005830815	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005830920	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_005831134	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831135	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P

ACC number	M	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18
YP_005831197	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831228	P	P	P	P	P	P	-	P	P	P	P	P	P	P	P	P	P	P	P
YP_005831292	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_005831398	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831454	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831559	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_005831568	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831620	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_005831631	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831639	P(?)	-	-	-	P	P	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831663	-	P	P	P	P	P	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831818	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831889	P	P	P	P	P	P	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831922	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831954	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YP_005831978	-	-	-	-	-	-	-	-	-	-	P	P	P	P	P	P	P	P	P
YP_005831982	-	P	P	P	P	P	-	P	P	P	P	P	P	P	P	P	P	P	P

Table A.1: Tree analysis of all proteins of Wolbachia endosymbiont of Culex quinquefasciatus Pel and Francisella tularensis subsp. tularensis NE061598 with ELD score equal to or greater than 4 and a bit-score ratio of equal to or greater than 0.5 using different PHAT search strings. The ACC numbers are given; M stands for manually evaluated; P stands for positive for eukaryotic origin whereas - stands for negative for eukaryotic origin. In the cases of added question marks it was not possible to evaluate the assign origin without a doubt.

Appendix B

Variation of the ELD-Score and its impact on the χ^2-square test

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
Pseudomonas aeruginosa DK2	5863	4	360	5503	17	5846	4	13	356	5490	6.1754	0.01295
		5	68	5795			2	15	66	5780	8.7351	0.003121
		6	24	5839			1	16	23	5823	2.6808	0.1016
		7	6	5857			1	16	5	5841	13.4402	0.0002463
		8	4	5859			1	16	3	5843	20.6406	0.000005541
		9	2	5861			1	16	1	5845	42.253	0.0000000000802
		10	2	5861			1	16	1	5845	42.253	0.0000000000802
Staphylococcus aureus subsp. aureus 71193	2620	4	40	2580	3	2617	0	3	40	2577	NA	NA
		5	23	2597			0	3	23	2594	NA	NA
		6	19	2601			0	3	19	2598	NA	NA
		7	14	2606			0	3	14	2603	NA	NA
		8	14	2606			0	3	14	2603	NA	NA
		9	11	2609			0	3	11	2606	NA	NA
		10	11	2609			0	3	11	2606	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
Chlamydophila	1051	4	23	1028	62	989	2	60	21	968	0.0164	0.898
pneumoniae CWL029		5	22	1029			2	60	20	969	0.0342	0.8533
		6	21	1030			2	60	19	970	0.0597	0.807
		7	19	1032			2	60	17	972	0.1388	0.7095
		8	18	1033			2	60	16	973	0.1955	0.6584
		9	18	1033			2	60	16	973	0.1955	0.6584
		10	18	1033			2	60	16	973	0.1955	0.6584
Mycoplasma	737	4	4	733	14	723	0	14	4	719	NA	NA
gallisepticum		5	3	734			0	14	3	720	NA	NA
NC06_2006.080-5-2P		6	1	736			0	14	1	722	NA	NA
		7	1	736			0	14	1	722	NA	NA
		8	1	736			0	14	1	722	NA	NA
		9	1	736			0	14	1	722	NA	NA
		10	1	736			0	14	1	722	NA	NA
Helicobacter pylori	1491	4	16	1475	2	1489	0	2	16	1473	NA	NA
PeCan18		5	14	1477			0	2	14	1475	NA	NA
		6	14	1477			0	2	14	1475	NA	NA
		7	14	1477			0	2	14	1475	NA	NA
		8	12	1479			0	2	12	1477	NA	NA
		9	11	1480			0	2	11	1478	NA	NA
		10	6	1485			0	2	6	1483	NA	NA
Mycoplasma wenyonii	651	4	9	642	14	636	0	14	9	627	NA	NA
str. Massachusetts		5	9	642			0	14	9	627	NA	NA
		6	9	642			0	14	9	627	NA	NA
		7	9	642			0	14	9	627	NA	NA
		8	9	642			0	14	9	627	NA	NA
		9	9	642			0	14	9	627	NA	NA
		10	9	642			0	14	9	627	NA	NA
Francisella tularensis	1529	4	36	1493	40	1489	2	38	34	1455	0.3479	0.5553
subsp. tularensis SCHU		5	28	1501			1	39	27	1462	0.0772	0.7811
S4		6	22	1507			1	39	21	1468	0.0103	0.919
		7	21	1508			0	40	21	1468	NA	NA
		8	21	1508			0	40	21	1468	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		9	12	1517			0	40	12	1477	NA	NA
		10	3	1526			0	40	3	1486	NA	NA
Streptococcus pyogenes MGAS315	1858	4	36	1822	0	1858	0	0	36	1822	NA	NA
		5	30	1828			0	0	30	1828	NA	NA
		6	18	1840			0	0	18	1840	NA	NA
		7	15	1843			0	0	15	1843	NA	NA
		8	13	1845			0	0	13	1845	NA	NA
		9	13	1845			0	0	13	1845	NA	NA
		10	13	1845			0	0	13	1845	NA	NA
Escherichia coli CFT073	5324	4	155	5169	2	5322	0	2	155	5167	NA	NA
		5	78	5246			0	2	78	5244	NA	NA
		6	47	5277			0	2	47	5275	NA	NA
		7	26	5298			0	2	26	5296	NA	NA
		8	3	5321			0	2	3	5316	NA	NA
		9	1	5323			0	2	1	5321	NA	NA
		10	1	5323			0	2	1	5321	NA	NA
Streptococcus mutans UA159	1952	4	29	1923	2	1950	0	2	29	1921	NA	NA
		5	17	1935			0	2	17	1933	NA	NA
		6	14	1938			0	2	14	1936	NA	NA
		7	4	1948			0	2	4	1946	NA	NA
		8	2	1950			0	2	2	1948	NA	NA
		9	2	1950			0	2	2	1948	NA	NA
		10	2	1950			0	2	2	1948	NA	NA
Yersinia pestis CO92	3882	4	64	3818	6	3876	0	6	64	3812	NA	NA
		5	22	3860			0	6	22	3854	NA	NA
		6	19	3863			0	6	19	3857	NA	NA
		7	18	3864			0	6	18	3858	NA	NA
		8	15	3867			0	6	15	3861	NA	NA
		9	14	3868			0	6	14	3862	NA	NA
		10	9	3873			0	6	9	3867	NA	NA
Coxiella burnetii RSA 493	1810	4	26	1784	35	1775	7	28	19	1756	74.0123	<2.2E-016
		5	24	1786			7	28	17	1758	81.1264	<2.2E-016

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		6	22	1788			6	29	16	1759	62.4857	2.684E-015
		7	22	1788			6	29	16	1759	62.4857	2.684E-015
		8	22	1788			6	29	16	1759	62.4857	2.684E-015
		9	21	1789			5	30	16	1759	42.5813	0.0000000000678
		10	21	1789			5	30	16	1759	42.5813	0.0000000000678
Rickettsia typhi str. Wilmington	836	4	10	826	45	791	3	42	7	784	7.6476	0.005685
		5	7	829			3	42	4	787	12.7515	0.0003557
		6	7	829			3	42	4	787	12.7515	0.0003557
		7	7	829			3	42	4	787	12.7515	0.0003557
		8	7	829			3	42	4	787	12.7515	0.0003557
		9	7	829			3	42	4	787	12.7515	0.0003557
		10	7	829			3	42	4	787	12.7515	0.0003557
Haemophilus influenzae R2846	1631	4	23	1608	3	1628	1	2	22	1606	5.0318	0.02489
		5	19	1612			1	2	18	1610	6.2729	0.01226
		6	12	1619			1	2	11	1617	10.4444	0.00123
		7	10	1621			1	2	9	1619	12.7112	0.0003635
		8	10	1621			1	2	9	1619	12.7112	0.0003635
		9	8	1623			1	2	7	1621	16.1128	0.00005968
		10	8	1623			1	2	7	1621	16.1128	0.00005968
Onion yellows phytoplasma OY-M	721	4	89	632	12	709	7	5	82	627	19.7269	0.000008934
		5	80	641			6	6	74	635	14.9277	0.0001117
		6	52	669			0	12	52	657	NA	NA
		7	33	688			0	12	33	676	NA	NA
		8	33	688			0	12	33	676	NA	NA
		9	33	688			0	12	33	676	NA	NA
		10	33	688			0	12	33	676	NA	NA
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	3662	4	157	3505	97	3565	14	83	143	3422	22.5193	0.00000208
		5	88	3574			6	91	82	3483	4.5346	0.03322
		6	71	3591			6	91	65	3500	7.2964	0.006909
		7	44	3618			5	92	39	3526	9.919	0.001636
		8	42	3620			5	92	37	3528	10.7182	0.001061
		9	40	3622			5	92	35	3530	11.6025	0.0006586
		10	35	3627			2	95	33	3532	0.3672	0.5445

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
Chlamydia trachomatis D/UW-3/CX	895	4	22	873	60	835	2	58	20	815	0.0005	0.9827
		5	21	874			2	58	19	816	0.0066	0.9351
		6	20	875			2	58	18	817	0.0207	0.8855
		7	18	877			1	59	17	818	0.078	0.7801
		8	17	878			1	59	16	819	0.1245	0.7242
		9	17	878			1	59	16	819	0.1245	0.7242
		10	17	878			1	59	16	819	0.1245	0.7242
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	2931	4	62	2869	74	2857	11	63	51	2806	53.449	0.0000000000002654
		5	28	2903			10	64	18	2839	113.2875	<2.2E-016
		6	23	2908			9	65	14	2843	111.6755	<2.2E-016
		7	21	2910			9	65	12	2845	123.7909	<2.2E-016
		8	14	2917			5	69	9	2848	50.1432	0.000000000001429
		9	11	2920			5	69	6	2851	66.1034	4.279E-016
		10	11	2920			5	69	6	2851	66.1034	4.279E-016
Staphylococcus aureus RF122	2509	4	40	2469	2	2499	0	2	40	2459	NA	NA
		5	23	2480			0	4	23	2476	NA	NA
		6	19	2484			0	4	19	2480	NA	NA
		7	14	2489			0	4	14	2485	NA	NA
		8	14	2489			0	4	14	2485	NA	NA
		9	11	2492			0	4	11	2488	NA	NA
		10	11	2492			0	4	11	2488	NA	NA
Bartonella quintana str. Toulouse	1139	4	1	1138	1	1138	0	1	1	1137	NA	NA
		5	0	1139			0	1	0	1138	NA	NA
		6	0	1139			0	1	0	1138	NA	NA
		7	0	1139			0	1	0	1138	NA	NA
		8	0	1139			0	1	0	1138	NA	NA
		9	0	1139			0	1	0	1138	NA	NA
		10	0	1139			0	1	0	1138	NA	NA
Mycoplasma hyopneumoniae 232	691	4	7	684	6	685	0	6	7	678	NA	NA
		5	3	688			0	6	3	682	NA	NA
		6	3	688			0	6	3	682	NA	NA
		7	1	690			0	6	1	684	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		8	1	690			0	6	1	684	NA	NA
		9	1	690			0	6	1	684	NA	NA
		10	1	690			0	6	1	684	NA	NA
Legionella pneumophila str. Lens	2920	4	67	2853	64	2856	12	52	55	2801	71.7078	<2.2E-016
		5	27	2893			7	57	20	2836	60.8711	6.094E-015
		6	22	2898			6	58	16	2840	53.792	0.0000000000002229
		7	22	2898			6	58	16	2840	53.792	0.0000000000002229
		8	14	2906			6	58	8	2848	90.2919	<2.2E-016
		9	12	2908			6	58	6	2850	107.0527	<2.2E-016
		10	12	2908			6	58	6	2850	107.0527	<2.2E-016
Legionella pneumophila str. Paris	3147	4	89	3058	72	3075	14	58	75	3000	67.9736	<2.2E-016
		5	63	3084			13	59	50	3025	88.6054	<2.2E-016
		6	43	3104			8	64	35	3040	44.7829	0.00000000002201
		7	35	3113			2	70	33	3043	0.6326	0.4264
		8	27	3121			2	70	25	3051	1.3017	0.2539
		9	18	3130			2	70	16	3060	2.9612	0.08528
		10	15	3133			0	72	15	3061	NA	NA
Campylobacter lari RM2100	1544	4	20	1524	2	1542	0	2	20	1522	NA	NA
		5	8	1536			0	2	8	1534	NA	NA
		6	7	1537			0	2	7	1535	NA	NA
		7	5	1539			0	2	5	1537	NA	NA
		8	5	1539			0	2	5	1537	NA	NA
		9	5	1539			0	2	5	1537	NA	NA
		10	5	1539			0	2	5	1537	NA	NA
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1	2074	4	28	2046	9	2065	0	9	28	2037	NA	NA
		5	20	2054			0	9	20	2045	NA	NA
		6	18	2056			0	9	18	2047	NA	NA
		7	13	2061			0	9	13	2052	NA	NA
		8	7	2067			0	9	7	2058	NA	NA
		9	6	2068			0	9	6	2059	NA	NA
		10	5	2069			0	9	5	2060	NA	NA
Xanthomonas campestris pv. vesicatoria str. 85-10	4628	4	104	4524	33	4595	2	31	102	4493	0.7992	0.3713

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		5	68	4560			2	31	66	4529	2.1724	0.1405
		6	29	4599			2	31	27	4568	8.1971	0.004196
		7	17	4611			2	31	15	4580	15.8536	0.00006844
		8	17	4611			2	31	15	4580	15.8536	0.00006844
		9	14	4614			2	31	12	4583	19.8398	0.000008421
		10	9	4619			0	33	9	4586	NA	NA
Campylobacter fetus subsp. fetus 82-40	1715	4	2	1713	6	1709	0	6	2	1707	NA	NA
		5	1	1714			0	6	1	1708	NA	NA
		6	1	1714			0	6	1	1708	NA	NA
		7	1	1714			0	6	1	1708	NA	NA
		8	1	1714			0	6	1	1708	NA	NA
		9	0	1715			0	6	0	1709	NA	NA
		10	0	1715			0	6	0	1709	NA	NA
Chlamydophila pneumoniae LPCoLN	1105	4	25	1080	58	1047	2	56	23	1024	0.029	0.8647
		5	24	1081			2	54	22	1025	0.07	0.7913
		6	23	1082			2	56	21	1026	0.0765	0.7821
		7	21	1084			2	56	19	1028	0.1544	0.6944
		8	20	1085			2	56	18	1029	0.2075	0.6487
		9	20	1085			2	56	18	1029	0.2075	0.6487
		10	20	1085			2	56	18	1029	0.2075	0.6487
Borrelia recurrentis A1	977	4	3	974	20	957	0	20	3	954	NA	NA
		5	1	976			0	20	1	956	NA	NA
		6	1	976			0	20	1	956	NA	NA
		7	1	976			0	20	1	956	NA	NA
		8	1	976			0	20	1	956	NA	NA
		9	0	977			0	20	0	957	NA	NA
		10	0	977			0	20	0	957	NA	NA
Lactococcus garvieae Lg2	1938	4	29	1909	1	1937	0	1	29	1908	NA	NA
		5	19	1919			0	1	19	1918	NA	NA
		6	13	1925			0	1	13	1924	NA	NA
		7	10	1928			0	1	10	1927	NA	NA
		8	10	1928			0	1	10	1927	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
Bacteroides vulgatus ATCC 8482	3983	9	10	1928	8	3975	0	1	10	1927	NA	NA
		10	10	1928			0	1	10	1927	NA	NA
		4	302	3681			0	8	302	3673	NA	NA
		5	182	3801			0	8	182	3793	NA	NA
		6	130	3853			0	8	130	3845	NA	NA
	3983	7	118	3865			0	8	118	3857	NA	NA
		8	79	3904			0	8	79	3896	NA	NA
		9	61	3922			0	8	61	3914	NA	NA
		10	56	3927			0	8	56	3919	NA	NA
Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	5497	4	177	5320	13	5484	0	13	177	5307	NA	NA
		5	68	5429			0	13	68	5416	NA	NA
		6	25	5473			0	13	25	5459	NA	NA
		7	6	5491			0	13	6	5478	NA	NA
		8	4	5493			0	13	4	5480	NA	NA
		9	2	5495			0	13	2	5482	NA	NA
		10	1	5496			0	13	1	5483	NA	NA
Polynucleobacter necessarius subsp. necessarius STIR1	1507	4	2	1505	5	1502	0	5	2	1500	NA	NA
		5	2	1505			0	5	2	1500	NA	NA
		6	2	1505			0	5	2	1500	NA	NA
		7	2	1506			0	5	1	1501	NA	NA
		8	2	1506			0	5	1	1501	NA	NA
		9	2	1507			0	5	0	1502	NA	NA
		10	2	1507			0	5	0	1502	NA	NA
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476	4758	4	105	4653	6	4752	1	5	104	4648	1.0449	0.3067
		5	57	4701			1	5	56	4696	2.5841	0.1079
		6	40	4718			0	6	40	4712	NA	NA
		7	8	4750			0	6	8	4744	NA	NA
		8	5	4753			0	6	5	4747	NA	NA
		9	5	4753			0	6	5	4747	NA	NA
		10	0	4758			0	6	0	4752	NA	NA
Streptococcus gordonii str. Challis substr. CH1	2050	4	51	1999	3	2047	0	3	51	1996	NA	NA
		5	19	2031			0	3	19	2028	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		6	19	2031			0	3	19	2028	NA	NA
		7	17	2033			0	3	17	2030	NA	NA
		8	12	2038			0	3	12	2035	NA	NA
		9	10	2040			0	3	10	2037	NA	NA
		10	8	2042			0	3	8	2039	NA	NA
Cryptobacterium curtum DSM 15641	1355	4	22	1333	6	1349	0	6	22	1327	NA	NA
		5	9	1346			0	6	9	1340	NA	NA
		6	6	1349			0	6	6	1343	NA	NA
		7	3	1352			0	6	3	1346	NA	NA
		8	3	1352			0	6	3	1346	NA	NA
		9	3	1352			0	6	3	1346	NA	NA
		10	3	1352			0	6	3	1346	NA	NA
Marinomonas posidonica IVIA-Po-181	3489	4	40	3449	20	3469	0	20	40	3429	NA	NA
		5	24	3465			0	20	24	3445	NA	NA
		6	10	3479			0	20	10	3459	NA	NA
		7	9	3480			0	20	9	3460	NA	NA
		8	5	3484			0	20	5	3464	NA	NA
		9	4	3485			0	20	4	3465	NA	NA
		10	4	3485			0	20	4	3465	NA	NA
Francisella tularensis subsp. tularensis NE061598	1709	4	42	1667	40	1669	5	35	37	1632	13.2088	0.0002786
		5	30	1679			1	39	29	1640	0.0607	0.8054
		6	26	1683			1	39	25	1644	0.0201	0.8872
		7	22	1687			0	40	22	1647	NA	NA
		8	22	1687			0	40	22	1647	NA	NA
		9	15	1694			0	40	15	1654	NA	NA
		10	3	1706			0	40	3	1666	NA	NA
Streptococcus mutans NN2025	1893	4	24	1869	0	1893	0	0	24	1869	NA	NA
		5	12	1881			0	0	12	1881	NA	NA
		6	7	1886			0	0	7	1886	NA	NA
		7	3	1890			0	0	3	1890	NA	NA
		8	1	1892			0	0	1	1892	NA	NA
		9	1	1892			0	0	1	1892	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		10	1	1892			0	0	1	1892	NA	NA
Candidatus Liberibacter asiaticus str. psy62	1102	4	18	1084	18	1084	1	17	17	1067	0.1492	0.6993
		5	10	1092			1	17	9	1075	0.7119	0.3988
		6	7	1095			1	17	6	1078	1.3309	0.2486
		7	1	1101			1	17	0	1084	14.5729	0.0001348
		8	1	1101			1	17	0	1084	14.5729	0.0001348
		9	0	1102			0	18	0	1084	NA	NA
		10	0	1102			0	18	0	1084	NA	NA
Staphylococcus aureus subsp. aureus TCH60	2693	4	45	2648	5	2688	0	5	45	2643	NA	NA
		5	25	2668			0	5	25	1663	NA	NA
		6	21	2672			0	5	21	1667	NA	NA
		7	15	2678			0	5	15	1673	NA	NA
		8	15	2678			0	5	15	1673	NA	NA
		9	12	2681			0	5	12	1676	NA	NA
		10	12	2681			0	5	12	1676	NA	NA
Listeria monocytogenes HCC23	2972	4	114	2858	7	2965	1	6	113	2852	0.208	0.6483
		5	59	2913			1	6	58	2907	0.9593	0.3274
		6	50	2922			1	6	49	2916	1.2648	0.2607
		7	47	2925			1	6	46	2919	1.3943	0.2377
		8	18	2954			0	7	18	2947	NA	NA
		9	6	2966			0	7	6	2959	NA	NA
		10	1	2971			0	7	1	2964	NA	NA
Bifidobacterium animalis subsp. lactis DSM 10140	1565	4	14	1551	5	1560	0	5	14	1546	NA	NA
		5	12	1553			0	5	12	1548	NA	NA
		6	6	1559			0	5	6	1554	NA	NA
		7	4	1561			0	5	4	1556	NA	NA
		8	2	1563			0	5	2	1558	NA	NA
		9	1	1564			0	5	1	1559	NA	NA
		10	1	1564			0	5	1	1559	NA	NA
Pseudomonas aeruginosa LESB58	5902	4	374	5528	19	5883	3	16	371	5512	1.4942	0.2216
		5	71	5831			2	17	69	5814	7.1818	0.007365
		6	25	5877			1	18	24	5859	2.2032	0.1377

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		7	8	5894			1	18	7	5876	8.7731	0.003057
		8	7	5895			1	18	6	5877	10.1612	0.001434
		9	1	5901			1	18	0	5883	76.9217	<2.2E-016
		10	1	5901			1	18	0	5883	76.9217	<2.2E-016
Pectobacterium wasabiae WPP163	4431	4	102	4329	10	4421	5	5	97	4324	81.248	<2.2E-016
		5	30	4401			1	9	29	4392	2.7853	0.09513
		6	13	4418			1	9	12	4409	7.5898	0.00587
	4431	7	8	4423			1	9	7	4414	12.9174	0.0003256
		8	8	4423			1	9	7	4414	12.9174	0.0003256
		9	5	4426			1	9	4	4417	21.2381	0.000004056
		10	2	4429			1	9	1	4420	54.5397	0.0000000000001523
Bifidobacterium longum subsp. infantis 157F	1975	4	38	1937	7	1968	0	7	38	1930	NA	NA
		5	22	1953			0	7	22	1946	NA	NA
		6	14	1961			0	7	14	1954	NA	NA
		7	5	1970			0	7	5	1963	NA	NA
		8	4	1971			0	7	4	1964	NA	NA
		9	0	1975			0	7	0	1968	NA	NA
		10	0	1975			0	7	0	1968	NA	NA
Acidaminococcus intestini RyC-MR95	2383	4	70	2313	7	2376	2	5	68	2308	8.4193	0.003713
		5	54	2329			2	5	52	2324	11.6403	0.0006454
		6	19	2364			2	5	17	2359	37.7811	0.0000000007915
		7	11	2372			2	5	9	2367	67.1719	2.488E-016
		8	7	2376			2	5	5	2371	107.0722	<2.2E-016
		9	3	2380			2	5	1	2375	253.3926	<2.2E-016
		10	3	2380			2	5	1	2375	253.3926	<2.2E-016
Wolbachia endosymbiont of Culex quinquefasciatus Pel	1161	4	66	1095	80	1081	24	56	42	1039	89.9378	<2.2E-016
		5	60	1101			24	56	36	1045	102.732	<2.2E-016
		6	60	1101			24	56	36	1045	102.732	<2.2E-016
		7	60	1101			24	56	36	1045	102.732	<2.2E-016
		8	54	1107			24	56	30	1051	118.4272	<2.2E-016
		9	54	1107			24	56	30	1051	118.4272	<2.2E-016
		10	54	1107			24	56	30	1051	118.4272	<2.2E-016

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
Mycobacterium	3814	4	142	3672	16	3798	1	15	141	3657	0.016	0.8992
africanum GM041182		5	84	3730			0	16	84	3714	NA	NA
		6	65	3749			0	16	65	3733	NA	NA
		7	60	3754			0	16	60	3738	NA	NA
		8	36	3778			0	16	36	3762	NA	NA
		9	35	3779			0	16	35	3763	NA	NA
		10	35	3779			0	16	35	3763	NA	NA
Bifidobacterium	1569	4	14	1555	5	1564	0	5	14	1550	NA	NA
animalis subsp. lactis		5	12	1557			0	5	12	1552	NA	NA
V9		6	6	1563			0	5	6	1558	NA	NA
		7	4	1565			0	5	4	1560	NA	NA
		8	2	1567			0	5	2	1562	NA	NA
		9	1	1568			0	5	1	1563	NA	NA
		10	1	1568			0	5	1	1563	NA	NA
Arcobacter nitrofigilis	3125	4	15	3110	34	3091	0	34	15	3076	NA	NA
DSM 7299		5	1	3124			0	34	1	3090	NA	NA
		6	1	3124			0	34	1	3090	NA	NA
		7	1	3124			0	34	1	3090	NA	NA
		8	1	3124			0	34	1	3090	NA	NA
		9	1	3124			0	34	1	3090	NA	NA
		10	1	3124			0	34	1	3090	NA	NA
Escherichia coli IAI1	4326	4	110	4216	2	4324	0	2	110	4214	NA	NA
		5	51	4275			0	2	51	4273	NA	NA
		6	31	4295			0	2	31	4293	NA	NA
		7	16	4310			0	2	16	4308	NA	NA
		8	7	4319			0	2	7	4317	NA	NA
		9	5	4321			0	2	5	4319	NA	NA
		10	1	4325			0	2	1	4323	NA	NA
Helicobacter pylori B38	1367	4	14	1353	2	1365	0	2	14	1351	NA	NA
		5	12	1355			0	2	12	1353	NA	NA
		6	12	1355			0	2	12	1353	NA	NA
		7	10	1357			0	2	10	1355	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		8	8	1359			0	2	8	1357	NA	NA
		9	7	1360			0	2	7	1358	NA	NA
		10	0	1367			0	2	0	1365	NA	NA
Actinobacillus succinogenes 130Z	2077	4	22	2055	8	2069	0	8	22	2047	NA	NA
		5	9	2068			0	8	9	2060	NA	NA
		6	8	2069			0	8	8	2061	NA	NA
		7	4	2073			0	8	4	2065	NA	NA
		8	4	2073			0	8	4	2065	NA	NA
		9	4	2073			0	8	4	2065	NA	NA
		10	1	2076			0	8	1	2068	NA	NA
Sinorhizobium fredii NGR234	6260	4	846	5414	18	6242	3	15	843	5399	0.0022	0.9629
		5	275	5985			2	16	273	5969	0.6673	0.414
		6	142	6118			2	16	140	6102	2.9952	0.08351
		7	98	6162			2	16	96	6146	5.3657	0.02054
		8	88	6172			2	16	86	6156	6.2507	0.01241
		9	64	6196			0	18	64	6178	NA	NA
		10	6	6254			0	18	6	6236	NA	NA
Shewanella putrefaciens 200	4064	4	34	4030	9	4055	0	9	34	4021	NA	NA
		5	18	4046			0	9	18	4037	NA	NA
		6	6	4058			0	9	6	4049	NA	NA
		7	4	4060			0	9	4	4051	NA	NA
		8	4	4060			0	9	4	4051	NA	NA
		9	2	4062			0	9	2	4053	NA	NA
		10	1	4063			0	9	1	4054	NA	NA
Streptococcus dysgalactiae subsp. equisimilis ATCC 12394	2050	4	44	2006	0	2050	0	0	44	2006	NA	NA
		5	34	2016			0	0	34	2016	NA	NA
		6	29	2021			0	0	29	2021	NA	NA
		7	29	2021			0	0	29	2021	NA	NA
		8	29	2021			0	0	29	2021	NA	NA
		9	0	2050			0	0	0	2050	NA	NA
		10	0	2050			0	0	0	2050	NA	NA
Treponema pallidum subsp. pallidum str. Chicago	980	4	4	976	9	971	0	9	4	967	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		5	1	979			0	9	1	970	NA	NA
		6	1	979			0	9	1	970	NA	NA
		7	1	979			0	9	1	970	NA	NA
		8	1	979			0	9	1	970	NA	NA
		9	0	980			0	9	0	971	NA	NA
		10	0	980			0	9	0	971	NA	NA
Corynebacterium pseudotuberculosis 1002	2090	4	10	2080	1	2089	0	1	10	2079	NA	NA
		5	4	2086			0	1	4	2085	NA	NA
		6	4	2086			0	1	4	2085	NA	NA
		7	1	2089			0	1	1	2088	NA	NA
		8	1	2089			0	1	1	2088	NA	NA
		9	1	2089			0	1	1	2088	NA	NA
		10	1	2089			0	1	1	2088	NA	NA
Helicobacter mustelae 12198	1398	4	3	1395	4	1394	0	4	3	1391	NA	NA
		5	3	1395			0	4	3	1391	NA	NA
		6	3	1395			0	4	3	1391	NA	NA
		7	3	1395			0	4	3	1391	NA	NA
		8	3	1395			0	4	3	1391	NA	NA
		9	3	1395			0	4	3	1391	NA	NA
		10	3	1395			0	4	3	1391	NA	NA
Rothia mucilaginosa DY-18	1903	4	9	1894	0	1903	0	0	9	1894	NA	NA
		5	7	1896			0	0	7	1896	NA	NA
		6	6	1897			0	0	6	1897	NA	NA
		7	3	1900			0	0	3	1900	NA	NA
		8	3	1900			0	0	3	1900	NA	NA
		9	3	1900			0	0	3	1900	NA	NA
		10	3	1900			0	0	3	1900	NA	NA
Corynebacterium diphtheriae 31A	2352	4	19	2333	3	2349	0	3	19	2330	NA	NA
		5	9	2343			0	3	9	2340	NA	NA
		6	8	2344			0	3	8	2341	NA	NA
		7	5	2347			0	3	5	2344	NA	NA
		8	5	2347			0	3	5	2344	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		9	5	2347			0	3	5	2344	NA	NA
		10	5	2347			0	3	5	2344	NA	NA
Novosphingobium sp. PP1Y	4613	4	225	4388	31	4582	6	25	219	4363	11.1324	0.0008483
		5	117	4496			6	25	111	4471	29.1913	0.00000006557
		6	90	4523			6	25	84	4498	40.6821	0.0000000001791
		7	70	4543			4	27	66	4516	19.9462	0.000007965
		8	37	4576			1	30	36	4546	0.2579	0.6116
		9	37	4576			1	30	36	4546	0.2579	0.6116
		10	37	4576			1	30	36	4546	0.2579	0.6116
Bacillus thuringiensis BMB171	5327	4	323	5004	5	5322	1	4	322	5000	0.1362	0.7121
		5	162	5165			0	5	162	5160	NA	NA
		6	65	5262			0	5	65	5257	NA	NA
		7	47	5280			0	5	47	5275	NA	NA
		8	26	5301			0	5	26	5296	NA	NA
		9	8	5319			0	5	8	5314	NA	NA
		10	8	5319			0	5	8	5314	NA	NA
Erwinia amylovora ATCC 49946	3563	4	20	3543	4	3559	0	4	20	3539	NA	NA
		5	14	3549			0	4	14	3545	NA	NA
		6	5	3558			0	4	5	3554	NA	NA
		7	4	3559			0	4	4	3555	NA	NA
		8	3	3560			0	4	3	3556	NA	NA
		9	2	3561			0	4	2	3557	NA	NA
		10	2	3561			0	4	2	3557	NA	NA
Herbaspirillum seropedicae SmR1	4732	4	106	4626	7	4725	0	7	106	4619	NA	NA
		5	81	4651			0	7	81	4644	NA	NA
		6	4	4728			0	7	4	4721	NA	NA
		7	1	4731			0	7	1	4724	NA	NA
		8	1	4731			0	7	1	4724	NA	NA
		9	1	4731			0	7	1	4724	NA	NA
		10	1	4731			0	7	1	4724	NA	NA
Rothia dentocariosa ATCC 17931	2212	4	24	2188	4	2208	0	4	24	2184	NA	NA
		5	20	2192			0	4	20	2188	NA	NA

Taxonomic name	Sum	ELD-Score	ELDP	nELDP	pPHAT	nPHAT	pPHAT+ pELDP	pPHAT+ nELDP	nPHAT+ pELDP	nPHAT+ nELDP	χ^2	p-value
		6	16	2196			0	4	16	2192	NA	NA
		7	10	2202			0	4	10	2198	NA	NA
		8	7	2205			0	4	7	2201	NA	NA
		9	6	2206			0	4	6	2202	NA	NA
		10	6	2206			0	4	6	2202	NA	NA
Prevotella denticola F0289	2380	4	36	2344	3	2377	1	2	35	2342	4.6304	0.03141
		5	19	2361			0	3	19	2358	NA	NA
		6	16	2364			0	3	16	2361	NA	NA
		7	12	2368			0	3	12	2365	NA	NA
		8	12	2368			0	3	12	2365	NA	NA
		9	5	2375			0	3	5	2371	NA	NA
		10	5	2375			0	3	5	2372	NA	NA

Table B.1: Variation of the ELD-Score and its impact on the χ^2-square test. χ^2-square test was only performed if at least one positive signal for the pPHAT+pELDP group was detected - this is indicated by 'NA' in the χ^2-square and p-value columns.

Appendix C

Details for 146 domain trees obtained from Legionella pneumophila str. Paris which resulted in a positive PHAT signal and having a proposed eukaryotic origin

UniProtKB AC/ID	RefSeq Protein	Pfam-Domain	ELD-Score
Q5X7D6	YP_123007.1	PF13380	-
PUR7	YP_123965.1	PF01259	-
Q5X7D8	YP_123005.1	PF00999	-
Q5X6L0	YP_123283.1	PF09265	7
Q5X3M5	YP_124323.1	PF01474	-
Q5X960	YP_122383.1	PF04389	-
Q5X1V3	YP_124945.1	PF00657	4
Q5X8D3	YP_122660.1	PF00271	-
Q5X0Y2	YP_125266.1	PF13806	-
Q5X8V6	YP_122487.1	PF01323	-
Q5X127	YP_125221.1	PF01207	-
Q5X603	YP_123491.1	PF00485	-
Q5X3N1	YP_124317.1	PF05221	-
Q5WS62	YP_122239.1	PF01243	-
Q5X6D3	YP_123361.1	PF01150	48
Q5X4N8	YP_123956.1	PF09471	4
Q5X425	YP_124173.1	PF13091	-
Q5X407	YP_124191.1	PF08007	-
Q5X8C7	YP_122666.1	PF04632	-
Q5X5D3	YP_123711.1	PF08031	-
CLPX	YP_124146.1	PF10431	-

UniProtKB AC/ID	RefSeq Protein	Pfam-Domain	ELD-Score
Q5X650	YP_123444.1	PF00328	4
Q5X6W6	YP_123177.1	PF01648	-
Q5X7H5	YP_122968.1	PF02852	-
NUOD	YP_125138.1	PF00346	-
Q5X896	YP_122697.1	PF00324	-
HGD	YP_123572.1	PF04209	-
Q5X2D1	YP_124767.1	PF13847	-
Q5X123	YP_125225.1	PF01242	-
Q5X5M4	YP_123620.1	PF08245	-
Q5X3A0	YP_124448.1	PF14765	-
Q5X3L1	YP_124337.1	PF00288	-
Q5X289	YP_124809.1	PF01382	7
Q5X4Z8	YP_123846.1	PF09084	-
Q5X8P1	YP_122552.1	PF01156	-
Q5X1D7	YP_125111.1	PF00884	-
KAD	YP_123690.1	PF05191	-
Q5X4V2	YP_123892.1	PF05175	-
Q5X503	YP_123841.1	PF02817	-
Q5X197	YP_125151.1	PF06276	-
Q5X3V2	YP_124246.1	PF01369	10000
Q5X366	YP_124482.1	PF00848	-
LUBX	YP_125189.1	PF04564	10000
Q5X6M5	YP_123268.1	PF00188	-
Q5X8K3	YP_122590.1	PF13450	-
Q5X8V1	YP_122492.1	PF03446	-
Q5X3S3	YP_124275.1	PF04488	10
Q5X155	YP_125193.1	PF12836	-
Q5X6N2	YP_123261.1	PF13766	-
Q5WS17	YP_122284.1	PF02958	-
RL30	YP_122752.1	PF00327	-
Q5X1L0	YP_125038.1	PF02826	-
Q5X4S2	YP_123922.1	PF14360	8
Q5X3S4	YP_124274.1	PF01965	-
Q5X5N8	YP_123606.1	PF14114	-
Q5X424	YP_124174.1	PF07859	-
Q5X4R4	YP_123930.1	PF00326	-
Q5X7W6	YP_122827.1	PF00723	-
Q5X342	YP_124506.1	PF00925	5
Q5X114	YP_125234.1	PF00710	-
Q5X4D9	YP_124055.1	PF13621	-
Q5X2E3	YP_124755.1	PF00122	-
Q5X4N5	YP_123959.1	PF14701	-
Q5X7M5	YP_122918.1	PF00248	-
Q5X7L0	YP_122933.1	PF00890	-
Q5X6V9	YP_123184.1	PF00254	-
Q5X6L5	YP_123278.1	PF01553	-
Q5X566	YP_123778.1	PF01433	4
Q5X5J9	YP_123645.1	PF00291	-
Q5X4W8	YP_123876.1	PF14378	-
Q5X341	YP_124507.1	PF07958	4
IF2	YP_125125.1	PF00009	-

UniProtKB AC/ID	RefSeq Protein	Pfam-Domain	ELD-Score
GCSH	YP_122481.1	PF01597	-
Q5X3L6	YP_124332.1	PF02535	-
Q5X203	YP_124895.1	PF14031	-
Q5X5N7	YP_123607.1	PF00314	7
Q5X3I2	YP_124366.1	PF08719	-
Q5X6S2	YP_123221.1	PF00743	-
Q5X2B7	YP_124781.1	PF13475	-
Q5X613	YP_123481.1	PF00205	-
DNAK	YP_124321.1	PF00012	-
Q5X1R7	YP_124981.1	PF07466	-
Q5X4N2	YP_123962.1	PF00551	-
Q5X315	YP_124533.1	PF01979	-
Q5X5B5	YP_123729.1	PF03171, PF14226	-
Q5X7U7	YP_122846.1	PF12608	-
Q5X8Y3	YP_122460.1	PF05761	5
Q5X3Y8	YP_124210.1	PF00150	-
Q5X4B9	YP_124079.1	PF00620	10000
Q5X5X8	YP_123516.1	PF10672	-
Q5X374	YP_124474.1	PF01894	-
Q5X2H2	YP_124726.1	PF09445	10
Q5X1I6	YP_125062.1	PF13410	-
FOLD	YP_123585.1	PF02882	-
Q5X4Z2	YP_123852.1	PF02265	-
Q5X795	YP_123048.1	PF03222	-
Q5X0M5	YP_125373.1	PF01400	-
SYH	YP_123825.1	PF13393	-
Q5X705	YP_123138.1	PF09360	-
Q5X373	YP_124475.1	PF00850	-
SYN	YP_124548.1	PF01336	-
Q5X4S0	YP_123924.1	PF07366	-
Q5X0W7	YP_125281.1	PF01179	-
Q5X493	YP_124105.1	PF00441	-
Q5X0S1	YP_125327.1	PF08123	14
Q5X1T4	YP_124964.1	PF13230	-
Q5X3Z9	YP_124199.1	PF01764	-
Q5X2M3	YP_124675.1	PF00690	-
Q5X330	YP_124518.1	PF00120	-
MNME	YP_125375.1	PF01926	-
Q5X7M7	YP_122916.1	PF05721	-
Q5X8A3	YP_122690.1	PF00617	10000
Q5X8G3	YP_122630.1	PF01126	-
Q5X8E8	YP_122645.1	PF04828	-
Q5X348	YP_124500.1	PF04116	
Q5X4J3	YP_124001.1	PF00023, PF00856	5
Q5X2U1	YP_124607.1	PF00724	-
Q5X2P9	YP_124649.1	PF01794	-
Q5X1S0	YP_124978.1	PF02463	-
Q5X5F0	YP_123694.1	PF00285	-
Q5X7D9	YP_123004.1	PF01327	-
Q5X874	YP_122719.1	PF11901	1000
Q5X634	YP_123460.1	PF08511	7

UniProtKB AC/ID	RefSeq Protein	Pfam-Domain	ELD-Score
Q5X8L7	YP_122576.1	PF00646	1000
Q5X377	YP_124471.1	PF08240	-
Q5X5L5	YP_123629.1	PF00171	-
Q5WS48	YP_122253.1	PF13637	-
Q5X615	YP_123479.1	PF01593	-
Q5X4Q3	YP_123941.1	PF02894	-
Q5X3A1	YP_124447.1	PF02668	-
Q5X978	YP_122365.1	PF02678	-
Q5X6W0	YP_123183.1	PF03313	-

Table C.1: Details for 146 domain trees which resulted in a positive PHAT signal and having a proposed eukaryotic origin. All domains present in the protein are shown.

Appendix D

Comprehensive table for results obtained for 70 prokaryots

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Pseudomonas aeruginosa DK2	152	H00313 - Pseudomonas aeruginosa infection	Human pathogen - secondary infection	Pseudomonas aeruginosa is the predominant microorganism in chronic lung infection of cystic fibrosis patients. Preceding to chronic infection is the intermittent colonization of the lung - after established infection the strains differ phenotypically from the intermittent strains. The most prominent changes are the switch to mucoidity (alginate overproduction) and loss of epigenetic regulation of virulence such as the Quorum Sensing	[Bjarnsholt *et al.*, 2010]	2
Output : pPHAT,pELD:4 - pPHAT,nELD:13 - nPHAT,pELD:356 - nPHAT,nELD:5490 Input: ELD: 363, nELD: 5521						
ACC	AI	Domains	Description			
YP_006483810	0.66	PF03088				
YP_006480122	1.14	PF06355				
YP_006483846	0.88	PF01989, PF04412				
YP_006485602	1.02	PF12710				
Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Staphylococcus aureus subsp. aureus 71193	54	/	Human pathogen - primary/secondary infection	During 1997-1999, S. aureus was reported as the most abundant cause of bloodstream, skin and soft tissue, and lower respiratory tract infections in the United States, Canada and Europe. Staphylococcus aureus subsp. aureus 71193 is community-associdated, methicillin-susceptible and infects otherwise healthy individuals. The phylum Staphylococcus aureus is the most abundant cause of bloodstream, soft tissue and lower respiratory tract infections in Europe and the United States.	[Deleo *et al.*, 2009]	3

Output: pPHAT+pELD:4, pPHAT+nELD:13, nPHAT+pELD:356, nPHAT+nELD:5490 - Input: ELD: 40, nELD: 2583

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Chlamydophila pneumoniae CWL029	55	/	Human pathogen - primary infection	C. pneumoniae is a natural pathogen of humans and causes pneumonia and bronchitis. Chlamydia are obligate intracellular eubacteria. Approximately 10% of pneumonia cases and 5% of bronchitis cases in the United States are attributed to C. pneumoniae infection. The possibility of chronic chlamydia infections may result following respiratory-acquired infection.	[Kalman *et al.*, 1999]	3

Output: pPHAT+pELD:2, pPHAT+nELD:60, nPHAT+pELD:21, nPHAT+nELD:968 - Input: ELD: 23, nELD: 1029

ACC	AI	Domains	Description
NP_224551	0.8	PF032119	ADP/ATP translocase
NP_224810	0.73	PF032119	ADP/ATP translocase

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Mycoplasma gallisepticum NC06_2006.080-5-2P	78	/	Animal pathogen - primary infection	Infection with Mycoplasma gallisepticum leads to slow onset, chronic respiratory disease in chickens, turkeys, game birds, pigeons and other wild birds. Further to note is that in common with other mycoplasmas, M. gallisepticum is minute in size with minimal genetic material and with a total lack of a bacterial cell wall. These properties are portrayed in a high degree of interdependence between the host animal and M. gallisepticum.	[Levisohn & Kleven, 2000]	3

Output: pPHAT+pELD:0, pPHAT+nELD:14, nPHAT+pELD:4, nPHAT+nELD:719 - Input: ELD: 4, nELD: 740

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Helicobacter pylori PeCan18	26	H00320 Helicobacter pylori infection - Hß1240 Immune thrombocytopenia	Human pathogen - primary infection	A strong correlation has been established between H. pylori - a gram-negative pathogenic bacterium that specifically colonizes in the gastric epithelium - and a wide spectrum of gastrointestinal diseases, comprising gastric and duodenal ulceration, gastric adenocarcinoma, mucosa-associated lymphoid tissue lymphoma and non-Hodgkin's lymphoma of the stomach. This findings resulted in the rating of H. pylori as a class 1 carcinogen for gastric cancer by the WHO and is considered as the most common etiologic agent of infection-related cancers.	[Noto & Peek, 2012b, Noto & Peek, 2012a]	3

Output: pPHAT+pELD:0, pPHAT+nELD:2, nPHAT+pELD:16, nPHAT+nELD:1473 - Input: ELD: 16, nELD: 1479

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Mycoplasma wenyonii str. Massachusetts	38	/	Animal pathogen - secondary infection	Mycoplasma wenyonii - an uncultivable, cell wall-less bacterium - is most commonly found in young, splenectomized or immuncompromised cattles in which it attaches itself to erythrocytes. The infection is mostly asymptomatic and rarely requires treatment. Symptoms which are associated with M. wenyonii infections are anemia, bacteremia, edema (scrotal, hind, and limb), lymphadenopathy, transient fever, anorexia, weight loss, infertility, and decreased milk production. M. wenyonii appears to have low morbidity and animals can become chronic, asymptomatic carriers.	[dos Santos *et al.*, 2012, Genova *et al.*, 2011]	2

Output: pPHAT+pELD:0, pPHAT+nELD:14, nPHAT+pELD:9, nPHAT+nELD:658 - Input: ELD: 9, nELD: 643

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Francisella tularensis subsp. tularensis SCHU S4	27	H00312 Tularemia	Animal pathogen - primary infection	Francisella tularensis is the etiologic agent of the zoonotic disease tularemia and - with an infectious dose between 10 and 50 CFU - one of the most infectious bacterial pathogens known to humans and animals. The known routs of infection are through the bite of an infected arthropod (ticks, deerflies and mosquitoes), by coming in contact with infected animal carcasses, through the ingestion of contaminated food or water and by inhalation of infective aerosols.	[Larsson *et al.*, 2005, Molins *et al.*, 2010]	4

Output: pPHAT+pELD:2, pPHAT+nELD:38, nPHAT+pELD:34, nPHAT+nELD:1455 - Input: ELD: 82, nELD: 1522

ACC	AI	Domains	Description
YP_170294	0.89	PF02931	ADP/ATP translocase
YP_170045	1.17	PF00328	ADP/ATP translocase

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Streptococcus pyogenes MGAS315	28	H00333 Streptococcal infection	Human pathogen - primary infection	Streptococcus pyogenes MGAS315 belongs to the group A Streptococcus (GAS) and an infection manifests with the following, possible symptoms: Streptococcal toxic shock-like syndrome, pharyngitis, cellulitis, sepsis, necrotizing fasciitis, and poststreptococcal acute rheumatic fever and shows unusually high rates of morbidity and mortality.	[Larsson *et al.*, 2005, Molins *et al.*, 2010]	3

Output: pPHAT+pELD:0, pPHAT+nELD:0, nPHAT+pELD:36, nPHAT+nELD:1822 - Input: ELD: 36, nELD: 1829

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Escherichia coli CFT073	276	H00279 Uropathogenic Escherichia coli (UPEC) infection	Human pathogen - primary infection	Uropathogenic Escherichia coli (UPEC) is the most abundant etiological agent of community-acquired urinary tract infections (UTI) accounting for about 80% of UTI infections worldwide. Uropathogenic Escherichia coli (UPEC) infection is a prevalent infectious disease with potentially severe complications (e.g., pyelonephritis).	[Dhakal *et al.*, 2008]	3

Output: pPHAT+pELD:0, pPHAT+nELD:2, nPHAT+pELD:155, nPHAT+nELD:5167 - Input: ELD: 155, nELD: 5214

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Streptococcus mutans UA159	8	H00297 Plague	Human pathogen - primary infection	Streptococcus mutans is the leading cause of dental caries worldwide - one of the most common infectious diseases afflicting humans and S.mutans is also considered to be the most cariogenic of all of the oral streptococci. Dental caries is one of the most common infectious diseases afflicting humans.	[Maruyama *et al.*, 2009, Suzuki *et al.*, 2011, Vickerman *et al.*, 2007]	3

Output: pPHAT+pELD:0, pPHAT+nELD:2, nPHAT+pELD:29, nPHAT+nELD:1921 - Input: ELD: 29, nELD: 1931

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Yersinia pestis CO92	123	/	Human pathogen, Animal pathogen - primary infection	Y. pestis is the etiological agent of plague - a deadly infectious disease. The reservoirs of Y. pestis are rodents and transmission to humans is accomplished by fleas as vectors. There are three major manifestations: bubonic, septicemic, and pneumonic plagues. Severe epidemics in the past include the Plague of Justinian in the sixth century, the Black Death in the fourteenth century, and the Third Pandemic that began in Central Asia in the nineteenth century.	[Parkhill *et al.*, 2001]	4

Output: pPHAT+pELD:0, pPHAT+nELD:6, nPHAT+pELD:64, nPHAT+nELD:3754 - Input: ELD: 159, nELD: 3907

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Coxiella burnetii RSA 493	45	H00310 Q fever	Human pathogen, Animal pathogen - primary infection	The etiological agent of Q-fever - Coxiella burnetii (a gram-negative obligate intracellular bacterium) - is a category B bioterrorism agent that is highly infective - the ID_{50} is one via inhalation - to both humans and livestock. Infection is commonly asymptomatic but may appear either in an acute form - manifested mainly by fever (self-limited flu-like disease, pneumonia, or hepatitis) or in a chronic form (mostly endocarditis).	[Kabir, 2011, Seshadri *et al.*, 2003]	4

Output: pPHAT+pELD:7, pPHAT+nELD:28, nPHAT+pELD:19, nPHAT+nELD:1765 - Input: ELD: 26, nELD: 1821

ACC	AI	Domains	Description
NP_820201	1.8	PF01222	Delta(24(24(1)))-sterol reductase
NP_819913	100	PF00086	Thyroglobulin type-1 repeat containing protein
NP_820468	1.1	PF08613	cyclin
NP_819085	0.75	PF00368	Hydroxymethylglutaryl-coenzyme A reductase
NP_820155	0.83	PF01222	Sterol delta-7-reductase
NP_819228	1.3	PF00961	Homing Endonuclease, 23 rRNA intron
YP_002332981	1.1	PF07671	HP
YP_170294	0.89	PF02931	ADP/ATP translocase
YP_170045	1.17	PF00328	ADP/ATP translocase

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Rickettsia typhi str. Wilmington	156	H00322 Typhus fever	Human pathogen - primary infection	The rickettsiae - gram-negative, aerobic, coccobacillary, α-proteobacteria - are obligate intracellular parasites with a life cycle that involves both vertebrate and invertebrate hosts. Rickettsia typhi is the etiological agent of typhus fever - infective louse and flea are the transmitting vectors to human. The disease is widely distributed in tropical and subtropical areas. Typhus fever present with constitutional symptom and maculopapular exanthema.	[Parola *et al.*, 2005]	3

Output: pPHAT+pELD:3, pPHAT+nELD:33, nPHAT+pELD:7, nPHAT+nELD:793 - Input: ELD: 10, nELD: 827

ACC	AI	Domains	Description
YP_067421	0.74	PF03219	ADP/ATP carrier protein 3
YP_067327	0.65	PF03219	ADP/ATP carrier protein 3
YP_067440	0.8	PF03219	ADP/ATP carrier protein 3

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Haemophilus influenzae R2846	14	H00304 Haemophilus influenzae infection	Human pathogen - primary infection	Haemophilus influenzae are subdivided into 6 serotypes according to their capsular polysaccharide composition. Serotype b (Hib) is the most virulent - in healthy children (below the age of 6) it is estimated that Hib was responsible for 8 million serious infections worldwide and 371,000 deaths in 2002. In young children, Hib is the most frequent cause of bacterial meningitis, with a fatality rate of 5% to 10% in developed countries. Up to one-third of survivors develop significant long-term neurologic complications. Hib is also by far the most important cause of epiglottitis, which accounted for about 15% of all invasive Hib infections.	[et al., 1995, Ladhani, 2012]	3

Output: pPHAT+pELD:1, pPHAT+nELD:1, nPHAT+pELD:23, nPHAT+nELD:1607 - Input: ELD: 23, nELD: 1613

ACC	AI	Domains	Description
YP_005829448	0.97	PF00312	30S ribosomal protein S15

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Onion yellows phytoplasma OY-M	62	/	Plant pathogen - primary infection	The genus Phytoplasma - members of this genus are without cell wall and inhabit typically phloem sieve elements - is an economically important group of bacterial phytopathogens transmitted by sap-sucking insect vectors. Phytoplasma infections can cause extensive phloem necrosis.	[Lee *et al.*, 2000]	3

Output: pPHAT+pELD:7, pPHAT+nELD:5, nPHAT+pELD:82, nPHAT+nELD:627 - Input: ELD: 99, nELD: 650

ACC	AI	Domains	Description
NP_950553	0.61	PF09419	
NP_950369	0.92	PF00004	ATP-dependent Zn protease
NP_950784	1.03	PF00004	
NP_950947	0.97	PF00004	ATP-dependent Zn protease
NP_950929	0.98	PF00004	
NP_950296	1.06	PF00004	
NP_950813	0.93	PF00004	ATP-dependent Zn protease

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	127	H00355 Leptospirosis infectious disease	Human pathogen, Animal pathogen - primary infection	Leptospirosis is an emerging zoonotic bacterial infection caused by spirochetes belonging to the genus Leptospira. Leptospira has a wide host range and is recognized as an important emerging global public health problem because of its possible epidemic features and increasing incidence in both developing and developed countries. Asymptomatic infections are common in endemic areas. However, small numbers of the patients may develop febrile illness with icterus, splenomegaly and nephritis, febrile illness with pulmonary hemorrhages or ocular manifestations.	[Ren *et al.*, 2003]	3

Output: pPHAT+pELD:14, pPHAT+nELD:69, nPHAT+pELD:143, nPHAT+nELD:3436 - Input: ELD: 157, nELD: 3510

ACC	AI	Domains	Description
YP_002595	0.89	PF10520	
YP_001985	0.52	PF01344	
YP_003198	1	PF12796	Ankyrin
YP_000337	0.74	PF01731	Paraoxonase
YP_002923	0.93	PF12796	Ankyrin
YP_001274	0.93	PF04768	Acetylglutamate Kinase
YP_003211	0.95	PF02493	
YP_002479	0.93	PF08713	
YP_001020	1	PF12796	Ankyrin
YP_000239	1.16	PF12796	Ankyrin
YP_000473	0.95	PF02493	
YP_003271	0.77	PF03088	Strictosidine Synthase
YP_002325	1.02	PF12796	Ankyrin
YP_003053	0.86	PF01210, PF07479	NAD(P)H-dependent glycerol-3-phosPHATe dehydrogenase

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Chlamydia trachomatis D/UW-3/CX	27	H00347 - Chlamydial urethritis, H00348 - Lymphogranuloma venereum, H00349 - Trachoma	Human pathogen, Animal pathogen - primary infection	Chlamydiaceae are obligate intracellular pathogens which undergo a unique biphasic life cycle that involves the inter-conversion between the extracellular infectious elementary body and the intracellular, replicative reticulate body. C. trachomatis causes several human diseases of medical significance. Trachoma is the major infectious reason of blindness worldwide. Infections with serovars A, B, Ba and C of Chlamydia trachomatis leads to chronic keratoconjunctivitis in children which often results in scarring and blindness in adults. Only in 2002 at least 1.3 million people in poor, rural communities in developing countries - especially in sub-Saharan Africa - were blind from trachoma.	[Stephens *et al.*, 1998]	3

Output: pPHAT+pELD:2, pPHAT+nELD:48, nPHAT+pELD:20, nPHAT+nELD:825 - Input: ELD: 22, nELD: 873

ACC	AI	Domains	Description
NP_219568	0.81	PF03219	ADP/ATP translocase
NP_219973	0.79	PF03219	SWIB (YM74) complex protein

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	162	H00311 Legionellosis Infectious disease	Human pathogen, primary/secondary infection	Legionella spp. - a gram-negative bacteria - is the etiological agent of Legionnaires disease. It can be nosocomial, community acquired or travel related. Legionella spp. colonize potable water systems and can be transmitted by inhalation of contaminated aerosols and aspiration. Two different forms of disease can be caused by Legionella spp. in humans: Legionaires disease - with an incubation period of 2-10 days - is a multisystem illness that involves the lungs, causing pneumonia and can result in neurological symptoms, diarrhea and high mortality (up to 50%). Legionella spp. can also cause Pontiac fever which presents itself with shorter incubation period (1-2 days) and results in acute, self-limited, influenza-like disease that does not cause pneumonia. Patients at risk for infection include the immunosuppressed, those with chronic lung disease, smokers and the elderly.	[Carratala & Garcia-Vidal, 2010]	3

Output: pPHAT+pELD:11, pPHAT+nELD:48, nPHAT+pELD:51, nPHAT+nELD:2821 - Input: ELD: 62, nELD: 2881

ACC	AI	Domains	Description
YP_095005	0.73	PF01150	Ecto-ATP diphosphohydrolase II
YP_095964	0.81	PF13516	

YP_095152	0.66	PF00328	Major Acid PhosPHATase
YP_095966	0.52	PF01369	Guanine nucleotide exchange protein
YP_094355	1	PF11901	
YP_095974	1.13	PF13516	
YP_096947	1.12	PF08123	
YP_095922	0.83	PF01150	Ectonucleoside triphosPHATe diphosphohydrolase I
YP_094149	0.57	PF05761	Cytosolic IMP-GMP specific 5-nucleotidase
YP_094149	0.57	PF13855	Leucine rich repeat

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Staphylococcus aureus RF122	57	/	Animal pathogen - primary infection - bovine mastitis-causing	Staphylococcus aureus - a highly adaptive and versatile gram-positive bacterium - infection results in a variety of serious diseases in multiple host species (including humans). In comparison with the sequenced S. aureus strains associated with human infection, it attracts attention that the allelic variation in bovine strain RF122 was high among virulence and surface-associated genes involved in host colonization, toxin production, iron metabolism, antibiotic resistance and gene regulation. Furthermore it should be noted that the majority of the RF122-unique genes were encoded by mobile genetic elements (Transposons, Plasmids, Group 2 introns).	[Herron-Olson *et al.*, 2007,Xue *et al.*, 2011]	3
Output: pPHAT+pELD:0, pPHAT+nELD:2, nPHAT+pELD:40, nPHAT+nELD:2461 - Input: ELD: 40, nELD: 2469						

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Bartonella quintana str. Toulouse	35	H00326 - Bartonellosis, H00327 - Trench fever	Human pathogen - primary infection	The genus Bartonella contains three major human pathogens which are facultative intracellular bacteria. Bartonella quintana is the causative agent of trench fever. Bartonella quintana infects humans via the human body louse, Pediculus humanus, which also serves as the vector for the typhus pathogen, Rickettsia prowazekii. B. quintana and B.henselae are unique among bacterial pathogens. They are able to induce tumor-like lesions of the skin (bacillary angiomatosis), the liver, and the spleen (bacillary peliosis) in immunocompromised individuals, predominantly AIDS patients. Not only the host immunity is responsible for the infection, other factors such as the virulence of the bacteria, the inoculation route, and the participation of vectors also play a crucial role.	[Alsmark *et al.*, 2004, Velho *et al.*, 2003]	3
Output: pPHAT+pELD:0, pPHAT+nELD:1, nPHAT+pELD:1, nPHAT+nELD:1137 - Input: ELD: 1, nELD: 1141						

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Mycoplasma hyopneumoniae 232	88	/	Swine pathogen - primary infection	Mycoplasma hyopneumoniae is the etiological agent of porcine enzootic pneumonia. Symptoms of this mild, chronic pneumonia of swine are often complicated by opportunistic infections with other bacteria.	[Minion *et al.*, 2004, Opriessnig *et al.*, 2011]	3

Output: pPHAT+pELD:0, pPHAT+nELD:6, nPHAT+pELD:7, nPHAT+nELD:678 - Input: ELD: 7, nELD: 684

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Legionella pneumophila str. Lens	154	H00311 Legionellosis Infectious disease	Human pathogen, primary/secondary infection	Already described for Legionella pneumophila subsp. pneumophila str. Philadelphia 1.	[Carratala & Garcia-Vidal, 2010]	3

Output: pPHAT+pELD:12, pPHAT+nELD:52, nPHAT+pELD:55, nPHAT+nELD:2801 - Input: ELD: 67, nELD: 2867

ACC	AI	Domains	Description
YP_125472	0.57	PF05761	
YP_127384	0.99	PF12796	
YP_125721	1	PF11901	
YP_126094	1.05	PF12796	
YP_127648	1.07	PF13516	
YP_127387	1.07	PF12796	
YP_127630	0.95	PF09445	
YP_127207	0.82	PF01150	
YP_127255	0.51	PF01369	RaIF protein, translocated into host cells by the Dot/Icm system
YP_126475	0.67	PF00328	Major Acid PhosPHATase Map (Histidine-acid phosPHATase)
YP_126358	0.73	PF01150	
YP_128208	1.14	PF08123	

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Campylobacter lari RM2100	10	H00321 Campylobacter infection	Human pathogen, animal pathogen - secondary infection	The clinical symptomatology of C. lari infections constitutes of gastroenteritis with abdominal pain, fever and diarrhea. Campylobacter lari has also been associated with bacteremia in immunocompromised and in patients with gastroenteritis. Serious sequelae of Campylobacter infection have been reported, including Guillian-Barre and Reiter syndromes.	[Jousimies-Somer, 1997, Miller *et al.*, 2008]	3

Output: pPHAT+pELD:0, pPHAT+nELD:2, nPHAT+pELD:20, nPHAT+nELD:1522 - Input: ELD: 20, nELD: 1525

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1	24	/	Plant symbiont	Polynucleobacter necessarius subsp. asymbioticus strain QLW-P1DMWA-1T is a planktonic freshwater bacterium which resides in the family Burkholderiaceae (class Betaproteobacteria). This strain earned attention because it represents a subspecies with cosmopolitan and ubiquitous distribution in standing freshwater systems.	[Meincke *et al.*, 2012]	0

Output: pPHAT+pELD:0, pPHAT+nELD:9, nPHAT+pELD:28, nPHAT+nELD:2037 - Input: ELD: 28, nELD: 2049

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Xanthomonas campestris pv. vesicatoria str. 85-10	104	/	Plant pathogen - primary infection	Xanthomonas campestris pv. vesicatoria - a gram-negative, rod-shaped proteobacterium with a high genomic G+C content - is the causative agent of bacterial spot disease in pepper and tomato plants. Xanthomonas campestris colonize - locally restricted - plant intercellular spaces which leads to macroscopically visible disease symptoms which include water-soaked lesions that later become necrotic. Pathogenicity of X. campestris pv. vesicatoria depends on the type III protein secretion system which is encoded by the chromosomal hrp gene cluster (hypersensitive response and pathogenicity) and translocates effector proteins into the plant cell.	[Thieme *et al.*, 2005, Gürlebeck *et al.*, 2006]	3

Output: pPHAT+pELD:2, pPHAT+nELD:31, nPHAT+pELD:102, nPHAT+nELD:4493 - Input: ELD: 143, nELD: 4583

ACC	AI	Domains	Description
YP_364278	0.64	PF04768	Acetylglutamate Kinase
YP_365163	1.04	PF03080	

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Actinobacillus succinogenes 130Z	20	/	Animal symbiont	A. succinogenes is a member of the Pasteurellaceae family. While most Pasteurellaceae are investigated for their pathogenic traits, A. succinogenes is studied for its industrially attractive metabolic trait of succinate production. Because A. succinogenes and has never been reported in association with any disease, Mckinlay et al. investigated its genome sequences for Pasteurellaceae pathogenicity genes and confirmed its potential for non-pathogenicity.	[Guettler *et al.*, 1999, Mckinlay *et al.*, 2010]	0

Output: pPHAT+pELD:0, pPHAT+nELD:8, nPHAT+pELD:22, nPHAT+nELD:2047 - Input: ELD: 22, nELD: 2057

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Campylobacter fetus subsp. fetus 82-40	3	H00321 - Campylobacter infection	Human pathogen, animal pathogen - secondary infection	Campylobacter infection is most commonly caused by Campylobacter jejuni and manifests itself as a diarrhoeal disease with the rare possibility of affecting the bloodstream. Serious sequelae of Campylobacter infection - including Guillian-Barre and Reiter syndromes - has been reported. The incidence of campylobacteriosis in HIV-infected patients is higher than in the general population. Deaths from C. jejuni infection are rare but may occur, primarily in infants, the elderly and patients with underlying illnesses.	[Schmidt *et al.*, 1980, Altekruse *et al.*, 1999]	3

Output: pPHAT+pELD:0, pPHAT+nELD:6, nPHAT+pELD:2, nPHAT+nELD:1707 - Input: ELD: 2, nELD: 1717

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Sinorhizobium fredii NGR234	102	/	Plant symbiotic	Sinorhizobium fredii NGR234 - a Gram-negative, motile and aerobic α-proteobacteria belonging to the order Rhizobiales - is classified as a plant symbiont and forms nitrogen-fixing nodules with more legumes than any other microsymbiont.	[Margaret *et al.*, 2011, Schmeisser *et al.*, 2009]	0

Output: pPHAT+pELD:3, pPHAT+nELD:15, nPHAT+pELD:843, nPHAT+nELD:5399 - Input: ELD: 867, nELD: 5495

ACC	AI	Domains	Description
YP_002824454	0.82	PF03328	4-hydroxy-2-oxovalerate aldolase
YP_002823667	0.98	PF04828	Glutathione-dependent formaldehyde-activating protein
YP_002826446	0.96	PF04828	Glutathione-dependent formaldehyde-activating protein

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Shewanella putrefaciens 200	10	/	Human pathogen - secondary infection - very rare	Shewanella sp is ubiquitous present in a wide range of natural and engineered habitats. It has the ability to conduct dissimilatory metal reduction which can is extensively studied for potential applications. Shewanella putrefaciens - a facultative anaerobe Gram-negative bacterium with the ability to reduce iron and manganese metabolically - has been isolated from marine environments as well as from anaerobic sandstone in the Morrison formation in New Mexico. There is no lead on any pathogenicity.	[Furukawa & Dale, 2013, Pagani *et al.*, 2003]	1

Output: pPHAT+pELD:0, pPHAT+nELD:9, nPHAT+pELD:34, nPHAT+nELD:4021 - Input: ELD: 41, nELD: 4138

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Chlamydophila pneumoniae LPCoLN	61	H00351 - Chlamydial pneumonia	Animal pathogen (Koala) - primary infection	Chlamydiaceae are obligate intracellular pathogens which undergo a unique biphasic life cycle that involves the inter-conversion between the extracellular infectious elementary body and the intracellular, replicative reticulate body. Chlamydophila pneumoniae is one of the most successful chlamydial species. It has successfully established a niche in a range of warm-blooded and cold-blooded hosts which include humans, horses, marsupials, frogs and reptiles and the koala. The koala (Phascolarctos cinereus) is commonly infected with C. pneumoniae, developing respiratory illnesses similar to that in humans with clinical symptoms such as sneezing, coughing, nasal discharge and chest congestion.	[Mitchell *et al.*, 2010, Myers *et al.*, 2009]	3

Output: pPHAT+pELD:2, pPHAT+nELD:56, nPHAT+pELD:23, nPHAT+nELD:1024 - Input: ELD: 25, nELD: 1080

ACC	AI	Domains	Description
YP_005661727	0.73	PF03219	ADP/ATP carrier protein family
YP_005661992	0.8	PF03219	ADP/ATP carrier protein family

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Lactococcus garvieae Lg2	12	/	Human pathogen, Animal pathogen - primary infection	Lactococcus garvieae is a major pathogen of fish and causes fatal haemorrhagic septicemia. The septicemia infection caused by L. garvieae is named lactococcosis and causes important economic losses in the fish farming industry. L. garvieae has also been isolated from buffaloes with mastitis and clinical specimens of human blood, urine, and skin. Based on this findings L. garvieae is considered an emerging zoonotic pathogen.	[Morita *et al.*, 2011]	3

Output: pPHAT+pELD:0, pPHAT+nELD:1, nPHAT+pELD:29, nPHAT+nELD:1908 - Input: ELD: 29, nELD: 1939

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Bacteroides vulgatus ATCC 8482	19	/	Human symbiont	Bacteroides vulgatus ATCC 8482 is derived from the distal gut of healthy human and shows no signs of pathogenicity.	[Xu *et al.*, 2007]	0

Output: pPHAT+pELD:0, pPHAT+nELD:8, nPHAT+pELD:302, nPHAT+nELD:3673 - Input: ELD: 307, nELD: 3759

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	38	/	Secondary infection	Nocardiopsis dassonvillei is of interest because of its ecological adaptabilities. Members of N. dassonvillei have been isolated from a large variety of natural environments such as soil and marine sediments, from different animal and plant tissues as well as from human patients presenting with conjunctivitis or cholangitis.	[Xu *et al.*, 2007]	2

Output: pPHAT+pELD:0, pPHAT+nELD:13, nPHAT+pELD:177, nPHAT+nELD:5307 - Input: ELD: 177, nELD: 5320

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Polynucleobacter necessarius subsp. necessarius STIR1	14	/	Plant symbiont	Polynucleobacter necessarius is an obligate endosymbiont of freshwater ciliates living in the cytoplasm of E. aediculatus.	[Hahn *et al.*, 2009]	0

Output: pPHAT+pELD:0, pPHAT+nELD:5, nPHAT+pELD:2, nPHAT+nELD:1500 - Input: ELD: 2, nELD: 1506

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476	210	H00113 Salmonellosis	Human pathogen - primary infection	In recent years, S. Heidelberg has been the third or fourth most common Salmonella serotype causing disease in the U.S., estimated at 84,000 cases of illness annually.Salmonellosis (non-typhoid) is an infectious disease caused by any serotype of bacteria in the genus Salmonella, other than Salmonella that causes typhoid and paratyphoid fever. Salmonellosis is one of the most frequently occurring food-borne diseases worldwide, usually acquired via the food chain. Foods prepared with contaminated raw eggs, egg products, and insufficiently heated poultry meat and pork have been identified as the primary sources of human Salmonella infections. Salmonellosis is characterized by the acute onset of fever, abdominal pain, nausea, diarrhea, and sometimes vomiting. The disease is usually self-limiting, lasting a few days. However, some cases have been reported with more serious complications following infection, especially in immunocompromised people, the elderly, pregnant women and children.	[Alvarez-Ordóñez *et al.*, 2011, Fricke *et al.*, 2011]	3

Output: pPHAT+pELD:1, pPHAT+nELD:5, nPHAT+pELD:104, nPHAT+nELD:4648 - Input: ELD: 105, nELD: 4675

ACC	AI	Domains	Description
YP_002047662	0.62	PF14269	PQQ enzyme repeat

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Streptococcus gordonii str. Challis substr. CH1	15	/	Human pathogen - primary infection	Originally isolated from human teeth plaque. Streptococcus gordonii is a primary colonizer of the multispecies biofilmmulti species on tooth surfaces forming dental plaque and a potential agent of endocarditis. Many oral streptococci, including Streptococcus gordonii, are naturally transformable, being able to take up exogenous DNA from the environment and incorporate it into their genome.	[Vickerman *et al.*, 2007]	3

Output: pPHAT+pELD:0, pPHAT+nELD:3, nPHAT+pELD:51, nPHAT+nELD:1996 - Input: ELD: 51, nELD: 2000

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Cryptobacterium curtum DSM 15641	8	/	Human pathogen - primary/secondary infections	The type strain 12-3T and a second strain of the species, KV43-B, both classified in C. curtum were isolated from a periodontal pocket sample of an adult patient and from necrotic dental pulp. C. curtum can also be isolated from human oral and dental infection like pulpal inflammations, advanced carries, dental abscesses or periodontitis.	[Kumar *et al.*, 2003, Nakazawa *et al.*, 1999, Mavrommatis *et al.*, 2009]	3

Output: pPHAT+pELD:0, pPHAT+nELD:6, nPHAT+pELD:22, nPHAT+nELD:1327 - Input: ELD: 22, nELD: 1335

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Marinomonas posidonica IVIA-Po-181	21	/	Plant symbiont	/	[Lucas-Elío *et al.*, 2012]	0

Output: pPHAT+pELD:0, pPHAT+nELD:20, nPHAT+pELD:40, nPHAT+nELD:3429 - Input: ELD: 40, nELD: 3451

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Francisella tularensis subsp. tularensis NE061598	27	H00312 Tularemia	Human pathogen - primary infection	Francisella tularensis - a highly pathogenic gram-negative coccobacillus - is the etiological agent of tularemia. The vast majority of infections is ulceroglandular in nature and transmitted by contact with an infected host (e.g., rabbit or cat) or vector (e.g., tick or mosquito). A minority of infections displace serious symptoms such as pneumonic tularemia and can be life-threatening. F. tularensis is considered a potential biowarfare agent.	[Nalbantoglu *et al.*, 2010]	4

Output: pPHAT+pELD:5, pPHAT+nELD:35, nPHAT+pELD:37, nPHAT+nELD:1632 - Input: ELD: 90, nELD: 1746

ACC	AI	Domains	Description
YP_005831663	0.82	PF02274	
YP_005831889	0.89	PF02931	

YP_005831982	0.97	PF11940	Aminopeptidase N
YP_005830479	0.79	PF02274	
YP_005831228	1.17	PF00328	

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Streptococcus mutans NN2025	9	/	Human pathogen - primary infection	Streptococcus mutans is the major pathogen of dental caries. Occasionally infection can also lead to endocarditis.	[Maruyama *et al.*, 2009]	3

Output: pPHAT+pELD:0, pPHAT+nELD:0, nPHAT+pELD:24, nPHAT+nELD:1869 - Input: ELD: 24, nELD: 1871

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Candidatus Liberibacter asiaticus str. psy62	17	/	Plant pathogen - primary infection	Citrus huanglongbing - caused by any of three uncultured species of Gram-negative α-Proteobacteria: Candidatus Liberibacter asiaticus, Ca. L. americanus, and Ca. L. africanus - is the most destructive disease of citrus worldwide. In consistence with its intracellular nature, Ca. L. asiaticus displays no type III and type IV secretion systems and also lacks extracellular degradative enzymes typical for free-living or plant-colonizing bacteria. Instead it appears to have all type I secretion system genes needed for toxin effector secretion and multidrug efflux.	[Duan *et al.*, 2009]	3

Output: pPHAT+pELD:1, pPHAT+nELD:17, nPHAT+pELD:17, nPHAT+nELD:1067 - Input: ELD: 20, nELD: 1089

ACC	AI	Domains	Description
YP_003064736	0.55	PF03219	

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Staphylococcus aureus subsp. aureus TCH60	54	/	Human pathogen - secondary infection - community acquired meticillin susceptible	Already described for Staphylococcus aureus RF122.	[Highlander *et al.*, 2007]	1

Output: pPHAT+pELD:0, pPHAT+nELD:5, nPHAT+pELD:45, nPHAT+nELD:2643 - Input: ELD: 45 nELD: 2648

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Listeria monocytogenes HCC23	15	H00332 Listeriosis	Avirulant strain - Human pathogen - primary infection	Listeriosis - caused by Listeria monocytogenes, a Gram-positive, facultative anaerobic bacteria - is a severe food borne infection with a high case fatality rate. Groups at high-risk for listeriosis include the elderly and immunocompromised patients, newborns, pregnant women and infants.	[Gilmour *et al.*, 2010]	3

Output: pPHAT+pELD:1, pPHAT+nELD:6, nPHAT+pELD:113, nPHAT+nELD:2852 - Input: ELD: 114, nELD: 2860

ACC	AI	Domains	Description
YP_002350681	0.9	PF08719	Swarming motility protein YbiA

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Bifidobacterium animalis subsp. lactis DSM 10140	13	/	Human symbiont	Bifidobacterium animalis subsp. lactis - a gram-positive lactic acid bacterium - is usually located in the guts of healthy humans and has been identified in the infant gut biota. It is a dominant genus considered beneficial to humans and includes probiotic strains.	[Barrangou *et al.*, 2009]	0

Output: pPHAT+pELD:0, pPHAT+nELD:5, nPHAT+pELD:14, nPHAT+nELD:1546 - Input: ELD: 14, nELD: 1552

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Pectobacterium wasabiae WPP163	144	/	Plant pathogen - primary infection	P. wasabiae isolates are pathogenic to potatoes causing soft rotting.	[Pitman & Harrow, 2010]	3

Output: pPHAT+pELD:5, pPHAT+nELD:5, nPHAT+pELD:97, nPHAT+nELD:4324 - Input: ELD: 102, nELD: 4335

ACC	AI	Domains	Description
YP_003260359	0.62	PF13343	Extracellular solute-binding protein
YP_003260359	0.62	PF13343	Extracellular solute-binding protein
YP_003259508	1.18	PF12796	Ankyrin
YP_003259880	0.67	PF13343	Extracellular solute-binding protein
YP_003258518	0.55	PF12796	Ankyrin
YP_003258335	1.09	PF06045	Rhamnogalacturonate lyase

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Bifidobacterium longum subsp. infantis 157F	9	/	Human symbiont	Bifidobacterium longum - a Gram-positive anaerobic branched rod-shaped bacterium - can be found in the intestines of infant humans. Inhibiting the growth of Gram-negative bacteria by producing lactic acid Bifidobacterium can be seen as beneficial to humans; they are also able to digest the complex sugars in human breast milk.	[Sela *et al.*, 2008]	0

Output: pPHAT+pELD:0, pPHAT+nELD:7, nPHAT+pELD:38, nPHAT+nELD:1930 - Input: ELD: 47, nELD: 1952

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Acidaminococcus intestini RyC-MR95	4	/	Human pathogen/symbiont - secondary infection	A. intestini - a Gram-negative bacterium - colonize the human gut flora as commensal. It has been occasionally associated to infective processes but always as part of polymicrobial infections and might act as a reservoir of antibiotic resistance mechanism.	[D'Auria *et al.*, 2011]	1

Output: pPHAT+pELD:2, pPHAT+nELD:5, nPHAT+pELD:68, nPHAT+nELD:2308 - Input: ELD: 70, nELD: 2331

ACC	AI	Domains	Description
YP_004897753	0.64	PF05049	hypothetical protein
YP_004897754	0.51	PF05049	hypothetical protein

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Wolbachia endosymbiont of Culex quinquefasciatus Pel	126	H01043 - Onchocerciasis, H01086 - Lymphatic filariasis	Endosymbiont of arthropoda - Human pathogen (indirect)	Human onchocerciasis is an infection caused by the filarial nematode Onchocerca volvulus transmitted by Simulium blackflies. Wolbachia bacteria colonize Onchocerca volvulus as mutualistic symbionts and are widely associated with affecting the nematode survival.	[Klasson *et al.*, 2008]	3

Output: pPHAT+pELD:24, pPHAT+nELD:56, nPHAT+pELD:42, nPHAT+nELD:1039 - Input: ELD: 73, nELD: 1202

ACC	AI	Domains	Description
YP_001976067	1.05	PF12796	Ankyrin repeat domain protein
YP_001975172	1.09	PF12796	Ankyrin repeat domain protein
YP_001975493	1	PF12796	Ankyrin repeat domain protein
YP_001975547	1.08	PF12796	Ankyrin repeat domain protein
YP_001975413	1.02	PF12796	Ankyrin repeat domain protein
YP_001975955	1.11	PF12796	Ankyrin repeat domain protein
YP_001975179	1.12	PF12796	Ankyrin repeat domain protein
YP_001976103	0.94	PF12796	Ankyrin repeat domain protein
YP_001975536	1.08	PF12796	Ankyrin repeat domain protein
YP_001975139	0.94	PF12796	Ankyrin repeat domain protein
YP_001975025	1.11	PF12796	Ankyrin repeat domain protein
YP_001975234	1.01	PF12796	Ankyrin repeat domain protein
YP_001976012	1.01	PF12796	Ankyrin repeat domain protein
YP_001975451	1	PF12796	Ankyrin repeat domain protein

YP_001975284	1.11	PF12796	Ankyrin repeat domain protein
YP_001975372	0.99	PF12796	Ankyrin repeat domain protein
YP_001975491	1	PF12796	Ankyrin repeat domain protein
YP_001975140	1.04	PF12796,PF00023	Ankyrin repeat domain protein
YP_001975417	1.1	PF12796,PF00023	Ankyrin repeat domain protein
YP_001975197	1.01	PF12796,PF00023	Ankyrin repeat domain protein
YP_001975764	1.02	PF12796	Ankyrin repeat domain protein
YP_001975619	0.99	YP_001975619	Ankyrin repeat domain protein
YP_001975403	1.03	PF12796,PF13637	Ankyrin repeat domain protein
YP_001975193	0.95	PF12796,PF00023,PF13637	Ankyrin repeat domain protein

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Mycobacterium africanum GM041182	51	H00342 - Tuberculosis	Human pathogen - primary infection	Mycobacterium africanum is responsible for up to half of human tuberculosis (TB) in West Africa.	[Jong *et al.*, 2010]	3
Output: pPHAT+pELD:1, pPHAT+nELD:15, nPHAT+pELD:141, nPHAT+nELD:3657 - Input: ELD: 142, nELD: 3688						
ACC	AI	Domains	Description			
YP_004724760	0.93	PF01329	Pterin-4-alpha-carbinolamine dehydratase			

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Arcobacter nitrofigilis DSM 7299	29	/	Plant symbiont.	/	[Pati *et al.*, 2010]	0
Output: pPHAT+pELD:0, pPHAT+nELD:34, nPHAT+pELD:15, nPHAT+nELD:3076 - Input: ELD: 15, nELD: 3111						

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Escherichia coli IAI1	260	/	Human symbiont.	/	[Touchon *et al.*, 2009]	0
Output: pPHAT+pELD:0, pPHAT+nELD:2, nPHAT+pELD:110, nPHAT+nELD:4214 - Input: ELD: 110, nELD: 4239						

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence

Helicobacter pylori B38	19	H00320 - Helicobacter pylori infection	Human pathogen - primary infection	Helicobacter pylori - a Gram-negative, microaerophilic bacterium - is found in approximately 50% of the human population and are associated with several inflammatory gastroduodenal diseases. There is also a strong correlation with two types of gastric cancers: gastric adenocarcinoma and gastric extra-nodal marginal zone B-cell lymphoma. H. pylori pathogenicity mechanisms include the cag-pathogenicity island (cagPAI) as a major pro-inflammatory actor, but its association with MALT lymphoma strains has yet to be verified.	[Thiberge *et al.*, 2010]	3

Output: pPHAT+pELD:0, pPHAT+nELD:2, nPHAT+pELD:14, nPHAT+nELD:1351 - Input: ELD: 14, nELD: 1368

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Streptococcus dysgalactiae subsp. equisimilis ATCC 12394	23	/	Human pathogen and symbiont	S. dysgalactiae subsp. equisimilis was long regarded only as a human commensal organism but is recently also recognized as an increasingly important human pathogen. It is associated with causing a spectrum of human diseases including cellulitis, peritonitis, septic arthritis, pneumonia, endocarditis, acute pharyngitis, bacteremia and toxic shock syndrome. Many of these infections are similar to those caused by Streptococcus pyogenes in human, and indeed this two organisms share many of the same virulence genes.	[Suzuki *et al.*, 2011]	3

Output: pPHAT+pELD:0, pPHAT+nELD:0, nPHAT+pELD:44, nPHAT+nELD:2006 - Input: ELD: 45, nELD: 2011

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Treponema pallidum subsp. pallidum str. Chicago	19	H00354 - Syphilis	Human pathogen - primary infection	Treponema pallidum subsp. pallidum is the etiological agent of Syphilis - one of the oldest recognized venereal infection. Its symptoms include genital ulceration, skin rash, and development of serious systemic disease.	[Cejková *et al.*, 2012]	3

Output: pPHAT+pELD:0, pPHAT+nELD:9, nPHAT+pELD:4, nPHAT+nELD:967 - Input: ELD: 4, nELD: 977

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Corynebacterium pseudotuberculosis 1002	9	/	Animal pathogen - primary infection	Corynebacterium pseudotuberculosis - a facultative intracellular pathogen - causes caseous lymphadenitis and mainly infects sheep and goats but can also lead to ulcerative lymphangitis in equines; superficial abscesses in bovines, pigs, deer and laboratory animals; arthritis and bursitis in ovines; pectoral abscesses in equines and - more rarely - in camels, caprines and deer. There are also reports of rare cases of human infections.	[Ruiz *et al.*, 2011]	3

Output: pPHAT+pELD:0, pPHAT+nELD:1, nPHAT+pELD:10, nPHAT+nELD:2079 - Input: ELD: 10, nELD: 2080

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Corynebacterium pseudotuberculosis 1002	16	/	Animal pathogen - primary infection	Helicobacter mustelae infection manifests itself with gastritis, ulcers and gastric cancer in ferrets and other mustelids. H. mustelae remains the only helicobacter besides H. pylori that causes gastric ulceration and cancer in its natural host.	[O'Toole *et al.*, 2010]	3

Output: pPHAT+pELD:0, pPHAT+nELD:4, nPHAT+pELD:3, nPHAT+nELD:1391 - Input: ELD: 3, nELD: 1395

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Rothia mucilaginosa DY-18	74	/	Human symbiont - secondary infection	Rothia mucilaginosa is a member of commensal bacterial flora in the oral cavity and has received attention as a potential opportunistic pathogen.	[Nambu *et al.*, 2013]	1

Output: pPHAT+pELD:0, pPHAT+nELD:0, nPHAT+pELD:9, nPHAT+nELD:1894 - Input: ELD: 9, nELD: 1895

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Corynebacterium diphtheriae 31A	10	H00343 - Diphtheria	Human pathogen - primary infection	Corynebacterium diphtheriae - a gram-positive, highly pleomorphic bacterium - is one of the most prominent human pathogens and the etiological agent of the communicable disease diphtheria.	[Trost *et al.*, 2012]	4

Output: pPHAT+pELD:0, pPHAT+nELD:3, nPHAT+pELD:19, nPHAT+nELD:2330 - Input: ELD: 20, nELD: 2361

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Novosphingobium sp. PP1Y	32	/	Possible human pathogen	Novosphingobium sp. is a gram-negative, strictly aerobic, ubiquitous bacterium that is found in soil, water, and coastal plain and is recently discussed as a potential human pathogen.	[Bogdanos & Vergani, 2009]	?

Output: pPHAT+pELD:6, pPHAT+nELD:25, nPHAT+pELD:219, nPHAT+nELD:4363 - Input: ELD: 239, nELD: 4425

ACC	AI	Domains	Description
YP_004532680	0.84	PF03080,PF14365	hypothetical protein
YP_004533567	1.36	PF03332	HAD family hydrolase
YP_004533187	0.51	PF02359,PF02933	ATPase AAA
YP_004533918	2.64	PF13577	hypothetical protein
YP_004534061	0.58	PF07110	hypothetical protein
YP_004533910	2.64	PF13577	hypothetical protein

Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Bacillus thuringiensis BMB171	57	/	Insect pathogen - primary infection	Insect pathogen which is used as a biopesticide displaying a high transformation frequency.	[He *et al.*, 2010]	3
Output: pPHAT+pELD:1, pPHAT+nELD:4, nPHAT+pELD:322, nPHAT+nELD:5000 - Input: ELD: 323, nELD: 5029						
Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Erwinia amylovora ATCC 49946	118	/	Plant pathogen - primary infection	Erwinia amylovora - a plant-associated member of the Enterobacteriaceae - is associated with fire blight, a devastating infection of rosaceous plants, especially pear and apple.	[Sebaihia *et al.*, 2010]	3
Output: pPHAT+pELD:0, pPHAT+nELD:4, nPHAT+pELD:20, nPHAT+nELD:3539 - Input: ELD: 20, nELD: 3545						
Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Herbaspirillum seropedicae SmR1	108	/	Plant symbiont	/	[Baldani & Baldani, 2005]	0
Output: pPHAT+pELD:0, pPHAT+nELD:7, nPHAT+pELD:106, nPHAT+nELD:4619 - Input: ELD: 106, nELD: 4629						
Name	ELP	KEGG	Pathogenicity	Pathogenicity-Long	Source	Virulence
Rothia dentocariosa ATCC 17931	34	/	Human pathogen - secondary infection	Rothia dentocariosa infections can - in rare cases - lead to endocarditis. This inflammation of the endocardium occurs most frequently in patients with prior heart conditions.	[Boudewijns *et al.*, 2003]	2
Output: pPHAT+pELD:0, pPHAT+nELD:4, nPHAT+pELD:24, nPHAT+nELD:2184 - Input: ELD: 24, nELD: 2193						

Table D.1: Comprehensive table for results obtained for 70 prokaryotes including alien index, phylogenetic tree reconstruction and analysis with PHAT, protein domains and ELDs.

Education and training

Dates	1992-1996
Name and type of organisation providing education and training	Volksschule Henndorf

Dates	1996-2000
Name and type of organisation providing education and training	Hauptschule Henndorf

Dates	2000-2005
Name and type of organisation providing education and training	Bundesoberstufen Realgymnasium mit naturwissenschaftlichem Schwerpunkt
Level in national or international classification	Matura

Dates	October 2006 - March 2010
Field of study	Bachelorstudium Genetik
Name and type of organisation providing education and training	University Salzburg
Title gained	Bachelor herum naturalium

Dates	Since March 2010 to presumambly December 2013
Field of study	Masterstudium Biologische Chemie
Title of Master thesis	Spectral characterisation of different glycerol diphytanyl ethers as pure substances and part of the cell membrane of archaea using time-of-flight secondary ion mass spectrometry (ToF-SIMS) and matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-MS)
Submitted and graded	March 2013
Grade	Sehr gut (1)
Supervised and grades by	o. Prof. Dr. Michael Wagner

Department/s the Master thesis was done in	Department of microbial ecology Fakultät für Lebenswissenschaften Universität Vienna Althanstr. 14 A-1090 Wien Institute of Chemical Technologies and Analytics (ICTA) Technical University of Vienna Getreidemarkt 9/164 A-1060 Vienna
Name and type of organisation providing education and training	University Vienna

Dates	From March 2010 to November 2013
Field of study	Masterstudium Molekulare Biologie (Schwerpunkt Molekulare Medizin)
Title of Master thesis	Bacterial effector proteins in the evolution of pathogenic and symbiotic bacteria
Submitted and graded	November 2013
Grade	Sehr gut (1)
Supervised and grades by	Prof. Dr. Thomas Rattei
Department/s the Master thesis was done in	Division of Computational Systems Biology Fakultät für Lebenswissenschaften Universität Vienna Althanstr. 14 A-1090 Wien
Name and type of organisation providing education and training	University Vienna