universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„The evolution of type III secretion effectors"

verfasst von / submitted by

Dagmar Tiefenbrunner, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 910
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Master's degree programme Computational Science
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Abstract

Type III secretion effectors (T3SE) are important factors for host cell interaction in a wide variety of gram-negative pathogens and symbionts. Due to their close contact with host cells and the host immune system, they are subjected to high selective pressure. Their high evolvability may partly be attributed to their modular structure, with an N-terminal signal sequence followed by the functional domains. N-terminal elongation and terminal-reassortment have been proposed as potentially common mechanisms of T3SE evolution. Nonetheless, the evolution of novel effectors remains poorly understood.

This study aimed to collect further evidence for these evolutionary pathways or to find alternative ones, using computational methods. Further, it tried to investigate if T3SEs may originate from eukaryotic proteins gained by horizontal gene transfer, as this is a known evolutionary mechanism in some other virulence factors.

Alignments and comparisons between effectors and to other homologs were used to deduce possible evolutionary pathways. The results support terminal-reassortment as a common mechanism of T3SE evolution. For N-terminal elongation, they stayed inconclusive, and N-terminal truncation most probably does not contribute in any meaningful way to the emergence of new effectors. Analyses utilizing the T3SE prediction tool EffectiveT3 seem to indicate that certain types of sequences tend to resemble a T3S signal more closely than others. In particular, frameshifted N-termini may be a convenient source of new secretion signals in some taxa. Firm evidence for the involvement of a T3SE in horizontal gene transfer with a eukaryote could only be obtained for one *Chlamydial* effector. However, it remained unclear if *Chlamydia* gained the effector from *Trypanosomatidae* or if *Trypanosomatidae* acquired the effector from *Chlamydia*.

Zusammenfassung

Typ III Sekretions Effektoren (T3SE) spielen bei einer Vielzahl gram-negativer Pathogene und Symbionten eine entscheidende Rolle in der Interaktion zwischen Bakterium und Wirt. Wegen ihres engen Kontakts mit den Wirtszellen und dem Immunsystem des Wirts sind sie einem hohen Selektionsdruck ausgesetzt. Zum Teil könnte ihre Fähigkeit schnell zu evoluieren auf ihre modulare Struktur mit einer N-terminalen Signalsequenz, gefolgt von funktionalen Domänen, zurückzuführen sein. N-terminale Elongation und terminales Reassortment wurden als möglicherweise häufige Mechanismen der Entstehung neuer T3SEen vorgeschlagen. Dennoch ist die Evolution neuer Effektoren nur unzureichend geklärt.

Diese Studie beabsichtigte, mit computerbasierten Methoden diese beiden evolutionären Wege zu untermauern oder Alternativen dazu zu finden. Darüber hinaus versuchte sie herauszufinden, ob manche T3SEen aus eukaryotischen Proteinen entstehen, die das Bakterium durch horizontalen Gentransfer übernommen hat, da dies bei einigen anderen Virulenzfaktoren ein bekannter evolutionärer Mechanismus ist.

Alignments zwischen Effektoren untereinander und mit anderen Homologen wurden benutzt, um mögliche evolutionäre Wege abzuleiten. Die Ergebnisse unterstützen, dass terminales Reassortment ein häufiger Mechanismus der Evolution von T3SEen ist. Über N-terminale Elongation konnten sie keine klare Aussage treffen und N-terminale Verkürzung spielt höchstwahrscheinlich keine nennenswerte Rolle bei der Entstehung neuer Effektoren. Es wurden Analysen mit EffectiveT3, einem Programm zur Vorhersage von T3SEen, durchgeführt, die darauf hindeuten, dass einige Sequenztypen öfter einem T3S Signal ähneln als andere. In einigen Taxa könnten N-terminale Frameshifts eine besonders günstige Quelle neuer Sekretionssignale darstellen. Nur für einen Effektor konnte ein klarer Hinweis darauf gefunden werden, dass er in horizontalen Gentransfer mit einem Eukaryoten involviert war. Allerdings konnte nicht geklärt werden, ob *Chlamydia* den Effektor von *Trypanosomatidae* übernommen hat oder umgekehrt.

Table of Contents

1. Introduction

1.1. Virulence factors and the type III secretion system

1.1.1. Bacterial virulence factors

Virulence factors are molecules, synthesized by pathogens, that can aid in the evasion or suppression of the host immune system, in dissemination within the host and colonization of a niche, in adhesion to, entry into or exit from host cells, in the transmission to new hosts, or in obtaining nutrients from the host, damaging the host in the process (Webband and Kahler, 2008). They can interact with, modify or cleave host proteins, lipids and other molecules, interfere in various cellular pathways and affect host gene regulation (Coburn et al., 2007; Sharma et al., 2017).

Virulence factors can be encoded on the chromosome or mobile genetic elements (Arnold et al., 2009). The genes encoding them are sometimes clustered together in pathogenicity islands. Some are subject to rampant horizontal gene transfer (HGT), which allows virulence mechanisms to more easily spread between different pathogens. While some pathogenic bacteria are opportunistic and only virulent in hosts with defective defenses, others can act as pathogens in otherwise healthy hosts (Webband and Kahler, 2008). Some bacterial species, such as *Escherichia coli*, have pathogenic and non-pathogenic strains (Stromberg et al., 2018). Virulence genes are usually tightly regulated and only expressed upon contact with or invasion of the host (Webband and Kahler, 2008).

Capsules and slime layers are virulence factors that consist of a usually negatively charged layer enveloping the bacterial cell. While slime layers can easily be washed off, capsules cannot. Most capsules are composed of polysacharides. They protect bacteria from recognition by the immune system, interfering with complement-mediated obsonization and preventing phagocytic killing. Some capsules possess structures similar to host molecules, evading immune recognition by molecular mimicry. They are found among gram-negative and gram-positive bacteria (Wen and Zhang, 2015).

Bacterial toxins are divided into endotoxins and exotoxins. Endotoxins are lipopolysaccharides of the gram-negative outer membrane. They are important for the function and integrity of the membrane and are the main surface antigen of gram-negative bacteria (Moran et al., 1996). However, they can also help with immune evasion via molecular mimicry, by looking similar to host surface molecules, or via antigenic variation, by altering the structures displayed on the surface of the cell (Moran et al., 1996; van der Woude and Bäumler, 2004). Endotoxins are secreted as part of vesicles but can be released in greater quantities upon destruction of the bacterial cell (Dufour et al., 2017; Kulp and Kuehn, 2010). This exposes the membrane-anchor of the lipopolysaccharide, the lipid A, which is recognized by toll like receptor 4 and elicits a strong immune response. It causes inflammation, induces common symptoms of disease, such as fever, and may in severe cases lead to septic shock (John et al., 2017; Sampath, 2018).

Gram-negative and gram-positive bacteria utilize exotoxins. Exotoxins are usually polypeptides, proteins or protein complexes. They are secreted and either stay attached to the bacterial membrane, are released into the extracellular space or are injected into host cells (Green and Mecsas, 2016; Sastalla et al., 2016). Some exotoxins, the AB-toxins, are secreted into the extracellular space but act as intracellular toxins. They contain a subunit that attaches to host cell membranes and allows the catalytic subunit to enter the cell (Cherubin et al., 2016). Many exotoxins mediate very specific interactions with the host, and they can cause a huge variety of symptoms. Among the best known and most dangerous exotoxins are the neurotoxic AB-toxins of *Clostridium tetani* and *Clostridium botulinum.* Both tetanospasmin and botolinum toxin enter neurons and cleave SNARE proteins, preventing the release of

neurotransmitters and stopping nerve signaling. Botolinum toxin inhibits activatory neurons, leading to flaccid paralysis, and tetanospasmin inhibits inhibitory neurons, leading to spastic paralysis (Binz et al., 2010).

1.1.2. Protein secretion

Several secretion systems, which channel virulence factors, other proteins or other substances through cell membranes into the periplasm or out of the cell, have been identified in bacteria. Some export proteins into the extracellular space or surrounding medium, while others inject them into host cells. Most secretion systems are protein complexes in the bacterial cell membranes, and they utilize a variety of different mechanisms. Some secretion systems are specific to gram-positive or gram-negative bacteria, while others occur in both groups. Many recognize an N-terminal signal sequence, but some use a C-terminal one. Signal sequences are cleaved off during transport by some secretion systems but not by others.

The general secretion (Sec) and twin arginine translocation (Tat) pathways are the most conserved mechanisms of protein secretion and are used by bacteria, archaea and eukaryotes. In gram-negative bacteria, most proteins transported by the Tat pathway stay in the periplasm, and most proteins transported by the Sec pathway remain in the periplasm or are inserted into the inner membrane. However, some proteins that are delivered to the periplasm by the Sec or Tat pathway are exported by other secretion systems. In gram-positive bacteria, proteins in the periplasm do not need to pass another membrane to leave the cell, and most Tat-secreted proteins of gram-positive bacteria are released extracellularly (Green and Mecsas, 2016).

The Sec system transports proteins in their unfolded state through a channel in the cytoplasmatic membrane, which is the inner membrane in gram-negative bacteria. It uses different mechanisms for the translocation of proteins that are inserted into the cytoplasmatic membrane and proteins that are secreted into the periplasm. Proteins meant for the periplasm are bound by a chaperone, which prevents them from folding, and are translocated post-translationally. The energy for the transport is provided by ATPases and by the proton motive force. Proteins meant to be inserted into the membrane are translocated co-translationally. Protein synthesis pushes them through the channel (Green and Mecsas, 2016; Tsirigotaki et al., 2017).

The Tat system transports folded proteins across the cytoplasmatic membrane, driven by the proton motive force. Translocation pores form upon substrate binding. Unlike the Sec system, it can translocate proteins that need to acquire post-translational modifications or need to form protein complexes while they are still in the cytoplasm (Lee et al., 2006).

Type II and type V secretion systems are specific to gram-negative bacteria and depend on the Sec or Tat pathway for secretion. The type II secretion system (T2SS) is a protein complex that transports proteins in a folded state through a β-barrel channel from the periplasm into the extracellular space. The secretion is powered by an ATPase. The type V secretion system (T5SS) consists of a β-barrel domain that inserts itself into the outer membrane. In its most simple version, it is an autotransporter. The functional domain passes through the β-barrel channel and is either cleaved off or remains attached to the outer membrane. In other cases, one protein forms the channel and another passes through.

The type I secretion system (T1SS) is another molecular machinery of gram-negative bacteria that releases its substrates into the extracellular space. However, it does not depend on the Sec or Tat pathway, as it channels its substrates through both bacterial membranes in one step. Proteins pass through the channel in an unfolded state, and export is driven by an ATPase. The type I secretion system closely resembles ABC transporters, which export small molecules, such as primary and secondary metabolites as well as drugs and antibiotics (Green

and Mecsas, 2016).

The type III, type IV and type VI secretions systems are protein complexes of gram-negative bacteria that inject their substrates directly into the host cell. They transport their substrates from the bacterial cytoplasm into the host in a single step and do not depend on the Sec or Tat pathway. The type III secretion system (T3SS) is related to the basal body of the bacterial flagellum but possesses a needle-filament instead of a flagellum. It is driven by ATP hydrolysis and proton motive force and secrets unfolded proteins (Green and Mecsas, 2016 ; Wagner et al., 2018). The type IV secretion system (T4SS) is related to the bacterial conjugation system and can transport proteins, protein-protein complexes and protein-DNA complexes. It can not only secrete its substrates into eukaryotic cells but also into other bacteria. Secretion is powered by ATPase (Green and Mecsas, 2016). The type VI secretion system (T6SS) is structurally similar to the tail spike of the T4 phage and may use a similar method of secretion. It injects proteins into eukaryotic and bacterial cells. The necessary energy is provided by ATPases (Green and Mecsas, 2016; Navarro-Garcia et al., 2019).

Gram-negative bacteria can also secrete proteins and other material in outer membrane vesicles. Outer membrane vesicles consist of outer membrane and periplasm and are enriched in substrates meant for secretion. Soluble proteins are entrapped in the periplasm of the vesicle or adhere to the outside of it. Unlike other mechanisms, outer membrane vesicles also allow the secretion of insoluble molecules, including membrane proteins and lipopolysaccharides. Adhesins on the vesicles can target them to other gram-negative or to gram-positive bacteria or to eukaryotic cells. Vesicles can lyse and release their content extracellularly, or they can fuse with a target cell or be taken up by endocytosis (Kulp and Kuehn, 2010).

Some gram-positive bacteria utilize two distinct Sec pathways with different substrate specificities. Often, no additional mechanism is needed to export proteins into the extracellular space, as many can diffuse through the peptidoglycan layer. However, as gram-positive bacteria do not have an outer membrane, they need to embed outer surface proteins into their cell wall. This is done by sortases, which covalently attach proteins to the cell wall after they have been transported across the cell membrane. Direct injection of proteins into host cells is rarer in gram-positive bacteria than in gram-negative bacteria, and most virulence factors that act inside host cells are self-translocating AB-toxins. However, some gram-positive bacteria possess an injectosome, a secretion apparatus that functions similarly to the gram-negative T3SS and T4SS but is not structurally related. Unlike the T3SS and the T4SS, it relies on the Sec pathway to transport substrates across the plasma membrane. Some gram-positive bacteria possess a type VII secretion system, which transports proteins through the cell membrane and the cell wall in a Sec-independent manner. This system is particularly important in *Mycobacteria* and *Corynebacteria*, as their cell walls are heavily lipidated, which prevents proteins from diffusing through them (Green and Mecsas, 2016).

1.1.3. The type III secretion system

The type III secretion system is among the biggest membrane-localized protein complexes, consisting of roughly 20 different proteins with copy-numbers from 1 to about 100. It can be found in many gram-negative bacteria, spans both bacterial membranes and channels effector proteins from the bacterial cytosol directly into the host cell. T3SS substrates are secreted in an unfolded conformation by an ATP- and a proton motive force-driven mechanism.

The T3SS and the bacterial flagellum share the same evolutionary origin and a structurally similar cytoplasmic component and export apparatus. The secretion mechanism corresponds to the mechanism of flagellar self-assembly. However, instead of a flagellum, the T3SS builds

a needle filament.

Proteins secreted by the T3SS can be divided into early, intermediate and late substrates. After assembly of the cytoplasmic components of the T3SS, secretion of early substrates starts. These early substrates are inner rod proteins, needle filament proteins and needle-length control proteins and are integrated into the nascent T3SS. Needle subunits are channeled through the needle conduit and inserted into the distal end. Once the needle has reached its full length, a substrate specificity switch takes place, and secretion resumes with intermediate substrates. They form the needle tip and complete the T3SS assembly.

Upon completion, the machinery is able to inject late substrates, the type 3 secretion effector proteins (T3SE), into host cells. The needle tip can sense contact to host cells, and a translocon pore is formed in the host cell membrane. Effector proteins are injected through the pore. In vivo, T3SEs are only secreted upon contact of the needle tip with the host cell membrane, but in vitro, secretion can be induced by chemical signals (Wagner et al., 2018).

A schematic representation of the structure of the T3SS is shown in figure 1.

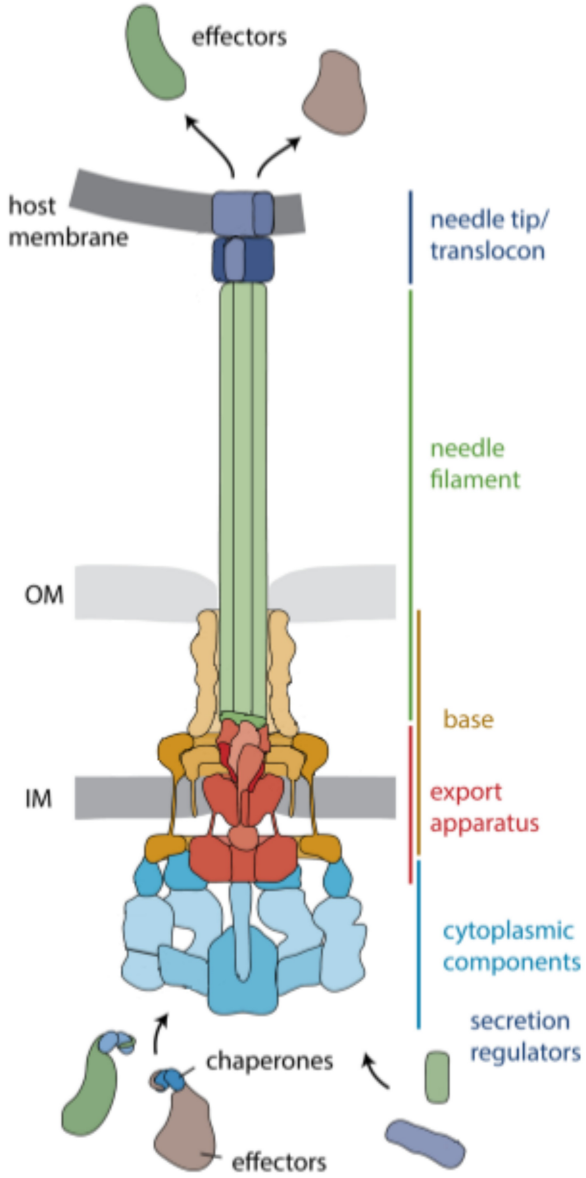

Figure 1: Structure of the T3SS, indicating the different structural units (color-coded) and the individual components. IM stands for inner membrane, and OM stands for outer membrane (figure adapted from Wagner et al., 2018)

T3SS proteins can be encoded throughout the chromosome, in genomic islands or on plasmids. They can be gained by HGT alongside some of the effectors they secrete (Arnold et al., 2009; Arnold et al., 2010).

1.1.4. Type III secretion effectors

T3SEs are modular proteins with an N-terminal secretion signal followed by one or several functional domains (Stavrinides et al., 2006). The secretion signal is not fully understood, as it is not a simple motive. The most important residues for secretion lie within the first 15 to 25 amino acids (aa), but later positions may contribute to the signal. These first positions turned out to be the ones with the most discriminatory power for several machine learning-based T3SE prediction tools (Arnold et al., 2009; Arnold et al., 2010; Samudrala et al., 2009; Wang et al., 2011; Wang et al., 2013). Fusion experiments of reporter genes with effector N-termini showed that the first 10 to 25 aa can be sufficient for secretion (Anderson and Schneewind, 1997; Subtil et al., 2005). The signal is taxonomically universal, as T3SE N-termini of one species can be used in another (Anderson et al., 1999; Subtil et al., 2005). When training the T3SE prediction tool EffectiveT3, exclusion of one taxon also hardly, if at all, diminished its performance (Arnold et al., 2009).

Initially, both an N-terminal amino acid signal or a signal in the underlying mRNA were taken into consideration and supported by evidence. The mRNA signal hypothesis is supported by experiments that showed the signal of some T3SEs to be robust to frameshift mutations, which drastically alter the protein but change the underlying mRNA sequence only a little. Additionally, some silent mutations reduce secretion, even though the amino acid sequence stays unaltered (Anderson and Schneewind, 1997; Arnold et al., 2010; Ramamurthi and Schneewind, 2003; Rüssmann et al., 2002).

However, synonymous replacements at 17 nucleotide positions within the first 10 codons of the yopE T3S signal does not abolish secretion. This is a huge change in the mRNA that does not affect the amino acid sequence. It supports the signal peptide hypothesis in a similar way the frameshift experiments support the mRNA signal hypothesis (Lloyd et al., 2001). Furthermore, frameshift mutations that do not abolish secretion tend to not alter the amino acid composition a lot (Arnold et al., 2009; Wang et al., 2011). The importance of aa composition in the secretion signal is backed up by experiments that enhanced the secretion of YopE by introducing more Serine into the N-terminus or by increasing its amphipathicity (Lloyd et al., 2002). Machine learning-based T3SE prediction tools tend to utilize aa sequences. Higher order features, like secondary structure or water accessibility, seem to have some predictive power as well. Since a lot of the information in the mRNA is no longer contained in the aa sequence or protein secondary structure, the usefulness of these features points towards an aa sequence-based signal (Arnoldi et al., 2009; Wang et al., 2013). Additionally, some T3SEs rely on chaperones for their secretion, which can only act on translated sequences. Translation before translocation has been shown for some T3SEs, and there is evidence that effectors can exist in a folded state in the bacterial cell and need to be unfolded to be secreted.

Some amino acids, like serine, are enriched in the signal sequence. Others, like leucine, are depleted. T3SE N-termini tend to be enriched in polar residues and amphipatic patterns and depleted in acidic and alkaline patterns (Arnold et al., 2010). T3SE prediction results suggest that the secretion signal is robust to random point mutations as long as the aa composition remains intact (Arnold et al., 2010). While the N-termini of many proteins tend to be buried, they are usually exposed to the solvent in T3SEs. T3SE N-termini tend to contain more coils and fewer strands and helixes than N-termini do on average, likely causing

them to be more flexible than the N-termini of most proteins. However, the mechanism of signal sequence recognition remains unclear (Wang et al., 2013).

Many T3SEs posses a 20 to 50 amino acids long chaperone binding site downstream of the N-terminal secretion signal. Chaperones are crucial for the secretion of some effectors, guide them to the T3SS and can prevent the degradation of effectors by cytosolic proteases (Wagner et al., 2018). They can confer secretion pathway specificity and prevent T3SEs from being recognized by the export mechanism used for flagellar self-assembly, as shown for a *Salmonella typhimurium* effector. *Salmonella typhimurium* contains two T3SSs, which are required for different stages of infection and translocate different proteins (Lee and Galán, 2004).

Additionally, many T3SEs contain localization signals to target the effector to specific components inside the host cell. They may contain transmembrane domains for insertion into a host membrane or membrane localization domains to attach them to the membrane as peripheral membrane proteins. In these cases, chaperone binding can prevent the erroneous insertion or attachment of T3SEs to the bacterial inner membrane (Wagner et al., 2018).

Some studies propose an additional mRNA secretion signal in the 5' untranslated region (Karlinsey et al., 2000). C-termini may be required for the secretion of some T3SEs (Dong et al., 2015). However, when trying to identify the residues with the most predictive power for the T3SE prediction tool EffectiveT3, the C-terminus did not allow to classify proteins into effectors and non-effectors.

Some T3SE genes are co-localized with genes encoding their chaperones or with T3SS apparatus genes. This is especially true in organisms that encode their T3SS on plasmids or in genomic islands on the chromosome, like *Salmonella*. However, in *Chlamydia*, which encodes its T3SS throughout the chromosome, most effector genes are not in close proximity to T3SS apparatus genes (Arnold et al., 2009). Many T3SEs are part of the accessory proteome, and the effector repertoire of different strains of a species can vary a lot (Stavrinides et al., 2006). Some effectors can be passed on via HGT together with T3SS apparatus proteins, as they are encoded on the same plasmid. T3SE genes are usually co-regulated with the T3SS or their chaperones. In some species, T3SS-specific promoters and transcription factor binding sites are known. However, these sequences are not identical in all bacteria that posses a T3SS, and in most, the regulatory sequences have not been identified (Arnold et al., 2010).

1.2. Bioinformatics

1.2.1. T3SE prediction tools

Several features have been used to find T3SE candidates. Effector candidates have been identified by searching homologs of confirmed T3SEs. However, this does not allow to find T3SEs that are unrelated to all known effectors. Putative effectors have been predicted based on co-regulation with the T3SS and transcriptional control. However, regulatory elements can differ between taxa and have only been identified in some. They are often difficult to detect with bioinformatics tools and are not necessarily specific to T3SEs but to virulence factors in general. Co-localization with predicted T3SS-related chaperones has been used to search for effector candidates. This approach is species independent, but many effectors are not co-localized with a chaperone or do not even need chaperones for secretion.

The N-terminal signal sequence is the feature most commonly used to predict T3SEs. As the signal is taxonomically universal and required in every T3SE, secretion signal-based methods can be applied to all T3SEs and do not come with the restrictions of other approaches (Arnold et al., 2010). Since the secretion signal is not fully understood, prediction tools are based on machine learning. They tend to use features derived from the primary

amino acid sequence. Some tools additionally utilize higher order features, such as secondary structure and water accessibility, or some of the other properties that are less universally applicable (Arnold et al., 2009; Dong et al., 2015; Li et al., 2020; Samudrala et al., 2009; Wange et al., 2011; Wang et al., 2013; Wang et al., 2013).

EffectiveT3 was among the first universal in silico T3SE prediction program. It uses a Naive Bayesian classifier, a machine learning-based, binary classification algorithm that learns the features of a positive and a negative set of training data. The features with the most discriminatory power, derived from the primary amino acid sequence as well as from two reduced alphabets, were determined and are given to the classifier as input. Experimentally confirmed effectors, discarding those that are too similar to another effector, were used as the positive training set. The negative training set was comprised of proteins randomly selected from the same taxa the confirmed T3SEs were taken from, excluding known effectors. To confirm the validity of the program, training was repeated with 5 different negative training sets, and for each 10-fold cross validation was done, always achieving comparable results. The program was retrained excluding individual taxa from the training data and tested with data from these taxa to show the taxonomic universality of the signal and that EffectiveT3 can be used for taxa not included in the training data. EffectiveT3 achieves a specificity of 85%, a sensitivity of 71% and an Area Under the Receiver Operating Statistics Curve (AUC) value of 0.86. Specificity is the proportion of negative predictions among those instances that should have resulted in a negative prediction. Sensitivity is the proportion of positive predictions among those instances that should have resulted in a positive prediction. The AUC value describes the performance of a classifier, considering the trade-off between sensitivity and specificity by varying the decision threshold. It is 1 for a perfect classifier and 0.5 for a random one. (Arnold et al., 2009).

EffectiveT3 model 2.0.1 works the same way and was trained in a similar manner to the initial version but with more data, slightly improving its performance. It achieved a specificity of 93% and a sensitivity of 73%. EffectiveT3 is available via EffectiveDB, which provides a variety of other tools, among them EffectiveCCBD to predict T3SEs based on chaperone binding sites, EffectiveS346 to predict complete T3SS, T4SS and T6SS in nearly complete genomes and tools to identify T4SEs, eukaryotic-like domains and subcellular localization of secreted proteins (Eichinger et al., 2016).

SIEVE is another of the earliest T3SE prediction tools. It uses a support vector machine to classify proteins based on N-terminal amino acid sequence and composition, GC content in the corresponding gene, protein evolutionary conservation and the phylogenetic distribution of related genes. SIEVE was trained on *Salmonella typhimurium* and *Pseudomonas syringae* proteins. Experimentally confirmed effectors were taken as the positive sample and the rest of the proteomes as the negative samples. To validate the approach, the classifier was trained on one of the organisms and the data from the other organism was used for testing. Additionally, leave-one-out cross validation was performed. SIEVE reached a specificity of 87%, a sensitivity of 90% and an AUC of 0.95 (Samudrala et al., 2009).

BPBAac also uses a support vector machine and makes its predictions based on position-specific amino acid composition profiles. It was trained with experimentally confirmed effectors from different organisms as a positive sample and proteins that are not confirmed effectors and are not homologous to confirmed effectors as a negative training set. If two proteins within a sample were too similar, one of them was removed. 5-fold cross validation was done to assess its performance, and the program was retrained excluding individual genera and tested on the excluded data to show its robustness. It achieved a specificity of 97%, a sensitivity of 91% and an AUC of 0.989 (Wange et al., 2011).

BEAN is a T3SE prediction tool that is based on a linear support vector machine model and takes profile-based amino acid pair information from protein N-termini as input. It was

trained with confirmed T3SEs as the positive training set and proteins sampled from pathogen proteomes as the negative set of training data. If two proteins within a data set were too similar to each other, one of them was excluded. The influence of different negative training samples was tested, and 5-fold cross validation was done to assess its performance. BEAN achieved a specificity of 96%, a sensitivity of 78% and an AUC of 0.97 (Dong et al., 2013).

BEAN 2.0 is an extended version of BEAN that was trained with larger data sets and makes use of other information, in addition to N-terminal signal sequences, to identify T3SEs. It aligns the query sequence to the proteins in the training data and classifies it accordingly if a very similar sequence is found. Next, it compares the domain composition of the query to those in the training data sets and classifies the protein if it contains a functional domain unique to the positive or the negative training set. The remaining proteins are classified with its support vector machine-based approach, using N-terminal signal sequences, the portion of the protein that may contain chaperone binding sites and C-terminal sequences. In addition to T3SE prediction tools, BEAN 2.0 provides functional analysis tools for putative T3SEs (Dong et al., 2015).

T3_MM uses a Markov Model to classify proteins into T3SEs and non-T3SEs based on amino acid composition probabilities conditional on the amino acid in the preceding position. It was trained with sets of T3SEs, excluding close relatives, and with proteins that are not known T3SEs from a variety of species. 5-fold cross validation was done, and the performance of the program was evaluated excluding individual taxa from the training data and testing on them. T3_MM obtained a specificity of 90% and a sensitivity of 84% (Wang et al., 2013).

T3SEpre is a T3SE prediction tool based on a support vector machine. It uses joint features of the amino acid composition, secondary structure and solvent accessibility in the N-termini. Other tools failed to increase their performance by including secondary structure and solvent accessibility variables as independent features. However, these features improve the predictive power of T3SEpre if amino acid composition, secondary structure and solvent accessibility are treated as co-variables dependent on each other. The support vector machine was trained with experimentally confirmed effectors and randomly selected proteins that are not known effectors from different genera. If two proteins within a training set were too similar, one of them was removed. The performance of T3SEpre was assessed with 5-fold cross validation. The classifier was also retrained on random sub-data sets or excluding individual genera and tested on the excluded data. It achieved a specificity of 98%, a sensitivity of 96% and an AUC of 0.995 (Wang et al., 2013).

ACNNT3 is a more recently developed T3SE prediction tool based on an attention convolutional neural network. The neural network takes one-hot encoded protein primary structure information and a position-specific scoring matrix of the first 100 N-terminal amino acids as input. The set of positive training data consisted of experimentally confirmed effectors. The negative training data was composed of non-T3SEs used in previous studies and of type I, II, IV, V, VI, VII and VIII secretion effectors. Proteins shorter than 100 amino acids and proteins that are too similar to other proteins in the same data set were removed. Trained with two different sizes of negative training data sets, ACNNT3 achieved an AUC of 0.95 and 0.98 in the 5-fold cross validation. Testing on two independent data sets, ACNNT3 obtained sensitivities between 91% and 99% and a wide range of specificities (Li et al., 2020). Table 1 provides an overview of the algorithms and features the T3SE prediction tools described above use as well as of their self-reported performances.

Just like ACNN3, some of the other tools were tested on independent data sets, and some T3SE prediction programs were compared to newer tools in the papers introducing these. Overall, the tests confirm the usefulness of the T3SE prediction tools, although some tools did not consistently perform as well as in their initial evaluation. This may be because, for some

tools, the algorithm was chosen and the parameters optimized based on the performance on the same data sets that were used for testing.

EffectiveT3 and EffectiveT3 2.0 had a similar performance to that given in their own papers with some other data sets. However, with others, it performed worse. Particularly, its sensitivity could be rather low with values between 50% and 60%, or rarely even below 50% (Dong et al., 2013; Dong et al., 2015; Li et al., 2020; McDermott et al., 2011; Wange et al., 2011; Wang et al., 2013; Wang et al., 2013). SIEVE had a higher sensitivity and AUC than EffectiveT3 in their respective papers, but it was trained and tested on relatively small positive data sets and only few genera (Arnold et al., 2009; Samudrala et al., 2009). On the data sets that were used to test both tools, they performed similarly (Dong et al., 2013; Wang et al., 2011). BPBAac often was more accurate than EffectiveT3 if tested on the same data set, but it usually did not do as well as in its initial publication. Its specificity was high on almost all test data. However, its sensitivity often was only around 50% or 60% (Dong et al., 2013; Dong et al., 2015; Li et al., 2020; Wang et al., 2013; Wang et al., 2013). As one of the newer tools, BEAN was not used for comparison by the authors of as many other tools. BEAN 2.0 did well for 2 out of 3 test sets, but in one its specificity was extremely low with only 8% (Eichinger et al., 2016; Li et al., 2020). T3_MM also was not used in many comparisons. It did well for the 2 independent data sets used in its own publication and slightly worse but still well for the data set used in the comparison with BEAN 2.0 (Dong et al., 2015; Wang et al., 2013). T3SEpre was not used for comparison by the authors of any of these other tools. Its own publication used some independent test sets. The specificity was high in both samples that included non-effectors, with 95% and 96%. 4 test sets included T3SEs and gave sensitivities of 93%, 91%, 83% and 59% (Wang et al., 2013).

On some data sets, most or all tools performed more poorly than they usually do. In one case, this is due to poorer quality in a data set, used to evaluate T3SEpre, BPBAAc and EffectiveT3, as the effectors were less thoroughly validated. In others, it may be due to chance or they may contain proteins that are harder to predict correctly (Wang et al., 2013).

Several prediction tools have been used to search for T3SEs in whole proteomes, and the proportion of positive predictions varies a lot. EffectiveT3 was applied to over 700 proteomes from gram-negative bacteria, gram-positive bacteria and archeae, classifying 0% to 12% of the proteins as putative effectors. Some proteomes that do not contain a T3SS, including some gram-positive ones, resulted in a high number of positive predictions (Arnold et al., 2009). EffectiveT3 2.0 was used on almost 1700 proteomes (Eichinger et al., 2016). T3SEpre was applied to a few proteomes, including gram-negative bacteria, gram-positive bacteria and yeast. In most gram-positive proteomes, it resulted in only few positive predictions. However, roughly 10% of the yeast proteome was predicted to be secreted. Out of 3 positively predicted yeast proteins that were tested experimentally, 2 could be secreted by *Salmonella* T3SS (Wang et al., 2013). BPBAAc and EffectiveT3 were used on the same *Ralstonia solanacearum* proteome, and EffectiveT3 predicted 9.6% of the chromosomal proteins to be secreted, while BPBAAc only predicted 1.4% to be secreted (Wange et al., 2011).

EffectiveT3 is available as an online tools and as a command line tool (Eichinger et al., 2016). BEAN also exists as an online tool and a downloadable version. However, the online tool restricts the number of input sequences to 50 per job (Dong et al., 2015). The online versions of SIEVE, T3SEpre, T3_MM and BPBAac appear to no longer be usable. However, BPBAac is available as a command line tool. ACNN3 never was an online tool but can be downloaded on github. While EffectiveT3 is not the best performing algorithm, it can most easily be used with a huge amount of data, as it is fast and conveniently accessible.

T3SE Prediction Tools	Algorithm	Features used for prediction	Specificity	Sensitivity
EffectiveT3	Naive Bayesian Classifier	Features derived from primary amino acid sequence and two reduced alphabets	85%	71%
EffectiveT3 2.0.1	Naive Bayesian Classifier	Features derived from primary amino acid sequence and two reduced alphabets	93%	73%
SIEVE	Support Vector Machine	N-terminal amino acid sequence; amino acid composition; gene GC content; protein evolutionary conservation; phylogenetic distribution of related genes	87%	90%
BPBAac	Support Vector Machine	Position-specific amino acid composition profile	97%	91%
BEAN	Support Vector Machine	Profile-based amino acid pair information from protein N-termini	96%	78%
BEAN 2.0	Classification based on alignments to proteins in training data and domain composition; Support Vector Machine	Protein sequence similarity; domain composition; profile-based amino acid pair information from protein N-termini; chaperone binding sites and C-termini	100%	86%
T3_MM	Markov Model	Amino acid composition probabilities conditional on the amino acid in the preceding position	90%	84%
T3SEpre	Support Vector Machine	Amino acid composition, secondary structure and solvent accessibility in the N-termini treated as co-variables dependent on each other	98%	96%
ACNNT3	Attention Convolutional Neural Network	One-hot encoded primary protein structure information and position-specific scoring matrix of the first 100 N-terminal amino acids	14%-97%	91%-99%

Table 1: T3SE prediction tools, algorithms and features they use for prediction and self-reported specificities and sensitivities. The listed specificities and sensitivities of different tools were assessed on different data sets by their respective authors and are not necessarily comparable. Some tools were also evaluated on other data sets, and results can vary (Arnold et al., 2009; Dong et al., 2013; Dong et al., 2015; Eichinger et al., 2016; Li et al., 2020; Samudrala et al., 2009; Wange et al., 2011; Wang et al., 2013).

1.3. Evolution

1.3.1. Virulence and the interaction between T3SEs and the host

The T3SS is a crucial factor in the virulence of a broad range of human pathogenic bacteria. Among others, it is used by enteropathogenic and enterohemorrhagic *E. coli* and some *Shigella*, *Salmonella*, and *Yersinia* species that cause intestinal diseases. It plays a role in enteric fever, induced by *Salmonella serovar Typhi*, whooping cough, caused by *Bordetella* species, and bacteremia and pneumonia, caused by *Burkholderia pseudomallei* . The opportunistic pathogen *Pseudomonas aeruginosa* uses its T3SS to induce pneumonia, urinary tract infections, wound infections, septicemia and endocarditis. *Chlamydia* is an obligatory intracellular bacterium that utilizes a T3SS. *Chlamydia trachomatis* is the leading bacterial cause of sexually transmitted disease. Ocular Chlamydial infections can lead to blindness, and *Chlamydia pneumoniae* infects the lungs, resulting in pneumonia. *Yersinia pestis*, the

causative agent of the plague, also uses a T3SS (Coburn et al., 2007). Additionally, the T3SS is of huge agricultural importance, utilized by pathogens affecting both livestock animals and crop plants (Eichinger et al., 2016). Some mutualistic bacteria, such as *Rhizobia* that form a symbiosis with leguminous plants, use a T3SS to manipulate their hosts as well (Fauvart and Michiels, 2008).

T3SEs interact in a variety of ways with host proteins, other molecules and cellular pathways to enable pathenogenesis. Several pathogens use T3SEs to promote colonization, adherence to or invasion of host cells, often by subverting components of the cytoskeleton, to create a convenient and safe niche for themselves. Enteropathogenic and enterohemorrhagic *E. coli* attach to the gut epithelium by injecting a receptor into the host cell that inserts into the host membrane, binds to the bacterium and leads to actin polymerization, pedestal formation and lesions in the intestinal epithelium (Coburn et al., 2007).

Chlamydia is not metabolically active outside the host cell, but injects T3SEs into cells and triggers its uptake by host actin reorganization and membrane deformation. It is internalized into endocytic vacuoles that form a specialized compartment at the microtubule-organizing center (Nans et al., 2015). The inclusion membrane, enclosing this compartment, is modified by the insertion of T3SEs, termed inclusion membrane proteins (Mital et al., 2013).

Salmonella and *Shigella* promote their uptake into intestinal epithelium cells by influencing the actin cytosceleton to induce membrane ruffles that engulf the bacterium. After their entry, they apply their T3SE repertoire to further manipulate the host cell. *Shigella* uses T3SEs to escape from the phagosome. It is not motile itself but can subvert the actin cytoskeleton to propel itself within and between cells, enabling its dissemination. *Salmonella* prevents phagosome maturation and changes it into a specialized vacuole in which it resides and replicates. It modulates cellular trafficking to gain access to nutrients and membrane material for its vacuole.

T3SEs can cause tissue damage by inducing apoptosis or necrosis of host cells or by disruption of tissue barriers. This enables the bacterium to gain access to nutrients and further its dissemination. In lung infections, *Pseudomnas aeroginosa* uses the T3SE ExoU to destabilize internal membranes, which leads to necrosis. *Yersinia pseudotuberculosis* utilizes its T3SEs to promote actin rearrangement, disrupt tight junctions and rob the epithelium of its function as a barrier to allow the bacterium to invade the epithelium. Some intestinal pathogens disrupt tight junctions in the intestinal epithelial barrier to promote diarrhea and thereby the spread of the bacterium to new hosts. Many pathogens employ additional T3SE-dependent strategies to cause diarrhea, for example *Salmonella enterica serovar Typhimurium* by modulating chloride secretory responses in epithelial cells or *Citrobacter rhodentium* by mislocalization of aquaporin water channels (Coburn et al., 2007).

However, T3SEs can not only be used to induce cell death but also to prevent it. As an obligatory intracellular bacterium, *Chlamydia* inhibits apoptosis of infected cells to avoid destruction of its niche (Mital et al., 2013). The plant pathogen *Pseudomonas syringae* suppresses the hypersensitivity response of its host. The hypersensitivity response is part of the defense of plants against infections and causes rapid localized cell death to kill infected tissue and thereby also the pathogen (Ma and Guttman, 2008).

Bacteria can use their T3SEs to interfere with the host immune response. *Yersinia*, enterophatogenic *E. coli* and *P. aeruginosa* employ T3SEs to prevent phagocytosis by macrophages and neutrophiles. If *Salmonella* is taken up by macrophages, it stops the delivery of enzymes that create reactive oxygen and nitrogen intermediates to the phagosome and renders it harmless to the bacterium. Some pathogens, like *Salmonella*, *Shigella*, *Yersinia* and *P. aeruginosa*, can induce apoptosis in macrophages. *Salmonella*, *Shigella* and *Yersinia* use T3SEs to inhibit or induce inflammatory gene expression by interfering with components in the NF-κB signaling cascade. *Shigella* down-regulates IL-12 in T-cells to steer them towards

the Th2 instead of the Th1 phenotype. Th1 helper cells lead to a cell-mediated immune response and are effective against intracellular pathogens, such as *Shigella*. Th2 helper cells increase humoral immune response, which is better suited for counteracting extracellular pathogens (Coburn et al., 2007).

1.3.2. High selective pressure imposed on T3SEs

The T3SS is of critical importance for pathogenicity, but the direct interaction of T3SEs with host molecules and pathways comes with the risk of exposing the pathogen to the host defense system and imposes strong selective pressure on the effector proteins. As part of the T3SS is exposed to the outside of the bacterium, it can be targeted by antibodies. Some T3SEs, like *Chlamydial* inclusion membrane proteins, can be presented to the immune system via the major histocompatibility complex class I pathway. In therapeutics, structural T3SS proteins and effectors have been considered as targets for passive or active immunization (Coburn et al., 2007).

The evolutionary arms race between pathogen and host, the resulting high selective pressure imposed on T3SEs and the rapid change of bacterial effector repertoires has been studied in plant pathogens, such as *Pseudomonas syringae*. Many T3SEs suppress pathogen-associated molecular patterns-triggered immunity, the basal defense of plants, that recognizes molecular patterns common to many microbes. However, some T3SEs can serve as tells to another component of the plant immune system and prompt effector-triggered immunity. The effectors themselves or the modification of their targets are recognized by resistance (R) proteins. Activation of R proteins can induce defense strategies, such as the strengthening of cell walls, the release of oxidative radicals, the expression of proteins to counteract infection, and hypersensitivity response, which leads to local programmed cell death. Consequently, some effectors can reduce or abolish the virulence of some pathogens in certain hosts, resulting in high host specificity. To infect previously resistant hosts, pathogens can lose a T3SE that activates R proteins if it is dispensable or redundant. The T3SE could also be modified by mutations or replaced by a functionally equivalent protein gained by HGT. Alternatively, a T3SE could be acquired that suppresses the R protein signaling pathway (Ma and Guttman, 2008). Figure 2 illustrates the methods, described above, the plant immune system and pathogen employ to counteract each other.

The variability of the effector repertoire of different bacterial strains is a result of the high selective pressure imposed by the arms race with the host. That many T3SEs are situated on plasmids allows them to be frequently exchanged by HGT, sometimes between distantly related taxa (Ma and Guttman, 2008; Stavrinides et al., 2006). In addition to the risk of recognition by the immune system, the expression of the T3SS and its effectors comes with a huge energetic cost and can double the generation time of *Salmonella enterica*. This is another reason why the T3SS and its effector proteins need to be tightly regulated (Sturm et al., 2011).

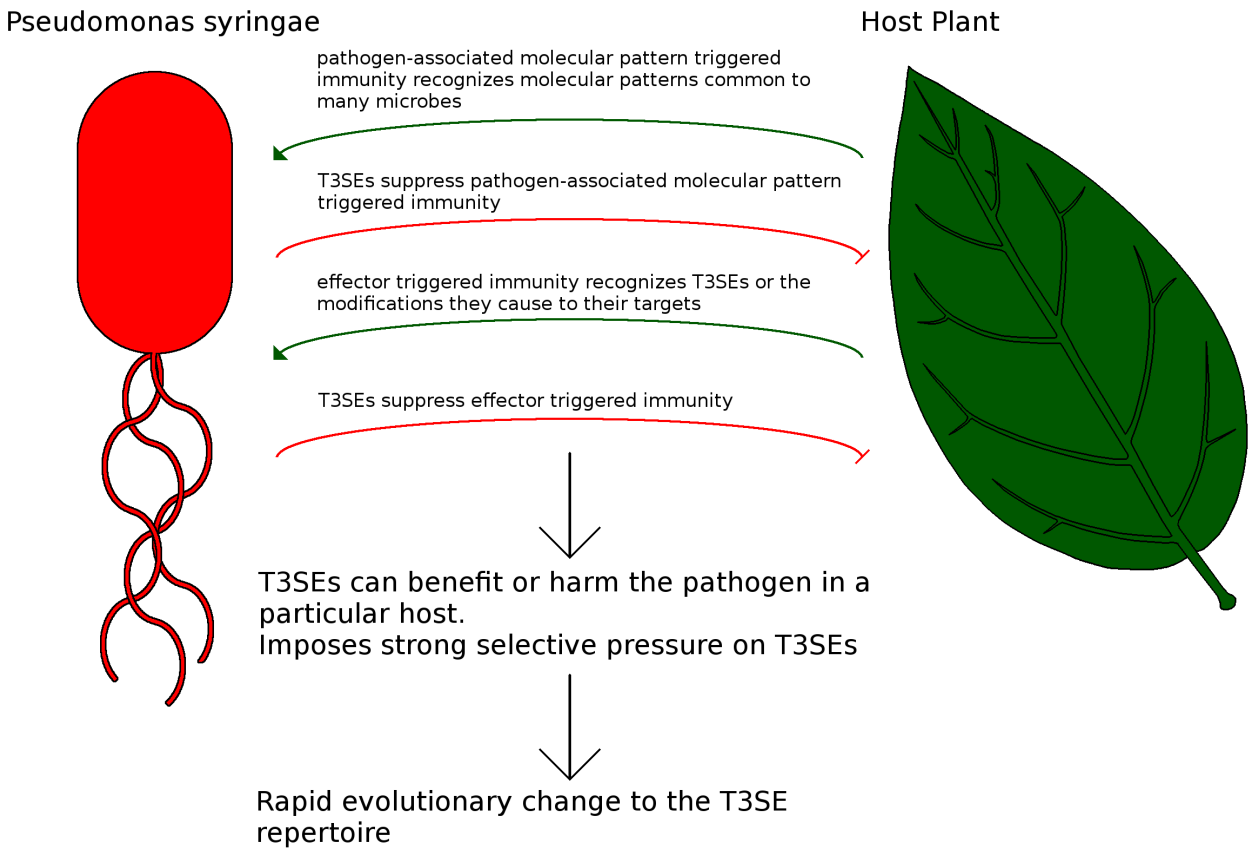

Figure 2: Schematic representation of the interactions between plant pathogens and the plant immune system, imposing strong selective pressure on T3SEs.

1.3.3. Evolution of novel T3SEs

While some bacterial species seem to be able to rapidly gain effectors by HGT, the evolution of completely novel effectors is less thoroughly studied but needs to take place to create the vast repertoire of T3SEs bacteria posses. The modular organization of T3SEs might facilitate their high evolvability, as it allows the signal sequence and the functional domains to change independent of each other.

One proposed mechanism of T3S signal evolution is the N-terminal elongation of proteins by shifting the start codon to an earlier position. Intergenic regions and effector N-termini have a similar amino acid composition, and some may happen to look like secretion signals or only take few point mutations to become T3S signal sequences (Arnold et al., 2010). However, when comparing T3SEs to orthologs in organisms without a T3SS, Arnold et al. did not find a clear pattern of N-terminal elongation. Effectors were often N-terminally elongated or truncated relative to their ortholog and sometimes the same length. Some effectors had longer, shorter and same-length orthologs.

Due to the absence of a clear evolutionary pattern, they suggest that all N-termini of proteins that are accessible to the T3SS by cellular localization and regulation may evolve towards or against a T3S signal. This is consistent with the lack of sequence homology between most known T3SEs (Arnold et al., 2009). The T3SE prediction tools EffectiveT3 and T3SEpre find more putative secretable proteins in some organisms that do not contain a T3SS than would be expected based on their false positive rates (Arnold et al., 2009; Wang et al., 2013). This may be caused by misannotated gene starts, which may be more likely to look like a signal sequence if they include an intergenic region. Alternatively, N-termini of some

organisms may tend to look more similar to effectors due to different amino acid compositions. The false positive rate could be higher in proteomes that look more similar to effectors on average. However, as there is not selection against a T3S signal in the absence of a T3SS, some of these proteins may be secreted if expressed in an organism with a T3SS. Wang et al. tested this for 3 positively predicted yeast proteins, and 2 of them could be secreted by *Salmonella* (Wange et al., 2013).

An earlier study created frameshifts in both alternative reading frames of the N-termini of 2 *Yersinia* T3SEs, YopE and YopN, and showed that 3 out of the 4 frameshifted proteins were secreted (Anderson and Schneewind, 1997). Another study did the same for the *Salmonella enterica* T3SE InvJ and demonstrated that both the +1 and -1 frameshift mutations were secreted (Rüssmann et al., 2002). Arnoldi et al. used EffectiveT3 to predict if frameshifts in both alternative reading frames of 74 experimentally verified and positively predicted effectors and 199 negatively predicted proteins could be secreted. They received a positive result for 10% of the effector frameshifts and 31% of the likely non-effector frameshifts (Arnold et al., 2009). Consequently, frameshifted protein sequences may be relatively likely to look like T3SE N-termini. The signal might not be very specific and may be more rare in protein N-termini than elsewhere throughout the genome.

T3SE N-termini of one effector can be reused for another via a mechanism called terminal reassortment. Terminal reassortment describes any process through which the terminus of one protein or open reading frame is pieced onto another protein or part of a protein, whether this occurs by deletion of DNA between two genes or by recombination. If the N-terminus is taken from a T3SE that only occurs once throughout the genome, the bacterium loses the ancestral effector to create a new one. However, if a protein is encoded on a multi-copy plasmid, or if an effector terminus is associated with mobile elements and occurs several times throughout the chromosome, a new effector can arise without loss of an existing effector. The new T3SE is a chimera of its ancestral proteins or open reading frames. Terminal reassortment allows a new effector to acquire its N-terminal signal sequence and upstream regulatory elements in one evolutionary step. The emergence of new proteins through terminal reassortment is more common among T3SEs than other proteins. Many T3SEs are chimeras of other effectors or of one effector and another protein. In some species, like *Pseudomonas syringae*, *Xanthomonas campestris* and *Salmonella entericia*, effector termini associated with mobile elements have been found (Stavrinides et al., 2006).

Figure 3 provides a graphic representation of possible mechanisms of T3SE evolution, including those that were proposed by previous studies and are supported by evidence as well as some that were not.

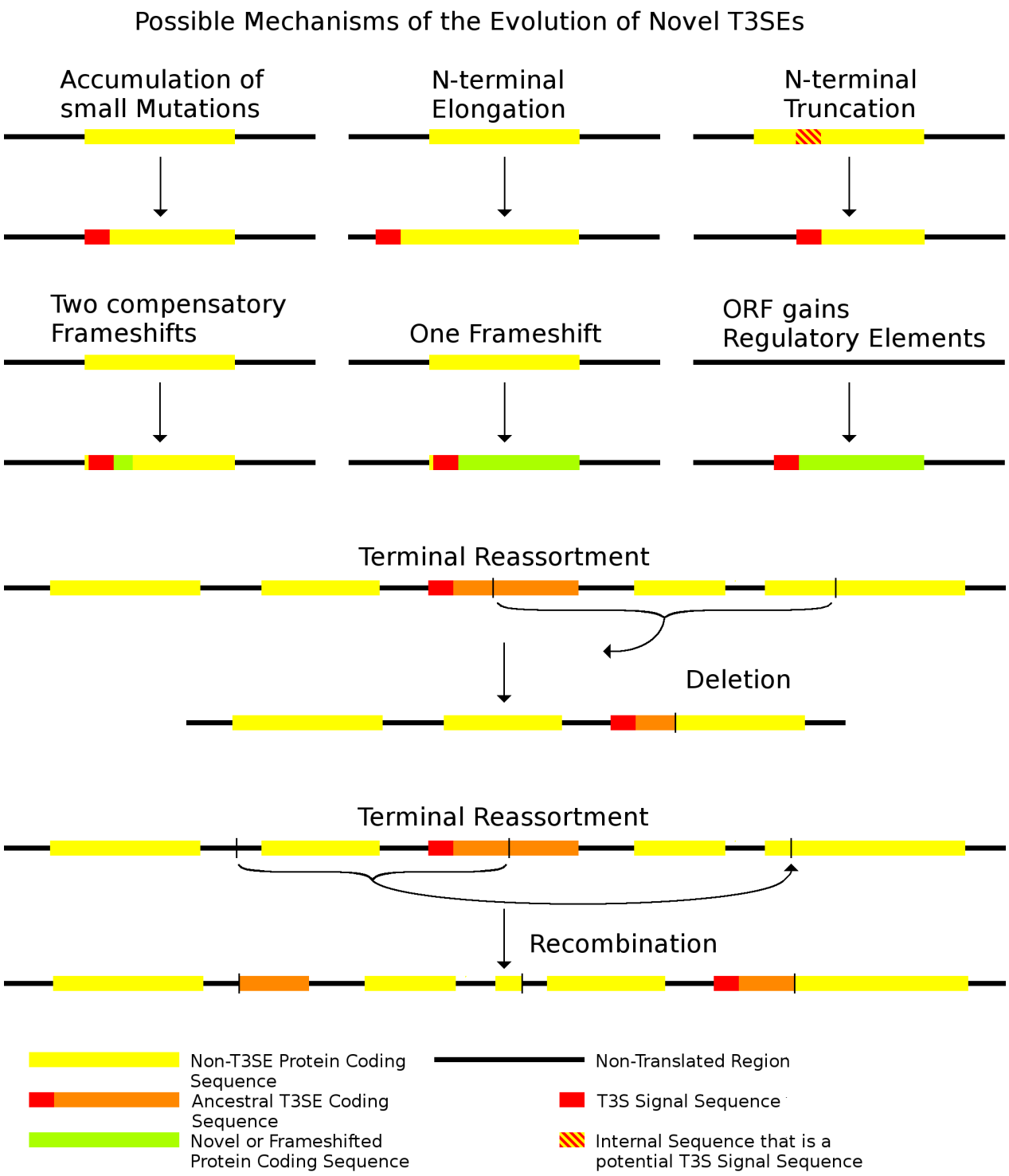

Figure 3: Schematic depiction of possible mechanisms of novel T3SE evolution. Secretion signals could arise by small mutations in a non-effector N-terminus. They could arise via N-terminal elongation, by a mutation that creates a new start codon upstream of the previously used start codon, if the new N-terminus resembles a T3S signal, or via N-terminal truncation, by losing the current start codon and using a start codon that was previously part of an internal sequence, if this internal sequence resembles a T3S signal. Frameshifts near the start of the coding sequence could create novel protein sequences that might contain a secretion signal. However, they would result in a completely novel protein unless a compensatory frameshift restores the original reading frame. Novel proteins, that may or may not be secreted, can also arise by an open reading frame gaining regulatory elements. Terminal reassortment can piece a secretion signal of one protein or open reading frame onto another protein or open reading frame by deleting the DNA sequence between the two pieces of the genes or open reading frames. Alternatively, terminal reassortment can combine the secretion signal of one protein or open reading frame with parts of another protein or open reading frame to a new effector by recombination.

1.3.4. Eukaryotic like virulence factors

Some virulence factors are similar to eukaryotic proteins or contain eukaryotic-like domains. This is a straightforward way to enable them to interfere with eukaryotic pathways and molecules. Eukaryotic-like domains can be a result of convergent evolution. However, phylogenetic analysis suggests that virulence factors can also evolve from proteins acquired from eukaryotes via HGT (Gomez-Valero et al., 2011). They seem to frequently be gained from the host species of the respective pathogen (Duplouy et al., 2013; Gomez-Valero et al., 2011).

Some T4SEs and T2SEs in *Legionella pneumophila* and *Legionella longbeachae* appear to have been transferred from its amoeba hosts to the bacterium. *L. pneumophila* and *L. longbeachae* are mainly environmental bacteria and infect amoeba. However, if inhaled, they can also infect human macrophages in the lungs, resulting in severe, often fatal pneumonia. They are not transmissible from human to human and do not seem to be directly adapted to macrophages. Instead, their ability to survive and multiply within phagosomes and to kill macrophages appears to be a side-effect of their adaptations to parasitizing amoeba. The genomes of some other amoeba symbionts have also been found to be enriched in proteins with eukaryotic domains (Gomez-Valero et al., 2011).

Another example of a bacterium that seems to have gained proteins from its host is *Wolbachia pipientis*. It is an insect symbiont, that can be a reproductive parasite or a mutualist, and is transmitted to offspring of female hosts. Different strains of *Wolbachia* modify the insect's reproductive system in various ways, including feminization of genetic males, cytoplasmic incompatibility and male-killing. Two recent HGT events between eukaryotes and *Wolbachia* have been identified. These genes are several thousand nucleotides long, which is unusual for *Wolbachia*, and fully align to insect genes. Phylogenetic analysis indicates that HGT took place from the insect host to *Wolbachia* rather than the other way around (Duplouy et al., 2013).

However, genes can also be transferred from bacteria to unicellular eukaryotes via endosymbiotic gene transfer (EGT) as well as HGT. Genes encoding metabolic enzymes are overrepresented among eukaryotic genes gained by HGT from prokaryotes and have been the focus of prokaryote to eukaryote HGT research. Events of prokaryotic to eukaryotic HGT of metabolic genes have been found in *Plasmodium*, *Theileria*, *Toxoplasma*, *Cryptosporidium*, *Leishmania*, *Trypanosoma*, *Phytophthora*, diatoms, *Ostreococcus* and *Saccharomyces* (Whitaker et al., 2009).

1.4. Aim of this study

This study aims to further investigate the evolution of the N-terminal T3S signal sequence, using computational methods. It intends to identify frequent modes of signal acquisition, considers the possibility that new effectors may be obtained from eukaryotes via HGT, with or without their signal sequence, and collects evidence on whether certain genomic regions look similar to a signal sequence and may easily evolve into one.

Previous research suggests secretion signal gain by terminal reassortment and N-terminal elongation. However, accumulation of small mutations, N-terminal truncation or two compensatory frameshifts could create a new signal sequence on a preexisting protein as well. T3S signals may also arise as part of completely novel proteins, either by frameshifting of an existing protein or by an open reading frame gaining regulatory elements, or they may evolve from proteins gained from distant taxa.

Effector N-termini are compared to those in effector and presumed non-effector homologs to infer frequently used mechanisms of signal acquisition. This should allow to distinguish

between accumulation of small mutations in the N-terminus, N-terminal elongation and truncation and mechanisms that lead to completely unrelated N-termini as well as to find effectors that are N-terminal homologs of each other as a result of terminal reassortment.

The taxonomic distribution of T3SE homologs is used to identify effectors that could originate from eukaryotic proteins gained via HGT. The portion of the effector and eukaryotic homolog that align to each other may indicate if the secretion signal was obtained alongside the protein or pieced onto it afterwards.

T3SE prediction is used to assess if random sequences, random genomic and intergenic regions, truncated proteins or frameshifted proteins might tend to resemble a T3S signal more closely than protein N-termini do on average.

Some of the analyses conducted are similar to ones done in the past. However, T3S signal evolution has barely been researched since around 2010, and a lot more data is available now. The study suggesting terminal reassortment as an important mechanism in T3SE evolution was done at a time when the signal sequence was more poorly understood. It considered proteins to be N-terminal homologs of effectors if the alignment between them started within the first 30 amino acids (Stavrinides et al., 2006). As it is now known that shorter sequences can be sufficient for the secretion of fusion proteins and that the strongest signal is in the first 15 to 25 amino acids, it might have treated some proteins as N-terminal homologs that do not contain a T3S signal. It is well known that some virulence factors are gained by HGT from eukaryotes (Duplouy et al., 2013; Gomez-Valero et al., 2011). However, this has not been investigated for T3SEs.

2. Results

2.1. N-terminal changes in T3SE evolution

2.1.1. Are T3SEs N-termini elongated, truncated or otherwise different to the N-termini of their non-effector orthologs?

The mechanisms of novel T3SE evolution that have been considered in particular and are supported by evidence are N-terminal elongation and terminal reassortment. Chimeric proteins that arose by terminal reassortment have been found to be overrepresented among T3SE families (Stavrinides et al., 2006). In a study, conducted in 2009, Arnold et al. could not find a clear pattern of N-terminal elongation when comparing T3SEs to their orthologs in organisms without a T3SS. However, the amino acid composition of T3SE N-termini and intergenic regions is similar, suggesting that N-terminal elongation may be a convenient mechanism of signal sequence gain (Arnold et al., 2010).

It was intended to further test these hypotheses. Aligning T3SEs to their ancestral non-effector proteins should allow to infer possible evolutionary pathways based on which alignment partner has the longer N-terminus, if they both have an unaligned N-terminus or if their N-termini align to each other. If the N-terminus of the ancestral protein aligns to the effector but the effector is N-terminally longer, this would indicate secretion signal gain by N-terminal elongation. Related C-termini but unrelated N-termini may hint towards terminal reassortment. Additionally, aligning effectors to their ancestors may indicate signal acquisition via accumulation of small mutations, if their N-termini fully align, or effector evolution by truncation, if the effector is the N-terminally shorter alignment partner.

However, it is not always possible to tell whether a protein is a T3SE or not. Experimentally confirmed effectors can be assumed to be true effectors with a high probability. T3SE prediction tools can help categorize T3SE homologs as effectors or non-effectors, but afflicted with considerable uncertainty.

Therefore, it was intended to align and compare effectors to orthologs in closely related

organisms that do not possess a functional T3SS. They should not be T3SEs, as a protein that functions as an effector would be of no use in an organism that cannot secrete it. Granted, an ortholog in a close relative without a T3SS does not need to be an ancestor. However, T3SEs that become useless or harmful if the T3SS is lost should quickly be lost or gain a new function. Thus, orthologs in close relatives without a T3SS should be a good source of ancestral proteins and related proteins with a different function.

This is similar to the analysis conducted by Arnold et al. in 2009. However, they included orthologs from more distantly related organisms, which may have had a lot of time to evolve independent of the T3SEs and become elongated or truncated themselves. Only including orthologs from close relatives may show a clearer pattern of signal evolution if they are ancestral and some of the T3SEs evolved recently.

It was intended to make alignments between pairs of confirmed effectors and their likely non-effector orthologs. Alignments were supposed to be sorted into categories depending on which termini align to the corresponding terminus of the alignment partner, an internal sequence of the partner or stay unaligned. It was intended to determine how common pairs were in which the N-termini align to each other, in which the effector is N-terminally longer than its ortholog, in which the effector is N-terminally shorter than its ortholog and in which they have different N-termini but the same C-terminus to deduce possible evolutionary mechanisms that may have given rise to the effector protein. The prevalence of pairs that share the same N-terminus but have different C-termini and pairs that differ in C-terminal length should also have been assessed.

E. coli proteins were chosen to conduct this analysis, as many *E. coli* effectors have been experimentally confirmed and only some *E. coli* strains contain a functional T3SS. To find proteomes where searching for orthologs seems useful, homologs of *E. coli* effectors from SecretEPDB were searched in RefSeq *E. coli* proteomes with BlastP, and PhenDB was used to predict if the *E. coli* proteomes contain a functional T3SS.

However, if a proteome was predicted to not have a functional T3SS, hardly any T3SE homologs could be found. There was one confirmed effector that had homologs of the same length and almost 100% identity in almost all proteomes. Since a highly conserved effector is not plausible in an organism that cannot secrete it, this protein is likely not a true effector. BlastP found some low identity hits for another T3SE in some proteomes. The remaining 118 out of 120 *E. coli* effectors did not have any homologs in strains without a T3SS. Therefore, no comparisons to orthologs could be made. The analysis that was supposed to be conducted is shown in figure 4.

Apparently, neofunctionalization of proteins common to a lot of strains is not how *E. coli* gains new effectors. *E. coli* does not seem to have recently evolved any new effectors from non-strain-specific proteins. It might have acquired all its T3SEs horizontally or vertically and may quickly have lost effectors in strains that lost their T3SS.

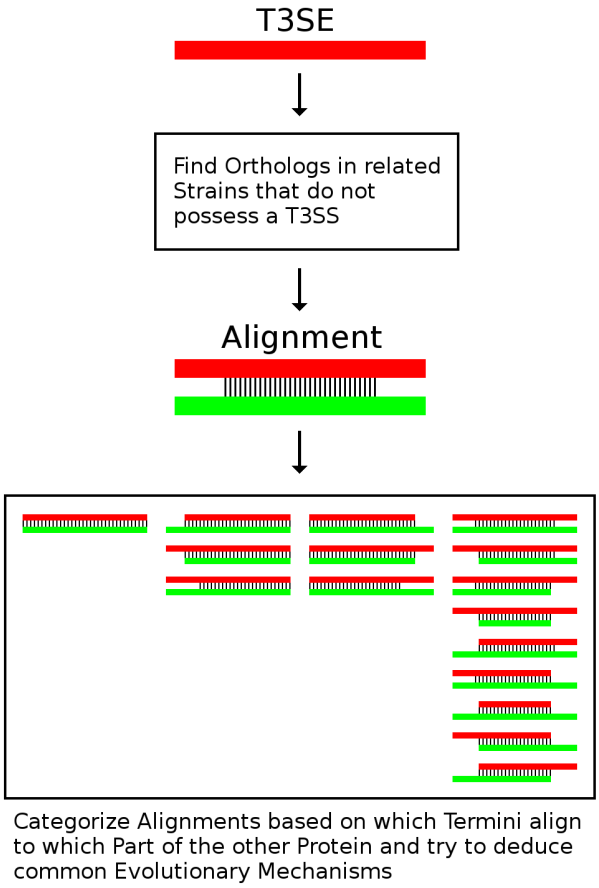

Figure 4: Schematic representation of the analysis that was meant to be carried out. Orthologs of T3SEs in strains of the same species that do not contain a T3SS were supposed to be identified and aligned to the respective T3SE. The alignments were supposed to be categorized based on which termini align to each other, to internal parts of the other protein or stay unaligned to try to infer common mechanisms of T3SE evolution. However, as no T3SE orthologs could be found in strains without a T3SS, the analysis could not be carried out.

2.2. Homologs in other taxa and HGT from eukaryotes

2.2.1. Taxonomic composition of T3SE homologs and searching for evidence of HGT from eukaryotes

The previous results left several options for the origin of *E. coli* T3SEs. They may be completely newly evolved proteins without any homologs. They may be acquired by HGT or passed on vertically from ancestors and quickly be lost in strains without a T3SS. If they have homologs, those do not seem to be proteins that are ubiquitous among bacteria, but they could be specific to pathogens, to bacteria with a T3SS, or they may be gained from very distant taxa, possibly from eukaryotes, as that is how some other virulence factors, for example in *Legionella* and *Wolbachia,* originated (Duplouy et al., 2013; Gomez-Valero et al., 2011).

The taxonomic composition of effector homologs may help to narrow down where the *E. coli* T3SEs come from. If they are recently evolved proteins without any ancestors, they may have few, if any, homologs. Older effectors gained by HGT may have homologs in several bacteria that possess a T3SS, some possibly being more distantly related. If an effector evolved from a eukaryotic protein that was acquired by HGT, it may have eukaryotic homologs but may lack homologs in a broad range of bacterial taxa or may be particularly closely related to its eukaryotic homologs.

Database searches were run to investigate where *E. coli* gains its effectors from. The same was done for *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia*, as they are

also important pathogens with a sufficient amount of experimentally confirmed effectors. The T3SEs from these taxa from SecretEPDB were searched against the eukaryotic and against the gram-positive part of the RefSeq database as well as against the prokaryotic part, excluding the genus of the respective effector, using BlastP.

All *E. coli* effectors had prokaryotic homologs outside their own genus, but only 7 out of 120 had gram-positive homologs. Only 1 *E. coli* T3SE had eukaryotic homologs, and it was the one that likely is not a true effector (Table 2). *E. coli* does not seem to have evolved completely new T3SEs recently enough for them to be confined to the species. It did not gain its confirmed effectors by HGT from eukaryotes.

For the vast majority of *E. coli* effectors, the best hit was either a protein from *Shigella boydii* or from *Citrobacter rhodentium*. The best hit often had a query coverage of 100% and usually an identity over 60%, often over 80%. Most *E. coli* T3SEs had closely related homologs in *Shigella* and *Citrobacter*. *Escherichia*, *Shigella* and *Citrobacter* are *Enterobacteriacea*. For 34% of *E. coli* T3SEs, no homologs were found outside *Shigella* and *Citrobacter*. Some other effectors were confined to relatively small taxonomic groups, like *Enterobacterales*. Others occurred in a broader taxonomic range. While *E. coli* does not seem to have recently evolved completely new T3SEs without ancestral proteins, its common ancestor with *Shigella* and *Citrobacter* may have at some point. *E. coli* may have acquired a substantial amount of its T3SEs vertically from its common ancestors with *Shigella* and *Citrobacter* and may have lost them in strains without a T3SS, or a lot of HGT may have taken place between these 3 genera.

Most, but not all, *Salmonella*, *Yersinia*, *Xanthomonas* and *Pseudomonas* T3SEs had homologs outside their own genus. In *Chlamydia*, only 38% of effectors had homologs outside its own genus (Table 2). Particularly, inclusion membrane proteins did not have homologs outside of *Chlamydia*, and *Chlamydia* possesses several inclusion membrane proteins that do not share sequence similarity with each other. While it is always possible that proteins were acquired horizontally from something not yet sequenced or lost in most organisms, these taxa, in particular *Chlamydia*, may have recently evolved T3SEs that are still confined to one genus or species and are not related to any proteins outside their own genus.

The best hit for *Salmonella* and *Yersinia* effectors was often found within closely related species, although their effector repertoire does not seem to be as similar to any other species or genus as that of *E. coli* is to *Shigella boydii* and *Citrobacter rhodentium*. The best hits for *Pseudomonas* and *Xanthomonas* effectors were usually found in other plant pathogens, and taxonomic relatedness seemed to matter less. Many of them were gained or passed on horizontally, also from and to distant taxa.

Salmonella, *Yersinia*, *Pseudomonas* and *Xanthomonas* had more T3SEs with eukaryotic than with gram-positive homologs. *Chlamydia* had 1 more effector with gram-positive than with eukaryotic homologs. All used taxa, except *E. coli*, had effectors that have related proteins within the eukaryotes but not the gram-positive bacteria (Table 2).

If an effector has homologs within the eukaryotes and the gram-positive bacteria, it may simply contain domains that are ubiquitous to life. However, if effectors were horizontally acquired from eukaryotes, it would be expected that a higher proportion of them may have eukaryotic but no gram-positive homologs than among random proteins.

To gain first evidence whether or not some T3SEs arose from eukaryotic proteins gained by HGT, 100 random proteins were selected from one proteome each of *E. coli*, *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia,* searched against the eukaryotic and the gram-positive part of the RefSeq database, and the proportion of proteins with eukaryotic but no gram-positive homologs was compared to that of the T3SEs. This comparison between effectors and random proteins is illustrated in figure 5 for ease of understanding.

In all taxa, the number of random proteins with gram-positive homologs was higher than the number of random proteins with eukaryotic homologs. Unlike among the T3SEs, most random proteins had homologs in both the eukaryotes and the gram-positive bacteria. Only 5 out of 600 random proteins had eukaryotic but no gram-positive homologs (Table 2). The proportion of effectors with eukaryotic but no gram-positive homologs was 7.2 times higher than among random proteins. Chi-Square test was used to test if the difference is significant and returned a p-value of $1.3 * 10^{-6}$.

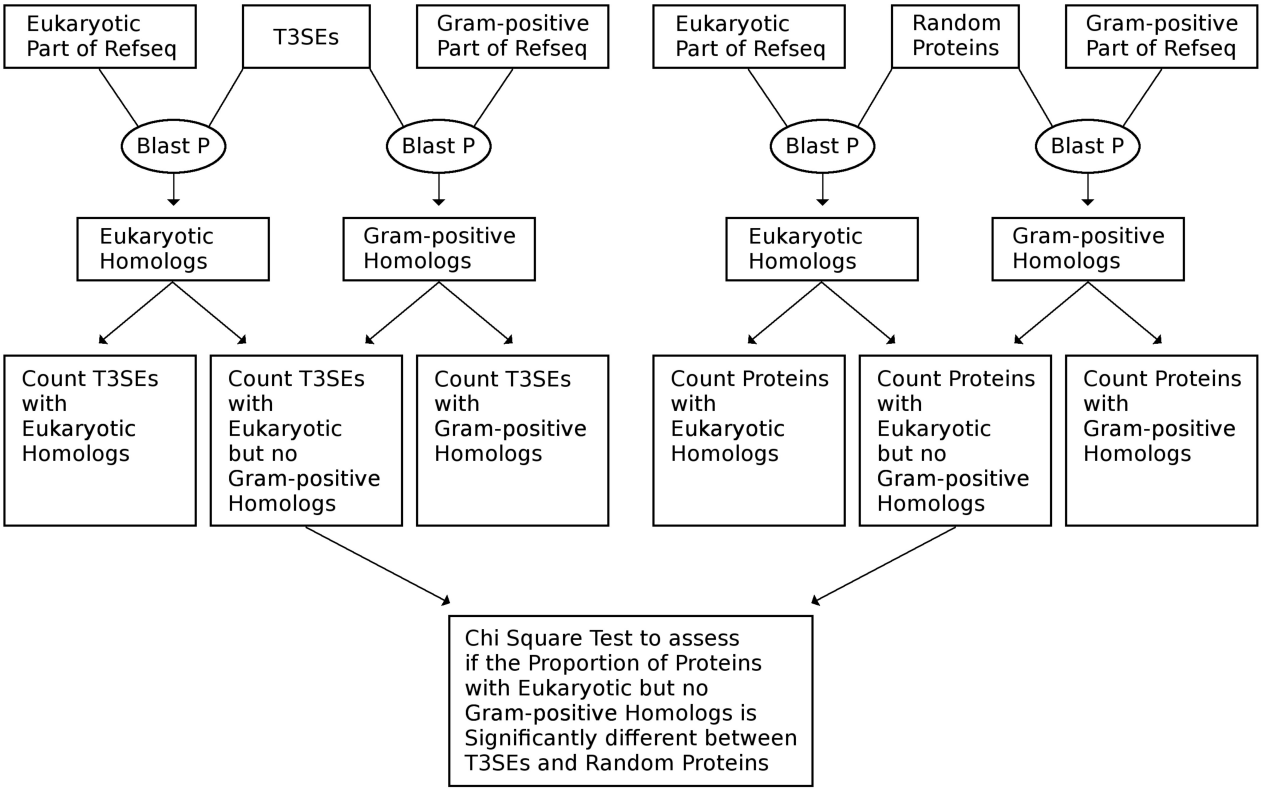

Figure 5: Schematic depiction of the analysis described above. T3SEs and random proteins from the same species were searched against the gram-positive part of RefSeq and the eukaryotic part of RefSeq with BlastP. The numbers of proteins with eukaryotic homologs, the number of proteins with gram-positive homologs and the number of proteins with eukaryotic but no gram-positive homologs were counted among these two groups. Chi-square test was used to assess if the proportion of proteins with eukaryotic but no gram-positive homologs was significantly different between T3SEs and random proteins.

Confirmed Effektors			
Eukayotic Relatives			
	Effektors	Relatives	%
Escherichia	120	1	0.83%
Salmonella	103	32	31.07%
Yersinia	46	14	30.48%
Psaudomonas	303	25	8.25%
Xanthomonas	32	12	37.50%
Chlamydia	74	9	12.16%

Gram+ Relatives			
	Effektors	Relatives	%
Escherichia	120	7	5.83%
Salmonella	103	16	15.53%
Yersinia	46	11	23.91%
Psaudomonas	303	23	7.59%
Xanthomonas	32	3	9.38%
Chlamydia	74	10	13.51%

Eukayotic but no Gram+ Relatives			
	Effektors	Relatives	%
Escherichia	120	0	0.00%
Salmonella	103	17	16.50%
Yersinia	46	4	8.70%
Psaudomonas	303	8	2.64%
Xanthomonas	32	11	34.38%
Chlamydia	74	1	1.35%

Prokaryotic Relatives of other Genera			
	Effektors	Relatives	%
Escherichia	120	120	100.00%
Salmonella	103	95	92.23%
Yersinia	46	44	95.65%
Psaudomonas	303	231	76.24%
Xanthomonas	32	30	93.75%
Chlamydia	74	28	37.84%

Random Proteins			
Eukayotic Relatives			
	Effektors	Relatives	%
Escherichia	100	77	77.00%
Salmonella	100	72	72.00%
Yersinia	100	64	64.00%
Psaudomonas	100	74	74.00%
Xanthomonas	100	59	59.00%
Chlamydia	100	61	61.00%

Gram+ Relatives			
	Effektors	Relatives	%
Escherichia	100	91	91.00%
Salmonella	100	88	88.00%
Yersinia	100	86	86.00%
Psaudomonas	100	85	85.00%
Xanthomonas	100	79	79.00%
Chlamydia	100	67	67.00%

Eukayotic but no Gram+ Relatives			
	Effektors	Relatives	%
Escherichia	100	0	0.00%
Salmonella	100	2	2.00%
Yersinia	100	1	1.00%
Psaudomonas	100	0	0.00%
Xanthomonas	100	0	0.00%
Chlamydia	100	2	2.00%

Table 2: Number of proteins, number of proteins with homologs in the respective taxonomic group and percentage of proteins with homologs in the respective taxonomic group. The upper 4 tables pertain to the confirmed effectors, the lower 3 to the randomly selected proteins. The 1st column lists the taxa. The 2nd column shows the number of confirmed effectors in the respective taxon. The 3rd column gives the number of effectors for which at least 1 homolog was found. The 4th column gives the percentage of effectors for which at least 1 homolog was found. The upper left tables show the results of the search against the eukaryotic part of the RefSeq database. The upper right tables show the results of the search against the gram-positive part of the database. The lower left tables show the numbers of proteins that have eukaryotic but not gram-positive homologs. The lower right table shows the results for the search against the prokaryotic part of the RefSeq database, excluding the genus the effector is from. This last search was only done for confirmed effectors and not for random proteins.

Confirmed T3SEs are no random sample, and many of them are related to each other. To ascertain that the found tendency is not due to very few proteins with many homologs within the confirmed effectors, 90% and 50% identity clustering were performed with CD-HIT, and the analysis was repeated, including only one representative sequence for each cluster. After identity clustering, the proportions of effectors with eukaryotic but no gram-positive homologs were only 5.6 and 4.2 times higher than among random proteins, but the results stayed significant on a 5% level, with p-values of $3*10^{-4}$ and $1*10^{-2}$, respectively (Table 3). This gives some evidence that HGT from eukaryotes may contribute to the evolution of T3SEs.

Eukaryotic but no Gram+ Relatives							
90% Identity Clustering				**50% Identity Clustering**			
	Effektors	Relatives	%		Effektors	Relatives	%
Escherichia	56	0	00.00%	Escherichia	29	0	0.00%
Salmonella	56	7	12.50%	Salmonella	46	2	4.35%
Yersinia	23	2	8.70%	Yersinia	19	2	10.53%
Psaudomonas	146	4	2.74%	Psaudomonas	78	2	2.56%
Xanthomonas	23	3	13.04%	Xanthomonas	19	2	10.53%
Chlamydia	66	1	1.52%	Chlamydia	64	1	1.56%

Table 3: The tables show the number of effectors, number of effectors with eukaryotic but no gram-positive homologs and percentage of effectors with eukaryotic but no gram-positive homologs. The 1st column lists the taxa. The 2nd column shows the number of confirmed T3SEs in the respective taxon. The 3rd column gives the number of effectors for which at least 1 homolog was found. The 4th column gives the percentage of effectors for which at least 1 homolog was found. The left table shows the results after 90% identity clustering and the right table after 50% identity clustering.

The same analysis as for the *E. coli*, *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia* T3SEs was repeated with *Legionella* T2SEs and T4SEs. This was done to see if there is a stark difference between *Legionella* and the other taxa, as *Legionella* is known to have eukaryotic-like effectors.

There were 310 *Legionella* effectors in SecretEPDB. 47 had eukaryotic homologs, 32 gram-positive homologs and 21 eukaryotic but no gram-positive homologs. 6.8% of *Legionella* effectors had eukaryotic but no gram-positive homologs. This number is higher than for some and lower than for some of the taxa containing a T3SS.

2.2.2. Identifying instances of T3SEs that originated from eukaryotic proteins gained by HGT

The above analysis provides evidence that some T3SEs might have originated from ancestors gained by HGT from eukaryotes. However, it is indirect evidence only, by showing a heightened prevalence of proteins with eukaryotic and no gram-positive homologs among effectors, and is not suitable to identify individual instances of T3SEs that evolved from eukaryotic ancestors. Therefore, it was intended to identify effectors that seem to have been acquired by horizontal transfer from eukaryotes or to rule out that they were to either confirm or reject the previous result.

The taxonomic distribution and number of homologs of the T3SEs that are related to eukaryotic proteins were inspected to find likely true events of HGT among the candidates or to exclude that individual effectors were acquired from eukaryotes. T3SEs that are confined to

a small taxonomic range among the bacteria but have eukaryotic homologs as well as effectors that are particularly closely related to their eukaryotic homologs were looked for. If an effector has homologs in a wide taxonomic range and the eukaryotic ones are not particularly closely related, it may simply contain common domains. If it has homologs in many bacteria, but its eukaryotic homologs are confined to a small taxon, HGT from bacteria to eukaryotes seems more likely than HGT from eukaryotes to bacteria.

The most interesting candidate that could be found is the *Chlamydial* protein WP_010883128.1. It is not related to bacterial proteins outside of *Chlamydia* but has eukaryotic homologs within the *Trypanosomatidae*, including *Trypanosoma* and *Leishmania*. The eukaryotic homologs themselves do not have eukaryotic homologs outside of *Trypanosomatidae* either. WP_010883128.1 is confined to a relatively small taxon within the bacteria and a relatively small taxon within the eukaryotes. Both in *Chlamydia* and in *Trypanosomatidae*, the homologs of this protein are roughly 700 amino acids long. Alignments of WP_010883128.1 with its prokaryotic homologs give 98% to 99% query coverage and 41% to 44% sequence identity. Alignments with its eukaryotic homologs have 88% to 99% query coverage and 33% to 37% identity. This results in e-values below 10^{-100} for all pairwise alignments between WP_010883128.1 and its homologs. The best eukaryotic hit, XP_827657.1, aligns with higher identity to all related eukaryotic proteins than to the prokaryotic ones. While it is highly likely that this protein was subject to HGT between eukaryotes and prokaryotes, the data does not allow to infer the direction of the transfer.

In this case, a virulence factor seems to have been gained from or passed on to another pathogen. *Chlamydia*, *Trypanosoma* and *Leishmania* are animal pathogens. *Chlamydia* is an intracellular pathogen, and some *Trypanosomatidae* spend part of their life cycle intracellularly. This might have allowed members of *Chlamydia* and *Trypanosomatidae* to meet in a confined space by infecting the same cell and may have facilitated one organism getting into contact with the DNA of the other. It is unclear if the homologs in *Trypanosomatidae* are secreted virulence factors as well, as their functions are unknown.

To test how common HGT between *Chlamydia* and *Trypanosomatidae* is, the whole proteome of one *Chlamydia pneumoniae* strain containing WP_010883128.1 was searched against the eukaryotic RefSeq database. Only one protein, WP_010883030.1, could be found with very similar relationship patterns to WP_010883128.1. It, too, only has homologs among *Chlamydia* and *Trypanosomatidae*, as do its eukaryotic homologs. However, in its case, the eukaryotic variants are a lot longer than the prokaryotic ones. The short description of WP_010883128.1 identifies it and its prokaryotic homologs as type III secretion chaperones.

There were some other C*hlamydial* proteins with homologs in the *Trypanosomatidae* but also in other taxa. However, *Trypanosomatidae* proteins were not the best hits, and there was no evidence that *Chlamydia* may have acquired them horizontally from this taxon.

Some confirmed effectors in other taxa may have been subject to HGT, but they are not as definite cases as the C*hlamydial* protein WP_010883128.1. Two T3SEs, unrelated to each other, were found in *Xanthomonas* that have several eukaryotic but no gram-positive homologs. They have multiple homologs among the confirmed *Xanthomonas* effectors, and one of them is related to confirmed effectors in *Pseudomonas*. Related prokaryotic proteins are almost exclusively found in plant pathogens. They may or may not all be T3SEs. One of them has eukaryotic homologs only among animals. The other has homologs among algae, moss and fungi, but they are not ubiquitous among these taxonomic groups.

Some other confirmed effectors have eukaryotic and only very few gram-positive homologs. For some other candidates, horizontal transfer could be excluded with almost certainty. They may not have true eukaryotic homologs, but the sequenced eukaryotic specimen may have been infected with some bacterium, and some genes may have wrongfully been assigned to the species. This seems to have happened for a few genes in the RefSeq

database in the fly *Lucilia cuprina*. Some of its genes are identical to prokaryotic entries but do not have insect or other eukaryotic homologs.

Effectors that do not only have eukaryotic but also many as close or more closely related gram-positive homologs probably contain conserved domains rather than having been acquired from eukaryotes. However, not all T3SEs have high query coverage with their best gram-positive or eukaryotic hits, and gram-positive homologs only serve as evidence against HGT from eukaryotes if they align to the same part of the T3SE.

To avoid missing any instances of HGT from eukaryotes, it was tested whether the region in the effector that aligns to the best eukaryotic hit and the region that aligns to the best gram-positive hit, excluding proteins that contain 'partial' in their short description, overlap for T3SEs with eukaryotic and gram-positive homologs. By how many amino acids the start and the end positions of the alignment with the eukaryotic and the prokaryotic protein differ in the effector was assessed as well.

The stretch of the effector that aligns to the best eukaryotic and the stretch that aligns to the best prokaryotic hit overlapped for all tested effectors. The start position of the alignment with the closest eukaryotic and the closest gram-positive homolog differed by 11 amino acids on average, the stop position by 25 amino acids. Often, they were almost identical, but there were a few proteins for which the alignment with one homolog was more than 100 amino acids longer than with the other.

If there are eukaryotic and gram-positive homologs, T3SEs seem to share the same domains with the eukaryotic and the gram-positive protein. Many of these effectors may contain domains that are ubiquitous to life. They do not happen to have a part in common with their gram-positive homolog and a different part with their eukaryotic homolog.

2.2.3. Do T3S signals arise from sequences shared with eukaryotic or gram-positive homologs?

While the previous results could not ascertain that any T3SEs evolved from eukaryotic proteins but indicate that it is probably rare, several effectors have homologs among distant taxa. It was intended to investigate if these proteins, whether they are conserved proteins or proteins gained by HGT, contribute to the evolution of T3S signal sequences or if signal sequences need to be added onto them when they become effectors.

The main proposed mechanisms of T3S signal evolution, terminal reassortment and N-terminal elongation, add a new N-terminus and do not require the N-terminus of the ancestral protein to be or become a secretion signal. However, if N-termini tend to evolve towards a secretion signal by accumulating point mutations and other small mutations, effectors may share an N-terminus with a homologous conserved protein or an ancestor that was gained horizontally from a distant taxon. Likewise, T3SE N-termini may correspond to an internal sequence if the effector was truncated or the homolog elongated. As an N-terminus that looks like a T3S signal in an organism without a T3SS would be inconsequential, proteins that were acquired from eukaryotes or gram-positive bacteria could be secretable by chance.

It was intended to test if T3SE N-termini are more, equally or less likely than T3SE C-termini to align to their eukaryotic homologs to gather evidence if the secretion signal evolves alongside the rest of the protein or needs to be added onto it. The same was done for their alignment partners to see if any observable tendency is specific to the effectors or holds true for both proteins. To get a first clue, the average length of the unaligned N-terminal and C-terminal ends of the T3SEs and their best hits among eukaryotic proteins, excluding those identified as partial proteins, were determined. As the strongest signal is in the first 15 to 25 aa, the number of effectors and best eukaryotic hits for which the alignment starts within the

first 15 aa were assessed and compared to the number of effectors and best eukaryotic hits for which the alignment ends within the last 15 aa.

On average, 171 aa of the effector N-terminus and 60 aa of the effector C-terminus remained unaligned. In the eukaryotic homologs, 185 N-terminal aa and 198 C-terminal aa were unaligned on average.

In 12 out of 92 T3SEs, the alignment started within the first 15 aa, and in 58 it ended within the last 15 aa. There were 4.8 times more alignments that ended within the last 15 aa than alignments that started within the first 15 aa. In the eukaryotic homologs, 34 alignments started within the first 15 aa and 36 ended within the last 15 aa. The proportion of alignments that ended within the last 15 aa to alignments that started within the first 15 aa was 1.1 (Table 4).

Chi-square test was used to assess if the different number of alignments that cover some of the first 15 N-terminal aa and alignments that cover some of the last 15 C-terminal aa is significant on the 5% level for effectors and for their eukaryotic homologs. The p-value for the effectors was $8.3*10^{-12}$, and for the eukaryotic homologs it was 0.88. While there is no evidence for a preference towards aligned C-termini or N-termini in the eukaryotic homologs, N-termini seem to be more likely to be unaligned in T3SEs. A graphic representation of how the analysis was carried out can be found in figure 6.

Since there are groups of related proteins among the effectors and, consequently, also among the best eukaryotic hits, the analysis was repeated after 90% and 50% identity clustering to exclude that the results are caused by one or a few big groups of related proteins of the same length.

After 90% identity clustering, the alignment started within the first 15 aa for 9 out of 48 effectors and ended within the last 15 aa for 26 effectors. The proportion of alignments that ended within the last 15 aa was 2.9 times higher than the proportion of alignments that started within the first 15 aa. The alignment started in 12 eukaryotic homologs within the first 15 aa and ended in 17 within the last 15 aa (Table 4). Chi-Square test gave a p-value of $5.9*10^{-4}$ for the effectors and a p-value of 0.37 for the related eukaryotic proteins.

After 50% identity clustering, the alignment started within the first 15 aa for 8 out of 33 T3SEs and ended within the last 15 aa for 18. The proportion of alignments that ended within the last 15 aa was 2.3 times higher than the proportion of alignments that started within the first 15 aa. The alignment started within the first 15 aa for 7 eukaryotic homologs and ended within the last 15 aa for 14 (Table 4). Chi-Square test gave p-values of 0.023 for the effectors and 0.11 for the eukaryotic homologs.

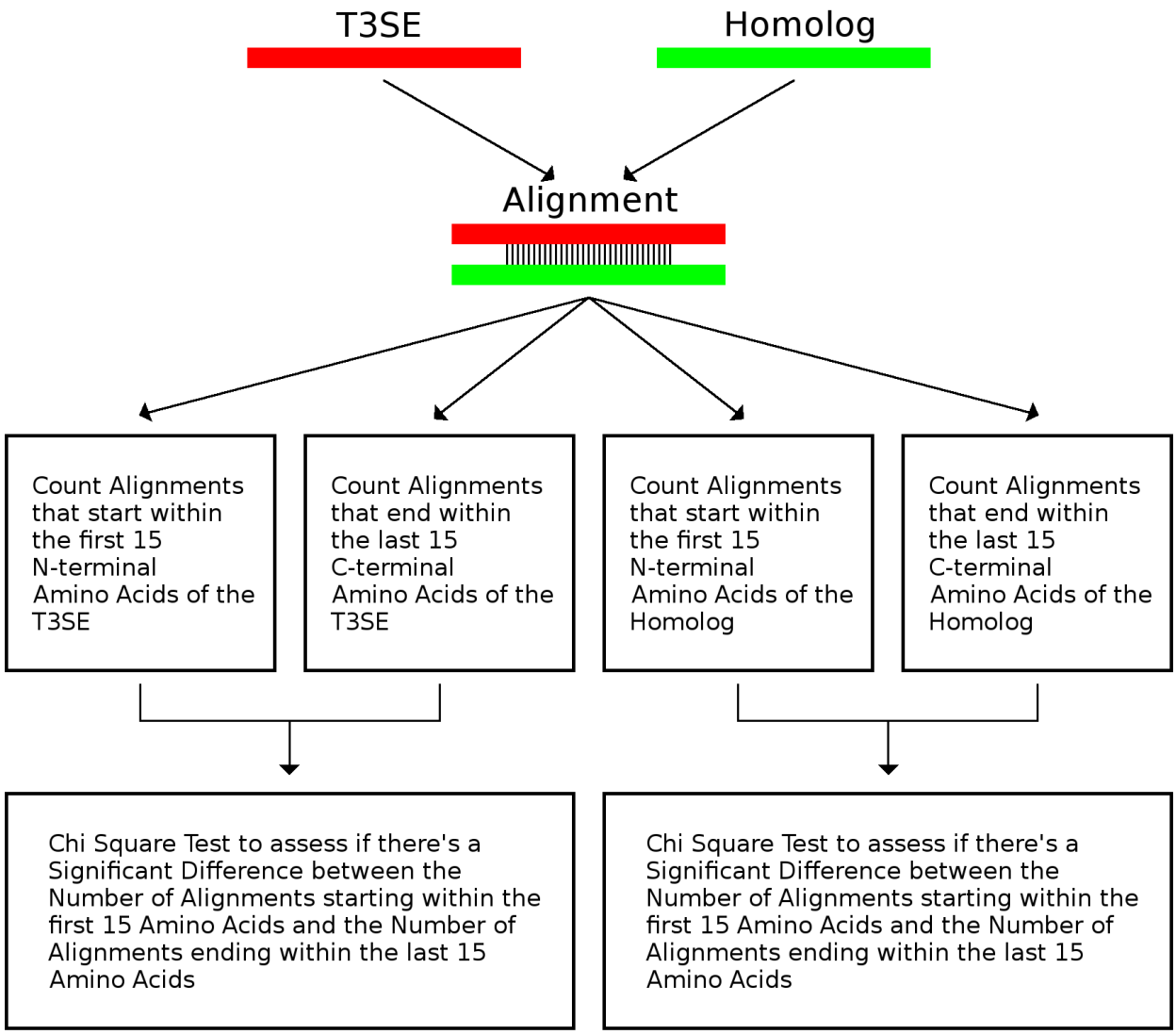

Figure 6: Schematic representation of the workflow of the analyses described in this section. T3SEs were aligned to their closest homologs that are not identified as partial proteins. The numbers of alignments that start within the first 15 N-terminal amino acids of the T3SE, that start within the first 15 N-terminal amino acids of the homolog, that end within the last 15 C-terminal amino acids of the T3SE and that end within the last 15 C-terminal amino acids of the homolog were counted. Chi-Square test was used to assess if the number of alignments starting and ending within the first and last 15 amino acids was significantly different for the T3SEs as well as the homologs.

	Eukaryotic Relatives Effector								
	all			90% Identity Clustering			50% Identity Clustering		
	Total	N-ter	C-ter	Total	N-ter	C-ter	Total	N-ter	C-ter
Escherichia	0	0	0	0	0	0	0	0	0
Salmonella	32	3	17	13	1	5	9	0	2
Yersinia	14	1	5	8	1	3	4	1	3
Pseudomonas	25	1	22	13	1	11	9	1	7
Xanthomonas	12	0	9	4	0	2	3	0	1
Chlamydia	9	7	5	8	6	5	8	6	5
Sum	92	12	58	46	9	26	33	8	18

	Eukaryotic Relatives Subject								
	all			90% Identity Clustering			50% Identity Clustering		
	Total	N-ter	C-ter	Total	N-ter	C-ter	Total	N-ter	C-ter
Escherichia	0	0	0	0	0	0	0	0	0
Salmonella	32	12	24	13	2	7	9	0	4
Yersinia	14	0	4	8	0	2	4	0	2
Pseudomonas	25	17	4	13	6	4	9	3	4
Xanthomonas	12	0	0	4	0	0	3	0	0
Chlamydia	9	5	4	8	4	4	8	4	4
Sum	92	34	36	46	12	17	33	7	14

Table 4: Number of alignments between T3SE and best eukaryotic hit that is not identified as a partial protein, number of effectors or alignment partners for which the alignment starts within the first 15 aa and number of effectors or alignment partners for which the alignment ends within the last 15 aa. The upper table shows the results for the T3SEs and the lower table for their alignment partners. The columns 'Total' give the number of alignments between T3SEs and best eukaryotic hits that do not contain 'partial' in their names. The columns 'N-ter' give the number of effectors or alignment partners for which the alignment starts within the first 15 aa. The columns 'C-ter' give the number of effectors or alignment partners for which the alignment ends within the last 15 aa. The numbers are given for all T3SEs and best hits that are not identified as partial proteins and after 90% identity clustering of the effectors and after 50% identity clustering of the effectors. The *E. coli* protein that was previously identified to be likely no true effector was excluded.

The tendency towards fewer T3SE N-termini than C-termini that align to the best eukaryotic hit persisted after identity clustering, but it was weakened. Identity clustering changed taxonomic composition, as the number of proteins was reduced a lot for all taxa but *Chlamydia*. *Chlamydia* was the only used taxon that did not show a tendency towards more effectors with aligned C-termini than N-termini. However, this may be by chance, as there is no known reason why different taxa should differ in this regard.

In all used taxa, except *Chlamydia*, alignments either started at the 1st aa, at the 2nd aa of the effector or at a later position than the 15th. The *Salmonella* effectors with aligned N-termini were all related to the same and only one eukaryotic gene. This gene was assigned to *Lucilia cuprina* and did not have any homologs among other eukaryotes. Similarly, the T3SE with aligned N-terminus in *Yersinia* was related to only 2 eukaryotic genes, and one of them belonged to *L. cuprina*. The *L. cuprina* gene was only related to 2 eukaryotic genes, had many prokaryotic homologs and was 100% identical to a *Pseudomonas* gene. The *Pseudomonas* effector with aligned N-terminus was only related to genes of 2 insects, one of them being *L. cuprina*. Its *L. cuprina* homolog was only related to 3 other eukaryotic proteins, had many bacterial homologs and was very similar to some of them. It is doubtful whether these

eukaryotic homologs of T3SEs are truly eukaryotic proteins or proteins assigned to the wrong taxon. Even if they were eukaryotic, they would appear to have been transferred from prokaryotes to eukaryotes instead of from eukaryotes to prokaryotes. N-termini were not aligned to the closest eukaryotic hit for any *Xanthomonas* effector. In *Chlamydia*, at least some effectors whose N-termini aligned to their best eukaryotic hit had enough eukaryotic homologs to assume that they are true homologs. Unlike in the other taxa, the alignments could begin at the start, in the middle or at the end of the first 15 aa.

In conclusion, N-terminal signal sequences usually do not seem to correspond to any sequence shared with eukaryotic proteins. The only used taxon where any examples of it could be found was *Chlamydia*, which is also the taxon for which the most convincing case of HGT was identified.

The same analysis was repeated with alignments between T3SEs and their best gram-positive hits that do not contain 'partial' in their short description to see if the results hold up for gram-positive homologs or only for eukaryotic ones.

Without identity clustering, the alignments started within the first 15 aa in 16 out of 64 effectors and ended within the last 15 aa for 31. They started for 25 gram-positive homologs within the first 15 aa and ended for 14 within the last. After 90% identity clustering, alignments started for 14 out of 42 effectors within the first 15 aa and ended for 21 within the last 15. They started for 15 gram-positive homologs within the first 15 aa and ended for 13 within the last. After 50% identity clustering, alignments started for 14 out of 34 effectors within the first 15 aa and ended for 18 within the last 15. They started for 11 gram-positive homologs within the first 15 aa and ended for 12 within the last (Table 5). Chi-Square test only returned a significant p-value, of 0.010, for effectors without identity clustering. All other differences were nonsignificant.

	Gram-positive Relatives Effector								
	all			90% Identity Clustering			50% Identity Clustering		
	Total	N-ter	C-ter	Total	N-ter	C-ter	Total	N-ter	C-ter
Escherichia	1	1	1	1	1	1	1	1	1
Salmonella	16	0	4	9	0	3	8	0	3
Yersinia	11	1	2	7	1	2	3	1	2
Pseudomonas	23	6	21	14	5	12	11	5	9
Xanthomonas	3	0	0	2	0	0	2	0	0
Chlamydia	10	8	3	9	7	3	9	7	3
Sum	64	16	31	42	14	21	34	14	18

	Gram-positive Relatives Subject								
	all			90% Identity Clustering			50% Identity Clustering		
	Total	N-ter	C-ter	Total	N-ter	C-ter	Total	N-ter	C-ter
Escherichia	1	0	1	1	0	1	1	0	1
Salmonella	16	2	1	9	2	1	8	2	1
Yersinia	11	1	1	7	1	1	3	1	1
Pseudomonas	23	16	8	14	7	7	11	3	6
Xanthomonas	3	0	0	2	0	0	2	0	0
Chlamydia	10	6	3	9	5	3	9	5	3
Sum	64	25	14	42	15	13	34	11	12

Table 5: Number of alignments between T3SE and best gram-positive hit that is not identified as a partial protein, number of effectors or alignment partners for which the alignment starts within the first 15 aa and number of effectors or alignment partners for which the alignment ends within the last 15 aa. The upper table shows the results for the T3SEs and the lower table for their alignment partners. The columns 'Total' give the number of alignments between T3SEs and best gram-positive hits that do not contain 'partial' in their names. The columns 'N-ter' give the number of effectors or alignment partners for which the alignment starts within the first 15 aa. The columns 'C-ter' give the number of effectors or alignment partners for which the alignment ends within the last 15 aa. The numbers are given for all T3SEs and best hits that are not identified as partial proteins and after 90% identity clustering of the effectors and after 50% identity clustering of the effectors. The *E. coli* protein that was previously identified to be likely no true effector was excluded.

E. coli only had one effector with gram-positive homologs that did not contain 'partial' in their descriptions, and in this protein, both termini were aligned. *Xanthomonas* did not have any T3SEs with aligned termini. *Salmonella* had effectors with aligned C-termini but none with aligned N-termini, and in *Yersinia* and *Pseudomonas,* more effectors had aligned C-termini than N-termini. In *Chlamydia*, this tendency was reversed. As before, identity clustering reduced the number of proteins a lot in all taxa but *Chlamydia* and *E. coli*.

Unlike in the case of the eukaryotic homologs, the gram-positive homologs that align to T3SE N-termini tend to have many gram-positive homologs and do not seem to be misassigned to their taxon. Effector termini can share similarity with proteins from gram-positive bacteria. This may be due to common bacterial sequences occasionally mutating into T3S signals or due to HGT in either direction. Either way, it only applies to a small minority of T3SEs.

2.3. Comparing T3SEs to homologs to infer method of signal sequence acquisition

2.3.1. Inferring evolutionary mechanisms based on how T3SEs differ from homologs in the same proteome

In the initial attempt to deduce potential mechanisms of novel T3SE evolution based on if effector termini align to the respective terminus of their possibly ancestral orthologs, to an internal sequence or stay unaligned, no effector orthologs could be found in *E. coli* strains without a T3SS. Therefore, a different source of possible ancestors was utilized to conduct a similar analysis with a similar aim. Homologs co-occurring in the same proteome as the T3SE were chosen to be compared to the effectors because evolution takes place within the context of the proteome, in spite of in some prokaryotes rampant HGT, and because the context of the proteome is needed to identify potential events of terminal reassortment without loss of the ancestral proteins.

Unlike in the earlier analysis, the homologs were collected from proteomes that possess a T3SS, and therefore, they may or may not be effectors themselves. Unless they happen to be confirmed effectors, it cannot be determined with high certainty if they are. Nonetheless, the T3SE prediction tool EffectiveT3 was used to categorize the remaining homologs, as different patterns of how effectors relate to their homologs are expected depending on if the homologs are effectors themselves or not. Homologs also may or may not be ancestral. As no strains without a functional T3SS were required for this analysis, it was not restricted to *E. coli* proteins.

Like in the intended earlier analysis, effector-non-effector pairs that have aligned N-termini may indicate secretion signal evolution by accumulation of small mutations. Effector-non-effector pairs that differ in N-terminal length would hint at signal gain by N-terminal elongation or truncation. Different N-termini but shared C-termini may suggest terminal reassortment.

While the lack of effector homologs in *E. coli* strains without a T3SS showed that *E. coli* effectors do not tend to originate from ancestral proteins common to its species, this could be different for other bacteria. A bacterium could also gain a protein that evolves into an effector, although there may be less reason to retain the ancestor. Unlike the initially intended analysis, this setup allows to compare homologous effectors and detect possible events of secretion signal gain by a mechanism of terminal reassortment without loss of the ancestral T3SE. This evolutionary pathway may result in pairs of effectors that share the same N-terminus but have different C-termini.

Homologs of effectors of the respective genus were searched in the RefSeq proteomes belonging to *E. coli* or *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia* species in which at least one of the SecretEPDB effectors was identified. Groups of effectors and all related proteins that could be found in the same proteome at least once were made. Pairwise alignments between T3SEs and their co-occurring homologs were sorted into categories, depending on which terminus of which partner was aligned to the other protein.

The categories the alignments were sorted into are shown in figure 7. They represent all possible combinations of termini being aligned to each other, the terminus of one protein being aligned to an internal sequence of the other protein or termini staying unaligned that do not require several partial alignments of the same two proteins. These categories were grouped into 4 bigger categories, one for full-length alignment of both proteins, one for pairs with C-termini that are aligned to each other but N-termini that are not, one for pairs with N-termini that are aligned to each other but C-termini that are not and one for pairs with no termini aligned to each other. Termini were counted as aligned if no more than 3 aa were unaligned. If BlastP returned several alignments for one protein-pair, that pair was excluded from the analysis. Since those cases were rare, no common mechanism of T3S signal

evolution could be missed by excluding them.

The genera were analyzed separately. All protein pairs of one genus were analyzed together, but they were also grouped into pairs containing homologs that are other confirmed effectors, pairs containing homologs that are predicted to be effectors by EffectiveT3 but are not confirmed effectors, pairs that contain homologs that likely are not effectors, according to EffectiveT3, and pairs with homologs that get an intermediate score by EffectiveT3.

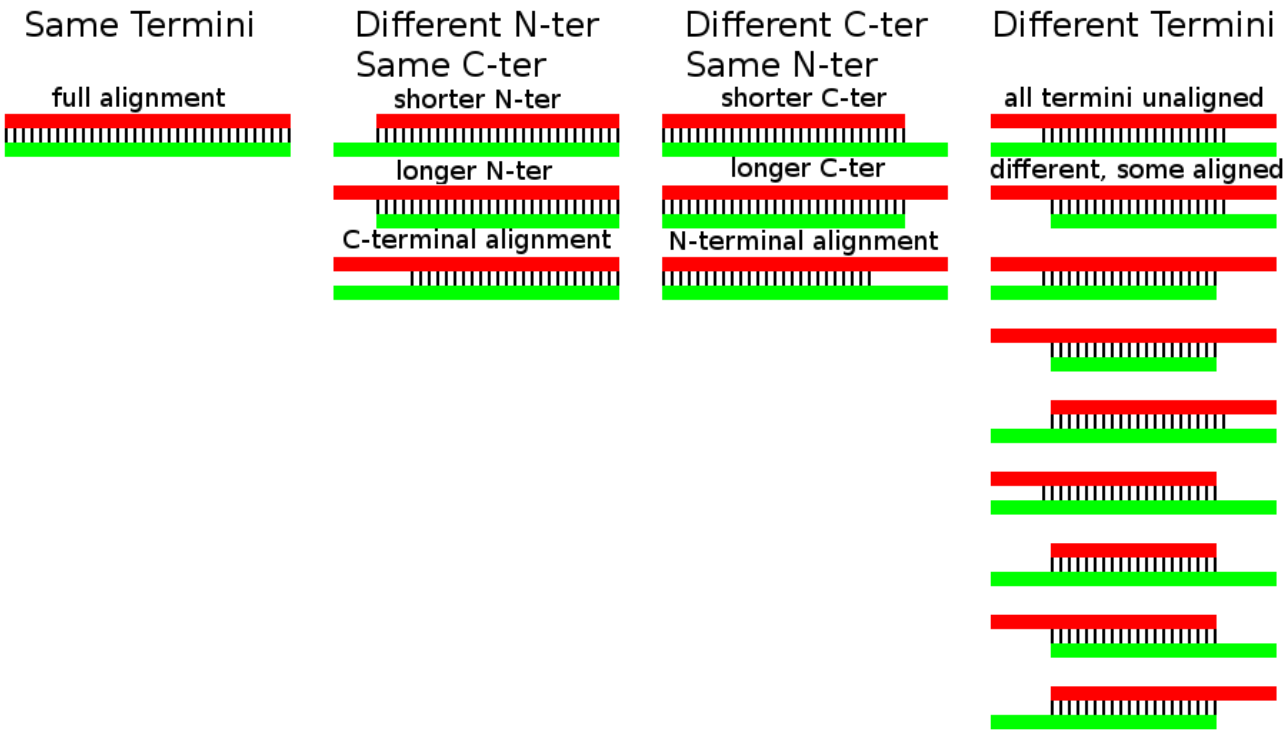

Figure 7: Schematic representation of the categories alignments were grouped into. Each column depicts one of the 4 bigger categories – both termini aligned to each other, C-termini aligned to each other but N-termini unaligned, N-termini aligned to each other but C-termini unaligned and no termini aligned to each other – and all the subcategories that belong to them. Red bars represent the confirmed effector and green bars the alignment partner.

Not all SecretEPDB effectors were found in RefSeq proteomes, and not all found effectors had any homologs within the same proteomes. Different taxa showed different tendencies. In *E. coli*, *Salmonella* and *Xanthomonas*, about 50% of found T3SEs had one or more homologs in one or several of the same proteomes the effector could be found in. In *Pseudomonas*, 26% of found effectors had homologs in the same proteome. In *Yersinia*, it was only 13% and in *Chlamydia* only 9%. The number of related proteins the T3SEs had varied a lot as well. Especially in *Chlamydia*, there were some effectors with a huge number of homologs (Table 6). Apparently, many effectors do not have homologs within the same proteome. They may have been gained by HGT and may or may not have originated from non-T3SEs within the context of a different proteome. They may have arisen as completely novel proteins, or the ancestral proteins may not have had homologs within the proteome and may have changed into an effector.

	Effectors	With Relatives	%	Pairs
Escherichia	77	41	53%	383
Salmonella	76	35	46%	825
Yersinia	30	4	13%	42
Pseudomonas	87	23	26%	103
Xanthomonas	15	7	47%	11
Chlamydia	65	6	9%	159

Table 6: The column 'Effectors' shows the number of SecretEPDB effectors that were found in the RefSeq proteomes. 'With Relatives' shows the number among them that co-occurred with at least one of their homologs in at least one proteome. '%' shows the percentage of effectors that co-occurred with at least one homolog among the T3SEs found in at least one proteome. 'Pairs' shows the number of effector-homolog pairs that could be formed and were not excluded from the analysis due to resulting in several partial alignments.

In *E. coli*, effectors more often had different N-termini and the same C-terminus as their homologs than different C-termini and the same N-terminus. This tendency was strongest when only looking at pairs with homologs that were predicted to not be effectors. Among these, more than 45% had different N-termini but the same C-terminus, and in most of the other pairs no termini align to each other. In roughly half of the cases with different N-termini but the same C-terminus, both proteins had an N-terminus that did not align to the other protein. However, if only one protein in the pair was a confirmed effector and one protein was a longer version of the other, the confirmed effector usually was the longer one. This would be in line with terminal reassortment without using the same N-terminus for several proteins and with N-terminal elongation. However, if those shorter homologs are indeed not effectors but fulfill a function independent of the T3SS, it is unclear why they can only be found in *E. coli* strains that have a T3SS but not in strains that do not. Perhaps, they are truncated effectors rather than ancestral proteins.

In *Salmonella*, about 38% of effector homologs had the same N-terminus but a different C-terminus. Among the homologs that were confirmed T3SEs themselves, the proportion was even higher, with 51%. In almost all of these pairs, both partners had an unaligned C-terminus, rather than one protein being a longer or shorter version of the other. This would be in line with methods of terminal reassortment that do not delete the ancestral protein.

In *Pseudomonas*, about 53% of related proteins fully aligned to the effector. If the homolog was a confirmed effector itself, the percentage was about 81%. Some had low sequence identity, and their functions may not be identical. In 17% of the protein pairs, only the C-termini aligned and in 14% only the N-termini. This allows for some terminal reassortment, especially since the respective terminus stayed unaligned in both partners for most of these pairs. In the original paper about the role of terminal reassortment in T3SE evolution, *Pseudomonas* was among the bacteria for which the most incidences of terminal reassortment could be found (Stavrinides et al., 2006). However, in this analysis, it did not stand out.

In *Chlamydia*, all T3SE homologs had different N-termini than the effector, even if they were confirmed effectors themselves. In roughly 25% of pairs, the C-termini were aligned to each other. In the rest, no termini were aligned to each other. However, there were few effectors with many related proteins and many effectors with none in *Chlamydia*.

In *Yersinia* and *Xanthomonas*, the numbers of effector-homolog pairs were low. In *Yersinia*, the percentage of effectors with homologs in the same proteome was low, and in both *Yersinia* and *Xanthomonas*, the number of effectors found in RefSeq proteomes was low. In *Xanthomonas*, there were no T3SEs with a huge number of related proteins. In both taxa, both termini were not aligned to each other for more than 80% of pairs. In *Yersinia*, some

proteins that shared the same N-terminus but had different C-termini were found (Table 7).

Escherichia	all	confirmed	predicted	non	undetermined
Full alignment	32.04%	53.28%	21.60%	5.13%	25.58%
Different N-ter	28.17%	18.25%	34.40%	46.15%	26.74%
Longer N-ter	8.79%	8.03%	7.20%	23.08%	5.81%
Shorter N-ter	5.17%	8.03%	6.40%	0.00%	1.16%
Partial alignment C-ter	14.21%	2.19%	20.80%	23.08%	19.77%
Different C-ter	7.49%	8.03%	10.40%	0.00%	5.81%
Longer C-ter	1.29%	0.73%	3.20%	0.00%	0.00%
Shorter C-ter	0.52%	0.73%	0.80%	0.00%	0.00%
Partial alignment N-ter	5.68%	6.57%	6.40%	0.00%	5.81%
Different termini	32.30%	20.44%	33.60%	48.72%	41.86%
Termini unaligned	27.13%	20.44%	24.00%	46.15%	33.72%

Salmonella	all	confirmed	predicted	non	undetermined
Full alignment	21.21%	19.37%	12.03%	10.71%	35.35%
Different N-ter	12.97%	10.41%	25.32%	10.71%	9.09%
Longer N-ter	2.30%	0.24%	1.27%	8.93%	5.56%
Shorter N-ter	0.24%	0.24%	0.63%	0.00%	0.00%
Partial alignment C-ter	10.42%	9.93%	23.42%	1.79%	3.54%
Different C-ter	37.70%	51.09%	40.51%	0.00%	18.18%
Longer C-ter	0.48%	0.00%	2.53%	0.00%	0.00%
Shorter C-ter	0.00%	0.00%	0.00%	0.00%	0.00%
Partial alignment N-ter	37.21%	51.09%	37.97%	0.00%	18.18%
Different termini	28.12%	19.13%	22.15%	78.57%	37.37%
Termini unaligned	19.76%	18.64%	12.03%	17.86%	28.79%

Yersinia	all
Full alignment	0.00%
Different N-ter	0.00%
Longer N-ter	0.00%
Shorter N-ter	0.00%
Partial alignment C-ter	0.00%
Different C-ter	19.05%
Longer C-ter	0.00%
Shorter C-ter	0.00%
Partial alignment N-ter	19.05%
Different termini	80.95%
Termini unaligned	23.81%

Xanthomonas	all
Full alignment	9.09%
Different N-ter	0.00%
Longer N-ter	0.00%
Shorter N-ter	0.00%
Partial alignment C-ter	0.00%
Different C-ter	0.00%
Longer C-ter	0.00%
Shorter C-ter	0.00%
Partial alignment N-ter	0.00%
Different termini	90.91%
Termini unaligned	36.36%

Pseudomonas	all	confirmed	predicted	non	undetermined
Full alignment	53.40%	80.65%	15.38%	0.00%	25.00%
Different N-ter	16.50%	9.68%	34.62%	0.00%	50.00%
Longer N-ter	2.91%	1.61%	7.69%	0.00%	0.00%
Shorter N-ter	0.97%	1.61%	0.00%	0.00%	0.00%
Partial alignment C-ter	12.62%	6.45%	26.92%	0.00%	50.00%
Different C-ter	13.59%	6.45%	34.62%	9.09%	0.00%
Longer C-ter	3.88%	1.61%	11.54%	0.00%	0.00%
Shorter C-ter	1.94%	1.61%	0.00%	9.09%	0.00%
Partial alignment N-ter	7.77%	3.23%	23.03%	0.00%	0.00%
Different termini	16.50%	3.23%	15.38%	90.91%	25.00%
Termini unaligned	5.83%	0.00%	11.53%	18.18%	25.00%

Chlamydia	all	confirmed	predicted	non	undetermined
Full alignment	0.00%	0.00%	0.00%	0.00%	0.00%
Different N-ter	24.53%	40.68%	26.67%	1.28%	60.00%
Longer N-ter	0.00%	0.00%	0.00%	0.00%	0.00%
Shorter N-ter	0.00%	0.00%	0.00%	0.00%	0.00%
Partial alignment C-ter	24.53%	40.68%	26.67%	1.28%	60.00%
Different C-ter	0.00%	0.00%	0.00%	0.00%	0.00%
Longer C-ter	0.00%	0.00%	0.00%	0.00%	0.00%
Shorter C-ter	0.00%	0.00%	0.00%	0.00%	0.00%
Partial alignment N-ter	0.00%	0.00%	0.00%	0.00%	0.00%
Different termini	75.47%	59.32%	73.33%	98.72%	40.00%
Termini unaligned	49.69%	18.64%	73.33%	82.05%	0.00%

Table 7: The tables show how many percent of the alignments between T3SEs and homologs that co-occur in the same proteome belong to which of the categories depicted in figure 7. The yellow rows correspond to the 4 bigger categories – protein pairs where both termini align to each other, pairs where the C-termini align to each other but the N-termini do not, pairs where the N-termini align to each other but the C-termini do not and pairs where no termini align to each other. The blue rows show the subcategories of the yellow ones above. For alignments where no termini align to each other, not all subcategories are shown. In the subcategory 'Longer N-ter', the T3SE has a longer N-terminus than the alignment partner. In the subcategory 'Shorter N-ter', the T3SE has a shorter N-terminus than the alignment partner. In the subcategory 'Partial alignment C-ter', both alignment partners have an unaligned N-terminus. The equivalent applies to the subcategories of 'Different C-ter'. In the subcategory 'Termini unaligned', no terminus of any alignment partner aligns to the other protein. The column 'all' shows the percentages with which these subcategories occur for all effector-homolog pairs. The column 'confirmed' only takes pairs between two confirmed effectors into account. 'predicted' corresponds to pairs with homologs that are not confirmed but predicted effectors. 'non' corresponds to predicted non-effectors, and 'undetermined' corresponds to pairs with homologs that get an intermediate score by EffectiveT3. For *Yersinia* and *Xanthomonas*, only the column 'all' is listed due to the small number of protein pairs.

If one protein in a pair was longer at the N-terminus than the other, the confirmed effector tended to be the longer one. However, termini can be uncertain in protein annotation, RefSeq might contain some degenerating proteins, and this analysis does not show which proteins are ancestral. Therefore, it cannot be concluded with certainty whether N-terminal elongation plays a role in T3S signal acquisition. Either way, it does not seem to be the most common mechanism. Additionally, in *E. coli*, the N-terminally shorter versions of effectors do not seem

to exist in proteomes that do not contain a T3SS, as almost no effector homologs could be found in them. This casts further doubt on them being ancestral proteins with a different function. There were very few cases in which a confirmed effector was the N-terminally shorter partner of a pair. Often, both partners were confirmed effectors in these rare instances. Therefore, N-terminal truncation probably plays almost no role in T3S signal evolution. More protein pairs with one terminus unaligned to each other, which may indicate terminal reassortment, were found than pairs where one partner had a longer terminus than the other. This is true for pairs that align N-terminally and pairs that align C-terminally.

Some T3SEs had more homologs than others, and some occurred in more proteomes. Homologs in different proteomes might be orthologous to each other. Some T3SEs contributed more to the results than others.

2.3.2. Prevalence of terminal reassortment among related effectors

To collect further evidence on whether terminal reassortment is a common mechanism of signal sequence acquisition, all used confirmed T3SEs within a genus were compared to each other. Otherwise unrelated effectors sharing the same N-terminus would indicate terminal reassortment. However, without the context of a proteome, it does not provide evidence if an evolutionary mechanism was used that resulted in the loss of the ancestral protein or if an N-terminus was reused for several proteins within the same organism.

Additionally, it was intended to determine how commonly related effectors that do not share the same N-terminus occur. In the previous analysis, shared N-termini and different C-termini were more common for pairs of confirmed effectors in *E. coli* and *Salmonella*, which were the taxa with the biggest sample sizes, but pairs that shared the same C-terminus and had a different N-terminus did occur among confirmed effectors. While the benefit of exchanging the C-terminus but keeping the N-terminus is clear, as it pairs a secretion signal with new functional domains, there is no obvious use to replacing the signal sequence of an already existing effector. Therefore, it might indicate a relative ease of signal sequence gain if one and the same non-effector gained a secretion signal several times independent of each other or if the secretion signal of an already existing effector was exchanged.

The same analysis as before was repeated for all pairs of confirmed T3SEs that are related to each other and belong to the same taxon. The number of used effectors and pairwise alignments per genus are listed in table 8. Pairwise alignments were grouped into the same 4 big categories as before, but there were fewer subcategories since all proteins were confirmed effectors and it did not matter which one was longer or shorter. Categories are shown in figure 8.

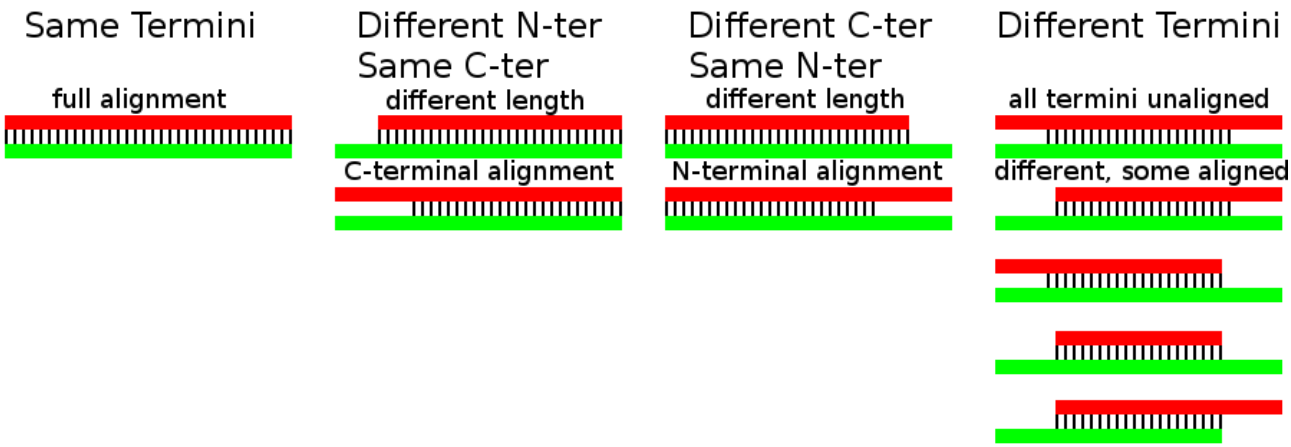

Figure 8: Schematic representation of the categories alignments were grouped into. Each column depicts one of the 4 bigger categories – all termini aligned to each other, C-termini aligned to each other but N-termini unaligned, N-termini aligned to each other but C-termini unaligned and no termini aligned to each other – and all the subcategories that belong to them. All proteins are confirmed effectors.

The tendencies that occurred for the individual taxa do not hold up for pairs of confirmed effectors without the context of the proteome. In *E. coli*, about 34% of effector pairs shared the same C-terminus but had a different N-terminus, and about 8% had the same N-terminus but a different C-terminus. In *Salmonella*, both cases occurred for a bit more than 10% of pairwise alignments. For both taxa, this is very different than among effector-homolog pairs that occur in the same proteome. *Pseudomonas* effector pairs had the lowest proportion of full alignments, with about 15%, unlike in the previous analysis where they had the highest number (Table 9).

	T3SEs	Pairs
Escherichia	120	515
Salmonella	103	232
Yersinia	46	68
Pseudomonas	303	1866
Xanthomonas	32	43
Chlamydia	74	36

Table 8: Number of unique SecretEPDB T3SEs in the respective taxa and number of alignments of related T3SEs among them.

	Escherichia	Salmonella	Yersinia	Pseudomonas	Xanthomonas	Chlamydia
Full alignment	54.56%	60.78%	54.41%	15.38%	32.56%	38.89%
Different N-ter	7.57%	10.78%	26.47%	27.44%	13.95%	5.56%
Different N-ter length	6.41%	5.17%	26.47%	22.29%	11.63%	0.00%
Partial alignment C-ter	1.17%	5.60%	0.00%	5.14%	2.33%	5.56%
Different C-ter	33.79%	12.93%	0.00%	11.68%	0.00%	2.78%
Different C-ter length	18.45%	0.86%	0.00%	7.45%	0.00%	0.00%
Partial alignment N-ter	15.34%	12.07%	0.00%	4.23%	0.00%	2.78%
Different termini	4.08%	15.52%	19.12%	45.50%	53.49%	52.78%
Termini unaligned	2.91%	11.21%	8.82%	4.88%	2.33%	41.67%
Longer and shorter p.	0.00%	2.59%	0.00%	19.45%	23.26%	0.00%

Table 9: The table shows how many percent of the alignments between T3SEs belong to which of the categories depicted in figure 8. The yellow rows correspond to the 4 bigger categories - protein pairs where both termini align to each other, pairs where the C-termini align to each other but the N-termini do not, pairs where the N-termini align to each other but the C-termini do not and pairs where no termini align to each other. The blue rows show the subcategories of the yellow categories above. 'Different N-ter length' are pairs where one N-terminus is aligned to the other effector but that other effector has a longer N-terminus. 'Partial alignment C-ter' are pairs where N-termini stay unaligned in both partners. The equivalent applies to the subcategories of 'Different C-ter'. Not all subcategories of 'Different Termini' are shown. 'Termini unaligned' corresponds to pairs where no terminus in any alignment partner is aligned to the other protein. 'Longer and shorter p.' corresponds to pairs in which one alignment partner is longer at both ends than the other and the shorter partner fully aligns to the longer partner. The columns show which percentage of SecretEPDB effector pairs of the respective taxon fall into which category.

Remarkably, effector pairs with the same C-terminus but a different N-terminus occurred more often than effector pairs with the same N-terminus but a different C-terminus in all used taxa but *E. coli* and *Salmonella*. The benefit of exchanging the N-terminus and keeping the C-terminus is not immediately obvious, as it exchanges the signal and may keep the functional domains. Some signal sequences may be stronger than others, or some functional domains may be gained or lost alongside the signal. Alternatively, it could be a neutral change not selected against. In *Pseudomonas* and *Xanthomonas*, effectors that are shorter at both ends than a related effector were also common.

However, in alignments with different N-termini but the same C-terminus, most pairs consisted of one shorter and one longer alignment partner, and in fewer, both N-termini stayed unaligned. Alignment partners that differ in length, rather than both having an unaligned terminus, could more easily occur due to entries in SecretEPDB that do not have the same length as the true protein. Among alignments with different C-termini but the same N-terminus, both termini stayed unaligned in most *Samonella* effector pairs, but *Escherichia* and *Pseudomonas* had slightly more pairs with one longer and one shorter alignment partner.

To get a better idea if the high number of pairs with different N-terminal length are due to shortened versions of T3SEs in the database, effectors were searched in the proteomes used before. While the occurrence of a protein in RefSeq proteomes does not prove that it truly exist at this length, it at least makes sure that the quality of data worked with is not poorer than RefSeq annotations.

In *E. coli*, 57% of effector pairs were left after excluding those not found in the used RefSeq proteomes. In *Salmonella*, 78% were left, in *Yersinia* 62% and in *Chlamydia* 72%. In *Xanthomonas*, only 12% were left and in *Pseudomonas* only 7% (Table 10). Some *Pseudomonas* and *Xanthomonas* species effectors were identified in were represented by few or no RefSeq proteomes. Excluding effectors not found in RefSeq probably excluded a lot of effectors that actually exist, especially in these two genera.

After the removal of pairs that include proteins not found in RefSeq, there were more pairs with different C-termini but the same N-terminus than pairs with different N-termini but the same C-terminus in *E. coli*, *Salmonella* and *Pseudomonas*. In *Yersinia* and *Chlamydia*, the number of pairs with different N-termini and the same C-terminus remained higher. However, the sample size was smaller for these two taxa than for *E. coli*, *Salmonella* and *Pseudomonas*. In *Xanthomonas*, no effector pair with one terminus aligned to each other and the other not aligned to each other was found, but the sample size was very small (Table 11).

	T3SEs	Pairs
Escherichia	77	292
Salmonella	76	182
Yersinia	30	42
Pseudomonas	87	136
Xanthomonas	15	5
Chlamydia	65	26

Table 10: Number of unique SecretEPDB T3SEs found in RefSeq proteomes in the respective taxa and number of alignments of related T3SEs among them.

	Escherichia	Salmonella	Yersinia	Pseudomonas	Xanthomonas	Chlamydia
Full alignment	56.16%	75.27%	78.57%	65.44%	60.00%	53.85%
Different N-ter	5.82%	7.14%	14.29%	5.15%	0.00%	3.85%
Different N-ter length	4.45%	0.55%	14.29%	3.68%	0.00%	0.00%
Partial alignment C-ter	1.37%	6.59%	0.00%	1.47%	0.00%	3.85%
Different C-ter	32.35%	10.99%	0.00%	22.79%	0.00%	0.00%
Different C-ter length	17.47%	0.55%	0.00%	13.97%	0.00%	0.00%
Partial alignment N-ter	15.07%	10.44%	0.00%	8.82%	0.00%	0.00%
Different termini	5.48%	6.59%	7.14%	6.62%	40.00%	42.31%
Termini unaligned	4.79%	6.04%	0.00%	3.68%	20.00%	30.77%
Longer and shorter p.	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 11: The table shows how many percent of the alignments between T3SEs belong to which of the categories depicted in figure 8, excluding all T3SEs not found in the RefSeq proteomes. The yellow rows correspond to the 4 bigger categories – protein pairs where both termini align to each other, pairs where the C-termini align to each other but the N-termini do not, pairs where the N-termini align to each other but the C-termini do not and pairs where no termini align to each other. The blue rows show the subcategories of the yellow categories above. 'Different N-ter length' are pairs where one N-terminus is aligned to the other effector but that other effector has a longer N-terminus. 'Partial alignment C-ter' are pairs where N-termini stay unaligned in both partners. The equivalent applies to the subcategories of 'Different C-ter'. Not all subcategories of 'Different Termini' are shown. 'Termini unaligned' corresponds to pairs where no terminus in any alignment partner is aligned to the other protein. 'Longer and shorter p.' corresponds to pairs in which one alignment partner is longer at both ends than the other and the shorter partner fully aligns to the longer partner. The columns show which percentage of the pairs of SecretEPDB effectors that were found in RefSeq proteomes of the respective taxon fall into which category.

Excluding proteins not found in RefSeq reduced the percentage of pairs with different N-termini but the same C-terminus in all used taxa, in some moderately, in others a lot. The percentage of pairs with the same N-terminus but different C-termini was slightly decreased in *E. coli*, *Salmonella* and *Chlamydia* and almost doubled in *Pseudomonas*. The percentage of effector pairs with different N-terminal length but the same C-terminus tended to be reduced

more than the percentage of effector pairs with both N-termini unaligned and the same C-terminus. If one effector of a pair was not found in any of the proteomes but the other was, the protein that was not found was more often the shorter one (Table 12). No pairs where one partner was an at both ends truncated version of the other stayed included.

	Escherichia	Salmonella	Yersinia	Pseudomonas	Xanthomonas	Chlamydia
Excluded longer	6	0	8	21	0	0
Excluded shorter	9	11	4	154	3	0
Both excluded	5	0	0	236	2	0

Table 12: Numbers of effector pairs that were excluded because one or both alignment partners were not found in RefSeq proteomes. The row 'Excluded longer' gives the numbers of pairs excluded because the longer alignment partner was not found in the RefSeq proteomes but the shorter partner was. 'Excluded shorter' gives the number of pairs excluded because the shorter alignment partner was not found in the RefSeq proteomes but the longer partner was. 'Both excluded' gives the number of pairs that were excluded because both proteins were not found in the RefSeq proteomes. The columns correspond to the used genera.

Nonetheless, some pairs with one N-terminally longer and one N-terminally shorter protein persisted. SecretEPDB might be more prone to contain partial effectors than the RefSeq proteomes, but it cannot be excluded that the different length pairs of effectors that could be found in the proteomes are due to wrongly annotated gene starts either. However, this happens less easily for pairs that both posses an N-terminus that is not aligned to the other protein. To not truly fall into the category of proteins with related C-termini but unrelated N-termini, both partners would need to be longer than the true protein. While not as prevalent as the analysis of all the SecretEPDB effector pairs may suggest, effectors that have different N-termini but are otherwise related may occur.

In roughly 15% of remaining *E. coli* effector pairs, 10% of remaining *Salmonella* effector pairs and 9% of *Pseudomonas* effector pairs, the N-termini aligned to each other but both C-termini stayed unaligned. Those are effectors that may have acquired the same N-terminus through some method of terminal reassortment. Their prevalence was higher than that of effector pairs with aligned C-termini but unaligned N-termini, with 1%, 7% and 1%, respectively. Neither case occurred in *Yersinia* and *Xanthomonas*, and one effector pair with unaligned N-termini but aligned C-termini occurred in *Chlamydia* after removal of proteins not found in the proteomes.

While the proportion of N-terminally homologous pairs with unrelated C-termini may not look huge, it is worth noting that 56% to 75% of the effector pairs in *E. coli*, *Salmonella* and *Pseudomonas* fully align to each other. Some of them share a very high sequence identity and may represent different versions of the same protein in different species or strains. If they were excluded, possible cases of terminal reassortment would make up a sizable portion of the remaining pairs, indicating that it may be an important mechanism of T3SE evolution.

3.4. Analyses based on T3SE prediction

3.4.1. Identifying mutations that may be convenient for T3S signal evolution

If a certain kind of mutation or genomic region is more likely to look like a T3S signal, it can more easily contribute to the evolution of a new T3SEs. To identify mutations that might more likely give rise to secretion signals, different types of sequences, most of which could become protein N-termini via potential mechanisms of T3S signal evolution, were generated, and EffectiveT3 was used to predict if they could be secreted. The proportion of positive predictions was compared between them.

Truncated proteins were generated by removing the 2nd to the 26th aa from 10 proteomes of

E. coli, one of *Salmonella*, one of *Yersinia*, 10 of *Pseudomonas*, one of *Xanthomonas* and one of *Chlamydia*. +1 and -1 frameshifts were created from the corresponding coding sequences files from RefSeq of the respective proteomes. However, if a frameshift led to a T3S signal, a compensatory frameshift would be needed to retain the rest of the protein, which might make this evolutionary pathway harder in practice. Random genomic regions and random intergenic regions were extracted from one RefSeq genome per genus. Intergenic regions may become protein N-termini if a gene becomes N-terminally elongated. 25 aa long random sequences, starting with methionine, assuming a GC content of 50% and taking into consideration that more codons correspond to some aa than to others, were generated. They do not resemble a specific kind of mutation. The original proteomes, the truncated proteomes, the frameshifted proteins, the genomic and the intergenic regions and the random sequences were given as input to EffectiveT3 (Figure 9).

However, T3SE prediction tools are imperfect and are black boxes. The current version of EffectiveT3 has a false positive rate of about 7% with standard settings. As it learned features of the positive and the negative set of training data, it might be even more unreliable for data that is fundamentally different to the training data (Eichinger et al., 2016). It cannot show how many percent of a given type of sequence would be secreted. At most, it may provide evidence for which kinds of sequences tend to look more similar to a T3S signal than other sequences, but even that is afflicted with uncertainty, and it cannot be considered strong evidence on its own.

18% of the randomly generated sequences were predicted to be secreted. On average, about 10% of whole and truncated proteins were predicted to be secreted. For the individual taxa, the percentages ranged from 8% to 13% for the whole proteins and from 7% to 11% for the truncated proteins. On average, 20% of the random genomic regions but only 12% of the intergenic regions were predicted to be secreted. Percentages ranged from 16% to 24% for the genomic regions and from 8% to 14% for the intergenic regions. About 24% of frameshifted proteins were predicted to be secreted (Table 13). However, there were huge differences between different taxa and between +1 and -1 frameshifts in *Pseudomonas*, *Xanthomonas*, *Salmonella* and *Chlamydia*. 60% of the -1 frameshifts were predicted to be secreted in *Pseudomonas* but only 16% of the +1 frameshifts. In *Xanthomonas* and *Salmonella*, the tendencies were similar but the numbers a bit lower. In *Chlamydia*, 18% of the +1 frameshifts but only 7% of the -1 frameshifts were predicted to be secreted. In *E. coli* and *Yersinia*, the proportion of positive predictions was between 15% and 20% for both frameshifts (Table 14). Different *E. coli* and *Pseudomonas* proteomes showed some variation in their proportions of positive predictions for proteins, truncated proteins and frameshifts, but strains or species of the same genus tended to be more similar to each other than they were to other genera. Especially frameshifted proteins received a very different percentage of positive predictions in different genera but not between strains or species of the same genus. The different aa frequencies and codon usages in the N-termini of individual taxa seem to result in very different probabilities of a frameshift leading to something that looks like a secretion signal to EffectiveT3.

To test if this only applies to the N-termini or holds true if a deletion is combined with a frameshift, -1 frameshifts followed by deletion of the 2nd to 26th amino acid of one previously used *Pseudomonas*, *Salmonella* and *E. coli* proteome were made. In *Pseudomonas* and *E. coli*, the ratio of positive predictions stayed similar to the prediction for -1 frameshifts without deletion. In *Salmonella*, it went down from 45% for only -1 frameshifts to 18% for the combination of frameshift and deletion. Thus, the amino acid frequency and codon usage in *Pseudomonas* is such that -1 frameshifts look particularly similar to an effector N-terminus throughout bigger parts of the protein, but in *Salmonella*, this is only true for the N-termini themselves. In *E. coli*, it is neither the case for the N-termini nor the 25 aa deletion.

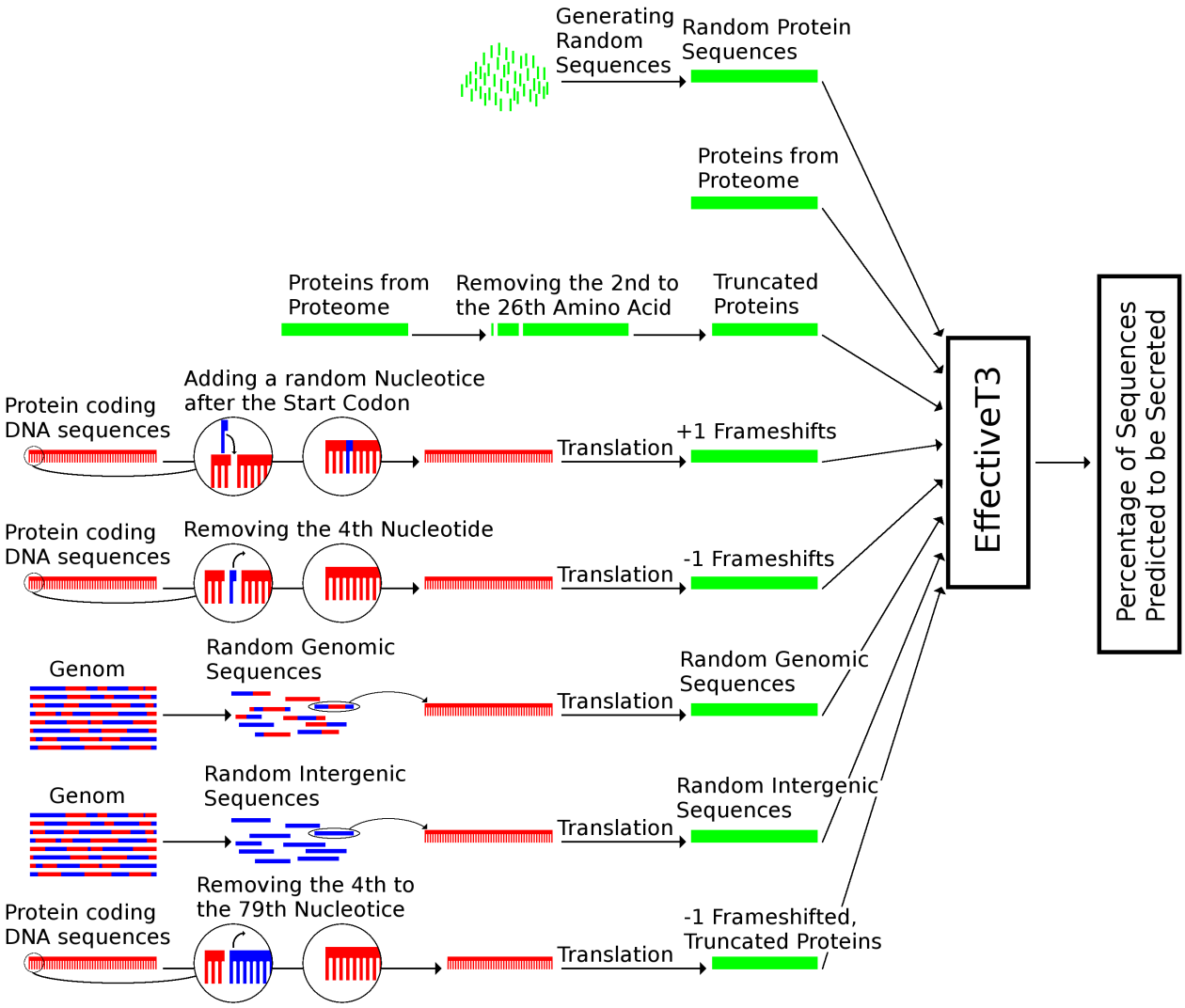

Figure 9: Schematic representation of the workflow of the analysis described in this section. Various protein sequences were collected or generated and given to EffectiveT3 as input. Random protein sequences were generated. Proteins from proteomes were used in an unaltered form and as truncated proteins, by removing the 2nd to 26th amino acid. Frameshifted proteins and a combination of frameshift and truncation were created from protein-coding DNA sequences by removing or adding random nucleotides after the start codon, followed by translation. Random genomic sequences and random intergenic sequences were extracted from genomes and translated. Amino acids and proteins are depicted in green, nucleotides and DNA in red or blue. The percentages of sequences predicted to be secreted were compared.

	Proteins	Truncated P.	Genomic	Intergenic	Frameshift	Random Seq.
Escherichia	9.83%	10.56%	20.44%	10.95%	17.21%	
Salmonelly	8.71%	9.52%	18.38%	12.41%	28.93%	
Yersinia	11.49%	11.36%	20.93%	13.61%	17.75%	
Pseudomonas	8.38%	6.80%	24.46%	13.71%	37.88%	
Xanthomonas	11.08%	8.89%	21.65%	13.83%	27.60%	
Chlamydia	13.18%	11.15%	16.49%	7.67%	12.53%	
Average	10.44%	9.71%	20.39%	12.03%	23.65%	18.24%

Table 13: Percentages of sequences predicted to be secreted by EffectiveT3. The column 'Proteins' shows the percentages for unaltered proteins from proteomes. The column 'Truncated P.' shows the percentages for the same proteins with the 2nd to 26th aa deleted. 'Genomic' shows the percentages for random genomic regions and 'Intergenic' for random intergenic regions. 'Frameshift' shows the percentages for +1 and -1 frameshifts averaged. 'Random Seq.' shows the percentages for randomly generated sequences. Each row corresponds to one of the used taxa, and the last row shows the average of all taxa. Random sequences are only included in this row, as they are not taxon-specific.

Frameshift Mutations	
Escherichia +1	17.02%
Escherichia -1	17.41%
Yersinia + 1	19.54%
Yersinia -1	15.96%
Pseudomonas +1	15.93%
Pseudomonas -1	59.84%
Xanthomonas +1	12.69%
Xanthomonas -1	42.51%
Salmonella +1	13.30%
Salmonella -1	44.56%
Chlamydia +1	17.77%
Chlamydia -1	7.28%

Frameshift Mutations -76	
Escherichia	16.64%
Pseudomonas	61.35%
Salmonella	18.31%

Table 14: The left Table shows the percentage of positive predictions by EffectiveT3 for +1 and -1 frameshifts for the used taxa. '+1' or '-1' after the taxon name specifies which frameshift it is. The right table shows the percentages of positive predictions for the combination of -1 frameshifts with 25 aa deletion for *E. coli*, *Pseudomonas* and *Salmonella*.

3. Discussion

Overall, the mechanism of T3S signal sequence acquisition best supported by evidence collected in this study was terminal reassortment, which is also one of the mechanisms suggested by previous research. N-terminal elongation, too, is a previously considered pathway of secretion signal gain, and some results of this study provide supporting evidence. However, others cast doubt on their meaningfulness. Additionally, frameshifted sequences might closely resemble secretion signals in some taxa.

Stavrinides et al. already collected various evidence for terminal reassortment as a mechanism of T3SE evolution in 2006. They discovered that many T3SEs are chimeras of an effector and another protein or of two effectors. Further, they found that the genomes of some organisms that posses T3SEs contain other open reading frames homologous to effector N-

termini. Both chimeric and truncated loci were overrepresented among T3SE families compared to non-T3SE protein families. The regions upstream of N-terminally homologous T3SEs tend to also be related and share ribosome-binding sites and promoters. Chimeric T3SEs and open reading frames containing effector N-termini are often associated with mobile genetic elements, facilitating the creation of new T3SEs and allowing terminal reassortment to take place without loss of the ancestral effector or effector terminus.

The results of this study agree on terminal reassortment as an important mechanism of T3SE evolution. Both the analysis aligning and comparing T3SEs to homologs co-occurring in the same proteome and the analysis aligning and comparing T3SEs within a genus to each other revealed patterns that are consistent with terminal reassortment being a common mechanism of secretion signal gain.

While the comparison between effectors and homologs co-occurring in the same proteome also provides some support for N-terminal elongation as a mechanism of T3S signal sequence acquisition, overall, the evidence for its role in T3SE evolution remained inconclusive. A decent number of effectors were N-terminally elongated relative to their homolog, particularly if the homolog was a predicted non-effector. In contrast, no instance of an effector that was N-terminally truncated compared to its predicted non-effector homolog was found. This is unlike the previous analysis by Arnold et al., who compared T3SEs to orthologs in organisms without a T3SS and could not identify a preference for N-terminal elongation over N-terminal truncation (Arnold et al., 2009). However, the lack of effector homologs in *E. coli* strains without a T3SS weakens the evidence collected in the current study. If these N-terminally shorter versions were ancestral proteins, they would be expected to fulfill a different function and exist in strains that do not contain a T3SS.

The initial reason why Arnold et al. proposed that N-terminal elongation may be a convenient pathway of T3S signal sequence acquisition was because they discovered that translated intergenic regions have a similar amino acid composition to the T3S signal and may easily evolve into one (Arnold et al., 2010). If the start codon is shifted upstream, an intergenic region can be turned into a new protein N-terminus. However, in this study, intergenic regions did not seem to resemble a secretion signal more closely than random proteins, according to EffectiveT3. Random genomic regions did, but via frameshifted gene-sequences and not due to intergenic regions. Likewise, truncated proteins were no more likely to be predicted to be secreted by EffectiveT3 than whole proteins.

To sum up, N-terminal elongation remains a feasible way of T3S signal sequence acquisition, albeit not supported by conclusive evidence, whereas the results strongly speak against N-terminal truncation as a relevant mechanism of T3SE evolution.

In the analysis using EffectiveT3 to identify types of mutations that may be convenient for T3S signal evolution, frameshifted gene starts resulted in an overwhelming proportion of positive predictions in some taxa. In *Pseudomonas*, this remained true even for a combination of N-terminal truncation and frameshift of the remaining sequence. However, this kind of evidence is inherently unreliable, and this study did not conduct further analysis to investigate if frameshifts majorly contribute to the evolution of new T3SEs. Furthermore, fameshifts would lead to a completely altered sequence instead of adding a novel N-terminus to a previously existing protein, unless a compensatory frameshift restoring the original reading frame or a frameshifted sequence of one gene was pieced together with the sequence of another gene via some sort of rearrangement.

Nonetheless, there are earlier studies that investigated a different question and coincidentally provide support for frameshifts as a possible way of T3SE evolution. Arnold et al. conducted a similar analysis with EffectiveT3 and also found a high rate of positive predictions among frameshifted proteins (Arnold et al., 2009). Their analysis is, of course, afflicted with the same uncertainties as the one in this study. However, there are studies that

provide direct experimental evidence for the secretion of frameshifted proteins. Anderson and Schneewind showed that 3 out of 4 frameshifts of the *Yersinia* proteins YopE and YopN could be secreted, and Rüssmann et al. showed that both frameshifts of *InvJ* could be secreted (Anderson and Schneewind, 1997; Rüssmann et al., 2002). As both studies investigated properties of effectors and were not trying to determine if frameshifts were a feasible pathway of T3SE evolution, all tested proteins were T3SEs. While the sample size is not big enough to draw reliable conclusions, frameshifts may be a potential mechanism of secretion signal gain worth further looking into.

It is also noteworthy that a lot of effectors do not have homologs in any organisms without a T3SS that are included in the RefSeq database and some are confined to a small taxonomic group. 34% of used *E. coli* T3SEs had no homologs outside of *Escherichia* and the closely related genera *Shigella* and *Citrobacter*. In *Chlamydia*, the proportion was even higher with 62% of its effectors not having any homologs in any genus other than *Chlamydia*. In particular, inclusion membrane proteins were not related to any proteins outside of *Chlamydia*, and *Chlamydia* possesses various different inclusion membrane proteins that do not share sequence similarity with each other. 24% of *Pseudomonas* effectors were not related to any proteins outside its own genus. This may indicate that many T3SEs might have arisen as completely novel proteins, rather than by attaching a secretion signal to a protein that fulfills a different function in organisms without a T3SS, although a total lack of homologs cannot be verified because the databases are incomplete.

While these effectors do not seem to have acquired any large part from a more widespread ancestral protein, the conducted analyses cannot exclude that some gained a tiny fragment at their N-terminus from another T3SE. The most important part of the signal can be small, and the cut-off e-value BlastP was set to was too low to detect very short homologous sequences. Alternatively, the high proportion of random genomic regions and frameshifts that resemble a signal sequence, according to EffectiveT3, may facilitate the secretion of newly arising proteins.

The lack of effector homologs in *E. coli* strains without a T3SS is in line with many T3SEs originating as novel proteins or at least from ancestral proteins that are not universal to bacteria and may already fulfill a function only needed in specific contexts, such as, perhaps, in some cases, being other virulence factors.

Another way of T3SE evolution that was considered is the acquisition of proteins from distant taxa without a T3SS that may evolve towards an effector. Of particular interest was HGT from eukaryotes, as this is a known evolutionary mechanism in some other virulence factors and a convenient way to gain proteins that have already been optimized for a function they carry out in the eukaryotic cell.

The T3S signal might not be very specific and depends on amino acid composition and other features in an, as-of yet, not well understood way rather than a distinct sequence motive. It has, therefore, been suggested that N-termini of proteins accessible to the T3SS may evolve towards or against a secretion signal (Arnold et al., 2009). Some proteins in organisms without a T3SS could be secretable because they are not subjected to the same evolutionary pressures and have no need to not contain a T3S signal. This could facilitate the evolution of a protein gained by HGT into an effector. Indeed, Arnold et al. and Wang et al. found more putative secretable proteins in some organisms that do not possess a T3SS with their respective prediction tools than would be expected based on their false positive rates (Arnold et al., 2009; Wang et al., 2013). Furthermore, Wange et al. experimentally verified that some proteins of organisms without a T3SS could be secreted if expressed in an organism with a T3SS. They tested this for 3 yeast proteins, which had been predicted to be secretable, in *Salmonella* and received a positive result for 2 (Wang et al., 2013).

Depending on the taxon, 6% to 24% of T3SEs used in the current study had homologs in

the gram-positive bacteria and 1% to 38% had homologs in the eukaryotes. This is a relevant minority of effectors that have homologs in distant taxa that certainly do not contain a T3SS.

However, no matter how convenient HGT from eukaryotes may seem as an evolutionary pathway for T3SEs, in practice, no definite instance of an effector that evolved from a eukaryotic protein acquired by HGT could be identified. In *Chlamydia*, one protein was found that either was gained from *Trypanosomatidae* or conferred to *Trypanosomatidae* by HGT, but, overall, HGT from eukaryotes seems to play at most a minor role in T3SE evolution.

Either way, the N-termini of none of the used effectors that have eukaryotic homologs aligned to any part of the eukaryotic proteins, except in *Chlamydia*, indicating that their secretion signal did not originate from a sequence shared with eukaryotes. A few instances of T3SEs that N-terminally aligned to their gram-positive relatives were found. However, it is unclear if HGT took place between the bacterium containing the T3SE and the gram-positive bacterium or in which direction. In rare cases, proteins with a coincidental signal sequence might be gained from distant taxa, but it does not seem to be a common mechanism of T3SE evolution with eukaryotic or gram-positive proteins.

During the alignment and comparison of SecretEPDB effectors to each other to detect instances of terminal reassortment or other meaningful patterns, it became apparent that several proteins included in SecretEPDB may have an inaccurate length. The analysis was repeated only using T3SEs found in RefSeq to ensure the quality of the data at least met RefSeq's standards. However, wrong-length proteins in SecretEPDB could have affected earlier parts of the study.

The effectors extracted from SecretEPDB were not used in the analysis that attempted to identify types of mutations that may be convenient for T3S signal evolution with EffectiveT3. Therefore, it stayed unaffected by the quality of the database.

The results of the comparison between effectors and homologs co-occurring in the same proteome should not have been majorly distorted by wrong-length proteins in SecretEPDB because only T3SEs found in RefSeq proteomes were used. It may partially explain why, in some taxa, only a small proportion of SecretEPDB effectors could be found in the proteomes, though.

Both, the analysis investigating the taxonomic composition of effector homologs to draw conclusions about the origin of T3SEs and the analyses to determine if any effectors evolved from proteins acquired from eukaryotes via HGT did not specifically deal with protein N-termini. Therefore, neither of them should have been severely affected by wrong-length proteins in the database.

However, the analysis trying to determine if T3SEs were more or less likely to share an N-terminus with a eukaryotic or gram-positive homolog than a C-terminus to investigate to what extent conserved domains or N-termini of proteins gained by HGT contribute to secretion signal evolution may have been. It directly used the termini of proteins in SecretEPDB without checking if they existed in a more reliable database.

T3SEs in SecretEPDB seem to be more often N-terminally truncated than elongated relative to the real protein. An N-terminally truncated effector could wrongly make it look as if the effector N-terminus aligns to the eukaryotic or gram-positive homolog when it, in fact, is an internal sequence that does. However, hardly any T3SEs were found that seemed to share an N-terminus with their eukaryotic homologs, anyway. Conversely, if a T3SE only N-terminally aligns to its homolog, the effector-homolog pair could be missed altogether. While it may distort the numbers of T3SEs that share an N-terminus with their homologs from a distant taxon, some should still be found if it was a common occurrence.

4. Materials and Methods

4.1. Data and tool settings used in several analyses

The confirmed T3SEs used in any of the analyses were taken from SecretEPDB (An et al., 2017). Only T3SEs from *E. coli*, *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia* were included. Since some effectors in SecretEPDB are identical to each other, duplicates within a genus were removed.

Proteome data was retrieved from RefSeq (O'Leary et al., 2016). Only the species of the genera *Escherichia*, *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia* at least one SecretEPDB effector belonged to were included in the analyses. The species that were represented in RefSeq proteomes and had effectors in SecretEPDB were *E. coli*, *Salmonella enterica*, *Salmonella bongori*, *Yersinia pestis*, *Yersinia pseudotuberculosis*, *Yersinia enterocolitica*, *Pseudomonas aeruginosa*, *Pseudomonas syringae*, *Pseudomonas fluorescens*, *Pseudomonas amygdali*, *Pseudomonas savastanoi*, *Pseudomonas avellanae*, *Xanthomonas oryzae*, *Xanthomonas campestris*, *Xanthomonas arboricola*, *Xanthomonas axonopodis*, *Xanthomonas vesicatoria*, *Xanthomonas euvesicatoria*, *Chlamydia trachomatis* and *Chlamydia pneumoniae*. All complete proteomes of these species that were in RefSeq as of November 2019 were used. Unless otherwise specified, these are the proteomes and confirmed effectors used in the conducted analyses.

If BlastP was used, the cut-off e-value was always set to 10^{-10} (Altschul et al., 1990). If the sequence comparison was done between SecretEPDB effectors or specific Refseq proteomes and no bigger database was used, the 2.8.1 or the 2.10.0 version of the NCBI Blast Plus command line tool was used. If searches were done against part of the RefSeq database, the online version was used at its current state as of late 2019, early 2020.

50% and 90% identity clustering were done with CD-HIT version 4.8.1 with word lengths of 5 for 90% identity clustering and word lengths of 3 for 50% identity clustering (Fu et al., 2012; Li and Godzik, 2006)

The current online version of EffectiveT3, as of late 2019, early 2020, was used for effector prediction with EffectiveT3 models 2.0.1 (Arnold et al., 2009; Eichinger et al., 2016).

R version 3.6.0 was used for statistical analysis (R Core Team, 2019).

4.2. Methods of Results 2.1.

4.2.1. Comparing T3SEs to orthologs that are not effectors

E. coli SecretEPDB effectors were searched against *E. coli* RefSeq proteomes with BlastP. PhenDB was used to predict which *E. coli* RefSeq proteomes contain a functional T3SS (Feldbauer et al., 2015). The balanced accuracy cut-off was set to 0.8. The results were used to look for proteomes that do not contain a functional T3SS but do contain proteins related to confirmed effectors. Since hardly any were found, the intended analysis could not be carried out.

4.3. Methods of Results 2.2.

4.3.1. Effector homologs in distant taxa

BlastP was used to search *E. coli*, *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia* T3SEs against the eukaryotic part of the RefSeq database, the gram-positive part of the RefSeq database and the prokaryotic part of the RefSeq database, excluding the genus the effector comes from. The numbers and proportions of effectors that have homologs among

eukaryotes, among gram-positive bacteria, outside their own genus and among eukaryotes but not gram-positive bacteria were determined for each genus.

100 random proteins were selected from one proteome of each of the used genera and searched against the eukaryotic and against the gram-positive parts of the RefSeq database. The numbers and proportions of effectors that have homologs among eukaryotes, among gram-positive bacteria and among eukaryotes but not gram-positive bacteria were determined for each genus. The proportion of proteins with eukaryotic but no gram-positive homologs was compared between effectors and randomly selected proteins. Whether the proportion of proteins with eukaryotic but no gram-positive homologs is significantly different between effectors and random proteins at a 5% level, was determined with Chi-Square test. All genera were combined, and proteins were grouped into those with eukaryotic but no gram-positive homologs and all others.

90% and 50% identity clustering was done for the T3SEs with CD-HIT. The analysis was repeated as described above, only using T3SEs that were selected as representative sequences by CD-HIT. The comparison with random proteins and significance test were done as before. No identity clustering was done for random proteins.

Legionella type IV secretion effectors and type II secretion effectors were taken from SecretEPDB and, after removal of duplicates, searched against the eukaryotic and against the gram-positive part of RefSeq. The proportions of effectors with eukaryotic, with gram-positive and with eukaryotic but no gram-positive homologs were assessed and compared to the results of *E. coli*, *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia* T3SEs.

The results of the BlastP searches of T3SEs against the different parts of the RefSeq database were inspected to find patterns among the taxonomic compositions of the homologs or the functions of proteins. How often each taxon showed up as the best hit for the effectors of each genus was assessed for the results of all three BlastP searches. The smallest higher order taxa to which both the organism the effector comes from and the organism the best hit was found in belong, according to NCBI taxonomy, were determined for the BlastP search against the prokaryotic part of the RefSeq database excluding the genus the effector comes from (Federhen, 2012). In *E. coli*, the number of T3SEs with homologs in *Shigella boydii* or *Citrobacter rhodentium* and of those that do not have homologs outside *Shigella boydii* or *Citrobacter rhodentium* were determined and the query coverage and sequence identities of the best hits investigated.

Effectors without homologs outside the genus they were identified in were searched against their own genus in RefSeq. In *Chlamydia*, a lot of them turned out to be inclusion membrane proteins, and it was checked whether they were related to each other and if any *Chlamydial* effectors with homologs outside of *Chlamyia* were also labeled as inclusion membrane proteins.

4.3.2. T3SEs that may be acquired from eukaryotes via HGT

The results of the BlastP searches of effectors against the eukaryotic, the gram-positive and the prokaryotic part of RefSeq, excluding the genus the effector belongs to, were inspected to find evidence for events of HGT from eukaryotes. In particular, effectors with eukaryotic but no gram-positive homologs and effectors with eukaryotic and only very few gram-positive homologs were looked at. The taxonomic distribution of the gram-negative homologs of the effectors was also taken into account, paying particular attention to those confined to a small taxonomic range or to an otherwise noteworthy range of taxa. For example, effectors that only had homologs in plant pathogens and eukaryotes or those that had few prokaryotic hits that were better than the best eukaryotic one were of interest.

Some eukaryotic homologs were searched against RefSeq or the eukaryotic part of RefSeq to determine how many eukaryotic homologs they had. If an effector had few eukaryotic homologs, this was done to assess if the homolog may not be truly a eukaryotic protein but a misassigned bacterial one or if it may be a protein that was transferred from prokaryotes to eukaryotes rather than the other way around. These eukaryotic homologs were searched against the prokaryotic part of RefSeq as well to see if they were almost identical to a prokaryotic protein. If an effector seemed to have been gained or transferred to a eukaryote by HGT, the search of the eukaryotic homolog against parts of Refseq was done to determine if alignments between eukaryotic homologs tended to get better scores than alignments between the effector and the best eukaryotic hit and if the eukaryotic proteins had more distant eukaryotic homologs that did not get a high enough score when aligned to the effector.

All proteins in the C*hlamydial* proteome GCF_000007205.1_ASM720v1_proteins.faa were searched against the eukaryotic part of the RefSeq database. Proteins that had homologs in *Trypanosoma*, *Leishmania* or *Leptomonas* among their best hits were searched against the prokaryotic and against the gram-positive part of RefSeq. The eukaryotic best hits from the *Trypanosomatidae* were searched against the eukaryotic and the prokaryotic parts of RefSeq as well. *Chlamydial* proteins that are most closely related to *Trypanosomatidae* proteins among the eukaryotes and not ubiquitous among large prokaryotic taxonomic groups or more closely related to *Trypanosomatidae* proteins than to most of their prokaryotic homologs were looked for.

All T3SEs that had both eukaryotic and gram-positive homologs in the previously done alignments were identified. The alignment of the T3SE against its best eukaryotic hit, that does not contain 'partial' in its name, and the alignment of the T3SE against its best gram-positive hit, that does not contain 'partial' in its name, were selected. The start positions in the effector sequence were determined for each alignment, and the average difference between the start position of the T3SE in the alignment with its eukaryotic homolog and the alignment with its gram-positive homolog was calculated. The same was done for the stop positions in the T3SEs.

Further, it was checked whether the alignments of each T3SE with the best eukaryotic and the best gram-positive hit, that do not contain 'partial' in their names, overlap. That is, whether one alignment starts between the start and stop or at the start position of the other. The one *E. coli* protein that was earlier identified to likely not be a real effector was excluded.

4.3.3. Overlap with eukaryotic and gram-positive proteins

The average length of the unaligned N-terminal and C-terminal ends of the T3SEs and their best hits among eukaryotic proteins, excluding those that contain 'partial' in their short descriptions, were calculated.

The number of T3SEs and the number of their best eukaryotic hits that are not identified as partial proteins for which the alignment starts within the first 15 aa and the number of T3SEs and the number of their best eukaryotic hits that are not identified as partial proteins for which the alignment ends within the last 15 aa were counted. How many times more common proteins for which the alignment ends within the last 15 aa were than proteins for which the alignment starts within the first 15 aa was calculated for the effectors and their alignment partners. Chi-Square test was used to determine whether the number of effectors for which the alignment starts within the first 15 aa and the number of effectors for which it ends within the last 15 aa is different on a 5% significance level. The same was done for the alignment partners.

90% and 50% identity clustering of the T3SEs was done with CD-HIT and the analysis

was repeated as before, only using the proteins selected as representative sequences by CD-HIT. The same analysis was also done for T3SEs and their best gram-positive hits that do not contain 'partial' in their description. It, too, was repeated after 50% and 90% identity clustering, only using effectors selected as representative sequences by CD-HIT and their alignment partners.

For the T3SEs for which the alignment to the best eukaryotic or gram-positive hit started within the first 15 aa, the number of eukaryotic or gram-positive alignment partners was checked. If that number was very low, the eukaryotic or gram-positive homologs were searched against RefSeq and against the eukaryotic or gram-positive part of RefSeq, respectively. It was determined how many homologs they had among the eukaryotic or gram-positive part of RefSeq as well as if there were any almost identical proteins in organisms more closely related to the genus the respective T3SE came from. This was done to see if the eukaryotic or gram-positive homolog might have been misassigned to its taxon or if HGT seems to have more likely transferred the protein from a bacterium with a T3SS to the eukaryote or gram-positive organism.

4.4. Methods of Results 2.3.

4.4.1. Comparing T3SEs to homologs in the same proteome

E. coli, *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia* SecretEPDB effectors were searched against the RefSeq proteomes of their respective genus with BlastP. Proteins in the proteomes were considered to be the effector if both proteins were fully aligned to each other and they shared at least 98% sequence identity. For each SecretEPDB effector that was found in the proteomes, all homologs that occurred in at least one of the same proteomes as the effector were collected.

Effector-homolog pairs that resulted in multiple alignments were excluded. The remaining pairwise alignments were sorted into the categories shown in figure 7. These are all combinations of how termini can be aligned to each other, aligned to other parts of their partner or stay unaligned. A terminus was considered to be aligned if no more than 3 aa stayed unaligned at the respective end of the protein. These categories were combined into 4 bigger categories – one for N-termini aligned to each other and C-termini aligned to each other, one for N-termini aligned to each other but C-termini not aligned to each other, one for C-termini aligned to each other but N-termini not aligned to each other and one for both termini not aligned to each other. The percentages of protein pairs that belong to each category were calculated for all genera separately.

The effector-homolog pairs were sorted into pairs where the homolog is a confirmed T3SE itself, pairs where the homolog is a predicted but not a confirmed effector, pairs where the homolog is a predicted non-effector and not a confirmed effector and pairs where the homolog belongs to neither of the other categories. EffectiveT3 was used to predict which proteins were T3SEs. The chosen cut-off value for predicted effectors was 0.9999 and for predicted non-effectors 0.0001. This sets the false positive rate for predicted effectors to about 7% and for predicted non-effectors to about 13%. The percentage of effector-homolog pairs that go into each category, based on which termini are aligned to each other or other parts of their partner, was determined for each of these groups.

4.4.2. Comparing T3SEs to other confirmed effectors

E. coli, *Salmonella*, *Yersinia*, *Pseudomonas*, *Xanthomonas* and *Chlamydia* SecretEPDB T3SEs were aligned to confirmed effectors of their own genus, using BlastP.

Effector pairs that resulted in multiple alignments were excluded. The remaining

alignments were sorted into the categories shown in figure 8. These are all combinations of how termini can be aligned to each other, aligned to other parts of their partner or stay unaligned. A terminus was considered to be aligned if no more than 3 aa stayed unaligned at the respective end of the protein. These categories were combined into 4 bigger categories – one for N-termini aligned to each other and C-termini aligned to each other, one for N-termini aligned to each other but C-termini not aligned to each other, one for C-termini aligned to each other but N-termini not aligned to each other and one for both termini not aligned to each other. The percentages of protein pairs that belong to each category were calculated for all genera separately.

The analysis was repeated, only including SecretEPDB effectors that could be found in the RefSeq proteomes. Proteins in the proteomes were considered to be confirmed effectors if both proteins were fully aligned to each other and shared at least 98% sequence identity.

Among the excluded alignments that consisted of one N-terminally longer and one shorter partner and shared the same C-terminus the numbers of pairs where both partners were not found in the proteomes, where the shorter partner was not found in the proteomes and where the longer partner was not found in the proteomes were counted.

4.5. Methods of Results 2.4.

4.5.1. T3SE prediction

10 000 sequences of 25 aa length were generated. They were started with methionine, and 24 random aa were added, taking the different numbers of codons for different aa into account and assuming a GC content of 50%.

One RefSeq proteome of *Salmonella*, *Yersinia*, *Xanthomonas* and *Chlamydia* and 10 RefSeq proteomes of *E. coli* and *Pseudomonas* were selected at random. Proteins shorter than 50 aa were excluded. Of the remaining proteins, truncated versions were made by removing the 2nd to the 26th aa. 888 to 6782 proteins and truncated proteins remained per proteome. Frameshifted proteins were generated from the '_cds_from_genomic.fna' files corresponding to the same genomes the proteome files for the whole and truncated proteins were taken from. +1 frameshifts were made by adding a random nucleotide after the 1st codon. -1 frameshifts were made by deleting the 4th nucleotide. The nucleotide sequences were translated, skipping stop codons. 906 to 7081 +1 and -1 frameshifts were gathered per proteome.

One RefSeq genome of each used genus was selected, and plasmids were removed. The opposite strands of the chromosomes were generated. From each strand, 5 000 sequences were selected by randomly choosing a position, except for the last 149 nucleotides (nt), and extracting a sequence of length 150 that starts at that position. These sequences were translated, skipping stop codons. From the same chromosomes, sequences that were not listed as genes in the associated '_feature_tables.txt' files were extracted. These sequences should be likely to be intergenic regions. The opposite strands of them were generated. If a sequence was at least 75 nt long, a random position between its 1st and 75th last nt was chosen. The sequence was extracted from the random position to its end, and 'AUG' was added to its front. This was done for the initial sequences and the opposite strand sequences. They were translated to proteins, skipping stop codons. 978 to 6630 sequences were extracted per genome.

EffectiveT3 was used to predict which sequences could be secreted with a threshold of 0.9999 for positive prediction. The percentages of sequences predicted to be secreted were calculated. For *E. coli* and *Pseudomonas*, the average of all used proteomes was taken for the truncated and full length proteins and the frameshifts. The percentages of positive predictions within these two taxa for individual proteomes were used to get some idea of the variation within one genus.

Combinations of deletions and -1 frameshifts were created from the previously used *Salmonella* cds file and one of the previously used *E .coli* and *Pseudomonas* cds files. Sequences shorter than 226 nt were excluded. The 4th to 79th nt were removed. The resulting sequences were translated, skipping stop codons, and the number of secreted sequences was predicted by EffectiveT3.

5. References

Altschul, SF., Gish, W., Miller, W., Myers, EW. & Lipman, DJ. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-410. doi: 10.1016/S0022-2836(05)80360-2

An, Y., Wang, J., Li, C., Revote, J., Zhang, Y., Naderer, T., Hayashida, M., Akutsu, T., Webb, GI., Lithgow, T., Song, J. (2017) SecretEPDB: a comprehensive webbased resource for secreted effector proteins of the bacterial types III, IV and VI secretion systems. *Sci Rep.* 23;7:41031. doi: 10.1038/srep41031

Anderson, DM., Fouts, DE., Collmer, A., Schneewind, O. (1999) Reciprocal secretion of proteins by the bacterial type III machines of plant and animal pathogens suggests universal recognition of mRNA targeting signals. *Proc Natl Acad Sci U S A* 96(22):12839-43. doi: 10.1073/pnas.96.22.12839

Anderson, DM., Schneewind, O. (1997) A mRNA Signal for the Type III Secretion of Yop Proteins by Yersinia enterocolitica. Science. 278(5340):1140-3. doi: 10.1126/science.278.5340.1140

Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, HW., Horn, M., Rattei, T. (2009) Sequence-Based Prediction of Type III Secreted Proteins. *PLoS Pathog.* 5(4):e1000376. doi: 10.1371/journal.ppat.1000376.

Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, HW., Horn, M., Rattei, T. (2009) Sequence-Based Prediction of Type III Secreted Proteins. *PLoS Pathog.* 5(4):e1000376. doi: 10.1371/journal.ppat.1000376.

Arnold, R., Jehl, A., Rattei, T. (2010) Targeting effectors: the molecular recognition of Type III secreted proteins. *Microbes Infect.* 12(5):346-58. doi: 10.1016/j.micinf.2010.02.003

Binz, T., Sikorra, S., Mahrhold, S. (2010) Clostridial Neurotoxins: Mechanism of SNARE Cleavage and Outlook on Potential Substrate Specificity Reengineering. *Toxins (Basel).* 2(4): 665–682. doi: 10.3390/toxins2040665

Cherubin, P., Garcia, MC., Curtis, D., Britt, CBT., Craft Jr, JW., Burress, H., Berndt, C., Reddy, S., Guyette, J., Zheng, T., Huo, Q., Quiñones, B., Briggs, JM., Teter, K. (2016) Inhibition of Cholera Toxin and Other AB Toxins by Polyphenolic Compounds. *PLoS One.* 11(11):e0166477. doi: 10.1371/journal.pone.0166477

Coburn, B., Sekirov, I., Finlay, BB. (2007) Type III Secretion Systems and Disease. *Clin Microbiol Rev.* 20(4):535-49. doi: 10.1128/CMR.00013-07

Dong, X., Lu, X., Zhang, Z. (2015) BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. *Database (Oxford).* 2015:bav064. doi: 10.1093/database/bav064

Dong, X., Zhang, YJ., Zhang, Z. (2013) Using Weakly Conserved Motifs Hidden in Secretion Signals to Identify Type-III Effectors from Bacterial Pathogen Genomes. *PLoS One.* 8(2):e56632. doi: 10.1371/journal.pone.0056632

Dufour, N., Delattre, R., Ricard, JD., Debarbieux, L. (2017) The Lysis of Pathogenic Escherichia coli by Bacteriophages Releases Less Endotoxin Than by β-Lactams. *Clin Infect Dis.* 64(11):1582-1588. doi: 10.1093/cid/cix184

Duplouy, A., Iturbe-Ormaetxe, I., Beatson, SA., Szubert, JM., Brownlie, JC., McMeniman, CJ., McGraw, EA., Hurst, GD., Charlat, S., O'Neill, SL., Woolfit, M. (2013) Draft genome sequence of the male-killing Wolbachia strain wBol1 reveals recent horizontal gene transfers from diverse sources. *BMC Genomics.* 14:20. doi: 10.1186/1471-2164-14-20.

Eichinger, V., Nussbaumer, T., Platzer, A., Jehl, MA., Arnold, R., Rattei, T. (2016) EffectiveDB—updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res.* 44(D1):D669-74. doi: 10.1093/nar/gkv1269

Fauvart, M., Michiels, J. (2008) Rhizobial secreted proteins as determinants of host specicity in the rhizobium-legume symbiosis. *FEMS Microbiol Lett.* 285(1):1-9. doi: 10.1111/j.1574-6968.2008.01254.x

Federhen, S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.* 40(Database issue):D136-43. doi: 10.1093/nar/gkr1178

Feldbauer, R., Schulz, F., Horn, M., Rattei, T. (2015) Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics.* 16 Suppl 14:S1. doi: 10.1186/1471-2105-16-S14-S1

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 28(23):3150-2. doi: 10.1093/bioinformatics/bts565.

Green, ER., Mecsas, J. (2016) Bacterial Secretion Systems: An Overview. *Microbiol Spectr.* 4(1):10.1128. doi: 10.1128/microbiolspec.VMBF-0012-2015.

Gomez-Valero, L., Rusniok, C., Cazalet, C., Buchrieser, C. (2011) Comparative and functional genomics of Legionella identified eukaryotic like proteins as key players in host–pathogen interactions. *Front Microbiol.* 2:208. doi: 10.3389/fmicb.2011.00208

John, CM., Phillips, NJ., Stein, DC., Jarvis, GA. (2017) Innate immune response to lipooligosaccharide: pivotal regulator of the pathobiology of invasive Neisseria meningitidis infections. *Pathog Dis.* 75(3). doi: 10.1093/femspd/ftx030

Karlinsey, JE., Lonner, J., Brown, KL., Hughes KT. (2000) Translation/Secretion Coupling by Type III Secretion Systems. *Cell.* 102(4):487-97. doi: 10.1016/s0092-8674(00)00053-2

Kulp, A., Kuehn, MJ. (2010) Biological Functions and Biogenesis of Secreted Bacterial Outer Membrane Vesicles. *Annu Rev Microbiol.* 64: 163–184.

doi:10.1146/annurev.micro.091208.073413

Lee, PA., Tullman-Ercek, D., Georgiou, G. (2006) The Bacterial Twin-Arginine Translocation Pathway. *Annu Rev Microbiol.* 60:373-95. doi: 10.1146/annurev.micro.60.080805.142212.

Lee, SH., Galán, JE. (2004) Salmonella type III secretion-associated chaperones confer secretion-pathway specificity. *Mol Microbiol.* 51(2):483-95. doi: 10.1046/j.1365-2958.2003.03840.x

Li, J., Li, Z., Luo, J., Yao, Y., (2020) ACNNT3: Attention-CNN Framework for Prediction of SequenceBased Bacterial Type III Secreted Effectors. *Comput Math Methods Med.* 2020:3974598. doi: 10.1155/2020/3974598

Li, W., Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 22(13):1658-9. doi: 10.1093/bioinformatics/btl158

Lloyd, SA., Norman, M., Rosqvist, R., Wolf-Watz, H. (2001) Yersinia YopE is targeted for type III secretion by N-terminal, not mRNA, signals. *Mol Microbiol.* 39(2):520-31. doi: 10.1046/j.1365-2958.2001.02271.x

Lloyd, SA., Sjöström, M., Andersson, S., Wolf-Watz, H. (2002) Molecular characterization of type III secretion signals via analysis of synthetic Nterminal amino acid sequences. *Mol Microbiol.* 43(1):51-9. doi: 10.1046/j.1365-2958.2002.02738.x

Ma W., Guttman DS. (2008) Evolution of prokaryotic and eukaryotic virulence effectors. *Curr Opin Plant Biol.* 11(4):412-9. doi: 10.1016/j.pbi.2008.05.001

McDermott, JE., Corrigan, A., Peterson, E., Oehmen, C., Niemann, G., Cambronne, ED., Sharp, D., Adkins, JN., Samudrala, R., Heffron, F. (2011) Computational Prediction of Type III and IV Secreted Effectors in Gram-Negative Bacteria. *Infect Immun.* 79(1):23-32. doi: 10.1128/IAI.00537-10

Mital, J., Miller, NJ., Dorward, DW., Dooley, CA., Hackstadt, T. (2013) Role for Chlamydial Inclusion Membrane Proteins in Inclusion Membrane Structure and Biogenesis. *PLoS One.* 8(5):e63426. doi: 10.1371/journal.pone.0063426

Moran, AP., Prendergast, MM., Appelmelk, BJ. (1996) Molecular mimicry of host structures by bacterial lipopolysaccharides and its contribution to disease. *FEMS Immunol Med Microbiol.* 16(2):105-15. doi: 10.1111/j.1574-695X.1996.tb00127.x

Nans, A., Ford, C., Hayward, RD., (2015) Host-pathogen reorganisation during host cell entry by Chlamydia trachomatis. *Microbes Infect.* 17(11-12):727-31. doi: 10.1016/j.micinf.2015.08.004

Navarro-Garcia, F., Ruiz-Perez, F., Cataldi, A., Larzábal, M. (2019) Type VI Secretion System in Pathogenic Escherichia coli: Structure, Role in Virulence, and Acquisition. *Front Microbiol.* 10:1965. doi: 10.3389/fmicb.2019.01965

O'Leary, NA., Wright, MW., Brister, JR., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B.,

Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, CM., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, VS., Kodali, VK., Li, W., Maglott, D., Masterson, P., McGarvey, KM., Murphy, MR., O'Neill, K., Pujar, S., Rangwala, SH., Rausch, D., Riddick, LD., Schoch, C., Shkeda, A., Storz, SS., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, RE., Vatsan, AR., Wallin, C., Webb, D., Wu, W., Landrum, MJ., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, TD., Pruitt, KD. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44(D1):D733-45. doi: 10.1093/nar/gkv1189

Ramamurthi, KS., Schneewind, O. (2003) Yersinia yopQ mRNA encodes a bipartite type III secretion signal in the first 15 codons. *Mol Microbiol.* 50(4):1189-98. doi: 10.1046/j.1365-2958.2003.03772.x

R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rüssmann, H., Kubori, T., Sauer, J., Galán, JE. (2002) Molecular and functional analysis of the type III secretion signal of the Salmonella enterica InvJ protein. *Mol Microbiol.* 46(3):769-79. doi: 10.1046/j.1365-2958.2002.03196.x

Sampath, VP. (2018) Bacterial endotoxin-lipopolysaccharide; structure, function and its role in immunity in vertebrates and invertebrates. *Agriculture and Natural Resources.* 52 115:120.

Samudrala, R., Heffron, F., McDermott, JE. (2009) Accurate Prediction of Secreted Substrates and Identification of a Conserved Putative Secretion Signal for Type III Secretion Systems. *PLoS Pathog.* 5(4):e1000375. doi: 10.1371/journal.ppat.1000375

Sastalla, I., Monack, DM., Kubatzky, KF. (2016) Editorial: Bacterial Exotoxins: How Bacteria Fight the Immune System. *Front Immunol.* 7:300. doi: 10.3389/fimmu.2016.00300

Sharma, AK., Dhasmana, N., Dubey, N., Kumar, N., Gangwal, A., Gupta, M., Singh, Y. (2017) Bacterial Virulence Factors: Secreted for Survival. *Indian J Microbiol.* 57(1):1-10. doi: 10.1007/s12088-016-0625-1

Stavrinides, J., Ma, W., Guttman, DS. (2006) Terminal reassortment drives the quantum evolution of type III effectors in bacterial pathogens. *PLoS Pathog.* 2(10):e104. doi: 10.1371/journal.ppat.0020104

Stromberg, ZR., Van Goor, A., Redweik, GAJ., Wymore Brand, MJ., Wannemuehler, MJ., Mellata1, M. (2018) Pathogenic and non-pathogenic Escherichia coli colonization and host inflammatory response in a defined microbiota mouse model. *Dis Model Mech.* 11(11):dmm035063. doi: 10.1242/dmm.035063

Sturm, A., Heinemann M., Arnoldini M., Benecke, A., Ackermann, M., Benz, M., Dormann, J., Hardt, WD. (2011) The cost of virulence: retarded growth of Salmonella Typhimurium cells expressing type III secretion system 1. *PLoS Pathog.* 7(7):e1002143. doi: 10.1371/journal.ppat.1002143

Subtil, A., Delevoye, C., Balañá, ME., Tastevin, L., Perrinet, S., Dautry-Varsat, A. (2005) A

directed screen for chlamydial proteins secreted by a type III mechanism identifies a translocated protein and numerous other new candidates. *Mol Microbiol.* 56(6):1636-47. doi: 10.1111/j.1365-2958.2005.04647.x

Tsirigotaki, A., De Geyter, J., Šoštaric, N., Spyridoula, AE. (2017) Protein Export Through the Bacterial Sec Pathway. *Nat Rev Microbiol.* 15(1):21-36. doi: 10.1038/nrmicro.2016.161. Epub 2016 Nov 28.

van der Woude, MW., Bäumler, AJ. (2004) Phase and Antigenic Variation in Bacteria. *Clin Microbiol Rev.* 17(3):581-611 doi: 10.1128/CMR.17.3.581-611.2004

Wagner, S., Grin, I., Malmsheimer, S., Singh, N., Torres-Vargas, C., Westerhausen, S. (2018) Bacterial type III secretion systems: a complex device for the delivery of bacterial effector proteins into eukaryotic host cells. *FEMS Microbiol Lett.* 365(19). doi: 10.1093/femsle/fny201

Wang, Y., Sun, M., Bao, H., Zhang, Q., Guo, D. (2013) Effective Identification of Bacterial Type III Secretion Signals Using Joint Element Features. *PLoS One.* 8(4):e59754. doi: 10.1371/journal.pone.0059754. Print 2013

Wang, Y., Zhang, Q., Sun, MA., Guo, D. (2011) High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics.* 27(6):777-84. doi: 10.1093/bioinformatics/btr021

Wang, Y., Sun, M., Bao, H., White, AP. (2013) T3_MM: A Markov Model Effectively Classifies Bacterial Type III Secretion Signals. *PLoS One.* 8(3):e58173. doi: 10.1371/journal.pone.0058173

Webband, SAR., Kahler, CM. (2008) Bench-to-bedside review: Bacterial virulence and subversion of host defences. *Crit Care.* 12(6): 234. doi: 10.1186/cc7091
Type III Secretion Systems and Disease.

Wen, Z., Zhang, JR. (2015) "Chapter 3 - bacterial capsules," in Molecular Medical Microbiology 2nd Edn, eds YW. Tang, M. Sussman, D. Liu, I. Poxton, and J. Schwartzman, (Boston: Academic Press), 33–53. doi: 10.1016/b978-0-12- 397169-2.00003-2

Whitaker, JW., McConkey, GA., Westhead, DR. (2009)The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. *Genome Biol.* doi: 10.1186/gb-2009-10-4-r36