

# **A Bioinformatics Analysis of Bacterial Type-III Secretion System Genes and Proteins**

By  
**Christopher Michael Bailey**

A thesis submitted to  
**The University of Birmingham**  
for the degree of  
**DOCTOR OF PHILOSOPHY**

**Division of Immunity and Infection  
The School of Medicine  
The University of Birmingham  
November 2010**

UNIVERSITY OF  
BIRMINGHAM

**University of Birmingham Research Archive**

**e-theses repository**

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

## Abstract

Type-III secretion systems (T3SSs) are responsible for the biosynthesis of flagella, and the interaction of many animal and plant pathogens with eukaryotic cells. T3SSs consist of multiple proteins which assemble to form an apparatus capable of exporting proteins through both membranes of Gram-negative bacteria in one step. Proteins conserved amongst T3SSs can be used for analysis of these systems using computational homology searching. By using tools including BLAST and HMMER in conjunction with phylogenetic analysis this thesis examines the range of T3SSs, both in terms of the proteins they contain, and also the bacteria which contain them. *In silico* analysis of several of the conserved components of T3SSs shows similarities between them and other secretion systems, as well as components of ATPases. Use of conserved components allows for identification of T3SS loci in diverse bacteria, in order to assess the different proteins used by different T3SSs, and to see where, in evolutionary space, these differences arose. Analysis of homology data also allows for comprehensive re-annotation of T3SS loci within *Desulfovibrio*, *Lawsonia* and *Hahella*, and subsequent comparison of these T3SSs with related Yersinia T3SSs, and also (in conjunction with *in vitro* assays) for identification of many novel effectors in *E. coli*.

## **Authors Contributions**

All the work presented here is the authors own with the exception of portions of Chapter 2 and Chapter 3. Within Chapter 2 initial survey work leading to the work on the relationship between b-subunit proteins of ATPases and SctL/FliH components was undertaken by Professor Mark Pallen, before being taken on by the Author. Within Chapter 3, the work presented was part of a collaborative effort between three separate lab groups: The Pallen group at the University of Birmingham, The Frankel group at Imperial College, London and the Hayashi group at Osaka University, Japan. Initial bioinformatics survey work was performed by Scott Beatson, all secretome, Flag and CyeA assays were performed by Toru Tobe, Hisaaki Taniguchi and Hiroyuki Abe at Osaka University.  $\beta$ -lactamase assays were prepared by the Author, Scott Beatson, Amanda Fivian, Rasha Younis and Sophie Matthews. The  $\beta$ -lactamase assay itself was performed under the supervision of Olivier Marches. Out of the complete set of all  $\beta$ -lactamase assays the author worked with half of the effector candidates from the stage of production of amplified DNA for effector candidates to production of source and destination gateway plasmids. Of this 50%, the author also and undertook the assay from start to finish for half (i.e. 25% of the all the effector candidates).

## **Dedication**

To my parents Dorrien and Michael

## **Acknowledgements**

The division of immunity and infection for funding my work. Prof Mark Pallen for his supervision, ideas, and support. Prof Charles Penn and Dr. Roy Chaudhuri for their stimulating insight. Dr Lori Snyder, Dr Scott Beatson and Dr Nick Loman for proof-reading, and last but not least, Dr. Dov Stekel for his mathematical knowledge, support and advice.

# Table of Contents

|   |    |
|---|----|
| Chapter 1 - Introduction.....   | 1  |
| 1.1. Molecular Evolution .....  | 1  |
| 1.1.1. Gene Gain, Gene Loss and Gene Mutation .....                           | 2  |
| 1.1.2. Vertical Gene Transfer.....  | 3  |
| 1.1.3. Horizontal Gene Transfer.....  | 4  |
| 1.1.3.1. Plasmids .....   | 6  |
| 1.1.3.2. Bacteriophage.....   | 6  |
| 1.1.4. Homology, Orthology and Parology.....                                  | 8  |
| 1.1.5. Detection of evolutionarily related sequences in silico .....          | 9  |
| 1.1.5.1. Simple sequence alignment algorithms .....                           | 10 |
| 1.1.5.2. Scoring alignments.....  | 14 |
| 1.1.5.3. Heuristic methods for searching large datasets: BLAST .....          | 16 |
| 1.1.5.4. Detecting distant homology .....                                     | 19 |
| 1.1.5.4.1. Alignment based methods: PSI-BLAST .....                           | 20 |
| 1.1.5.4.2. Markov model based methods: HMMER .....                            | 21 |
| 1.1.6. Sequence similarity as a predictor of structure and function.....      | 24 |
| 1.1.6.1. Divergence between sequence, structure and function .....            | 25 |
| 1.1.6.2. Conserved sequence, un-conserved structure .....                     | 26 |
| 1.1.6.3. Conserved sequence and structure, un-conserved function .....        | 27 |
| 1.1.6.4. Conserved sequence, unknown function.....                            | 27 |
| 1.1.6.5. Un-conserved sequence, conserved structure .....                     | 28 |
| 1.1.6.6. Un-conserved sequence and structure, conserved function .....        | 30 |
| 1.1.6.7. Implications for assignments based on sequence similarity alone..... | 31 |
| 1.1.7. Phylogenetics and phylogenomics .....                                  | 32 |
| 1.1.7.1. Methods for classifying proteins and larger units into trees .....   | 33 |
| 1.2. Protein Secretion by Bacteria.....                                       | 34 |
| 1.2.1. Secretion across the inner membrane .....                              | 35 |
| 1.2.2. Secretion across the outer membrane.....                               | 37 |
| 1.2.2.1. Type I Secretion.....  | 37 |
| 1.2.2.2. Type II secretion .....  | 39 |
| 1.2.2.3. Type IV secretion.....   | 42 |
| 1.2.2.4. Type V secretion .....   | 43 |
| 1.2.2.5. Type III secretion.....  | 45 |
| 1.3. Type III Secretion Systems .....   | 46 |
| 1.3.1. The bacterial Flagellum and Non-Flagellar secretion systems .....      | 46 |
| 1.3.2. The construction of type-III secretion systems .....                   | 47 |
| 1.3.2.1. Cytoplasmic and Inner Membrane Proteins.....                         | 47 |
| 1.3.2.2. Periplasmic and outer membrane proteins .....                        | 51 |
| 1.3.3. Directing secretion: T3S needle and translocon .....                   | 53 |
| 1.3.3.1. The needle and associated proteins .....                             | 53 |
| 1.3.3.2. The translocon - proteins that put the tip on the needle .....       | 56 |
| 1.3.4. Control of apparatus and effectors: Regulators and chaperones.....     | 57 |
| 1.3.5. Communicating with the host: effector proteins.....                    | 60 |

|   |     |
|---|-----|
| 1.4. Viewing Type-III Secretion in an Evolutionary Context.....                                 | 61  |
| 1.5. Sequence – Structure – Function Relationships in Type-III secretion proteins....           | 62  |
| 1.6. Aims .....   | 63  |
| Chapter 2 - Specific T3ss components.....   | 66  |
| 2.1. Introduction.....  | 66  |
| 2.1.1. NF-T3SS components shared with other systems .....                                       | 66  |
| 2.1.2. The machinery of ATP synthesis and utilisation .....                                     | 67  |
| 2.1.3. Aims .....   | 69  |
| 2.2. Methods.....   | 70  |
| 2.2.1. BLAST and PSI-BLAST .....  | 70  |
| 2.2.2. Alignments .....   | 71  |
| 2.2.3. Domain Searching.....  | 71  |
| 2.2.4. Phylogenetic trees .....   | 71  |
| 2.3. Results.....   | 71  |
| 2.3.1. Diversity of N-terminal domains in Secretin proteins .....                               | 71  |
| 2.3.2. Relationship between secretin C-terminus and pilot proteins.....                         | 77  |
| 2.3.3. The relationship between type-III secretion system and ATPase components .....           | 79  |
| 2.3.4. A common protein architecture between type-III secretion and Mycobacterial ATPases ..... | 81  |
| 2.4. Discussion .....   | 82  |
| 2.4.1. The diverse origins of type-III secretion system proteins.....                           | 82  |
| 2.5. Summary .....  | 86  |
| Chapter 3 - Additional T3ss Effectors .....   | 89  |
| 3.1. Introduction.....  | 89  |
| 3.1.1. Aims .....   | 90  |
| 3.2. Methods.....   | 91  |
| 3.2.1. Bioinformatics analysis.....   | 91  |
| 3.2.2. Proteomics analysis of culture supernatant .....   | 92  |
| 3.2.3. Preparation of candidate effectors.....  | 92  |
| 3.2.3.1. Prime design PCR amplification.....  | 92  |
| 3.2.3.2. Transfer into gateway entry vector .....   | 93  |
| 3.2.3.3. Transfer into gateway destination vector .....   | 95  |
| 3.2.4. Translocation assays .....   | 95  |
| 3.2.4.1. Cya translocation assay .....  | 96  |
| 3.2.4.2. FLAG-tagged translocation assay .....  | 96  |
| 3.2.4.3. $\beta$ -lactamase translocation assay .....   | 96  |
| 3.3. Results.....   | 97  |
| 3.3.1. Bioinformatics.....  | 97  |
| 3.3.2. Gene cloning and transfer .....  | 97  |
| 3.3.3. Secretion and translocation assays .....   | 99  |
| 3.4. Discussion .....   | 107 |
| 3.4.1. Recurrent domains and motifs .....   | 107 |
| 3.4.2. Pitfalls of screening and assays.....  | 110 |
| 3.5. Summary .....  | 112 |



|   |     |
|---|-----|
| Chapter 4 - Specific T3s Systems .....  | 113 |
| 4.1. Introduction.....  | 113 |
| 4.1.1. The Study of Model NF-T3SS systems .....   | 113 |
| 4.1.2. Novel T3SSs found in diverse bacteria.....   | 114 |
| 4.1.3. Annotation and analysis of novel T3SS systems .....  | 115 |
| 4.1.4. Aims.....  | 116 |
| 4.2. Methods.....   | 117 |
| 4.2.1. T3SS Region Prediction.....  | 117 |
| 4.2.2. Gene identification.....   | 118 |
| 4.2.2.1. HMMER searches .....   | 118 |
| 4.2.2.2. BLAST Searches.....  | 118 |
| 4.2.3. Analysis and annotation.....   | 121 |
| 4.3. Results.....   | 121 |
| 4.3.1. Comparison and re-annotation of the <i>Lawsonia intracellularis</i> T3SS regions .....                         | 121 |
| 4.3.2. Comparison and re-annotation of the <i>Hahella chejuensis</i> T3SS regions .....                               | 127 |
| 4.4. Discussion.....  | 132 |
| 4.4.1. Mis-annotation of T3SS Regions.....  | 132 |
| 4.4.2. Evolution through Paralogy.....  | 134 |
| 4.4.3. Stepwise evolution and conserved regulation in T3SSs .....   | 135 |
| 4.5. Summary .....  | 138 |
| Chapter 5 Conserved and Specific features of t3S systems .....  | 139 |
| 5.1. Introduction.....  | 139 |
| 5.1.1. The breadth of type-III secretion .....  | 139 |
| 5.1.2. Diverse gene complements in T3SSs.....   | 142 |
| 5.1.3. Mapping diversity to evolution.....  | 146 |
| 5.1.4. Aims.....  | 147 |
| 5.2. Methods.....   | 148 |
| 5.2.1. Locating T3SSs in completed genomes .....  | 148 |
| 5.2.2. Defining T3SS regions within genomes .....   | 149 |
| 5.2.3. Generating networks of related proteins .....  | 149 |
| 5.2.4. Defining the conserved and specific sets of T3SS .....   | 150 |
| 5.3. Results.....   | 150 |
| 5.3.1. Finding T3SS Loci.....   | 150 |
| 5.3.2. The distribution of flagellar and non-flagellar type-III secretion systems in sequenced bacterial genomes..... | 155 |
| 5.3.3. Phylogenetic groups of T3SSs .....   | 160 |
| 5.3.4. ‘Essential’ gene families .....  | 162 |
| 5.3.5. Partially conserved gene families .....  | 166 |
| 5.3.6. ‘Absent’ proteins within the result set .....  | 170 |
| 5.4. Discussion.....  | 175 |
| 5.4.1. Complexity in T3SS loci.....   | 175 |
| 5.4.2. Mapping the change in gene complement to sequence phylogeny.....   | 176 |
| 5.4.3. Issues with automated locus and protein family finding .....   | 179 |
| 5.4.4. Non-detection of known proteins.....   | 184 |

|  |     |
|--|-----|
| 5.5. Summary .....   | 194 |
| Chapter 6 - Discussion .....   | 196 |
| 6.1. The current view of type-III secretion .....                                | 196 |
| 6.2. Discovering new T3SSs .....   | 198 |
| 6.3. The role of bioinformatics in T3SS research .....                           | 200 |
| References .....   | 204 |
| Appendix 1 – Complete list of all genomes containing a T3SS .....                | 226 |
| Appendix 2 – Locus finding approaches applied to type-VI secretion systems ..... | 233 |

## List of Illustrations

|   |     |
|---|-----|
| Figure 1. Schematic over view of the type I secretion system.....   | 38  |
| Figure 2. Schematic representation of the type II, III and IV secretion systems .....                               | 41  |
| Figure 3. Schematic overview of the type V secretion systems .....  | 44  |
| Figure 4. Phylogenetic tree of T3SS secretin proteins. ....   | 75  |
| Figure 5. Phylogenetic tree of T2SS and T3SS secretin proteins .....  | 76  |
| Figure 6. Phylogenetic tree of T3SS secretins with C-terminal information overlaid. ....                            | 78  |
| Figure 7. Comparison of T3SS and ATPase systems .....   | 84  |
| Figure 8. Experimental flow chart for determining novel effectors .....   | 98  |
| Figure 9. PCR reaction products visualised on EtBr stained agarose gels .....                                       | 100 |
| Figure 10. NotI digests and M13 PCR amplification from pENTR plasmids visualised on EtBr stained agarose gels ..... | 101 |
| Figure 11. Results of $\beta$ -lactamase assay for selected effector candidates .....                               | 103 |
| Figure 12. Effector Locations and Phage effector loci in <i>E. coli</i> O157 Sakai.....                             | 106 |
| Figure 13. Comparison of Lawsonia T3SS loci to Desulfovibrio vulgaris T3SS locus. ....                              | 124 |
| Figure 14. Phylogenetic tree of Ysc type T3S systems .....  | 131 |
| Figure 15. Phylogenetic tree of Ysc type T3S systems, with groupings and evolutionary annotations .....             | 137 |
| Figure 16. Graph showing expansion in bacterial genome data .....   | 140 |
| Figure 17. T3SS families identified from Protein Homology Networks .....  | 144 |
| Figure 18. Heat maps of T3SS genes from <i>E. coli</i> and <i>Yersinia pestis</i> .....                             | 152 |
| Figure 19. Heat Map showing comparison of LEE system using 4 different BLAST settings .....                         | 153 |
| Figure 20. Screen shots from T3SS finder web program.....   | 156 |
| Figure 21. Taxonomic tree showing distribution of non-flagellar T3SSs .....   | 157 |
| Figure 22. Taxonomic tree showing distribution of flagellar T3SSs .....   | 158 |
| Figure 23. Phylogenetic tree of non-flagellar T3SSs .....   | 161 |
| Figure 24. Homology Network showing Secretin and HrpJ/InvE proteins .....   | 164 |
| Figure 25. Phylogenetic tree of T3SSs with highly conserved networks overlaid ....                                  | 167 |
| Figure 26. Phylogenetic tree of T3SSs with well conserved networks overlaid .....                                   | 168 |
| Figure 27. Phylogenetic tree of T3SSs with partially conserved networks overlaid .                                  | 169 |
| Figure 28. Graph of SepL and YopN protein homology networks.....  | 172 |
| Figure 29. Graphical representation of AraC protein homology network .....  | 178 |
| Figure 30. Graphical representation of SctL/HrpE protein homology network .....                                     | 182 |
| Figure 31. Graph of SctF and SctI networks joined by PSI-BLAST .....  | 185 |
| Figure 32. Graph of SctF networks showing PSI-BLAST homology to <i>D. vulgaris</i> network.....                     | 186 |
| Figure 33. Number of BLAST hits by e-value. ....  | 189 |

## List of Tables

|   |     |
|---|-----|
| Table 1. Pairwise alignment algorithms which utilise dynamic programming .....        | 13  |
| Table 2. Summary of the components of non-flagellar type-III secretions systems. ..   | 48  |
| Table 3. List of domain architectures for proteins containing a secretin domain ..... | 73  |
| Table 4. T3SS effectors in the <i>E. coli</i> O157:H7 genome .....                    | 105 |
| Table 5. List of PFAM domains used to search for T3SS loci .....                      | 120 |
| Table 6. List of genes present in <i>Lawsonia</i> T3SS locus one .....                | 125 |
| Table 7. List of genes present in <i>Lawsonia</i> T3SS locus two .....                | 126 |
| Table 8. List of genes present in <i>Hahella</i> T3SS locus one .....                 | 128 |
| Table 9. List of genes present in <i>Hahella</i> T3SS locus two .....                 | 129 |
| Table 10. Pfam Domains and the number of hits found in bacterial genomes .....        | 154 |
| Table 11. Largest protein networks containing T3SS proteins .....                     | 165 |

# **CHAPTER 1 - INTRODUCTION**

## **1.1. Molecular Evolution**

Prior to 1955 the sequence of proteins and DNA were entirely unknown. It was in this year that the first protein sequence (that of insulin) was determined by Frederick Sanger and his colleagues [1]. In that same year the sequences of Pig and Sheep insulin were also determined [2], a discovery which enabled for the first time comparative analysis of related proteins. In the decade that followed there was a rapid increase in the number of proteins for which there was a known sequence. This was particularly true for proteins such as haemoglobins [3-5] and cytochromes [5, 6]. This information in turn led to the development of techniques which have subsequently become commonplace in the field of molecular evolution such as molecular phylogenetics.

It was however, another discovery by Frederick Sanger, of the highly efficient (by standards of the time at least) chain terminal method of DNA sequencing [7] that has truly revolutionised the field of molecular biology. In 30 years since the discovery of this method of sequencing, we have gone from being able to sequence small viral genomes, through sequencing of the first bacterial genome in 1995 [8], to the sequencing of the 3 gigabases of the human genome, published in 2001 [9]. As of October 2008 there are now over 4100 ongoing or published genome projects according to GOLD [10], and nearly 195 gigabases of sequence deposited in Genbank [11]. Together this information has enabled a huge amount of work to be done in the field of molecular evolution. By analysing related genes from different genomes it is

possible to infer evolutionary relationships between species for which traditional approaches would have been unable to do, and draw conclusions about the evolution of organisms which from a phenotypic point of view appear to have little or even nothing in common.

### ***1.1.1. Gene Gain, Gene Loss and Gene Mutation***

Owing to the fact that the machinery that controls replication and repair of DNA is not perfect, it should logically follow that through time, changes in a given DNA sequence will occur. This introduction of changes can take a variety of forms, such as a simple change of base or the insertion or deletion of small numbers of bases. Where these errors take the form of base changes then the likelihood is that it will cause minor changes to the resultant protein such as the change of a single amino acid, or even no change in the resultant protein. When looking at DNA at a codon level there are 9 possible single base changes per codon, and 61 different sense (amino acid encoding) codons, resulting in 549 different potential mutations. Of those 549, over three quarters result in a change of the resultant amino acid sequence, however, only 23 will have the effect of shortening the protein (through the introduction of stop codons). The effect of adding or removing nucleotides can be much more dramatic, almost always leading to a premature stop in the sequence, although the strength of the effect is changed by the natural bias in the genome towards being A+T rich, or G+C rich. The relation between base compositional bias and the introduction of stop codons is a consequence of the 3 different possible codons utilised for encoding a stop in a DNA sequence, those codons being TAA, TAG and TGA. In any stretch of DNA, where the underlying rate of base composition is 50% G+C (and hence also 50% A+T) the chance of any stretch of 3 nucleotides encoding a stop codon is about 4.5%.

Compare this with a genome which contains only 33% G+C on average, where the chance is increased to just less than 7.5%.

There are also mechanisms which result in more gross changes in the DNA of a given cell such as the introduction of new domains within genes, or even whole genes by duplication. There are also mobile genetic elements such as insertion sequences and transposons, which can move about 'freely' within the genome. The result of which can be the disruption of genes where the mobile element inserts itself within a coding region. Finally of course there is the possibility that larger portions of DNA can be deleted.

Regardless of the mechanism of change be it small or large, its effect will not be felt in future generations of the cell where it survives to replicate and create daughter cells, and even then the future of the mutation is far from certain, and depends on and advantage (or disadvantage) that the mutation confers on the cell

### ***1.1.2. Vertical Gene Transfer***

Vertical gene transfer can be thought of as the classical method by which genes can occur in two separate species: Where both strains share a common ancestor which also contained the gene, and passed it on through direct duplication of its DNA to create daughter cells. Where mutation has occurred the ultimate fate of the new DNA molecule will be dependent on a number of factors, not least of which is the effect that the mutation has on the cell. Where the effect is deleterious to the cell, such as would be the case for a mutation which inactivates a crucial enzyme, then the mutation will most likely be lost. The result of other types of mutation depends on a series of factors centrally concerned with the overall 'fitness' of a mutant allele. The strength of the effect of any change in fitness is one of the key factors which

discriminate between several of the conflicting theories of evolution. In the classical neo-Darwinian theory, natural selection is the driving force in shaping the genetic makeup of populations [12], and very few mutations are seen as having a negligible effect on fitness. Conversely the neutral mutation hypothesis states that most alleles occur by random genetic drift, and do not have a significant effect on the ability of a protein to perform its function, as such most mutations can be thought of as neutral [13, 14]. The reality sits somewhere in the middle of the two, and so the eventual effect is one where mutations may be fixed into the population as a whole to be passed on to descendants, or lost, dependant either on random genetic drift (neutral theory), or through selection of advantageous characteristics (neo-Darwinian theory).

### ***1.1.3. Horizontal Gene Transfer***

By contrast to vertical gene transfer, horizontal gene transfer (HGT) implies a transfer of DNA from one cell to another which is not its offspring. This transfer can occur in a variety of fashions. There are three main ways in which DNA can be transferred between species in a non-parental manner. The first amongst these is transformation, a method known about for some considerable time [15]. In transformation DNA is taken up from the environment by competent cells, and can be thought of as a five step process [15]:

1. Release or appearance of DNA in environment
2. Induction of a competent state in the recipient host cell(s)
3. Interaction of cells and DNA
4. Entry of DNA and processing in cell: passage through membranes etc
5. Functional integration and expression of entering DNA into cell operations

This obviously begs the question as to how and in what form DNA is present in the



environment. Presumably this DNA arises as the result of cell lysis, be that spontaneous, or by the action of a specific source (e.g. bacteriophage). Some bacteria are naturally competent for the uptake of DNA, for example *Acinetobacter* are naturally competent for most of their growth cycle [15], whilst others (e.g. *E. coli*) can be rendered competent by chemical methods [12].

The other two forms of DNA transmission which can result in horizontal gene transfer are conjugation and transduction. Both of these methods follow a similar pattern to the five steps outlined above for transformation, with the exception that neither requires the presence of naked DNA in the environment. In the case of conjugation, single stranded DNA (ssDNA) is passed between bacteria through a conjugation system [16]. This conjugation apparatus consists of a cell membrane spanning pilus in Gram-negative bacteria, which is produced from a multimeric protein complex commonly belonging to the type-IV family of secretion systems [17, 18]. In conjunction with the secretion system there is also a relaxase which is responsible for processing of DNA into its single stranded form ready for transport, and a coupling protein which brings together the relaxase + DNA and the type IV secretion system [19].

In transduction, bacteriophages act as transfer agents for host DNA [15]. As part of the production process for bacteriophage, phage DNA must be packaged. This process is not perfect and fragments of the host genome are packaged instead of the phage DNA, resulting in a functional phage which contains no phage DNA [20]. Owing to the ‘modular’ nature of bacteriophage and the amount of recombination which occurs within them it is also common to see extra genes incorporated within their DNA [21]. This ‘more DNA’ or morons as they have been described [22], can

also act as agents of horizontal DNA transfer.

### 1.1.3.1. Plasmids

Plasmids are among the most commonly observed and well-studied DNA elements transferred by conjugation [16]. Transfer of plasmids can take one of two main forms. They are either self-transmissible (i.e. encode all the machinery required to conjugatively transfer themselves to another bacteria), or mobilisable (where a non-self-transmissible plasmid is transferred by the action of a conjugative plasmid) [23]. In either case the plasmid must contain an origin of conjugal transfer (*oriT*) in order to allow the binding of the relaxase to the DNA. Most conjugative plasmids have an extremely broad host range. For example the IncQ family of mobilisable plasmids have a host range that includes most Gram-negative bacteria, and several Gram-positive bacteria such as *Streptomyces*, *Actinomyces*, *Synechococcus*, and *Mycobacterium* [24]. Plasmids come a wide variety of sizes, from just 846 bases in the case of plasmid pRKU1 from *Thermotoga petrophila* [25], to over 2 Mb in the case of plasmid pGMI1000MP from *Ralstonia solanacearum* GMI1000 [26], and hence also greatly vary in the content and amount of DNA transferred. As such they can contain anything from simple antibiotic resistance genes (for examples see [27, 28]), to larger complete systems, such as type-III secretion systems [29-31].

### 1.1.3.2. Bacteriophage

The number of bacteriophage in the environment is truly astronomical. There are an estimated  $10^{31}$  tailed phage particles on Earth [32] (cf.  $7 \times 10^{22}$  stars in the observable universe [33]). These phage initiate  $10^{25}$  infections per second [34], resulting in  $2 \times 10^{16}$  gene transfer events into bacteria every second [20]. As mentioned above, this can be due to simple transduction of host DNA only into the phage, or by the

integration of more DNA (morons) into the packaged phage DNA. The type of transduction which occurs can either be generalised, where any gene from the bacterial host can be transferred, or specialised, where only genes located near the site of prophage integration into the host genome can be transferred [23]. The bacterial gene complement can also be reduced by the action of phages, caused by disruption of bacterial genes as a result of prophage integration into the bacterial genome [20].

Unlike other methods of horizontal gene transfer thus far described, the host-range and spread by bacteriophage is somewhat limited by the specificity of the interaction between the bacteriophage and the bacterial receptor site. This might seem to limit the role that bacteriophage have to play in the horizontal transfer of DNA, however, this is not the case. In fact in the pathogen *Escherichia coli* O157:H7 strain Sakai, no less than 16% of its genome is comprised of prophage [35]. There are also a large number of examples of fitness factors such as toxins that are encoded in prophage. This includes prominent virulence determinants such as the cholera toxin of *Vibrio cholerae* [36], the shiga toxin of enterohaemorrhagic *E. coli* [37], and the diphtheria toxin of *Corynebacterium diphtheriae* [38]. There are also numerous other types of fitness factors which have been found in prophages, including lipopolysaccharide-modifying enzymes [39], type-III effector proteins [40] and detoxifying enzymes [41, 42].

These fitness factors are commonly found as morons within prophages, with each moron only containing only a small number of genes, surrounded by a transcription promoter and terminator sequence, meaning that they can be transcribed independently from the rest of the prophage, even if the prophage is repressed [22]. Whilst several of the examples of moron encoded fitness factors can function alone,

there are also numerous examples where moron encoded genes will often not be of any value by themselves, the bacteria into which phage has lysed will also have to have the requisite complement of genes in order to take advantage of the phage's extra cargo. Subsequently these genes may become key components of the bacterium's fitness factors. For example within *Salmonella typhimurium*, around a quarter of the type-III effector proteins are encoded within prophage, or prophage remnants [20]. It remains to be seen why these morons are so commonly observed within phages, although their potential advantage to the host cell may provide some degree of positive selection to those phages which do carry them. It has also been hypothesised that prophages can be key in creating diversity within closely related species [43]. One such example is the various *Salmonella enterica* serovars, which show great diversity in their prophage complement, and also some degree of correlation between prophage complement and their specific lifestyle [43].

#### **1.1.4. Homology, Orthology and Paralogy**

The term homology was first defined, in a biological sense at least, by Richard Owen in 1843, as a term to designate “the same organ from different animals under every variety of form and function” [44]. Whilst Owen introduced the term in order to describe morphological features (e.g. the similar structure of extremities such as the bat's wing and the human hand), the term homology was used right from the start of the molecular era in biology to describe genes and proteins which had evolved from a common origin [45]. In order to clarify the different ways in which protein can evolve by descent, Walter Fitch added two additional terms both of which can be thought of as subsets of the larger groups ‘homologues’ [46]. These two key terms added to the nomenclature were orthologue and paralogue. These two terms were created to

describe genes derived in from different sources. Orthologues are genes derived from a single gene in the last common ancestor of the species being compared. As enunciated by Koonin [47], this implies two separate conditions, the first of which being that there can only be one possible gene in the ancestral strain from which the gene in the child strains was derived. The second condition is that the ancestral gene is present in the last common ancestor rather than some earlier ancestor. The definition of paralogue is somewhat looser. Paralogues can be defined as genes related by duplication, regardless of age of the duplication, and whether they lie in the same genome or not. To go with these terms there are several more specific terms which give more specific definitions. For example the age of a duplication event leading to paralogy can be defined by the terms inparalogues and outparalogues to separate duplication after, or prior (respectively) to a given speciation event.

#### ***1.1.5. Detection of evolutionarily related sequences in silico***

Whilst it may be easy to define homology in a biological sense, being able to make use of this definition in a way that can be utilised in combination with the large amount of sequence data available is somewhat more problematic. Evolutionarily related genes and proteins sequences should show a degree of similarity beyond that expected of unrelated sequences. As such any computational approach to determine the presence or absence of homology should be able to determine the likelihood of the two sequences sharing a common sequence by chance, or because they also share a common ancestor. In order to accomplish this we need a method which allows us to align two sequences and then score this alignment.

The most simplistic approach to doing this would be to simply try every possible combination of aligning sequence 1 with sequence 2 and score each individual

alignment to see which one was the optimal one. However such an approach becomes rapidly unfeasible as the size of the two sequences to align becomes longer. For example, if the two sequences to be aligned are of length 100 (which would actually constitute quite a short protein), then there would be approximately  $10^{59}$  different alignments, and for two proteins of length 1000, approximately  $10^{600}$ . This is obviously not the best way to approach this problem, and some sort of shortcut is required. This is particularly the case for situations where we wish to find related proteins in a large database.

#### **1.1.5.1. Simple sequence alignment algorithms**

Fortunately, such a shortcut is available through use of method known as dynamic programming. In order for a problem to be solvable by a dynamic programming approach the problem should show the properties of overlapping subproblems, and optimal substructure [48, 49]. In the case of sequence alignment, we have overlapping subproblems: Take for example an alignment of strings  $S$  and  $T$ . For all possible alignments of  $S$  and  $T$  there will be many where characters  $S_i$  and  $T_j$  will be aligned to each other. Sequence alignment can also have optimal substructure in that we can solve regions of the alignment at a time. Take for example, our strings  $S$  and  $T$ , of lengths  $n$  and  $m$ , which we wish to globally align. Given that we have determined the score of aligning all characters in  $S$  against all characters in  $T$ , we can solve the problem simply by working backwards from  $S_n$ ,  $T_m$  utilising along the way the optimal solution to each of our subproblems. This may not at first seem obvious, but by examining a dynamic programming algorithm which is able to optimally align two sequences, this second point should become clearer.

One of the first algorithms to utilise a dynamic programming approach in order to

align sequences was developed by Needleman and Wunsch [50]. In their paper they discuss an approach which enables the global alignment of two amino acid sequences. Global alignment entails the alignment of all characters of both strings with each other, such that for our two strings  $S$  and  $T$ , characters  $S_0$  and  $T_0$  align to each other, as do characters  $S_n$  and  $T_m$ . The Needleman-Wunsch algorithm, thanks to its dynamic programming approach to solving the problem, only requires  $n^2$  calculations to be performed (and also only  $n^2$  memory), where  $n$  is the length of the sequences to be aligned.

The Needleman-Wunsch algorithm is based on three separate steps: Initialisation, matrix-fill and traceback. In order to determine the optimal alignment, we begin by creating a matrix  $F$ , indexed by  $i$  and  $j$ , one index per sequence, and a scoring function  $\sigma(A,B)$ , which returns the score of aligning two characters, or a character against a gap. The initialisation is then:

$$\begin{aligned} F(0,0) &= 0, \\ F(i,0) &= F(i-1,0) + \sigma(S_i, '-'), \\ F(0,j) &= F(0,j-1) + \sigma('-', T_j) \end{aligned}$$

And the matrix fill (working from top left to bottom right) is done using the equation:

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + \sigma(S_i, T_j) \\ F(i-1,j) + \sigma(S_i, '-') \\ F(i,j-1) + \sigma('-', T_j) \end{cases}$$

As  $F(i,j)$  is filled in we also keep a pointer in each cell back to the cell or cells from which  $F(i,j)$  was derived. Once the matrix is filled, the score of the alignment is stored in bottom right cell of the matrix  $F(m,n)$ , and we can begin the traceback. Starting at

$F(m,n)$ , we use the stored pointers to work our way back to  $F(0,0)$ , at each stage moving from cell  $(i,j)$  to one or more of the cells  $(i-1,j-1)$ ,  $(i-1,j)$  or  $(i,j-1)$ , at the same time adding the pair of characters  $S_i$  and  $T_j$  any to the current alignment if the step was to  $(i-1,j-1)$ , the character  $S_i$  and a gap if the step was to  $(i-1,j)$  or a gap and the character  $T_j$  if the step was to  $(i,j-1)$ . If more a move in more than one direction through the scoring matrix  $F$  is possible, then we follow both directions and generate another optimal alignment.

One of the great advantages of the Needleman-Wunsch algorithm is that it can be easily adapted in order to fulfil different alignment requirements. Some algorithms that use a similar dynamic programming approach to Needleman and Wunsch are listed in Table 1. By adapting the algorithm we can implement a variety of alignments such as semi-global and local, or more complex alignments such as repeat matches (looking for repetitive regions in a sequence based on a pattern), or even more complex scoring models. The model as it stands only allows for the use of a linear penalty for gaps, that is the penalty for a gap of length  $k$ , is  $g(k) = \alpha k$ , where  $\alpha$  is a constant. Whilst this model may be simple and easy to implement it is not very representative of the underlying biology of amino acid sequences. The most representative model utilises a convex scoring model:  $g(n) = \alpha \log(n)$ , however this is a computationally expensive model requiring a potential  $2n$  matrices (where  $n$  is the length of the sequence to be aligned). However, the convex model can be approximated using a affine gap scoring model:  $g(k) = \beta + \alpha k$  [51]. In this model we have a penalty for the existence of a gap ( $\beta$ ), as well as a penalty for the length of the gap ( $\alpha k$ ). This model can be calculated using just 4 matrices [51, 52]. Given that we now have a system for aligning two sequences together, and can score gaps in a biologically relevant manner, it should follow that we now need a system for scoring.



| <b>Algorithm</b>      | <b>Needleman-Wunsch [50]</b>  | <b>Overlap[52]</b>  | <b>Smith-Waterman [53]</b>   |
|-----------------------|---|---|--|
| <i>Alignment Type</i> | Global  | Semi-Global   | Local  |
| <i>Initiation</i>     | $F(0,0) = 0$<br>$F(i,0) = F(i-1,0) + \sigma(S_i, '-' )$<br>$F(0,j) = F(i-1,0) + \sigma(' -', T_j)$                                      | $F(0,0) = 0$<br>$F(i,0) = 0$<br>$F(0,j) = 0$  | As for Overlap   |
| <i>Matrix Fill</i>    | $F(i,j) = \max \begin{cases} F(i-1,j-1) + \sigma(S_i, T_j) \\ F(i-1,j) + \sigma(S_i, '-' ) \\ F(i,j-1) + \sigma(' -', T_j) \end{cases}$ | As for Needleman-Wunsch   | $F(i,j) = \max \begin{cases} 0 \\ F(i-1,j-1) + \sigma(S_i, T_j) \\ F(i-1,j) + \sigma(S_i, '-' ) \\ F(i,j-1) + \sigma(' -', T_j) \end{cases}$ |
| <i>Traceback</i>      | Start at: $F(m,n)$<br>End at: $F(0,0)$  | Start at:<br>$F(i,j) = \max \begin{cases} F(m,k) & k = 0, \dots, n \\ F(k,n) & k = 0, \dots, m \end{cases}$<br>End at: $i = 0$ or $j = 0$ | Start at: Maximum value in $F$<br>End at: $F(i,j) = 0$   |

**Table 1. Pairwise alignment algorithms which utilise dynamic programming**

the alignment of two characters (amino acid residues) together.

#### **1.1.5.2. Scoring alignments**

Once we have our alignments we need a sensible method for scoring it, so that we can determine the optimal one. For DNA, this is fairly trivial, and can be based on a simple match/mismatch scheme. For example, score +3 for a match and -1 for a mismatch, sum the scores together, and the result is the score for the alignment. The situation is less clear for proteins. Whilst for DNA we can think of the bases as being more or less equivalent (i.e. one mismatch is no different than any other mismatch), this assumption does not hold true for proteins. If in one protein, at a particular position we have a leucine, and in another protein we have an isoleucine at the equivalent position, then the difference is unlikely to cause a change in the structure of the protein. Conversely if in the second protein the leucine was replaced by an aspartic acid, then the change charge at that position may well introduce a change in the structure of the protein, and its behaviour in equivalent physiochemical conditions. In order to take this into account we need a scoring mechanism which can score mismatches based on the type of amino acid change.

There are several methods through which a scoring system can be calculated. One such mechanism is by simple analysis of the physiochemical properties of each amino acid, such as charge, side chain type, or hydrophobicity [54]. In such an analysis scoring is often based on an adapted alphabet different to the 20 letter one used to describe the primary sequence of a protein (see [55] for an example of such an alphabet in use). There are however, more empirical methods available to calculate the likelihood of any amino acid changes. In order to do this it is necessary to analyse alignments of related proteins in order to compare the observed frequency of an

amino acid residue changing to any other amino acid. The two major scoring systems in use today utilise different, but related, methods to examine these differences and construct a matrix summarising the likelihood of any amino acid mutating to any other amino acid (or not mutating at all). The PAM (Point Accepted Mutation) matrix was created based on an examination of 71 different phylogenetic trees produced from protein families [56]. By analysing the changes between each branch, and tabulating the all the changes, it becomes possible to create a mutation probability matrix. The PAM matrix, in common with other matrices, is calculated based on simple frequency analysis of each of the amino acids along with the number of amino acid mutations.

Starting with  $p_a$  being the proportion of amino acid 'a' in all the trees, and  $f_{ab}$  being the frequency of amino acid 'a' mutating to amino acid 'b' (and also vice-versa, since directionality cannot be determined) Then the total number of mutations amino acid 'a' is involved in is :

$$f_a = \sum_{b \neq a} f_{ab}$$

And the frequency of all mutations ( $f$ ) is:

$$f = \sum_a f_a$$

Then the relative mutability (the observed versus expected rate of change) of amino acid 'a' is:

$$m_a = \frac{f_a}{100 \cdot f \cdot p_a}$$

We can then calculate the mutation matrix M. Where the amino acid stays the same then the value in the matrix becomes 1 minus the relative mutability, and the value for all other elements in the matrix become:

$$M_{ab} = \frac{f_{ab}}{f_a} m_a$$

And the final scoring matrix  $S$  is calculated as the log-odds of the probability of mutation versus the probability of random occurrence:

$$S_{ab} = 10 \log_{10} \left( \frac{M_{ab}}{p_b} \right)$$

The final PAM matrix (PAM1) represents a scoring system based on there being on average 1 mutation per 100 residues. Other versions of the PAM matrix, such as PAM250 are created simply by matrix multiplication of the original PAM1 matrix. In the case of PAM250, the PAM1 matrix is multiplied by itself 250 times.

The other scoring matrix in common use today is the BLOSUM matrix family [57]. The BLOSUM matrices are calculated using the same format of equations as the PAM matrices; however, the initial frequency data for mutations were obtained by analysis of the BLOCKS alignment database, which contained much more information: Over 2000 blocks of aligned sequences from more than 500 groups of related proteins. Mutation frequencies were then calculated by looking at the different residues in each column in each block. Together the PAM & BLOSUM matrices, and several others which have been developed along similar principles, are the mainstay of biological sequence alignment and are utilised in a huge variety of bioinformatics programs.

### **1.1.5.3. Heuristic methods for searching large datasets: BLAST**

With the advent of modern sequencing methods the number of protein sequences we have available to us has grown exponentially. This creates a problem, as even a dynamic programming algorithm such as Needleman-Wunsch would require a large amount of computer time and memory in order to align a query protein sequence

against a database of all known proteins. In order to accomplish this we need a shortcut, which in this case is provided by a heuristic called the basic local alignment search tool (or BLAST) [58, 59]. BLAST was first developed in 1990, before the growth in the protein search space brought about by whole genome sequencing projects. However, it is in this post-genomic era that BLAST has really come of age as a tool for finding homology in large datasets.

Without any further shortcuts a simple Smith-Waterman approach to searching the complete Genbank database for a protein of length 300 would require over  $10^{13}$  computational operations in order to locate matches within the database. At a rate of 100 million calculations per second, it would take  $10^5$  seconds (or just under 28 hours) to complete the search. It is for this reason, amongst several others, that has led to BLAST becoming one of the best known and utilised bioinformatics applications available today.

BLAST calculates alignments between sequences in three separate stages, the final stage of which is very similar to the local alignment algorithm proposed by Smith and Waterman. It is the first two steps though which enable BLAST to produce alignments in a much shorter period of time. In the first of these stages BLAST takes the query sequence and splits it into a series of overlapping words of length  $W$  (the default for  $W$  is 3 for proteins and 11 for nucleic acids). Take for example the String  $S = \text{'MVIDGETS'}$ , then the overlapping words will be: 'MVI', 'VID', 'IDG' ... 'GET' and 'ETS'. These words are then used to calculate a set of neighbourhood words. Neighbourhood words are calculated by taking each of the overlapping words and obtaining by use of a scoring matrix, all related words which score greater than a cut-off  $T$ . If we take the example of the word 'MVI', and  $T=11$ , then the neighbourhood

words are ‘MII’, ‘MIV’, ‘MVI’, ‘MVL’ and ‘MVV’. This list of neighbourhood words is used to search a preformatted sequence database. This database contains all the raw sequences in the original database (a requirement for the final step of the BLAST algorithm), but also information on the location of each of the possible words in the database. Once the location of all matching words has been found in the database, BLAST then extends the hits using a dynamic programming approach in both directions until the score for a region drops below a cut-off at which point the extension is terminated.

As well as being able to produce alignments of similar sequences, BLAST is also able to produce a statistical evaluation of the quality of an alignment between two sequences [58, 60]. For alignment scores, the score of a random sequence is the sum of a series of random variables (the scores for aligning each character), and so should be well approximated by a normal distribution (from central limit theorem). Given this fact, then the distribution of the maximum for the same series will be approximated by an extreme value distribution (EVD) [61].

If we begin by calculating the number of unrelated match with score greater than  $S$ :

$$E(S) = K m n e^{-\lambda S}$$

where  $K$  and  $\lambda$  are constants, and  $m$  and  $n$  are the size of the sequences (i.e. the search space). The result of this equation is that a doubling of the search space will result in a doubling of the number of hits of a given score, whilst the relationship between score and number of hits is exponential. Taking the above equation, in order to calculate the probability of there being a match of score greater than  $S$  is:

$$P(x > S) = -e^{-E(S)}$$

Note that this equation follows the form of a type of EVD known as the Gumbel distribution [61] which has the general form:

$$P(x) = \exp(-e^{-(x-\mu)/\sigma})$$

By utilising the information on the score of the alignment along with the size of the database it is possible for BLAST to produce a statistical estimate as to the likelihood of two sequences being related by chance alone.

#### **1.1.5.4. Detecting distant homology**

Whilst BLAST and its relatives such as FASTA [62] perform well in returning relevant hits to large databases of sequence data, there is information in the literature pointing to the gap in sensitivity between these heuristics and full dynamic programming algorithms [63-66]. In most cases this is not a problem, as the search will still return the vast majority of hits found by a more sophisticated approach, so long as the correct initial parameters are used [63, 64]. Using one protein sequence is however not the only way to locate homologous hits within a database. Before the advent of BLAST there were several examples of researches using profiles built on multiple protein sequences for locating more distant homologues [67-70]. After its invention BLAST was also used by researchers as a profile searching tool [71], whilst others used specific profile based approaches to determine the extent of protein families [72-75]. In attempt to present a simple to use interface to these profile methods several software packages have become available which enable profile based homology searching. One is based on the original principles of BLAST, with an additional iterative element which allows for the generation of position specific scoring matrices (Position Specific Iterated or PSI-BLAST), and other based on the principle of Profile based hidden Markov models (HMMER).

#### **1.1.5.4.1. Alignment based methods: PSI-BLAST**

PSI-BLAST was first released in 1997, at the same time as the second version of the standard BLAST program [59]. PSI-BLAST works by taking the output of a BLAST run and using the output to construct a position specific scoring matrix (PSSM). A modified BLAST program then reads this PSSM and use it place of a simple query.

For the first iteration, a standard matrix (e.g. BLOSUM62) is used in order to compile a list of homologous proteins. This data set is purged of any hits identical to the query sequence, and only one copy of hits that are >98% identical are kept. The remaining hits are then used to create a multiple alignment, based solely on their alignment to the query sequence, rather than to each other as would be the case for a true multiple alignment. Each sequence is then reweighted in order to reduce the effect of multiple similar sequences overwhelming the information provided by more divergent sequences, using a distance measure based on position specific differences in amino acid residues, as described by Henikoff and Henikoff [76]. Similarly each column is also weighted in accordance to the amount of information it provides, based on a simple measure of the number of different residues present in the alignment column.

Once the alignment has been reweighted, then the scores for each residue per position is calculated as the sum of the counts of each residue, along with a pseudocount based on the expected amount of each residue (based on the residue frequencies implied in the scoring matrix used initially). For subsequent iterations of PSI-BLAST the PSSM is used in place of the query and standard matrix. This approach allows for the location of more distant homologies by allowing patterns to be developed through the information provided by closer homologues, and then utilising these patterns in order to find more distant members of the family.



#### **1.1.5.4.2. Markov model based methods: HMMER**

The basis of the approach taken by HMMER in order to locate distant homologies is in some ways very similar to the taken PSI-BLAST. Both are based on the principle of their being more information in an alignment of multiple related sequences than there is in a single sequence, and both utilise this property in order to generate a scoring system which is specific to the individual alignment in hand. However, the actual methodology of the two programs is somewhat different. Instead of using an initial homology search in order to prime further searches using alignments, HMMER is reliant on prebuilt alignments in order to generate models for searching, in other words the researcher must already have a family of proteins to hand before HMMER can be used. The source of the proteins, and the method used to align them is essentially unimportant to the functioning of HMMER (although both will, perhaps unsurprisingly, have an effect on the output from the program). Given a multiple alignment, HMMER takes each column of the alignment and creates a hidden Markov model (HMM) based on it, which encodes the information of the likelihood of encountering each residue at each position, as well as likelihoods for the insertion and deletion of bases at each position. Whilst a complete explanation of Markov models and hidden Markov models is beyond the scope of this introduction (for an excellent overview of the use of Markov models in biological sequence analysis see [52]), in essence a Markov model can be thought of as a series of states, connected together by a series of arrows representing the probabilities of moving between those states. For example in a protein sequence we would have twenty states, one for each amino acid, and the arrows would represent the probabilities of moving from one residue to another as you move along the protein sequence. The difference between Markov models and Hidden Markov models is that there is no longer a direct relationship

between what we observe and the states within the model. In HMMs the things we can observe are called symbols, and they are linked to the states within the model in a similar way to the transition arrows, except that the arrows linking states to symbols represent what are known as emission probabilities rather than transition probabilities. Again using our amino acid analogy, an example of a HMM would be a model to detect signal peptide regions (Such a model exists and is widely used as part of the SignalP package [77]). In such a model the states would be whether we were in a signal peptide domain or not, and the symbols would be the amino acids of the sequence.

Given a position specific scoring matrix it is possible to generate a profile HMM which encapsulates that information giving an emission probability based on the frequency of each type of base at each position, along with transition probabilities between each state (i.e. aligning the next character against the model, as an insert relative to the model, marking a deletion of states in the model relative to the sequence). Once such a model has been produced it is possible to then use it to align a query sequence to the model. Due to the number of connections available within a profile HMM it is impossible to analyse every possible route from the start to the finish of the model. If we simplify a profile HMM by ignoring the complexity added by emission probabilities and looping to allow arbitrary length inserts then a model of length 100 has approximately  $10^{35}$  paths through the model. More generally, the number of paths ( $P_n$ ) through a network of length  $n$ , is  $P_n = 2P_{n-1} + P_{n-2} - P_{n-3}$ . Fortunately it is a general property of Markov models that the transition from one state to the next depends only on the previous state, and not on all the states prior to it. It is this property which makes analysing hidden Markov models amenable to dynamic programming approaches. The two algorithms used by the HMMER package

are the Viterbi and forward algorithm. Both algorithms are designed to calculate the most probable state path through a model based on a series of observations. These two algorithms are similar in their methodology, the only difference being that the Viterbi algorithm calculates the probabilities of the state path based only on the most likely path only, whereas the forward algorithm calculates it based on the sum of all the possible state paths that could have produced the observations being tested against the model. Whilst the forward algorithm would seem to be more thorough in its calculations of the probability of the best path through the model, the assumption that the optimal path through the model is the only significant one is a surprisingly good generalisation [52], and so often there is little difference between the output of the two algorithms.

Like PSI-BLAST, HMMER also provides for the weighting of input sequences so that divergent sequences are not drowned out by large numbers of similar sequences. In the case of HMMER, it implements a slightly more complex method than PSI-BLAST based on tree-based weighting scheme proposed by Gerstein et al [78], which calculates a sequence weight based on a measure of its proportion of the branch lengths between the leaf on which it resides and the root of the tree. HMMER also allows for empirical calibration of the model by testing it against a set of randomly generated sequences (5000 sequences with a mean length of 350 by default) in order to derive parameters which describe the location and scale parameters of the extreme value distribution which best fits the scores of random sequences to the model. These advantages are possible with HMMER as the model only has to be prepared once, rather than for every iteration, as is the case for PSI-BLAST. However, as previously mentioned, in order to be able to build a meaningful model for use within the HMMER software it is necessary to have a prior idea of the protein family you want

to search with, something which is not a requirement for use of PSI-BLAST.

### ***1.1.6. Sequence similarity as a predictor of structure and function***

The ultimate aim of homology searching tools is not only to locate proteins with a similar primary structure but to also make predictions/assignments of function based on observed similarities. In this regard there then has to be one major assumption made: That proteins with similar primary sequence will fold to form proteins which also share a common tertiary structure, and that as a result of also sharing a common structure, two sequentially homologous proteins will also have the same function.

Given that protein sequence is being used as an analogue of protein structure and function, we are presented with several questions:

1. Why not determine whether two proteins are homologous by direct computational calculation and comparison of a proteins structure?
2. If (1) is not possible, how accurate a predictor of structure is a protein's sequence?
3. How strong a predictor of function are both sequence and structure?

In answer to question one, if the assumption is that the folding of a protein is determined by the conformation in which it is in its lowest free energy state then an algorithmic approach to solving a proteins structure computationally is NP-hard [79], a class of computationally complex problems which are not possible to solve using current computing technology. There are, however, several heuristic methods which seek to produce structural models of proteins through computational analysis of the physical properties of a protein's constituent atoms. Such techniques have shown a good degree of success, albeit only with small proteins or domains (for example the

albumin-binding domain [80]). More recently work using physics based models of protein folding have been able to determine the structure of proteins with up to 100 residues [81]. However these approaches still require massive amounts of time to compute, and often require large compute clusters or distributed computing facilities (for example folding@Home [82, 83]) in order to resolve structures. Thus while it is possible to make assignments of homology by direct calculation of a protein's structure, it is not yet a feasible technique.

#### **1.1.6.1. Divergence between sequence, structure and function**

In order to answer question 2 posited above it is necessary to determine the correlation between particular sequences and the structures which they form. In particular, what are the proportions of similar sequences adopting different structures/folds (how many different structures can a sequence be related to), and how much sequence diversity is there in proteins/domains which share a common structure (how many different sequences can a structure be related to).

The nature of the sequence similarity between two proteins will also strongly affect the likelihood of them being functionally analogous. For example research has shown that local short sequences are not a predictor of structure [84-86]. This situation also applies to much larger amino acid sequences, such as domains. The presence of multiple domains within a protein can also be a trap for the unwary when assigning annotation based on homology, when only one of the domains is the region identified as being homologous.

When examining whole domains and proteins at the global level then there is much evidence to suggest that stronger degrees of similarity indicate an increased likelihood of function also being conserved [87]. For example enzymes showing 70% or greater

sequence similarity across the whole length of the protein will show 90% conservation of enzyme activity based on them being members of the same EC group (all four parts of the EC number) [88]. Similarly, protein sequence similarity is a predictor of structure [89], despite the difference in size of the sequence and structure spaces [90, 91]. This is not always the case though: immunoglobulins and cytokines are both examples of protein families which show little to no sequence homology, but do have readily identifiable structural similarities [91-93].

#### **1.1.6.2. Conserved sequence, un-conserved structure**

It has been generally held that the sequence of a protein specified a single structure [94]. Thus one would expect that identical or nearly-identical proteins will only form one particular structure. However a class of proteins held responsible for a range of neurodegenerative diseases, namely prions, has shown that this need not be the case. Prions are capable of existing in two stable structures: The normal structure which is nearly half  $\alpha$ -helix, with nearly no  $\beta$ -sheet, and the modified (disease) structure which shows over half  $\beta$ -sheet, but only 30%  $\alpha$ -helix [95]. These two structural conformations, despite being substantially different, are identical in sequence, and are not caused by any form of posttranslational modification [96, 97].

More recently there has also been evidence that in other groups of proteins, only small alterations in the sequence of the protein can lead to substantial alterations in its structure. Alexander *et al* have demonstrated that by starting with two proteins which show 77% sequence identity, but bind to two different proteins, it is possible to elucidate the minimum number of differences in amino acid sequence required to change the structure and function of a protein [98]. Through gradual reduction of the number of non-identical residues between the two proteins they were able to show

that a single amino acid substitution was able to alter the proteins structure from all  $\alpha$ -helix to 4  $\beta$ -sheet, 1  $\alpha$ -helix.

#### **1.1.6.3. Conserved sequence and structure, un-conserved function**

Even when both the sequence and structure of proteins are conserved it does not necessarily follow that function will also be conserved. Subtle changes in small areas of a protein's sequence will not alter the overall picture of sequence homology, nor will it necessarily change the structure of the protein, but it may change the way in which the protein functions. One such example of this is the  $\alpha$ - and  $\beta$ -subunits of the F1 portion of ATP synthases. Both proteins are sufficiently similar that they have only one model in domain databases (for example PFAM: ATP-synt\_ab), and are folded almost identically [99]. However, whilst both proteins are capable of binding ATP, only the  $\beta$ -subunit is actually catalytically active, whilst the  $\alpha$ -subunit functions in a regulatory capacity [99-101].

A similar situation can be observed in several other enzymes, where small changes in amino acid sequence do not alter the overall pattern of sequence and structural conservation, but do alter the proteins function by changing its enzymatic specificity. One such example of this is dehydrogenase enzymes. Members of the malate dehydrogenase (MDH) and lactate dehydrogenase (LDH) enzyme families share sequence and structural similarity [102]. But by altering just one residue in these proteins it is possible to change an LDH protein into an MDH one, and *vice-versa* [103].

#### **1.1.6.4. Conserved sequence, unknown function**

Beyond the issues surrounding predicting structure purely from assessment of a

proteins sequence and structure, one must also examine its environment in order to make a further assessment of its likely role within a cell. One such example of this is human protein kinases. The human genome contains over 500 protein kinase genes [104], which through splice site variations produce over 900 separate protein kinase proteins [105]. Whilst many of these proteins contain different domains, many are readily identifiable as members of the same family through sequence homology [104]. However it is not only the sequence and structure of the protein which will define the substrate or substrates with which it will interact. Where the protein is expressed in the human body will also have an effect on final function of the protein by determining the range of proteins available to interact with the kinase [106].

Also within the field of protein kinases, there is the example of the SctD family of proteins within type-III secretion systems (see section 1.3.2.1 for more details). This protein contains an FHA domain, a domain responsible for phosphoprotein recognition [107]. Normally FHA domains interact with serine/threonine protein kinases and phosphatases (STPK/STPP), however such a role for SctD proteins may not be the case. For example, there are genomes which contain a T3SS which do not contain any kinases or phosphatases: *Candidatus protochlamydia* is one such example of this [108]. In such cases it is hard to determine precisely what function this protein will fulfil, as despite any sequence or structural homology the absence of any STPKs or STPPs more or less precludes SctD from fulfilling its expected function.

#### **1.1.6.5. Un-conserved sequence, conserved structure**

Given a difference in the size of the sequence and structure spaces for proteins [90, 91], it is an inevitable conclusion that there will be proteins which share similar structures without sharing similar sequences. This class of proteins presents an



interesting question for those involved in the study sequence and structural homology. Namely, is the observation of a conserved structure, but un-conserved sequence the result of divergent or convergent evolutionary processes? By the very virtue of the lack of obvious sequence homology this question is very difficult to answer for any individual case in the absence of any other lines of evidence.

Immunoglobulin domain containing proteins have long been known to show little sequence homology to each other [109]. The characteristic  $\beta$ -sheet fold found in all sub-types of the immunoglobulin domain is conserved in proteins which show less than 10% identity to each other [110]. Within immunoglobulin domain members as a whole, the conformation of the central four  $\beta$ -sheets are highly conserved, with the folding being defined by the presence of a hydrophobic core [109, 110]. Across members of this domain family however there are no residues which can consistently be said to form part of this hydrophobic core [110]. Whilst the lack of obvious sequence similarity between immunoglobulin proteins makes for difficult analysis in the absence of structure, comparisons of immunoglobulins in the light of structural knowledge does demonstrate some correlation between certain residue changes or insertion/deletion events and membership of certain subclasses of the immunoglobulin family [110].

Within the field of type-III secretion, there are also examples of proteins which have a broadly conserved structure in the absence of obvious sequence homology. Type IB chaperones, a group of proteins which interact with multiple T3SS effectors within a particular secretion system, are just such an example. Examination of this class of proteins reveals a conserved structural motif, which when altered results in destabilisation of the chaperone-effector complex [111].

Although this motif contains both conserved and variable structural regions the overall configuration of the motif is retained across multiple chaperones which show little sequence similarity to each other. In particular the conserved interaction pocket, into which effectors bind, folds to result in a consistent three-dimensional location for the key binding residues in all proteins containing this pocket structure. When the solved structures for this class of proteins are aligned to each other the overarching shape of the domain and location of key binding residues/regions is very easy to observe [111].

#### **1.1.6.6. Un-conserved sequence and structure, conserved function**

Where it was the case for immunoglobulins that family members could be identified by conserved structure if not by conserved sequence, the same cannot be said for a group of bacterial proteins which interact with them. Several proteins have been found in bacteria which bind to the Fc region of type-G immunoglobulins (IgG), this includes protein A from *Staphylococcus aureus*, protein G and protein H from *Streptococcus* sp [112, 113]. Like immunoglobulins, these proteins lack any identifiable sequence homology in the region responsible for binding to IgG [114]. Unlike immunoglobulins however, they demonstrate a lack of homology at the structural level as well [115]. Interestingly, directed mutation of these proteins has resulted in the creation of two proteins with nearly 60% sequence identity whilst retaining the corresponding proteins retaining their original structure (all  $\alpha$  for protein A,  $\alpha+\beta$  for protein G) [116].

Within vertebrates, there is another example of proteins which show no sequence or structural homology to each other, but still perform an identical function. Crystallins are found in eye lenses and form the bulk of the protein content within the lens [117].

There are multiple types of crystallins which have been co-opted from other functions, typically as enzymes [118]. In fact in several birds  $\delta$ -crystallin proteins are still enzymatically active, and function as arginosuccinate lyases [119]. Similarly  $\alpha$ -crystallins are related to heatshock proteins,  $\beta$ - and  $\gamma$ -crystallins are related to calcium binding proteins, and  $\epsilon$ -subunits retain enzymatic functionality as lactate dehydrogenases [119]. It would seem that the requirement to produce large amounts of protein in order to create the right refractive properties was the overriding force which drove the co-option of these diverse proteins into a common role, and thus the major selection criteria was controllable up-regulated production of a stable protein, rather than any more specific structural properties of the protein in question [118, 120].

#### **1.1.6.7. Implications for assignments based on sequence similarity alone**

In an ideal world it would be possible to make all annotation of genes based on multiple lines of evidence including analysis of the sequence, structure and function and known interactions with other proteins, co-factors and molecules. The reality however is somewhat different, often lack of supporting evidence, and the time implications for genome annotation projects have led to much annotation being based on sequence homology to other proteins and domains alone.

Anfinsen's dogma that protein structure is solely determined by amino acid sequence [94], is both well demonstrated (see [121] for a recent example), and on the face of it would seem to support the case for annotation by sequence analysis alone. However, care needs to be taken in parsing this statement, as whilst it may be true to state that sequence alone is enough to determine a proteins structure, the lack of a 1:1 relationship between entities in protein sequence space versus structure space means

that the reverse will not necessarily be true.

In addition the examples given above demonstrate the caveats that should be applied when using sequence similarity tools and measures to make inferences about structure and function. Not all proteins which are homologous at a sequence level will necessarily fold into the same structure, and those that do will not necessarily perform the same function. Similarly, an absence of evidence for homology at the sequence level does not preclude that those two proteins will fold into the same structure, or perform the same function.

### ***1.1.7. Phylogenetics and phylogenomics***

With the availability of a large amount of sequence data and the ability to search it in order to locate homologous proteins within this data set, it becomes beneficial to have some method which allows us to compare families of homologous proteins to each other. In order to do this the techniques of molecular phylogenetics can be used. Phylogeny (or phylogenesis) is defined as the pattern of historical relationships between species or other groups resulting from divergence during evolution [122], and molecular phylogenetics is the study of phylogeny through the use of DNA or amino acid sequence data. There are several methods available for the reconstruction of a phylogenetic tree based on estimation of the true tree given the information provided by sequence data. It is this key issue of reconstruction that is the main problem when considering phylogenetic trees. For a sample of 10 different taxa there are over 34 million possible topologies which the phylogenetic tree may take, only one of which will be the correct topology. As such certain optimisations have to be performed in order to locate the optimal topology from amongst the massive set of alternatives.

### **1.1.7.1. Methods for classifying proteins and larger units into trees**

When examining individual proteins/genes, or even small numbers of genes then traditional molecular phylogenetic approaches can be used to estimate the phylogenetic tree. These methods can be broadly broken down into three separate categories: Distance methods, maximum parsimony methods, and maximum likelihood methods. In distance methods evolutionary distances are calculated for all pairs of taxa, and the tree topology is calculated by an examination of each of those distances. Maximum parsimony methods function by calculating a series of correct topologies and then choosing the one which requires the smallest number of changes in sequence in order to be correct. Finally, maximum likelihood methods function by calculating the likelihood of observing a given set of sequence data for each topology based on a given substitution model, and the topology with the maximum likelihood is chosen as the best. Each category of method has within it a series of different algorithms which implement the principle of the method in different ways. For example the unweighted pair-group method using arithmetic averages (UPGMA) [123], Least Squares (LS) [124], Minimum Evolution (ME) [125], and Neighbour Joining (NJ) [126] methods are all examples of techniques which employ a distance method approach to estimating the correct tree topology.

Where more than one protein is to be phylogenetically examined, then it is possible to create a phylogenetic tree using alignment based approaches as mentioned above, simply by concatenating together the alignments of each protein and then creating the tree based on the concatenated alignment. There are however, several other methods of creating phylogenetic trees which do not require the presence of an alignment. These methods can broadly be classified into four different groups: Alignment-free genome trees based on properties of the complete genome, gene content trees based

on the presence/absence of certain gene sets, gene order trees based on the synteny of genes within the genome, and genome trees based on average sequence similarity. These methods range from the simple, such as the alignment free approaches which approximate distance between genomes simply by counting the frequency of words within the genome (e.g. the count of each type of DNA or amino acid sequence of length  $n$ ) [127], to more complicated approaches involving analysis of each gene in each genome being examined in order to calculate distances based both on the number of shared genes but also the similarity between genes conserved in different species [128, 129].

Together these methods allow us to examine the evolutionary relationship between anything from individual genes, to whole genomes, in order to better understand their origin and diversity, along with allowing us to analyse the role of events such as horizontal gene transfer, through incongruencies between trees of genes suspected of horizontal gene transfer, and those for which transfer is known to only have occurred only through vertical transfer.

## **1.2. Protein Secretion by Bacteria**

If a bacterium is to interact with its environment then it is a requirement that it should be able to export elements from its cytosol into the external milieu and *vice versa*. Until around forty years ago it was assumed that protein secretion by bacteria was a rare phenomenon, and where it occurred it happened in a protein specific manner [130]. This assumption has been dispelled by the discovery of numerous systems dedicated to the export of proteins through the cytoplasmic membrane, and also in the case of Gram-negative bacteria, the periplasm and outer membrane. Together these systems function to export proteins to either be anchored on the outer surface of the

bacterial cell, or exported into the external environment or in some cases even directly into the cytoplasm of eukaryotic cells.

### ***1.2.1. Secretion across the inner membrane***

The first impediment any protein will encounter when trying to exit a bacterial cell will be the inner membrane, regardless of whether the bacterium is Gram-positive or negative. There are three main systems involved in the transport of proteins through the inner membrane alone: The Sec (general secretory, or GSP), SRP (signal-recognition particle) and Tat (twin-arginine translocation) pathways.

The Sec pathway is produced via the interaction of several different proteins, which are conserved across both prokaryotes, and eukaryotes (where it is known as the Sec 61 complex, and is involved in transport across the membrane of endoplasmic reticulum [131]). Several of these proteins are also common to the SRP system. Those proteins include SecYEG which together produce a heterotrimeric molecule which forms in the inner membrane [132], as well as SecA, which interacts with SecY as a dimer, energising the system through its ATPase activity [133]. It is at this point where the mechanics of the two systems diverge. Within the Sec system the general chaperone SecB binds proteins both co- and posttranslationally [134], and then delivers them to the Sec machinery through its binding with SecA [135]. In contrast the SRP system functions by the integration of the signal recognition particle with newly synthesized membrane proteins in a co-translational manner. The SRP then binds to the ribosome-nascent chain (RNC) complex [136]. SRP+RNC complex then binds to the protein SRP receptor FtsY which in turn directs it to the SecYEG machinery [137, 138].

The more recently recognised Tat system would seem to consist of no more than three

components: TatA, TatB and TatC. All three of these proteins are required in order for *E. coli* to produce a functional Tat export system [139]. This is not always the case however, as some Tat systems do not encode a TatB protein (e.g. *Staphylococcus aureus* and *Rickettsia prowazekii*) [140]. Conversely some systems encode multiple copies components of the Tat system. For example *Bacillus subtilis* has two copies of *tatC* and three copies of *tatA* [141]. In the Tat export system all three proteins would seem to interact together in the inner membrane to form the functional machinery [142, 143]. Proteins destined for export by the Tat system are targeted to the machinery based on an interaction between TatBC and the protein to be exported, the protein is then directed to the pore formed by TatA, and exported [144].

In all the systems described above, there is a characteristic signal sequence which is contained within the N-terminal region of the peptide to be exported. This signal sequence allows the protein to be targeted to the correct system for its export. In the sec pathway the signal sequence consists of a 15-30 amino acid N-terminal peptide, which lacks a simple consensus sequence, but consists of three generalised regions: An N-terminal positively charged region (n-region), a hydrophobic region of at least six residues (h-region) and an C-terminal regions of polar uncharged residues (c-region) [145]. The SRP system employs a similar signal sequence, with the pathway the protein is directed to being dependant on the hydrophobicity of the central region (h-region) of the signal sequence. If the region is more hydrophobic then it will be directed down the SRP pathway rather than the sec pathway, and *vice versa* [146]. Finally the Tat pathway utilises a more conserved, but none the less related, signal sequence. Again the signal sequence has an n-region, h-region, c-region arrangement; however there is a conserved motif which occurs in the signal sequence at the end of the n-region and start of the h-region. This motif has the form Ser/Thr-Arg-Arg-X-



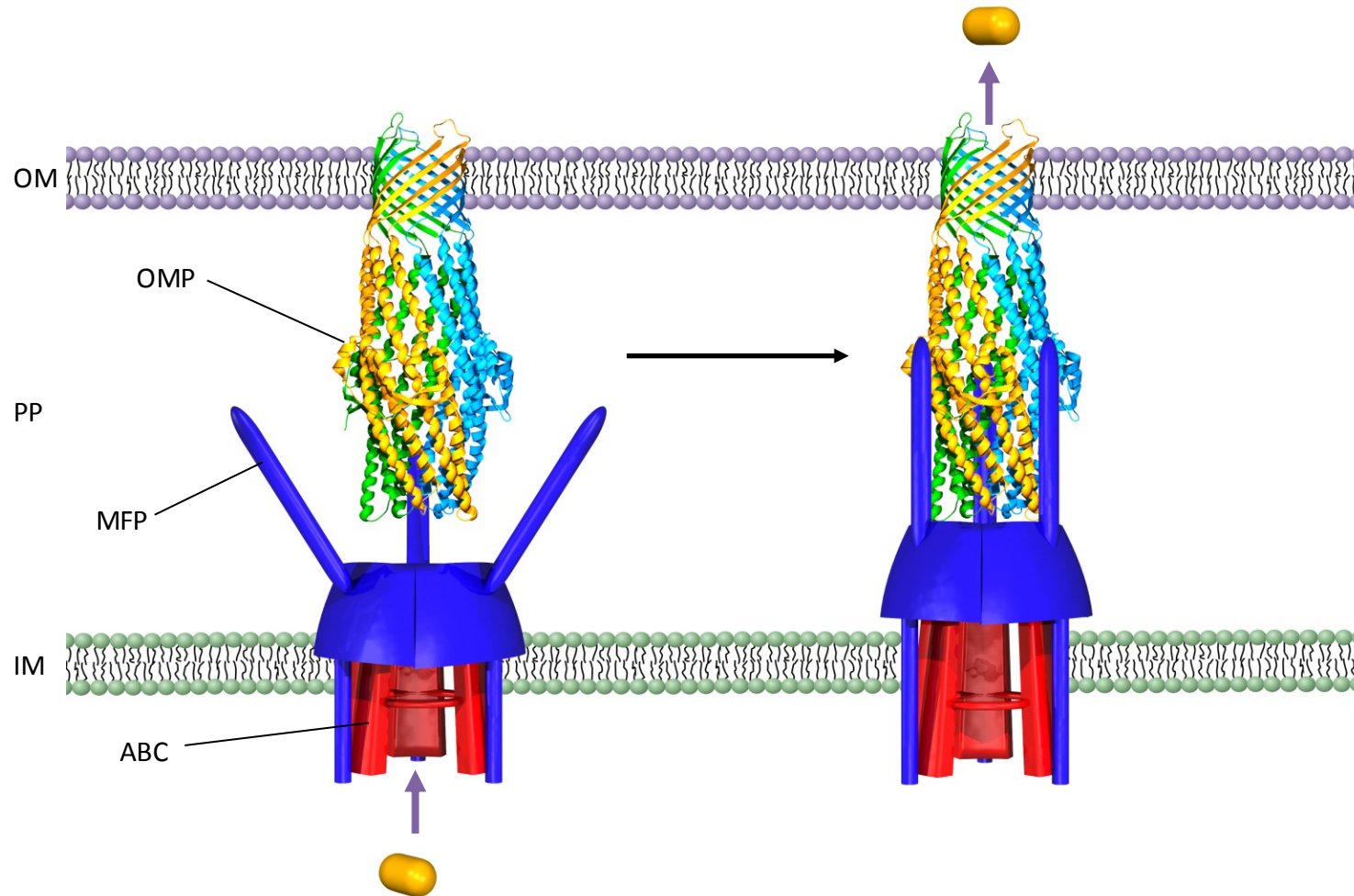
Phe-Leu-Lys (where X is any polar amino acid) [140, 147]

### ***1.2.2. Secretion across the outer membrane***

Secretion across the outer membrane is by its very nature a process only undertaken by Gram-negative bacteria. In order to accomplish this task Gram-negative bacteria have evolved a series of mechanisms which allow them to either export proteins as a two-step process, utilising one of the methods mentioned above to export the protein across the inner membrane, or as a one step process, where the protein is exported from the cytosol to outside of the cell without any intermediate steps. The mechanisms available to accomplish this task are named, for better or worse in a simple numerical manner. There are at present five major systems that are well described: The type I-V secretion systems. The following sections briefly describe each of those systems.

#### **1.2.2.1. Type I Secretion**

Type-I secretions systems (T1SSs) allow for the movement of proteins from the cytoplasm to outside of the cell in a one-step manner, utilising a simple system of just three proteins. These three proteins are an ATP-binding cassette (ABC) transporter, a membrane fusion protein (MFP) and an outer membrane pore forming protein (OMP) [148] (See Figure 1). The ABC protein consists of a cytoplasmically located nucleotide binding domain, and a transmembrane domain produced from six  $\alpha$ -helices [149]. ABC proteins typically function as homodimers or trimers in producing a functional pore through which the secreted protein can traverse [149, 150]. It is also the role of the ABC protein to provide substrate specificity to the secretion machinery [151]. The MF proteins interact in a trimeric fashion with the ABC proteins in order to provide a periplasmic channel through which the secreted protein can travel [152].



**Figure 1. Schematic over view of the type I secretion system**

The position of the outer membrane (OM), periplasm (PP), inner membrane (IM) and the major components of the T1SS are shown: ABC – ATP Binding Cassette, MFP – Membrane Fusion Protein, OMP – Outer Membrane Protein. The structure of the OMP is that of TolC (PDB entry 1EK9). The process shown is the secreted molecule (shown in orange) binding to the ABC, and causing a conformational change in the MFP leading to its interaction with the OMP, and subsequent translocation of the secreted molecule to the external environment

It has been suggested that the binding of substrates to the ABC protein leads to a conformational change in the MFP, such that it interacts with the OMP to complete the channel to the external environment [153]. The exact mechanistics of this interaction remain unclear though [154]. There is also evidence to suggest that such an interaction can also exist in a substrate independent manner [155]. The OMP, as typified by the TolC protein from *E. coli* exists as a trimer anchored in the outer membrane by a  $\beta$ -barrel structure [156]. It has been shown that the ABC and OM proteins can interact together directly, although this interaction *in vivo* requires the presence of the MFP [155].

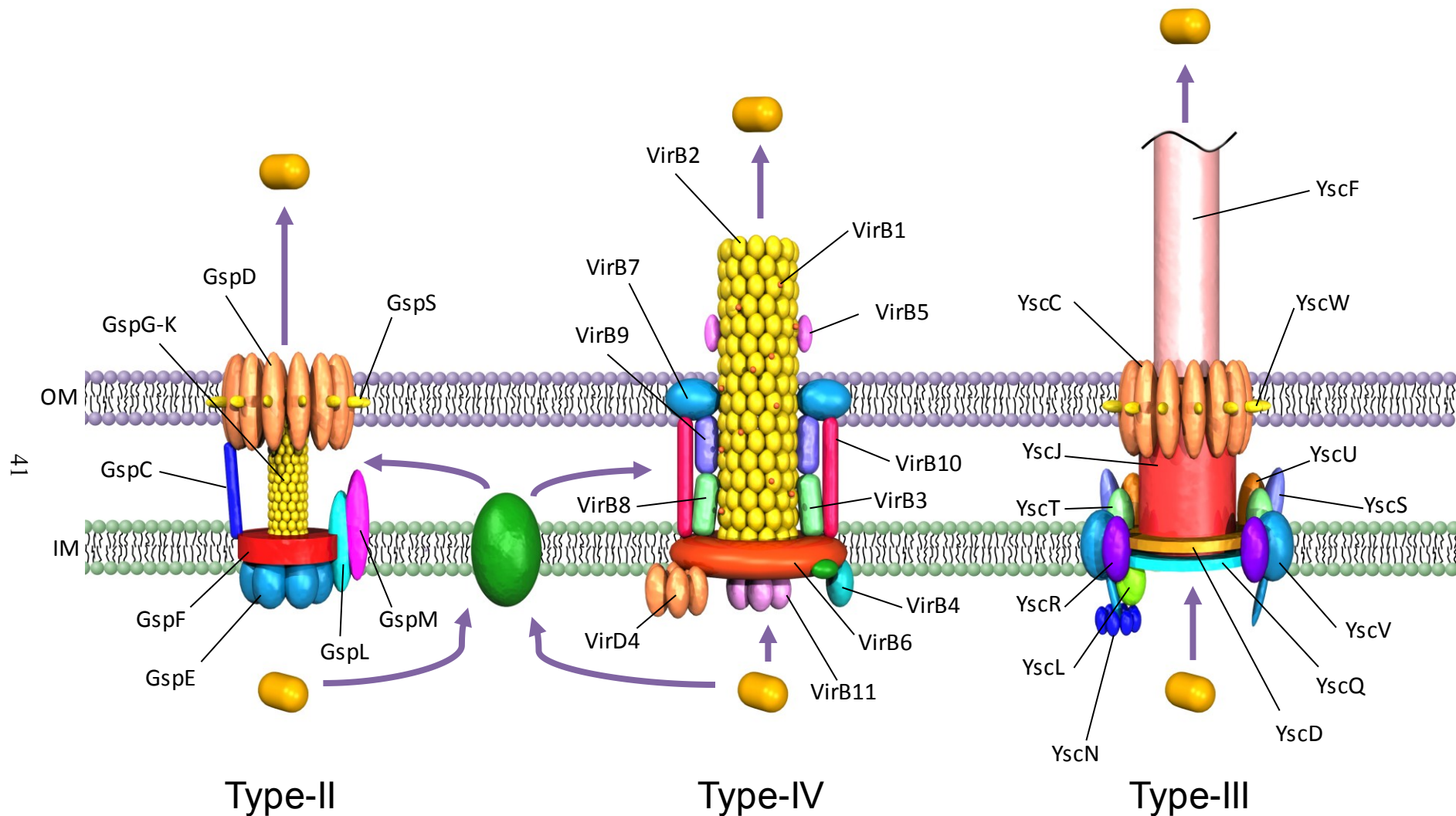
There are a wide variety of proteins exported by T1S machinery, from enzymes to toxins and adhesins [148]. These proteins also vary dramatically in size from several hundred amino acids (HasA from *Serratia marcescens*: 188aa [157]), to several thousand (LapA from *Pseudomonas putida*: 8682aa [158]). Proteins secreted by T1SSs contain a C-terminal secretion system, most likely in the terminal 15-30 amino acids [159-161], implying that the molecule must be secreted in a post-translational fashion. There is no specific consensus for this signal sequence, although it would seem that there is a bias towards certain amino acids (LDAVTSIF) [148].

#### **1.2.2.2. Type II secretion**

Compared to T1SSs, type II secretion systems (T2SSs) are considerably more complex. This is despite the fact that T2SSs only traffic proteins across the outer membrane. In order to traverse the inner membrane a separate secretion system is required. Generally this is done by the Sec pathway [162], however there is also evidence of there being type-II secreted proteins exported through the inner membrane via the Tat system [163].

T2SSs utilise 12 to 16 proteins in order to effect secretion through the outer membrane. Perhaps surprisingly though only a couple of these are actually located in the outer membrane [162, 164]. The remainder of the proteins are located in the cytoplasmic membrane or in the periplasmic space (See Figure 2). There is also a cytoplasmic protein GspE, which interacts with ATP [165], and presumably acts as an ATPase, providing energy to drive the export of proteins [166, 167]. The inner membrane located proteins include several (GspL and M) which anchor the ATPase to the apparatus [165, 168, 169], and GspF, which may function to provide a pore for the translocation of pseudopilins into the periplasmic space [170]. The pseudopilins themselves are a group of proteins (GspG-K) which come together to form a large multimeric structure called the pseudopilus. Based on evidence from several experiments it has been hypothesised that the pilus may grow in order to push secreted molecules through the outer membrane complex, or alternatively as a cork to close off the outer membrane channel when not required [162].

The outer membrane complex of T2SSs consists of two components, and pore forming protein GspD, which exists as a multimer of 12-14 copies and forms the pore in the outer membrane, and a lipoprotein GspS. GspD and GspS proteins iterate in a 1:1 stoichiometry [171], and GspS serves to aid the localisation of the GspD multimer into the outer membrane [172, 173]. The pore formed by GspD is approximately 95Å in diameter, a size large enough for proteins to pass through T2SSs in a folded state [171].



**Figure 2. Schematic representation of the type II, III and IV secretion systems**

The type-II is exemplified by the pullinase secretion in *Klebsiella oxytoca*, the type III secretion system by the Yops secretion in *Yersinia*, and the type IV system by the VirB/VirD system of *Agrobacterium tumefaciens*. IM = Inner Membrane, OM = Outer Membrane

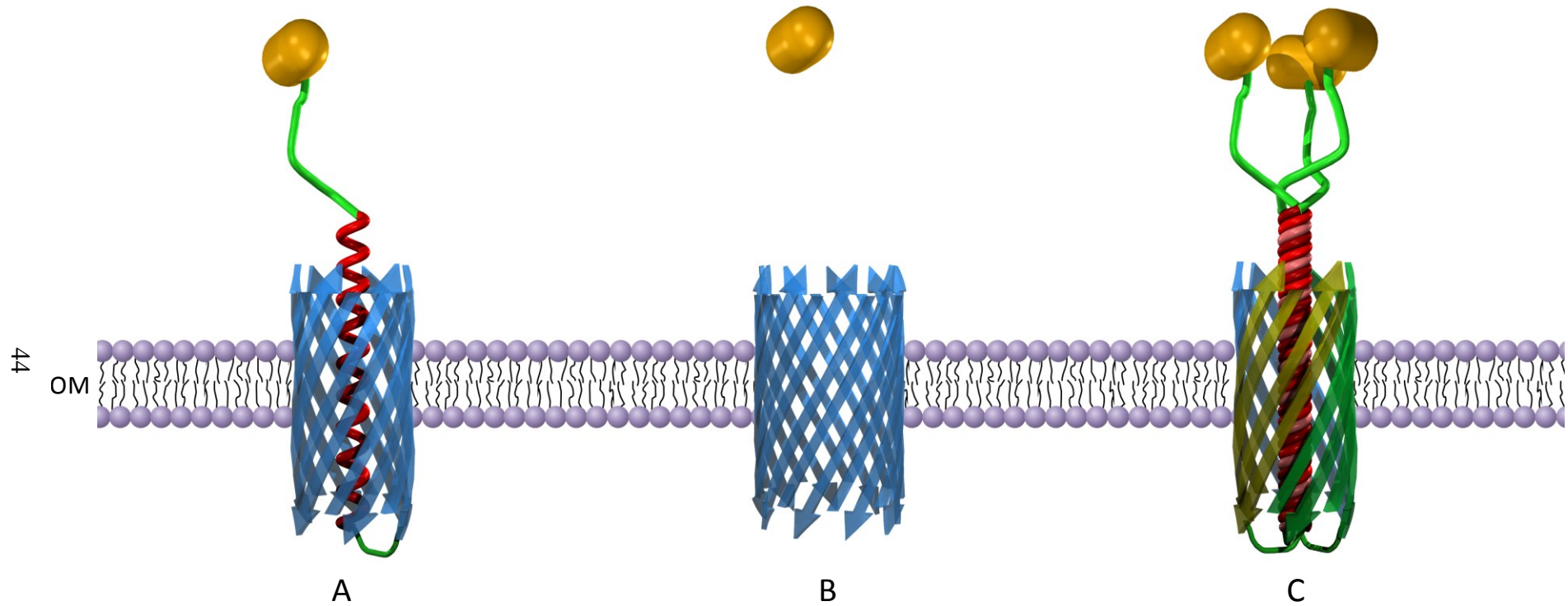
### 1.2.2.3. Type IV secretion

Type IV secretion systems (T4SSs) are unique amongst the systems characterised to date in that they can translocate proteins both in a one-step and two-step manner. Furthermore, T4SSs are not just limited to the transport of proteins; they can also function to transport DNA. One of the best studied T4SSs is that of *Agrobacterium tumefaciens*, which utilises a T4SS to export transfer-DNA (t-DNA) into dicotyledonous plants, where the single stranded DNA which is transferred contains oncogenes, resulting in tumour formation in the plant [174]. The *A. tumefaciens* T4SS is encoded on a plasmid which contains two operons named *virB* and *virD* [174], comprising eleven and five genes respectively. The proteins encoded in the *virB* operon (VirB1-11) function as part of the main secretion apparatus, whilst the *virD* operon consists of four genes which encode proteins (VirD1,2,3 and 5) involved in the processing of the t-DNA and one (VirD4) which couples the t-DNA to the T4SS [175]. As with T2SSs, the VirB proteins are located in various positions between the cytoplasm of the bacteria and the outer membrane (See Figure 2). VirB4 and VirB11 both contain nucleotide binding domains and exhibit ATPase activity [176-178], suggesting that these components provide energy for the translocation of proteins through the secretion system. The VirD4 proteins also contain a Walker A nucleotide-binding motif [177]. Of the remaining proteins VirB6, VirB8 and VirB10 all lie in the inner membrane. Whilst the precise function of these three proteins is unclear, they all show a propensity to interact with several other members of the secretion system suggesting that they act as a bridge anchoring the components of the system together [175]. VirB8 for example has been shown to interact with VirB1, VirB4, VirB8, VirB9, VirB10 and VirB11 [179, 180]. All the remaining six *virB* operon proteins have characteristic GSP type signal sequences, and have been shown

to localise to the periplasm and outer membrane. VirB2 is the pillin subunit, and assembles together into the T4SS pilus [181]. VirB3 and VirB5 are thought to interact with the VirB2 pilus, with there being evidence that VirB5 interacts with VirB2 in a manner dependant on several other VirB proteins [182]. VirB7 and VirB9 interact together , and help form the outer membrane pore for translocation [183, 184], with VirB9 forming the pore and VirB7 lipoprotein stabilising the complex [185], in a situation similar to GspD and GspS in T2SSs. Finally VirB1 has been shown to localise to the periplasm where it functions as a transglycosylase, and is postulated to aid the biogenesis of the T4SS by creating a hole in the periplasm [186], as well as interacting with the pilus proteins VirB2 and VirB5 [187].

#### **1.2.2.4. Type V secretion**

Type V secretion systems (T5SSs) can be grouped into three main categories: Type Va – the autotransporters (AT), type Vb – the two partner system (TPS) and Vc – the oligomeric coiled-coil (Oca) system, also known as the AT-2 system (See Figure 3). Each of these three sub-systems have several identical characteristics, which include the presence of a N-terminal GSP signal sequence for transport through the inner membrane [188-190], formation of periplasmic intermediates and formation of a  $\beta$ -barrel pore in the outer membrane to permit secretion into the external environment [188, 190, 191]. However beyond those points there are a number of differences between the 3 varieties of T5SS. For example in the case of the AT and Oca systems the signal sequence,  $\beta$ -barrel and secreted molecule are all within in a single protein [188, 190]; TPS systems by contrast encode the  $\beta$ -barrel and secreted molecule as two separate proteins [192]. Furthermore, the AT and TPS systems produce a complete  $\beta$ -barrel through one protein [130], whilst Oca system must produce a homotrimer in the



**Figure 3. Schematic overview of the type V secretion systems**

Secretion across the outer membrane (OM) is shown for each of the three forms of T5SS: Autotransporter (A), two-partner (B) and oligomeric coiled-coil (C).



periplasm in order to direct translocation of the secreted molecule through the outer membrane [190].

These differences present a series of obstacles for the export of proteins via these systems. In the case of the TPS systems the separation of the pore and exported protein means that they are spatially separated. In order to direct the protein to its pore it is necessary to have to some sort of signal recognition event between the two. In order for the proteins to associate the secreted protein contains what is termed a TPS domain in its N-terminal region which interact specifically with the pore forming protein to initiate translocation through the outer membrane [193, 194]. Oca systems also require regions within the protein to allow for assembly of the monomers into the trimer necessary for formation of a complete pore in the outer membrane [190, 195].

#### **1.2.2.5. Type III secretion**

Type-III secretion systems (T3SSs) are possibly the most complex of all the secretion systems thus far described. They accomplish secretion of proteins from the bacterial cytoplasm to the cytoplasm of other cells in a one step process by utilising a needle like appendage with a 'tip' on the end which allows for a hole to be made in the cell membrane of the cell into which the secreted protein will be translocated. The following sections describe the core apparatus which comprises the type III secretion apparatus, along with the accessory components involved, and the diverse range of bacteria which possess the system.

## 1.3. Type III Secretion Systems

### ***1.3.1. The bacterial Flagellum and Non-Flagellar secretion systems***

Type-III secretion systems fulfil two significantly different roles with bacteria. The first is to act as an assembly and export system in the production of the bacterial flagellum (Flagellar- or F-T3SSs). The second is to translocate effector proteins into the cells of plants and animals. Whilst the majority of these T3SS secrete pathogenicity factors, this not always the case and as such are termed here Non-Flagellar- or NF-T3SS rather than pathogenic-T3SS. As with certain other secretion systems the apparatus spans both membranes and the periplasm creating an apparently continuous channel through from the cytoplasm to the external environment [196, 197]. This channel is made from a series of components which form the channel itself and a series of accessory components which are required for the function of the secretion system. Between F- and NF-T3SS there are at least ten conserved proteins. Each system also has its own set of unique genes which allow it accomplish its role within the bacterium. This is especially true of the flagellar system which contains a large number of additional proteins which are not directly related to the type-III system, but are essential for the formation of a functional flagellum.

Whilst there are over 2200 hits in the PubMed database to the term “type-III secretion” surprisingly little is known about the type-III secretion apparatus itself. This is particularly true of non-flagellar systems, where much information has been inferred from homology between flagellar and non-flagellar T3SS proteins. This situation is now improving with resolution of protein structures for several components of the type-III secretion apparatus [198-201], along with protein interaction data [202-205], which have both added a wealth of additional information

about some of the better known NF-T3SSs, and their protein components

In the following sections both empirical and homology data are used to draw parallels between flagellar- and non-flagellar-T3SSs in order to define shared characteristics and functions between the two systems, and also with some of the proteins found in other secretion systems discussed above. In order to avoid some of the problems created by the inconsistent nomenclature used within NF-T3SS, the Sct (for SeCretion and Translocation) naming convention as proposed by Hueck [196] will be used when referring to NF-T3SS proteins in general. An overview of the proteins involved in the formation of the type-III secretion apparatus is presented in Table 2.

### ***1.3.2. The construction of type-III secretion systems***

#### **1.3.2.1. Cytoplasmic and Inner Membrane Proteins**

As with most of the multi-component systems mentioned above T3SSs have a protein capable of binding ATP and hydrolysing it, using the classic nucleotide binding motifs Walker boxes A and B [206]. Whilst there is no evidence to suggest exactly how this protein (SctN) powers the system, it has been shown that the protein, and in particular a function nucleotide binding domain is required for a functional T3SS [206, 207]. The SctN proteins from the various T3SSs including the flagellar ATPase FliI, show similarity to the  $\alpha/\beta$ -subunits of the  $F_0F_1$  proton translocating ATPase [208-210]. F-type ATPases and their close relatives (V-type and A-type) all possess a hexameric complex (in the form  $\alpha_3\beta_3$ ) which is responsible for the processing of ATP [99, 211]. On this basis it has been posited that FliI/SctN also form a homo-hexamer in a similar fashion [212]. More recent studies have shown that SctN/FliI proteins do indeed require oligomerisation to optimally couple ATP hydrolysis to translocation [213-215] and form hexamers, or in some cases dodecamers formed from two stacked

| <b>Protein Family</b> | <b>Flagellar Homologue</b> | <b>Location</b>    | <b>Function</b>             | <b>Notes</b>  |
|-----------------------|----------------------------|--------------------|-----------------------------|---|
| SctN                  | FliI                       | Cytoplasm          | ATPase                      | Homologous to F-type ATPases $\alpha/\beta$ subunit           |
| SctL                  | FliH                       | Cytoplasm/<br>I.M. | ATPase regulator            | Interacts with SctN and SctQ                                  |
| SctD                  | None                       | I.M.               | ? Secretion Regulator       | Contains an FHA domain  |
| SctQ                  | FliN                       | I.M.               | ATP system (SctL/N) anchor  | Binds multiple proteins in I.M.                               |
| SctR                  | FliP                       | I.M.               | ?                           | Multiple transmembrane domains                                |
| SctS                  | FliQ                       | I.M.               | ?                           | Multiple transmembrane domains                                |
| SctT                  | FliR                       | I.M.               | ?                           | Multiple transmembrane domains                                |
| SctU                  | FlhB                       | I.M.               | Substrate specificity       | Interacts with needle length regulator SctP                   |
| SctV                  | FlhA                       | I.M.               | ?                           | Large number of protein-protein interactions                  |
| SctJ                  | FliF                       | Periplasm          | Periplasmic pore            |   |
| SctC                  | None                       | O.M.               | Outer membrane pore         | Homologous to T2SS and filamentous phage secretins            |
| SctW                  | None                       | O.M.               | Stabilising SctC            | Not conserved throughout T3SSs                                |
| SctF                  | FlgE*                      | External to cell   | Extracellular Needle        | } Structural similarities between SctF, EspA and FliC         |
| EspA/HrpA             | FliC                       | Distal to SctF     | Needle Extension            |   |
| SctP                  | FliK*                      | ?                  | Needle length controller    | ? Molecular Ruler   |
| YopB/D<br>EspB/D etc  | None                       | E.M.               | Host cell pore (translocon) | Some translocators (e.g. EspB) can also function as effectors |

**Table 2. Summary of the components of non-flagellar type-III secretions systems.**

For more information on each protein see text. I.M. - Inner Membrane, O.M. - Outer Membrane, E.M. - Eukaryotic Membrane. \* These proteins are not homologous but perform an analogous function.

hexamers [215-217]. It has also been suggested that it is the ATPase which is responsible for recognition of substrates to be type-III secreted [218].

In the Yersinia Ysc T3SS there are two other conserved proteins which are known to interact with the ATPase. These proteins are YscL and YscQ (members of the SctL and SctQ families respectively) [204]. The YscL protein is homologous the flagellar protein FliH and shows a series of conserved residues shared between both families [219]. Evidence suggests that FliH interacts with the N-terminal region of flagellar ATPase FliI, and in doing so inhibits the ability of FliI to hydrolyse ATP [220, 221]. Studies of YscN and YscL have also shown that YscL can act in a similar fashion [222]. Yeast two-hybrid/three-hybrid and other interaction studies have shown a direct interaction occurs between YscQ and YscL [204, 205, 223]. SctQ's flagellar homologue FliN is known to form the major part of the C-ring within the flagellum [224], and interacts with FliH anchoring it and FliI to the flagellum [225, 226]. FliN also interacts with several other flagellar components: FliG and FliM [225], however neither of these components exists within NF-T3SS systems. Instead the Shigella SctQ protein (Spa33) interacts with several basal body components MxiG and MxiJ (part of the SctJ family) [205], suggesting that the role of SctQ, like FliN, is also to anchor the ATPase to the rest of the apparatus.

Other proteins which are known to occur in the inner membrane are SctD, SctR, SctS, SctT, SctU and SctV. The SctRST protein families and their flagellar counterparts (FliP, FliQ and FliR respectively) show the highest levels of sequence homology between the flagellar and non-flagellar T3SSs [227]. Together with several other proteins within the flagellum, including FlhA, FlhB and FliO, they form the central pore of the flagellar type-III export machinery [227, 228], and are all required for

functional export of flagellar proteins [229]. FlhPQR all have multiple transmembrane domains as predicted by hydrophobicity analysis of their protein sequences [230, 231], and have been shown to be associated with FliF (homologue of SctJ), which forms the flagellar MS ring [232].

SctU is homologous to the flagellar protein FlhB, which is known to have a role in controlling substrate specificity of the T3SS of which it is a member [233, 234]. Several years after the discovery of this function for FlhB, studies of the Yersinia SctU protein (YscU) showed that YscU acts as a coordinated regulator of NF-T3SS substrate specificity along with another protein, YscP [235, 236].

The SctD protein is a requirement for a functional T3SS [237]. However its role in the apparatus remains enigmatic. From domain analysis it has become clear that SctD proteins have cytoplasmic Fork Head Associated (FHA) domain, present in the N-terminal 120 amino acids [107, 219]. FHA domains are found both in eukaryotic and bacterial domains of life and act as phosphoprotein recognition domains, showing a particular preference for phosphothreonine containing proteins [238]. The domain's role appears to be regulatory in nature, interacting with phosphorylated proteins modified by serine/threonine protein kinases and phosphatases [238]. Whilst no such role has as yet been ascribed to SctD proteins, an FHA domain protein involved in a type VI secretion system in *Pseudomonas aeruginosa* (Fha1) has been shown to regulate secretion through the system dependent on the presence and activity of a kinase/phosphatase pair [239].

The final protein present in the inner membrane component of T3SSs is SctV. SctV proteins contain eight transmembrane domains, all of which are located within its N-terminal sequence, the C-terminus containing a large hydrophilic domain which

extends into the cytoplasm [240]. There is a degree of complementability between SctV proteins from different T3SSs. For example when the SctV protein from *Salmonella typhimurium* (InvA) is knocked out, function can be restored to the T3SS by using a homologous protein from *Shigella flexneri*, MxiA [241]. In the same experiment a chimeric protein containing the N-terminal domain of YscV (which is also known as LcrD) from *Yersinia pseudotuberculosis* and the C-terminal domain of InvA was also able to complement the knock out. However, the complete YscV protein could not, suggesting that it is the C-terminal cytoplasmic domain that is important in determining the specificity of the SctV proteins to individual T3SSs. The flagellar homologue of SctV, FlhA has been shown to interact with a large number of proteins within the flagellum including FlhA, FliF, FliO, FliP, FliQ and FliR [228], suggesting that it plays a role in anchoring proteins to the flagellar complex.

#### **1.3.2.2. Periplasmic and outer membrane proteins**

Beyond the inner membrane there are two main proteins which span from the outer edge of the inner membrane to the outer edge of the outer membrane. The first of these is SctJ which serves as the periplasmic spanning protein. The structure of the *E. coli* SctJ protein, EscJ has been solved using nuclear magnetic resonance spectroscopy [242], and also through crystallographic methods [200]. EscJ contains two subdomains, D0 and D1, each containing 3  $\beta$ -sheets, flanked on either side by  $\alpha$ -helices [242]. The protein also contains a linker region in between the two domains, which is required for a functional EscJ protein [242]. Analysis of the crystal packing of EscJ showed a superhelical structure containing 24 EscJ monomers per helical turn [200], suggesting that EscJ exists as a 24mer within the T3S needle complex. Stoichiometric analysis of the *Salmonella typhimurium* needle complex also supports

this fact [200]. Portions of flagellar protein FliF are homologous to SctJ. FliF is the protein responsible for the formation of the MS-ring within the flagellum, and exists as a 26mer ring structure [243], thus giving the FliF and SctJ protein complexes a similar degree of rotational symmetry. However, FliF is very much larger than SctJ proteins (*Salmonella typhimurium* FliF: 560aa, *E. coli* EscJ 190aa). By *in silico* analysis of the two proteins it is possible to see that many of the domains essential to FliF are absent from the SctJ, including the C terminal region which form the M-ring, and also mediate its interaction with FliG [244, 245]. This is not an entirely surprising result since there is no FliG homologue present in NF-T3SS systems, and there is also evidence of the interaction between the inner membrane complex and SctJ occurring via SctQ [205].

The final protein found 'inside' the cell is SctC. SctC forms the pore in the outer membrane, and is the only major component of NF-T3SSs for which there is no flagellar homologue. Instead SctC is part of the secretin family which also contains the T2SS protein GspD, and proteins involved in filamentous phage assembly [196, 246]. This protein forms a multimeric complex which anchors into the outer membrane of the bacterial cell [171, 198, 247]. The number of multimers within the complex depends on the individual secretin. Within the *Salmonella typhimurium* SPI-1 T3SS scanning electron microscopy revealed 20 and 21 fold rotational symmetry within the basal components of the secretion system, including InvG (the secretin) [197]. However within *Yersinia enterocolitica* plasmid encoded T3SS, the YscC proteins show a 13 fold angular symmetry [198]. The PulD secretin of the type II secretion system from *Klebsiella oxytoca* shows a 12 fold symmetry [171], a feature which it shares in common with the type IV pilli secretin of *Neisseria meningitidis* [247, 248]. As is the case for the secretins of T2SSs, some T3SS secretins require an



associated pilot lipoprotein in order to stabilise themselves in the outer membrane. Thus far only three T3SS secretin associated pilot proteins have been investigated those being YscW, MxiM and InvH from *Yersinia enterocolitica*, *Shigella flexneri* and *Salmonella typhimurium* respectively [249-251]. The exact mechanism of interaction between the secretin and its pilot protein remains unclear. For several secretins from both T2SSs and T3SSs, the interaction between secretin and pilot is mediated by the C-terminal portion of the secretin [172, 173, 199, 251]. In contrast C-terminal deletions in YscC did not inhibit its interaction with YscW [249]. In the absence of interaction (or absence of the pilot altogether) the secretin oligomerises and localises in a much slower fashion [249-251].

### **1.3.3. Directing secretion: T3S needle and translocon**

#### **1.3.3.1. The needle and associated proteins**

Outside of the cell there are several components required for the function of T3SSs. In contrast with proteins found within the cell the degree of similarity between homologous proteins is much lower. This may have something to do with their extracellular location making them exposed both to host immune systems but also to other organisms which may seek to exploit the proteins as receptor molecules for infection (e.g. bacteriophage). These proteins include the needle protein SctF, which forms a large multimer, some 80nm long in *Yersinia* [252]. The type-III secretion needle sits external to the bacterial cell outer membrane, and acts as a channel through which proteins destined for the target cell travel. The needle formed by SctF proteins may also function as a regulator of the secretion apparatus. Recent work on the NF-T3SSs of *Yersinia* and *Shigella* demonstrate a possible role for the needle as a signal transducer and controller of secretion [253, 254]. In both cases mutagenesis of the

needle protein (YscF in *Yersinia pestis* and MxiH in *Shigella flexneri*) led to the production of T3SSs which did not secrete effector molecules in a normal manner.

In common with most NF-T3SSs, the *Y. pestis* NF-T3SSs, once assembled, does not constitutively secrete molecules into the external milieu, instead it requires certain conditions to be present before secretion will commence. In the case of *Y. pestis* and other *Yersinia* species, the key condition required for secretion is the presence of low levels of calcium, a fact discovered over twenty years ago [255]. No sensor for the external condition has ever been determined. However, certain mutants of YscF required much higher concentrations of calcium than would normally be required in order to inhibit secretion [256]. Similarly there were also mutants for which secretion was constitutively on, and also mutants which did not secrete under any circumstances [253]. A comparable situation was observed with MxiH, where some mutants were secreting effectors constitutively, but increases could still be induced using the artificial activator of *Shigella* type-III secretion, Congo red [254]. Other mutants became unresponsive to Congo red, and either constitutively expressed effectors or were non-secreting [254]. Both of these examples suggest that the needle senses external stimuli ( $\text{Ca}^{2+}$  concentration in the case of YscF, Congo red in the case of MxiH), and transduce this information to the secretion apparatus, causing the apparatus to 'switch on' and begin secretion of effector molecules.

The assembled needle filament constructed from SctF family proteins shows a helical structure [257, 258]. Analysis of the packing of the monomers (MxiH) into the *Shigella flexneri* T3SS needle also showed that there were extensive interactions formed between subunits within the needle, and that it was this interaction which may be able to mediate transduction of the signal through the needle [258]. In support of

this theory mutations which affect signalling in *Shigella* and *Yersinia* needles are located in the region of the SctF proteins responsible for their putative interaction [253, 254, 258].

Whilst for some T3SSs the SctF is the only protein which extends from the basal apparatus, there are other systems which use additional proteins to create a longer needle for translocation of effector proteins. This includes EspA and its homologues, a filament forming protein found in *E. coli* and closely related species [259]; and HrpA and its homologues, a major pilus unit protein which forms the Hrp pilus [260, 261], a structure found in many of the NF-T3SSs of phytopathogenic bacteria. Bioinformatics analysis has shown homology between EspA and regions of flagellin [219], and structural analysis has shown homology between these two proteins and the needle protein MxiH. All three proteins form hollow tubes with a similar helical architecture [258], suggesting that each of these proteins shared a common ancestor which diverged to fulfil the specific functions required of each protein.

Where no additional filament proteins are to be found attached to the end of the needle, then there is an additional protein which functions to tightly regulate the length of the needle. This regulatory function is performed by the SctP proteins. As with SctF proteins there is little similarity between members of the SctP family, and their assignment to this protein group has been mostly inferred from functional analogy, rather than sequence homology. In all cases, mutations within the protein causes a deregulation of the length of the needle [262-264], or in the case of the flagellum, where the protein FliK performs the analogous function, the length of the flagellar hook [265]. The means by which they achieve this process may well be different between the flagellar and non-flagellar systems.

Within NF-T3SSs SctP proteins seem to act as molecular rulers, determining the length of the finished filament during its assembly, switching off the export of its monomeric component once the needle has reached a desirable length [264, 266, 267]. In support of this hypothesis, truncations in the YscP protein of *Yersinia enterocolitica* produce a shorter YscF filament [264]. In contrast, truncations within FliK produce a longer hook filament [265], however longer FliK proteins also create longer hooks [268]. Regardless of their method of length regulation Both SctP and FliK proteins appear to have a conserved binding domain for the substrate specificity determination protein FlhB/SctU [234, 235, 269]. This interaction is determined by the C-terminal region of the SctP/FliK proteins [269].

Recent work by Cornelis *et al* has led to the proposition of a Type-III Secretion Substrate Specificity Switch (T3S4), within the last 120 or so amino acids of the protein, as determined by deletion analysis of the YscP protein [270]. They also suggest some degree of conservation of this region within FliK, and possibly even Spa32 & InvJ [270]. However, a chimeric protein of YscP with its T3S4 domain replaced with that of FliK, when expressed in a *yscP*- strain was unable to complement the knock-out [270].

#### **1.3.3.2. The translocon - proteins that put the tip on the needle**

On the top of the needle sits the translocation apparatus, the proteins which form the pore within the host cell, and allow for the final stage of a proteins journey from the bacterial cytosol into the target cell. There is little homology to be found between members of this group of proteins, and even using PSI-BLAST it is difficult to find any homology between the translocation proteins of even closely related T3SSs. One common theme that has recently emerged for proteins which form the translocon of

NF-T3SSs is the requirement of cholesterol in the target cell's membrane [271]. Studies on the related translocation proteins SipB from *Salmonella*, and IpaB from *Shigella*, show that they bind cholesterol with high affinity [272], and in its absence the T3SSs of which they are members are unable to translocate effector molecules into the host cell [272]. Similar effects were found for the translocation apparatus of enteropathogenic *E. coli* (EPEC) [272], however, the protein within in the EPEC translocon which binds cholesterol has not yet been determined.

In *Yersinia* there are three main proteins which form the translocon: YopB, YopD and LcrV [273]. LcrV is required for the secretion of YopB and YopD to form the tip of the translocation apparatus [274]. The structure of LcrV is known, and shows a degree of similarity to needle proteins [258, 275]. The data also suggest that LcrV oligomerises at the tip of the needle interacting with the tip of the needle formed by SctF. However, the LcrV protein is not found in all systems, and so it is interesting to note that in its absence homologues of YopB and YopD are able to bind to the tip of the needle in order to create a functional translocon [276].

#### ***1.3.4. Control of apparatus and effectors: Regulators and chaperones***

The regulators of type-III secretion are, unsurprisingly, very important for expression of system at the right time. As an important factor in eukaryotic cell interaction, T3SSs should be rapidly activated when conditions are favourable, and conversely as a system that has high energy requirements, should not be expressed when there is no need, this is especially important where there are other processes or organelles within the bacterium that are highly energy dependant. One such example of this is within bacterial species where there is more than one T3SS present. In such cases there is

often cross talk between the two systems to ensure that they are not expressed at the same time [277].

One of the key themes in the regulation of T3SSs is the necessity to respond to environmental cues. Obviously, since the cell only wants to express the secretion system at times which are beneficial to it, responding to the right environmental cues is very important. As such there are a wide variety of stimuli used by different pathogens. Some of the recurring common stimuli include temperature, cation concentration (in particular  $Mg^{2+}$  and  $Ca^{2+}$ ), acidity, presence of bacteria from the same species (quorum sensing), and host cell contact [245].

Within the cell these changes in external environment affect proteins involved in transcriptional regulation, and often many members of different regulational families will act together to influence the transcription of NF-T3SS genes. Even within one T3SS this may involve proteins representing: Two component regulators, AraC like transcriptional activators, nucleoid-binding proteins and even molecular chaperones [278-281].

Chaperones function in a multitude of roles which ensure the delivery of proteins to the completed secretion apparatus. In fulfilling this role chaperones must be able to stabilise proteins, keep them from interactions with other proteins and molecules within the cytoplasm, and maintain them in a secretion-competent state. The chaperones of the T3SS fall into three main classes [282]. Class I chaperones are generally chaperones to effector proteins, and can be subdivided into two different subgroups, those which chaperone only one effector (class IA), and those which chaperone two or more effectors (class IB). Class II chaperones are usually chaperones to the translocators. Finally there are the class III chaperones, which

chaperone proteins secreted by the flagellar T3SS. Despite the fact that chaperones within each group (IA, IB and II) tend to have a distinct function and also genomic context (for example type IA chaperones are almost always encoded by a gene adjacent to that which encodes the effector they chaperone [283]) there are exceptions to this classification. For example, LcrH, a class II chaperone, binds the translocation apparatus proteins YopB and YopD within *Yersinia* [284, 285]. However, within *Chlamydia* LcrH interacts with the effector YopN, which in *Yersinia* is chaperoned by a complex of SycN and YscB [202]. This suggests that the interactions between chaperones and effector/translocator proteins is not a simple one, and that there is possibly some degree of functional redundancy within each class of chaperones which allows them to act as chaperones to proteins not normally associated with the class to which they belong.

Despite the important role that chaperones provide in protecting and trafficking proteins to the secretion apparatus, they are not essential to the functioning of the whole secretion system. There are, for example, several proteins which are exported by T3SSs which do not require a chaperone, such as the effector protein YopM of *Yersinia pestis* [286]. Other proteins meanwhile do not need to be maintained in an unfolded state for export through the T3SS machinery [287]. As Parsot *et al* [283] comment, if chaperones are not essential for export of some proteins, why should they be for any. In response to this question they hypothesise that chaperones may provide a hierarchy for secretion, ensuring that the right proteins are delivered to the secretion apparatus at the right time. For example, there would be no point exporting effector proteins through the secretion apparatus before the translocon has been formed. Hence, it would make sense for translocation proteins to be secreted first, and only after this event, to allow the export of effector proteins.

### ***1.3.5. Communicating with the host: effector proteins***

The final class of proteins required for a NF-T3SS to have an effect on eukaryotic cells is effectors. Effectors are the proteins that are transported through the T3SS apparatus into the target cell. As such it should come as no surprise that the effectors of NF-T3SS show a massive range of diversity, since each effector is designed to fulfil a role within each bacterium's ecological niche. This diversity in function means that there is little similarity to be found between effector proteins, with the exception of those which are known to perform the same function in host cells.

The search for effectors within bacterial genomes is also hampered by the fact that the genes encoding can also be found outside of the locus encoding the structural components of the apparatus [20, 43, 288-292]. There is also no known consensus signal sequence which targets effector proteins to the T3SS. There has been some suggestion that the signal is mRNA based [293], however, this is at odds with other observations that suggest an N-terminal amino acid signal sequence, within the first 10-15 amino acids, is responsible [294, 295]. To date there have been several hypotheses put forward as to what exactly constitutes a type-III signal sequence. This includes a requirement for high numbers of serine and low numbers of aspartate within the first 50 residues [288], the requirement for an amphipathic sequence of residues at the N-terminus [296], or simply the need for an unordered N-terminal sequence, to allow for recognition by chaperones in a similar manner to GroEL [296]. The range of effects that can be brought about by T3SS effector molecules is wide ranging, but unified by their efforts to interact with host cells in an attempt to hijack host cells processes and machinery in order to gain an advantage. In general T3SS effector molecules can be thought of as fulfilling one of several main functions:



Cytoskeletal alteration, immune system subversion and vesicular trafficking.

## 1.4. Viewing Type-III Secretion in an Evolutionary Context

Type III secretion systems are found in a wide variety of bacteria. Non-Flagellar systems have thus far been identified in a wide range of Proteobacteria, and Chlamydia [108, 297-301]. Flagellar systems meanwhile have been located in no fewer than six bacterial phyla: Aquificae, Firmicutes, Planctomycetes, Proteobacteria, Spirochaetes and Thermotogae [302]. Many of the NF-T3SSs identified to date lie in single pathogenicity islands, either within the bacterial chromosome(s) or on a plasmid, making them amenable to rapid identification and easy analysis. However, there are several systems where the system is broken in two (e.g. ETT2 from *Escherichia coli* and CPI-1 of *Chromobacterium violaceum* [303, 304]), or even more pieces (as is the case for all the Chlamydial T3SSs thus far identified).

With the exception of Chlamydial NF-T3SSs there is also evidence that the T3SSs are of foreign origin, such as differing GC content & codon bias from the rest of the genome, and absence in closely related bacterial species [305]. This evidence points to T3SS being horizontally transferred into the hosts in which they now lie. Attempts to reconcile differences in the 16s rRNA trees and trees of T3S proteins by looking for evidence supporting horizontal gene transfer events demonstrate that such events took place on multiple (at least six) occasions, and mostly on internal branches on the tree [306]. T3SSs also seem to cluster into distinct groups when looked at phylogenetically. There are at least five major groups of T3SSs as defined thus far, each with its own prototypical member [297]. The five groups (with prototypical

members show in brackets) are: Ysc (*Yersinia pestis* plasmid system), Inv/Mxi/Spa (*Salmonella* SPI-1 system), Esc (*E. coli* LEE system), Hrp1 (*Pseudomonas syringae*) and Hrp2 (*Xanthomonas campestris*). T3SS effectors have also been seen to transfer between bacteria horizontally in methods separate from the transfer of the main apparatus [20, 43].

The degree of similarity between NF-T3SS systems vary greatly, as does the similarity between Flagellar and Non-Flagellar systems. There is also evidence of recruitment of proteins which fulfil related and distinct roles within the bacterium, such as the ATPase, which is related to the F-type ATPase required for ATP generation from proton gradients, and the Secretin which is found in other secretion systems beyond type-III. There are also those proteins which show little similarity between T3SS systems, and proteins which are a requirement of secretion in one T3SS but are absent from other functional T3SSs. All of these aspects along with the strong evidence for horizontal transfer in the spread of the apparatus and its associated effector molecules make this system an ideal target for examination using the wide variety of computer based techniques available for analysing the large number of whole bacterial genome sequences available in public databases.

## **1.5. Sequence – Structure – Function Relationships in Type-III secretion proteins**

As discussed earlier in section 1.1.6 the mapping between a proteins sequence, the structure it forms and the function it fulfils are not precise. Within T3SS proteins there are numerous examples just such incongruences. The vast majority of proteins which form the core structure of T3SSs show a good degree of similarity across NF-

and F-T3SSs to the extent that they can be readily identified by standard homology searches [307]. However, identification of homology alone can cloud the picture, as there are several multi-domain proteins within T3SSs (e.g. SctC), non-complementary protein families (SctV), and proteins with no-observable sequence homology which fulfil identical functions (SctP) [207, 241, 270]. Such examples serve to show that homology does not necessarily imply replaceability or identical functionality.

As such it is necessary to apply an understanding of the structure of the proteins which form the T3SS apparatus. By assessing sequence similarity in conjunction with knowledge of domain architecture, protein structure and functional motifs improved assessments and assignments of homology can be made. For example, whilst no structures currently exist for the needle length regulator proteins (SctP family) it has been posited that this family retains a conserved secondary structure in the domain which controls needle subunit secretion (and thus also needle length), which is not detectable by standard homology searching techniques [270]. Similarly several needle tip (e.g. LcrV and IpaD) and needle extension (e.g. EspA) proteins show a good degree of structural similarity despite of their very low sequence similarity [308]. Solved structures also provide a great deal of information which cannot be provided by analysis of a protein's sequence alone. For example: They can provide data on the interaction regions which hold proteins within the T3SS apparatus together, such as the T3SS needle, or chaperone binding sites on the ATPase [216, 309]. They can also show the macromolecular structure formed from such interactions, and the number of subunits present within the apparatus [197, 200, 310].

## **1.6. Aims**

T3SSs can be thought of as a collection of a series of proteins, some of which are

conserved throughout different systems (outside of T3SSs), others which can be identified as having clear homologues just within T3SSs, and others which are just conserved within a few or even just an individual T3SS. Given that there are a good number of structural components of NF-T3SSs which can be identified as having homologues both within other non-flagellar systems and within flagellar systems, *in silico* homology searching tools present themselves as an ideal candidate to survey T3SSs. These tools, in combination with the ever increasing body of genome sequence data available, present an opportunity to answer a number of questions about type-III secretion.

With the range of phyla containing a flagellar T3SS being much larger than that for non-flagellar systems, are Proteobacteria and Chlamydia the only two phyla that actually contain NF-T3SSs, or is this an artefact of a sampling bias in experimental work and genome sequencing? Allied to this question is that of the diversity of T3SSs, how many T3SSs are there, and where are they located? Within the range of proteins which have clear homologues throughout T3SS there are also those with homologues to proteins in other systems. The same homology searching techniques combined with phylogenetic analysis allow for an analysis of the differences between proteins found in T3SSs and their homologues in these other systems. Such analysis can also allow estimations to be made as to possible evolutionary events which lead to them being adopted or lost by different systems.

Whilst homology searching using conserved proteins amongst NF-T3SSs allow us to survey breadth, they can also provide insight into diversity. By exploiting the fact that most of the major components of NF-T3SSs are encoded in single or very few loci, all the proteins which form a particular NF-T3SSs can be located by finding these loci

through the presence of the highly conserved components within them. By taking this approach the degree of conservation of various protein families amongst NF-T3SSs can be determined, along with the complete breadth of proteins involved in NF-T3SSs as a totality.

Finally, homology searching and genome sequence data also allows for analysis of the proteins which are least well conserved amongst NF-T3SSs: effectors. Their lack of conservation makes the job of locating them by homology searches harder, and given the evidence that they can be found not only within loci which encode the T3SS apparatus but also elsewhere on the chromosome, the search space cannot be narrowed down to just small regions of genomes. In the case of effectors, however we can also supplement homology searches with lab-based assays to confirm or refute any data obtained from *in silico* sources alone. Through such analyses we can help determine the true repertoire of effector proteins present in particular bacteria, and through examination of their genomic locale posit ideas as to how these effectors were inherited by the bacterium.

## CHAPTER 2 - SPECIFIC T3SS COMPONENTS

### 2.1. Introduction

#### ***2.1.1. NF-T3SS components shared with other systems***

There are numerous components of NF-T3SSs which are shared with F-T3SSs. However, there are very few components found in T3SSs for which homologues are known within other systems, with two fairly well characterised exceptions: The ATP binding component of T3SSs SctN, which is homologous to the  $\alpha/\beta$ -subunit of ATP synthases [208, 210, 311], and the outer membrane secretin SctC, which is homologous to proteins with known equivalent function within Type-II secretion systems, and the closely related Type-IV pilus system [246].

The secretins in Type-II/Type-III secretion systems form highly stable multimeric complexes which resist breakdown even in boiling sodium dodecyl sulphate [171, 312]. These multimers anchor in the outer membrane to form a pore within it in order to facilitate export of proteins [171, 198, 247]. The number of monomers within the complex vary between different systems, with multimers consisting of between 12 and 21 subunits being so far described in the literature [171, 197].

There are other differences between different secretins aside from different monomer counts within the assembled multimer, such as the requirement for a pilot lipoprotein to assist formation and localisation of the secretin multimer [173, 249, 251]. The pilot protein is not universal to all secretins and there several examples of systems, such as the LEE system in *E. coli*, where no pilot protein has been found associated with the secretin.

All secretins show a common C-terminal domain, which is thought to be responsible for multimerisation [246, 313], but have a unique N-terminal region which confers subject/system specificity [313, 314]. However, little is known about the differences between the different N-terminal regions of secretins and how this affects the specificity of the protein, and also to see if this region of the protein could play a role in the requirement for stabilisation by a pilot protein. With the presence of a common domain with which to locate secretins from different secretion systems it becomes possible to apply a bioinformatics approach to determine the breadth of different secretins in sequenced genomes and from there discern the differences in N-terminal regions, domain architectures, and suggest models to describe the evolution and inheritance of secretin proteins by different systems.

### ***2.1.2. The machinery of ATP synthesis and utilisation***

As stated above SctN is homologous to the catalytic subunits of the ubiquitous ATP synthase. The F-type ATP synthase is a membrane bound complex, which in *Escherichia coli* is assembled from eight different proteins:  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ , a, b and c [315]. These eight proteins come together to form two major parts:  $F_o$  (consisting of a, b and c subunits) and  $F_1$  (consisting of  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$  and  $\epsilon$  subunits). The  $F_o$  part is membrane bound, and all its subunits contain transmembrane domains anchoring them into the membrane [315]. Conversely the  $F_1$  part is located in the cytoplasm, and contains the proteins responsible for ADP/ATP binding and catalysis [99, 316, 317]. In the F-type ATP synthase the proton gradient between the two sides of the membrane in which the  $F_o$  lies allows for movement of protons through the a- and c-subunits, causing a rotation of the c-subunits, (which exist in a stoichiometry of  $c_{9-12}$  within the complex), relative to the stator provided by the a-subunit [318-320]. The  $F_1$

part is anchored to the  $F_0$  part by the interaction of the b-subunit from  $F_0$  and the  $\delta$ -subunit from  $F_1$ . The  $\gamma$ -subunit links to the centre of the c-subunits and rotates as well, altering the conformation of the static  $\alpha$ - and  $\beta$ -subunits as it turns [317]. This change in conformation is as part of the three step process: Open – where molecules of the ATP/ADP +  $P_i$  can be released bound, Loose – where molecules are bound to the catalytically inactive subunit and Tight – where the protein becomes catalytically active and the bound molecule is converted [99, 316]. The  $\alpha$ - and  $\beta$ - subunits exist in a  $\alpha_3\beta_3$  stoichiometry and produce 3 molecules per rotation of the  $\gamma$ -subunit, by the action of the  $\beta$ -subunits alone (the  $\alpha$ -subunits can bind ADP/ATP, and function in a regulatory capacity, but remaining catalytically inactive) [99-101]. The final component of the F-type ATPase is the  $\epsilon$ -subunit, which is also required for coupling of the proton motive force to ATP synthesis, both through its action in linking the  $\gamma$ - and c-subunits, but also through conformational changes which it triggers within the  $\gamma$ -subunit itself [321-323].

However, the first role ascribed to the  $\epsilon$ -subunit was in regulation of the F-type ATPase. When in contact only with  $F_1$  subunit components it acts as a potent inactivator of ATP catalysis, this function is then counteracted by the binding of the  $\epsilon$ -subunit to  $F_0$  components, resulting in activation of the whole F-type synthase [324-329]. The  $\epsilon$ -subunit also has differential effects in inhibiting ATP synthesis versus ATP hydrolysis, suggesting that its regulatory role extends to controlling switching between the two modes of function available to F-type ATPases (ATP synthesis /  $Na^+/H^+$  Transport) [330, 331].

Within the related vacuolar type (V-type) ATPases there are also a number of additional components which form the functional system. In the case of V-type



systems, the system functions only as hydrogen or sodium ion pump, and not as a generator of ATP through the use of proton motive force. These additional components found only in V-type ATPases include the H-subunit, which is required for the anchoring the E-subunit to the complex [332]. The E-subunit in turn binds the G-subunit, which is homologous to the b-subunit of F-type ATPases [333, 334]. By comparison the b-subunit is bound to the F-type ATPase complex via it's interaction with the a-subunit [335, 336].

### ***2.1.3. Aims***

The bacterial flagellum and non-flagellar T3SSs share numerous common components. As a result of this it is possible to make inferences on the function of NF-T3SS proteins where information is known about the flagellar homologue. In the case of the two examples outlined above it is also possible to examine the proteins in light of the evidence available from other systems. With this in mind this chapter sets out to examine what information could be gleaned from an examination of T3SS proteins in tandem with those from other related systems. There is already some evidence of homology existing between FliH/SctL proteins and other (non-catalytic) subunits of F-type ATPases [204, 337], and thorough use of homology searching tools should be able to place statistical backing behind these claims. Based on the evidence that there are two separate F-type ATPase components involved in T3SSs there is a strong likelihood that there are further comparisons to be made with homologous components of the V-type and A-type ATPases and FliI/FliH. Secondly, within proteins containing a secretin domain there exists an opportunity to examine where T3SS secretin proteins sit in relation to similar proteins from other systems. With the number of different types of systems which contain proteins with secretin domains,

the assumption is that different domain architectures define the ability of a protein to function within a particular system. As such the methods employed here set out to survey the relationship between domain architecture of secretin proteins and the systems in which they function, to see how well the two correlate. The data obtained will also be used to examine patterns of evolutionary changes and inheritance which may be associated with changes in domain architecture. These changes in domain architecture may also alter the interactions of these proteins, and so one may also expect to see sequence and/or domain differences between those secretins where a pilot protein has been shown to be required compared to those where no such protein has been identified.

## **2.2. Methods**

### ***2.2.1. BLAST and PSI-BLAST***

Searches were performed using NCBI BLAST version 2.2.14 for Linux, and searched using the Non-Redundant (NR) database of proteins also available from the NCBI (Downloaded November 2005 from <ftp://ftp.ncbi.nih.gov/blast/db/>). For searches performed on the NCBI website, returned hits were filtered to Bacteria only using the phrase “Bacteria [orgn]”. BLASTS were performed with the filter off, and using the BLOSUM62 matrix. PSI-BLASTS were performed using the same starting conditions as BLAST searches, and were run until convergence, or for a maximum of ten iterations, where convergence had not been reached previously (approx 3% of cases). These fairly relaxed search criteria were chosen due to the relatively small data set being used, and the expected low number of results expected from the searches. Given this the search results can be examined manually to remove false-positives.

### ***2.2.2. Alignments***

Alignments were performed using T-coffee using the default parameters, or where domain information from PFAM [338] was available using the hmalign program, and presented using the CHROMA package.

### ***2.2.3. Domain Searching***

Domain searches were performed with domains from release 20 of the PFAM dataset. Including models for the ATPsynthase b- and  $\delta$ - subunits (PFAM accessions PF00430 and PF00213 respectively), and the secretin domain (PF00263). Pfam\_ls and Pfam\_fs versions of these domains were searched against a database of bacterial proteins assembled from the NCBI prokaryotic dataset downloaded on November 2005, using hmmsearch from the HMMER package version 2.3.2 [339]. Additional domains were located within secretin proteins using the complete PFAM dataset (release 20) [338] and hmmpfam from the same version of the HMMER package as above.

### ***2.2.4. Phylogenetic trees***

Phylogenetic trees were created using clustalw (which utilises a neighbour-joining method) from alignments created as per section 2.2.2, ignoring gapped columns in the alignment and bootstrapping using 1000 replicates. Trees were drawn using the MEGA 4 package [340].










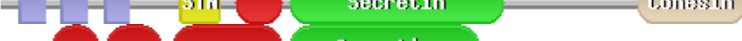










## **2.3. Results**

### ***2.3.1. Diversity of N-terminal domains in Secretin proteins***

A search of bacterial proteins using the PFAM secretin domain reveals hits to a total of 365 proteins with an e-value < 0.05. These proteins were then searched for any

other domains using the complete PFAM dataset to locate all domains within these proteins so that a set of unique domain architectures could be found. After filtering out domain fragments and overlapping domains (domains which do not encompass the full domain model, and domains which lie within other smaller e-value domains respectively) a list of 20 unique secretin containing domain architectures were left (See Table 3).

When the secretin domains of each of these proteins are aligned and used to produce a phylogenetic tree, there is little congruence between the domain architecture and location within the tree. Little can be read into this, as the bootstrap values on many of the internal branches of the tree are very low, and so the distribution of domain architectures around the tree could be just be artefactual. In order to simplify the analysis proteins were grouped into those belonging to Type-III secretion systems, those belonging to Type-II secretion systems/Type-IV pilli, and those belonging to neither. All assignments were done based on the genomic locale of the protein in question. Of the 365 secretin proteins, 61 could be identified as belonging to type-III secretin systems, and 217 could be identified as belonging to type-II secretion/type-IV pillus systems. Those identified as belonging to type-III secretion systems could be group into three distinct domain architectures, with the number of each architecture given in brackets: Secretin\_N – Secretin (1), Secretin\_N – Secretin\_N – Secretin (50), Secretin\_N – Secretin\_N – Secretin\_N – Secretin (10). The one protein containing a Secretin\_N - Secretin domain architecture (gi: 76581903) is the result of a frameshift in the genomic sequence (*Burkholderia pseudomallei* 1710b chromosome II), and comparison of the DNA sequence to *B. pseudomallei* strain K96243 reveals that the region containing the secretin proteins is otherwise identical, and without the frameshift would produce a protein with a Secretin\_N – Secretin\_N – Secretin domain architecture.

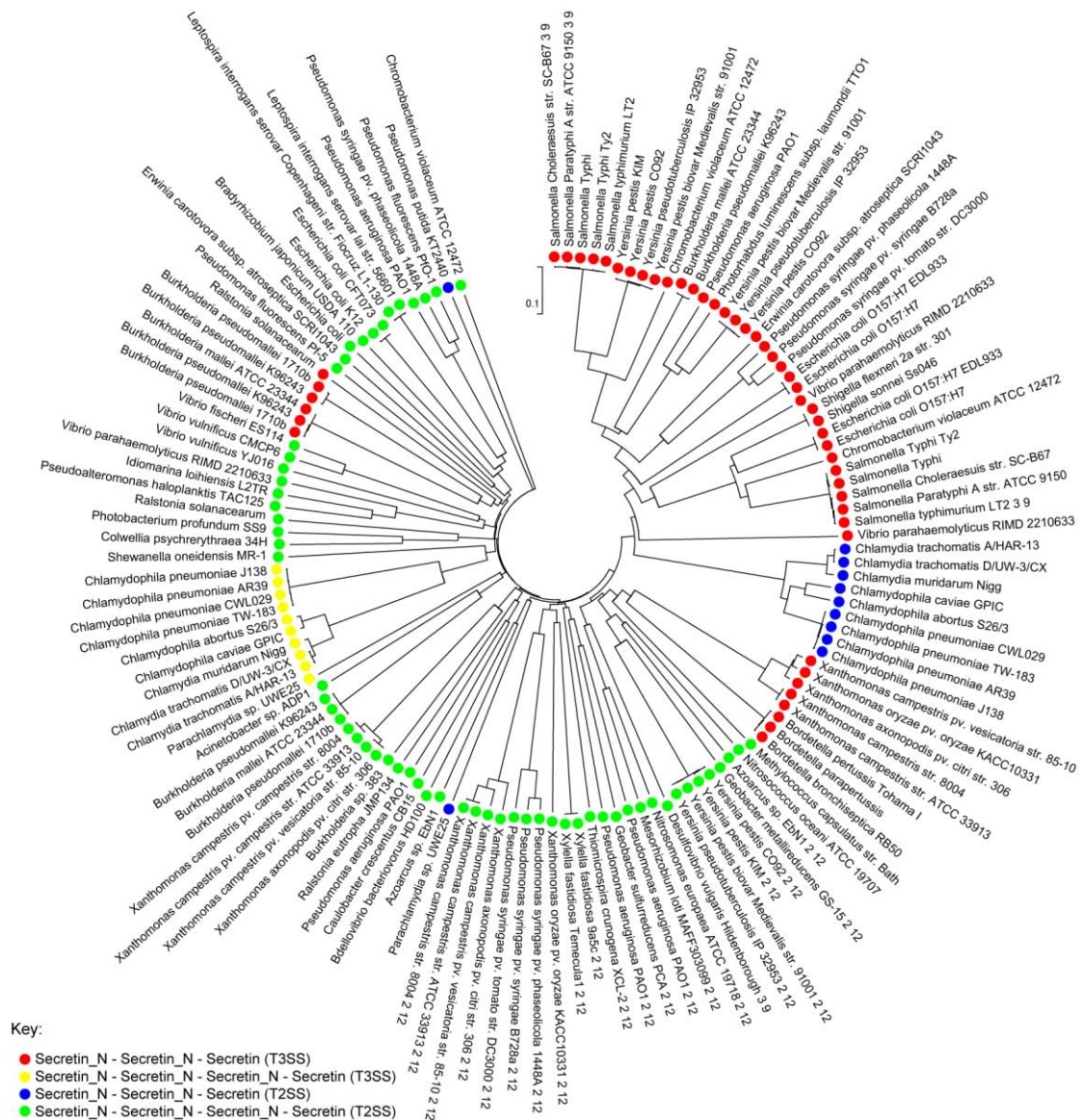
| Domain Architecture  | N  | Example   |
|--|----|---|
| STN – Secretin_N – Secretin                                | 83 |    |
| (Secretin_N) <sub>3</sub> – Secretin                       | 67 |    |
| (Secretin_N) <sub>2</sub> – Secretin                       | 64 |    |
| BON – Secretin   | 46 |    |
| Secretin   | 39 |    |
| STN – Secretin_N_2 – Secretin                              | 17 |    |
| Secretin_N – Secretin                                      | 12 |    |
| Secretin_N_2 – Secretin                                    | 10 |    |
| (TPR) <sub>n</sub> – STN – Secretin_N – Secretin           | 7  |    |
| (TPR) <sub>n</sub> – STN – Secretin_N – Secretin – Cohesin | 4  |    |
| SPOR – (Secretin_N) <sub>3</sub> – Secretin                | 3  |    |
| (Secretin_N) <sub>4</sub> – Secretin                       | 3  |    |
| STN – (Secretin_N) <sub>3</sub> – Secretin                 | 2  |    |
| STN – Secretin   | 2  |    |
| STN – (Secretin_N) <sub>2</sub> – Secretin                 | 1  |   |
| (STN – Secretin_N_2) <sub>2</sub> – Secretin               | 1  |  |
| (Secretin_N) <sub>7</sub> – Secretin                       | 1  |  |
| STN – STN – Secretin_N – Secretin                          | 1  |  |
| Zot – Secretin   | 1  |  |
| TPR_2 – Secretin   | 1  |  |

**Table 3. List of domain architectures for proteins containing a secretin domain**

List is organised by occurrence of domain architecture, figures were drawn based on representative members of the architecture using the PFAM[338] domain image generator. Each rectangle/lozenge represents a domain in the order as it appears in the domain architecture description. Domains of the same type are coloured identically. Domains not named in images: Red – Secretin\_N, blue/cyan – Secretin\_N\_2, purple – TPR repeat, brown – Zot

The type-II secretion/type-IV pilli systems encompass most (18 out of 20) of the different domain architectures represented in the above table. The spread of domain architectures between those annotated as belonging to type-II versus type-IV pilli shows some interesting differences. The STN – Secretin\_N – Secretin architecture is commonly seen in proteins annotated as type-IV pilli, but rarely seen in those annotated as belonging to type-II secretion. Conversely, the domain architecture Secretin\_N – Secretin\_N – Secretin\_N – Secretin is common to type-II secretion annotated proteins. As with the phylogenetic tree produced for all secretins there is little congruence between branch order and domain architecture for a tree of all T2/T3SS proteins, but again the bootstrap values on many of the branches were very low. What was possible however, was production of a phylogenetic trees consisting of type-III secretion proteins, and type-II/type-IV pilli proteins with domain architectures limited those found in type-III secretion proteins (i.e. Secretin\_N – Secretin\_N – Secretin and Secretin\_N – Secretin\_N – Secretin\_N – Secretin). The bootstrap values on such trees are much better, and the domains and secretion system types to cluster together well. For the tree of type-III secretion systems the two types of domain architecture cluster into two monophyletic groups (See Figure 4). The group containing all the Secretin\_N – Secretin\_N – Secretin\_N – Secretin domain architecture are all from the Chlamydiae phylum, whilst all other T3SS secretin proteins which are from Proteobacteria have only two Secretin\_N domains. What is more interesting however, is the comparison of type-II and type-III secretin proteins (Figure 5), which when placed on the same tree leads to clustering occurring between proteins of the same architecture rather than by the secretion system to which they belong. This analysis also shows that the Chlamydiae proteins demonstrate a reversal in domain architectures between type-II and type-III secretion systems compared to





**Figure 5. Phylogenetic tree of T2SS and T3SS secretin proteins**

Tree drawn by clustal from a HMMER alignment of the secretin domain of each protein. In this phylogenetic tree clustering occurs between proteins of similar domain architecture, rather than those belonging to the same type of secretion system. This tree suggests multiple events leading to the difference in protein domain architectures seen between T2- and T3SSs from Proteobacteria and Chlamydia.



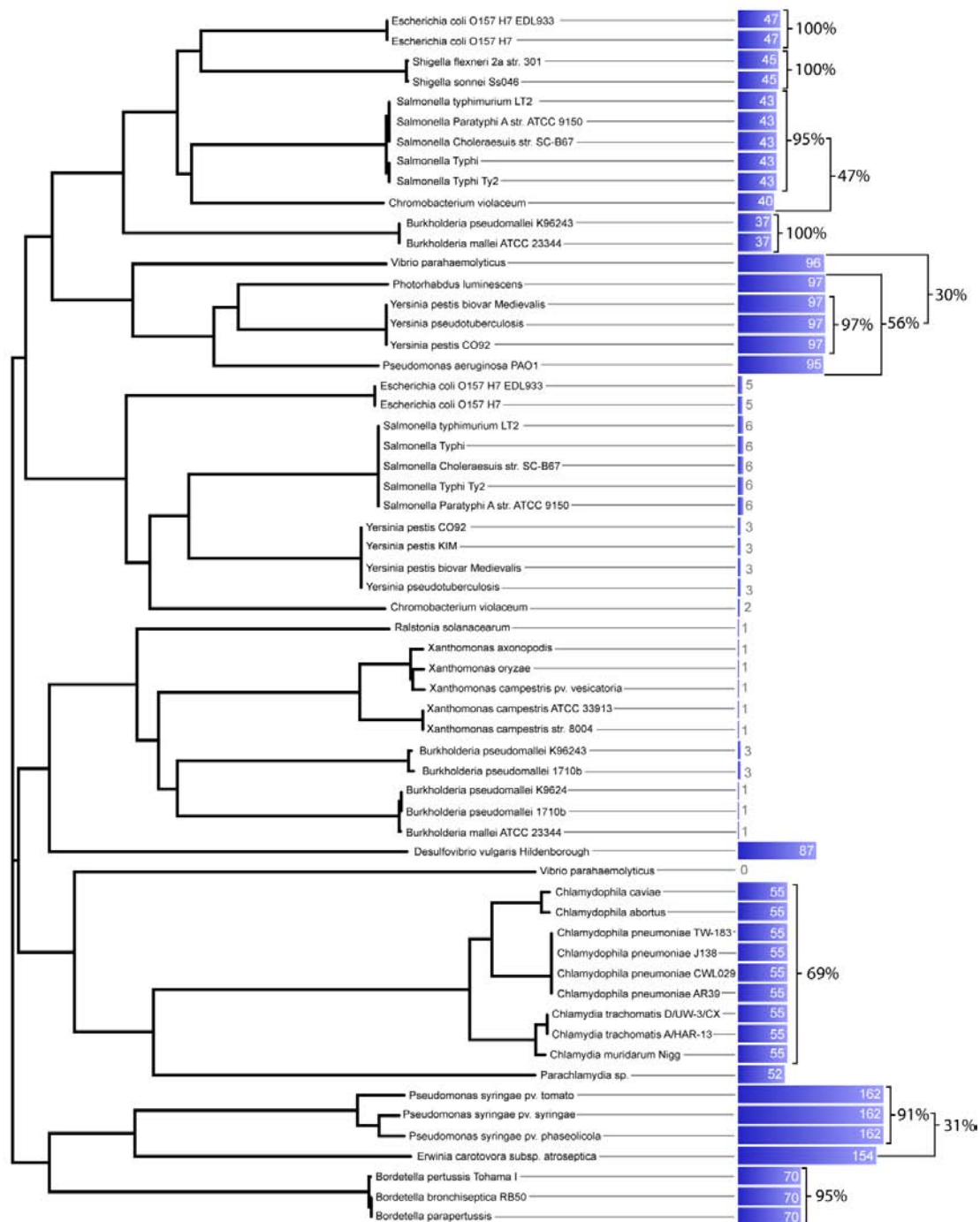
nearly all other systems. Of the four groups of proteins which exist within the tree the proteins originate from the following organisms:

- Type-II secretion proteins with 2 Secretin\_N domains: Chlamydiae phylum
- Type-II secretion proteins with 3 Secretin\_N domains: Gram negative bacteria except Chlamydiae
- Type-III secretion proteins with 2 Secretin\_N domains: Proteobacteria phylum
- Type-III secretion proteins with 3 Secretin\_N domains: Chlamydiae phylum

### ***2.3.2. Relationship between secretin C-terminus and pilot proteins.***

By examination of Secretin proteins belonging to type-III secretion systems a range of different locations for the secretin domain relative to the C-terminus are revealed. There are some proteins where the secretin domain lies immediately adjacent to the C-terminus, i.e. at the end of the protein, whilst other proteins have a region of anywhere between 37 and 162 amino acids between the end of the secretin domain and the C-terminus.

Searches of PFAM reveal that there are several PFAM-B domains which relate to this region of secretin proteins in type-III secretion systems. This suggests it may be important to the function of certain proteins to conserve this region, but not essential to the overall function of the protein, otherwise such C-terminal regions would be found in all T3SS secretin proteins, not just a subset. Mapping the number of amino acids between the end of the secretin domain and the end of the protein onto a phylogenetic tree drawn on the secretin domain (Figure 6), shows the presence/absence and length of the C-terminal region associates well with the



**Figure 6. Phylogenetic tree of T3SS secretins with C-terminal information overlaid**

Tree drawn by clustalw using an alignment created by t-coffee using the full length secretin protein. Length of blue bars proportional to the distance between the end of the secretin domain and the end of the protein. Number in/by the blue bar is the same distance in amino acid residues. Black brackets denote homologous C-terminals, percentage is the minimum percentage similarity between c-terminal regions.

There are a large number of separate C-terminal regions which show no detectable homology to each other. However, these regions are only found in groups of T3SSs which also have an identified pilot lipoprotein.

phylogenetic distribution of the secretin proteins. There are two major groups of T3SS secretins for which there is no C-terminal region: The Esc/Ssa group of T3SSs, and the Hrp2 group of T3SSs. For those proteins with C-terminal regions the levels of similarity between the C-terminal regions are much lower than the similarity between the secretin domains. Searches with any secretin domain from a T3SS protein will find all other T3SS secretin domains, but the same is not true for the C-terminal regions, where there are no fewer than ten groups of C-terminal regions which show no observable homology to each other. Previous studies have shown that the C-terminal region of secretin proteins is responsible for the binding of pilot proteins [172, 249]. With this in mind it is interesting to note that there is both little similarity between different pilot proteins, in common with the C-terminal regions to which they bind, and that the proteins which have no C-terminal region belong to systems which have no known pilot protein.

### ***2.3.3. The relationship between type-III secretion system and ATPase components***

A PSI-BLAST starting with the SctL protein YscL from the plasmid encoded NF-T3SS system of *Yersinia enterocolitica* (SwissProt entry YSCL\_YEREN) revealed a total of 121 hits in the first iteration, all of which are hits to either members of the YscL family, including several proteins annotated as NoIV in the rhizobial systems and HrpB5 in *Xanthomonas*, or FliH proteins from various flagellar systems. Of the 121, 68 hits fell above the e-value inclusion threshold (e-value > 0.005) for inclusion in the PSI-BLAST matrix. In iteration two a further 501 hits were found (220 of which were over the inclusion threshold). Of these further hits, many were also members of the FliH family and SctL family (including proteins annotated as HrpE/F

in *Pseudomonas*, *Ralstonia* and *Erwinia*). However, the most interesting hits in this iteration were to b-subunit proteins of F-type ATPases, and E-subunits of V-type ATPases originating from four different bacterial phyla: Green-sulfur bacteria, Gram-positives, chloroflexi, and spirochetes. The top hit to a b-subunit was from the bacterium *Prosthecochloris aestuarii* (e-value  $1e^{-05}$ ), while the top E-subunit hit was from *Methanocaldococcus janeschii* (e-value  $9e^{-06}$ ). A third iteration revealed hits to many more b- and E-subunits, but rendered any further iterations unhelpful due to the inclusion of several keratin related proteins in the hit list, and their effect on the PSI-BLAST matrix.

Examination of the alignments of YscL to E-subunit proteins reveals that the homology is full length between both proteins. However, the alignment of YscL to b-subunits is full length for the b-subunit, but only encompasses the N-terminal region of YscL. In order to examine the C-terminal (i.e. non-homologous to the b-subunit region) of the YscL protein in isolation, PSI-BLASTs were repeated using the C-terminal region of YscL alone (residues 115 to 223 of SwissProt entry YSCL\_YEREN), with compositional based statistics turned off, and in the first iteration 129 hits were found to a range of SctL and FliH family proteins. However, there was also plausible similarity shown between the C-terminal of YscL and several  $\delta$ -subunits from F-type ATPases from several members of the Cyanobacteria phylum, although the top hit (to  $\delta$ -subunit from *Trichodesmium erythraeum*) had an e-value of only 0.086. On the second iteration the hits to  $\delta$ -subunits from Cyanobacteria disappeared but several other new  $\delta$ -subunits appeared in the hit list. However at no point did any  $\delta$ -subunits reach a sufficient level of significance to be included in the PSI-BLAST matrix.

In order to support any conclusion of homology between the C-terminus of YscL and  $\delta$ -subunits the search was performed in reverse using the  $\delta$ -subunit from *Trichodesmium erythraeum*. Within one iteration many significant hits to other  $\delta$ -subunits from other bacteria, and the OSCP (oligomycin sensitivity conferral protein) from mitochondria were found, but more importantly also to several SctL proteins from members of the *Yersinia* genus, albeit with unimpressive e-values, thus confirming the reciprocal nature of the similarity.

#### ***2.3.4. A common protein architecture between type-III secretion and Mycobacterial ATPases***

A search of FusionDB [341] using the b-subunit as the query term (COG id COG0711), reveals a total of 19 hits to proteins which contain a fusion of the b-subunit with a different COG entry, of which well over half (12) are fusions of the b-subunit to the  $\delta$ -subunit (COG id COG0712). These hits are to three separate organisms: The two Mycobacterial genomes present in their database (*Mycobacterium tuberculosis* H37Rv and *Mycobacterium leprae* strain TN), and *Methanosarcina acetivorans* str. C2A, an Archaeal organism. A further search of all *Mycobacteria* thus far sequenced reveals that this fusion exists in all mycobacterium, but not in the genomes of other bacterium outside of the *Mycobacterium* genus. The fusion gene in *Methanosarcina acetivorans* is localised to itself and only one other member of the *Methanosarcina* genus: *Methanosarcina barkeri* str. Fusaro. The other two genomes within the *Methanosarcina* genus for which there is a known sequence show no evidence of possessing such a fusion protein.

The region encoding the fusion gene with *Mycobacterium* (all named *atpH*) also contains another copy of a gene encoding a b-subunit. This gene (*atpF*) is encoded in

an ORF adjacent to *atpH*. This produces a situation similar to the vacuolar type ATPases which contain one subunit which is homologous to the b-subunit alone: The G-subunit, and one subunit which is homologous to the b-subunit in its N-terminus, but homologous to the  $\delta$ -subunit in its C-terminus: The E-subunit.

## 2.4. Discussion

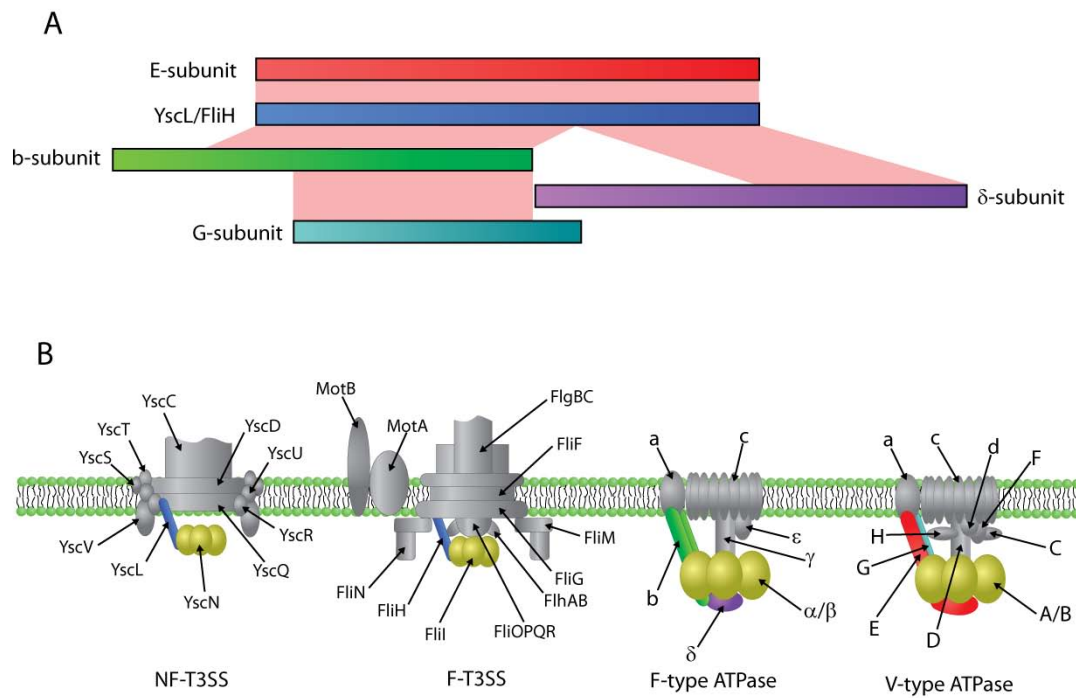
### ***2.4.1. The diverse origins of type-III secretion system proteins***

Through careful use of a range of simple bioinformatics tools it is possible to show the different origins of proteins which go towards producing a fully functional T3SS. The sequence analysis presented above demonstrates the relationship between the outer membrane secretin proteins of type-II and type-III proteins is complex. Phylogenetic trees show the possible methods of inheritance of the secretin proteins by the two groups of secretin systems is not as obvious as one might assume. Normally it would be expected that proteins from different systems would fall into separate monophyletic groups (such as is the case when comparing, for example flagellar and non-flagellar proteins), if they were inherited once per system. But in this case we see that type-II and type-III secretin proteins occur on the same branch of phylogenetic trees drawn on their sequences. Instead the two major branches of the tree are differentiated by the overall domain architecture of the proteins. This suggests multiple inheritance events have occurred for either one or both systems in order to acquire a secretin protein. The potential source of these proteins is unclear, as proteins with the domain architectures of type-II and type-III secretion systems are unique to those systems. In either case it becomes clear that the secretins of Chlamydial T3SS secretins form a special case, as the secretin domain clusters closer to those of T2SSs, and have a different domain architecture compared to secretins from all other known

T3SSs. The same is also true for Chlamydial T2SSs which cluster close to T3SSs

Analysis of the location of the secretin domain within secretin proteins, and in particular the presence of any region between the end of the secretin domain and the C-terminus of the protein gives us clues as to the potential presence or absence of a pilot protein within the T3SS. The interaction between the C-terminus of the secretin protein and pilot proteins has been established in several secretins from type-II and – III secretion systems [249, 312]. There is little similarity between pilot proteins from different T3SSs and detecting homology through sequence analysis, even for proteins which are known to fulfil the same function is often not possible, and so locating pilot proteins in new T3SSs through tools such as BLAST may not be an applicable approach. However, through use of the presence/absence of a C-terminal region in the secretin we can hypothesise whether there should be a pilot protein encoded within the T3SS locus, and so can use other information sources such as synteny and elimination of proteins which have detectable homology to other types of protein can be used to narrow down the search.

The homology shown between stator proteins of ATPases and T3SS proteins, provides evidence of the complex interplay between proteins originating from different cellular processes and systems in creation of a functional T3SS. It also provides additional information on the possible way in which proteins interact to form the assembled T3SS (Figure 7). The sequence analysis presented above adds support to previous studies which have shown similarities between YscL, FliH, b- and E-subunits. In addition to homology data, it is also known that FliH, b-subunits and E-subunits form a similar extended nonglobular structure, and all form dimers [333, 342, 343]. However, YscL/FliH lack the N-terminal transmembrane domain present



**Figure 7. Comparison of T3SS and ATPase systems**

A: Comparison of homologous subunits from different systems, red lines between proteins represent homology. B: Schematic representation of Non-flagellar and Flagellar T3SSs, F-type and V-type ATPases. Coloured proteins are the catalytically active proteins (coloured in yellow), and the homologous stator proteins (coloured in the same scheme as panel A)

Each of these homologous proteins has a role to play in interacting with a separate homologous protein or itself, and the ATP-binding subunits of all of these systems. These proteins then bind to the membrane bound components of their respective systems in order to anchor the ATP-binding proteins to their systems.



in b-subunits [344], suggesting that other proteins must be involved in mediating association of YscL/FliH to the membrane. In flagellar systems FliH has been shown to attach to the C-ring protein FliN [225, 226, 345], and the chaperone like protein FliJ, which in turn interacts with FliM, another C-ring component [225, 343, 346]. In the Yersinia non-flagellar T3SSs YscL interacts with YscQ, a homologue of FliN, a situation mirrored in the Shigella T3SS system where MxiN (YscL homologue) interacts with Spa33 (YscQ homologue) [204, 205, 223].

The similarities shown above between YscL/FliH and  $\delta$ -subunits (albeit with borderline statistical significance), presents a second interesting angle to this group of proteins. It is the  $\delta$ -subunit which is responsible for binding the  $\alpha/\beta$ -subunits to the stator in F-type ATPases, and so one can presume that the C-terminal of the homologous T3SS proteins perform a similar function. In support of this hypothesis both types of proteins bind to helical structures found in N-terminal of ATPases [337].

In F-type ATPases the genes encoding the b- and  $\delta$ - subunit are encoded by adjacent genes, and the presence of a b- $\delta$ -subunit fusion protein in Mycobacterial proteins provides an interesting example of how YscL/FliH proteins could have originated from gene(s) encoded by F-type ATPase genes. It is also interesting that in the case of the Mycobacterial ATPases there is no evidence of a separate  $\delta$ -subunit, but there is copy of the *atpF* gene, meaning that a functional F-type ATPase in this system will have a similar stator arrangement to V-type ATPases, where the stator is formed from and E-, G-subunit heterodimer, rather than the more typical b<sub>2</sub>- $\delta$ -subunit arrangement found in other F-type ATPases.

Whilst this study demonstrates the probable role of YscL/FliH in anchoring the

ATPase to the remainder of the T3SS, through a similar arrangement to the stator subunits of F-type/V-type ATPases, the actual function of the FliH remains unclear, especially in light of the fact that the flagellum remains functional in the absence of the protein [347]. This situation is not mirrored in non-flagellar T3SS, where YscL homologues such as HrpB5 (from *Xanthomonas*) and Orf5/EscL (from *E. coli*) are required for the function of the apparatus [290, 348].

This study illustrates a series of interesting concepts in analysis of proteins using genome sequence data. The availability of large numbers of sequences for ATPase proteins and T3SS proteins allows for comparison of related proteins in both systems, and enables more complex analyses to be performed which would not have been available. For example it is likely that without the large and diverse set of  $\delta$ -subunits present in protein datasets no homology would have been found between them and the C-terminus of YscL. By being able to assign homology between hitherto unrelated proteins it becomes possible to flesh out different evolutionary scenarios regarding the ancestry of these now related proteins. Similarly, it also allows predictions to be made regarding the structure and function of homologous proteins. Finally this study demonstrates that primary sequence analysis is not the only technique available for analysing protein data. The correlation between pilot proteins and C-terminal regions of secretin proteins allows further predictions about novel T3SSs to be made (i.e. the presence/absence of a pilot protein) independent of assignment by homology.

## **2.5. Summary**

The data presented in this chapter demonstrate the inheritance of several components common to multiple bacterial systems by T3SSs. Secretin proteins, found in multiple different secretion systems, including non-flagellar T3SSs contain a series of different

domain architectures. NF-T3SS secretins have two separate domain architectures, which they also share with the secretins of type-II secretion systems. Closer examination of these different domain architectures reveals that Proteobacterial NF-T3SSs contain secretins with a different architecture to Chlamydial NF-T3SS (a situation shared with type-II secretion systems).

Phylogenetic comparison of secretins (based on domains shared amongst all proteins), when overlaid with the domain architecture demonstrates that it is likely that multiple inheritance events took place amongst and possibly between type-II and type-III secretion systems, rather than domain duplication in secretins following adoption by the last common ancestor of type-II or type-III secretion systems. Examination of the region C-terminal to the secretin domain also reveals that this region is required to binding to pilot lipoproteins. The requirements for binding would appear to be quite loose based on the lack of similarity between these regions amongst T3SS secretins. Presence/absence of this region may also be of predictive value in determining whether a T3SS may contain a pilot lipoprotein.

Analysis of FliH/SctL adds another element to the list of proteins shared amongst multiple bacterial systems that exist within T3SSs. The similarities shared amongst E, G-, b- and  $\delta$ -subunits and FliH/SctL is another example of the formation of, and changes in, multi-domain proteins. The stator used in F- and V-type ATPases show different arrangements ( $b_2$ - $\delta$  versus E-G respectively), where the E-subunit fulfils the role of one b-subunit and the  $\delta$ -subunit. Since FliH/SctL is homologous to the E-subunit (and hence also to the b- and  $\delta$ -subunits) it is reasonable to assume that it can fulfil a similar function, and serves to anchor the ATP binding component of T3SSs (FliI/SctN), to the main apparatus through its interaction with FliN/SctQ.

Furthermore, evidence of fusions of *atpF* and *atpH* within Mycobacterial genomes suggests a method through which E-subunits and FliH/SctL may have formed in the past to produce these different components.

## CHAPTER 3 - ADDITIONAL T3SS EFFECTORS

### 3.1. Introduction

The traditional model of NF-T3SSs was that there were only a few effectors translocated through the T3SS apparatus for each individual system, and that those effectors were for the most part encoded in the same locus as the structural components of the secretion system (see for example the review of T3S in [196]). One of the only early examples of a T3SS with effectors known to lie outside of T3S apparatus locus was *Salmonella enterica* serovar Typhimurium where several type-III effectors had been identified within prophage or prophage remnants [20], including the gene *sopE* which is present in two copies in two separate prophage within the *Salmonella* chromosome [40, 349, 350].

Owing to their diversity and the lack of consensus signal sequence amongst T3S effector proteins it is difficult to identify novel effectors in bacteria, however a seminal paper in 2002 by Guttman *et al* [288], demonstrated the existence of a potential 38 separate effectors present within the genome of *Pseudomonas syringae*. By taking a modular approach, and assuming that the signal for type-III dependant secretion is N-terminal based, they attached the N-terminal region of candidate effector proteins to a reporter protein, the C-terminal portion of AvrRpt2. AvrRpt2 is a protein known to elicit the hypersensitivity response in plants [351]. By creating a transposon containing the C-terminal region of AvrRpt2, insertions of the transposon into proteins containing T3S N-terminal signal sequences should result in translocation of the fusion protein into host cells in a T3S dependant manner. Using this approach they were able to locate a total of 13 effector proteins in the bacterium

*P. syringae* pv. *maculicola* strain ES4326. Examination of the N-terminal regions of the secreted fusions revealed certain amino-acid biases compared to the rest of the protein (more serine residues and fewer aspartic acid, leucine and lysine residues), and this property was used to identify other related proteins in a bioinformatics screen. This screen identified a total of 38 proteins, and from this set two previously uncharacterised proteins were chosen and were both shown to be secreted in a type-III dependant manner.

In a similar study, a transposon based genome wide scan was undertaken on the bacterium *Ralstonia solanacearum*, showing a total of 30 potential effector genes present outside of the T3SS apparatus locus [289]. In this experiment, regulation by the T3SS regulator HrpB was taken as evidence of the gene encoding an NF-T3SS effector. Studies of the bacterium *Citrobacter rodentium* revealed that knocking out the gene *sepL* resulted in deregulated secretion of T3SS effectors [290]. Analysis of the secretome of this  $\Delta sepL$  strain of *Citrobacter rodentium* revealed a total of seven novel effectors encoded outside of the T3SS locus.

### **3.1.1. Aims**

These previous studies suggest that our understanding of NF-T3SS effectors can be greatly enhanced by no longer looking just at those proteins encoded within the T3SS locus. Enterohaemorrhagic *E. coli* (EHEC) and enteropathogenic *E. coli* (EPEC) both contain a NF-T3SS locus, known as the locus for enterocyte effacement (LEE), which is responsible for causing the attaching and effacing phenotype seen in both of these organisms. This locus is also present in *C. rodentium* [352-354]. Given the identification of additional effectors within *C. rodentium*, it seems likely that there may be additional effectors within the genomes of EHEC and EPEC. Previous studies

have shown that both bioinformatic and *in vitro* techniques can be used to identify novel effector candidates. This chapter sets out to see if there are additional novel effectors in the enterohaemorrhagic *E. coli* (EHEC) strain RIMD 0509952 (also known as the “Sakai strain”). This bacterium has an available genome sequence, and so bioinformatics techniques can be used in order to locate candidate effectors within the genome. Additionally, these assignments based on homology can be confirmed using *in vitro* techniques. These techniques can be used to test each candidate effector to see whether they can be secreted and/or translocated via the T3SS. This *in vitro* data can be used to test and validate the use of tools such as BLAST to correctly identify effectors and also to conclusively determine the breadth of effectors present in EHEC. It is unlikely that effector genes have been lying dormant in the genome of this bacterium waiting for the arrival of a T3SS, and so many effectors will likely show evidence of horizontal gene transfer, and so any identified effectors will be examined to identify any markers of such events taking place.

## **3.2. Methods**

### ***3.2.1. Bioinformatics analysis***

Over 300 proven or predicted effectors were collated from recent type-III secretion literature and the peptide sequences used to create a query library which was then used to search the *E. coli* Sakai genome using both BLASTP and TBLASTN, with filtering and compositional based statistics off. Relaxed search criteria were chosen based on the fact that effectors are known to show low levels of similarity to each other, and that a small database means that manual examination of the search results is feasible.

The percentage GC content was calculated for all genes in the Sakai genome, and then genes were sorted by rank order to calculate percentiles. Each identified candidate effector was localised within the genome and was compared to the coordinates of prophage boundaries as published in the Sakai genome paper [355], as well as the coordinates of O-islands as published by Perna *et al* [356].

### **3.2.2. Proteomics analysis of culture supernatant**

Two mutants of *E. coli* O157 Sakai (RIMD 0509952) were constructed,  $\Delta sepL$  and  $\Delta sepL \Delta escR$ , using the Datsenko and Wanner method [357]. The culture supernatant for  $\Delta sepL$  (constitutively on for effector secretion), and  $\Delta sepL \Delta escR$  (type-III secretion negative) were examined by Liquid Chromatography-Tandem Mass (LC-MS/MS) and a database of *E. coli* O157:H7 Sakai proteins.

### **3.2.3. Preparation of candidate effectors**

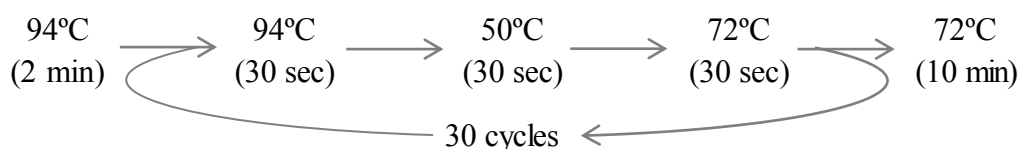
#### **3.2.3.1. Prime design PCR amplification**

Primers were designed to allow genes identified by our bioinformatics screening of the genome, to be inserted into the pENTR/D-TOPO vector (Invitrogen, California, USA). This involves adding the sequence CACC to the 5' end of the forward primer sequence, and ensuring that the 5' end of the reverse primer does not match any more than 2 bases out of the sequence GTGG, thus preventing the PCR product from cloning into the entry vector in the opposite orientation. The reverse primer also did not include the stop codon of the gene of interest, since we wished to make a C-terminal fusion to the gene. As such the primers were created as: CACC + first 18 nucleotides of the gene for the forward primer, and last 18 nucleotides of the gene (without the stop codon), unless this created a primer which could allow opposite



orientation cloning of the gene, in which case the primer was shifted back along the gene 1 codon at a time until this was no longer a problem. Primers were then tested for formation of hairpins, melting temperature etc using Primer3 [358]

Genes were amplified using 10ul each of 2nM forward and reverse primers, 0.25ul Ex-Taq DNA polymerase, 5ul 10x Ex Taq Buffer (TaKaRa, Shiga, Japan), 5ul 2.5nM dNTPs (Invitrogen), and 4ul 20x diluted DNA template. DNA was prepared from a shiga toxin cured strain of *E. coli* O157:H7 Sakai. The reaction mixture was then made up to 50ul with sterile water. Thermal cycling conditions were:



Products were then checked to be of the right size by running on a 1.5% agarose gel, and DNA concentrations determined by comparison with the Hyperladder I DNA ladder (BioLine, London, UK)

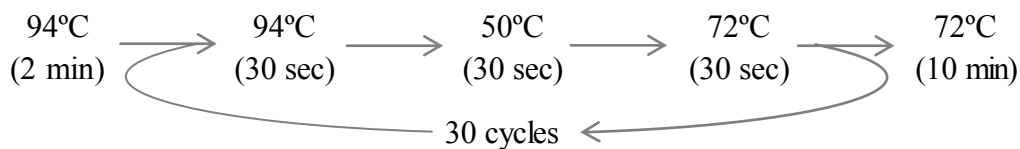
### 3.2.3.2. Transfer into gateway entry vector

PCR products were transferred into the gateway entry vector pENTR/D-TOPO using the pENTR directional TOPO<sup>®</sup> Cloning kit (Invitrogen). 0.5 – 4ul of PCR product were used to achieve a total DNA amount of between 10 and 40ng. This was added to 1ul of Salt solution (1.2M NaCl, 0.06M MgCl<sub>2</sub>), and sterile water added to make a total volume of 5ul. Finally 1ul of pENTR/D-TOPO vector was added, and the mixture left for 5 minutes at room temperature before being put on ice.

Vectors were then transformed into chemocompetent TOP10 *E. coli* cells. 2ul of vector mixture was added to 25ul of TOP10 cells in solution on ice. The mixture was

then heat-shocked at 42°C for 30s, before being transferred back to ice. 250µl of SOC medium was added, and the cells shook at 200 RPM at 37°C for 1 hour. 50µl of the mixture was then plated on LB agar plates containing 50µg/ml Kanamycin to positively select for cells into which the plasmid has been transformed, and left overnight at 37°C. Colonies were picked from the plates and suspended in LB broth containing 50µg/ml Kanamycin, and left overnight at 37°C. Cells were pelleted from the solution by centrifugation for 10 minutes at 4000 RPM, and the plasmid DNA prepared using the QIAprep Miniprep kit (Qiagen, Germany), as per the manufacturer's protocol, with the DNA being eluted into a final volume of 100µl.

To check whether the plasmids contained the insert, and in the correct orientation, restriction digests were run using 10 units of NotI (New England Biosciences) (10µl plasmid DNA, with 5µl NE Buffer 3, 0.5µl BSA, 0.2µl NotI, 34.4µl dH<sub>2</sub>O), at 37°C for 3 hours. Digested pENTR plasmids were run on 1% agarose gels. Plasmid DNA was also sequenced by PCR amplification followed by capillary sequencing using M13 primers (20µl of 2µM), 0.4µl Taq polymerase, 5µl of 10x Taq Buffer (New England Biolabs), 5µl of 2.5mM dNTPs, and sterile water to a total volume of 50µl. Thermal cycling conditions were:



Products from the restriction digest and PCR were run on a 1.5% agarose gel, to determine the insert size, and the concentration of plasmid DNA obtained from the miniprep. Plasmids were sequenced using the M13 forward primer (3.2µl) and 1.6µl Plasmid DNA, with dH<sub>2</sub>O to make a 10ml reaction mixture.

### **3.2.3.3. Transfer into gateway destination vector**

Candidate effector genes were transferred from the pENTR gateway entry plasmid into a gateway compatible version of the pCX340 plasmid [292], pCX340gw. Genes were transferred using the gateway LR clonase as per the manufacturer's instructions. 0.2ml of the pENTR plasmid containing the candidate gene was added to 0.2µl of pCX340gw, 0.4µl LR clonase buffer, 0.4µl LR clonase and 0.8µl dH<sub>2</sub>O. The mixture was left for 2 hours at 25°C before being terminated with 0.2µl of Proteinase-K and incubation at 37°C for 10 mins.

The transformation method used to transfer the plasmids into these cells was the same as used in 3.2.3.2, but 10µg/ml Tetracycline was added instead of 50mg/ml Kanamycin to the LB broth/agar. Destination vectors were extracted from the TOP10 cells using GeneElute plasmid miniprep kit (Sigma), with the DNA being eluted into a final volume of 100µl. Insertion of the gene was confirmed by DNA sequencing as per the method used in 3.2.3.2.

The destination vector was finally transformed into the E22 strain of rabbit enteropathogenic *E. coli* (EPEC) [359], as a negative control the vector was also transformed into an *escN::Kan* mutant E22 strain [360], where the *escN* gene is interrupted by a kanamycin resistance gene. The E22 strains were rendered chemocompetent by washing in CaCl<sub>2</sub> solution,

### **3.2.4. Translocation assays**

Three independent methods based on translational fusion plasmids were used to assay type-III secretion dependent translocation from *E. coli* into eukaryotic cells. Fusion plasmids for each gene were constructed from PCR products encompassing the full

gene length or the first approximately 300 nucleotides as determined by the *E. coli* O157:H7 (Sakai) gene predictions. In all cases, fusion plasmids without DNA inserts produced negative results. Type-III-secretion deficient mutants were also tested with each plasmid to ensure that any observed translocation was dependent on the type-III secretion system.

#### **3.2.4.1. Cya translocation assay**

N-terminal translational fusions of CyaA were constructed using pTB101-*cyaA*, which encodes the N-terminal region of *Bordetella pertussis* CyaA toxin. An enteropathogenic *E. coli* strain was transformed with the *cyaA*-fusion plasmid, and used to infect Caco-2 cells. Cell extract from the Caco-2 cells was then obtained by centrifugation, and cAMP concentration in the extract was measured using the Cyclic AMP EIA Kit (Cayman Chemical)

#### **3.2.4.2. FLAG-tagged translocation assay**

C-terminal FLAG fusions were constructed using pFLAG-CTC. The enterohaemorrhagic *E. coli* Sakai strain was transformed with the FLAG-fusion plasmid, and used to infect Caco-2 cells. Cells were incubated for 2 hours, and the Caco-2 cells were fixed with paraformaldehyde. FLAG-tagged proteins were visualised with anti-FLAG antibody (Sigma), following attachment of AlexaFluor484-conjugated anti-mouse antibody (Molecular Probes).

#### **3.2.4.3. $\beta$ -lactamase translocation assay**

Rabbit EPEC strains containing C-terminal  $\beta$ -lactamase fusions were produced as described above. These transformed strains were then used to infect HeLa or Hep2 cells. After infection cells were incubated with the fluorescent substrate CCF2-AM.

Cleavage of CCF2-AM by the TEM-1  $\beta$ -lactamase was indicated by blue fluorescence after illumination by UV light at 409nm. Conversely green fluorescence under the same conditions indicates uncleaved CCF-2AM, and hence no translocation.

### **3.3. Results**

#### ***3.3.1. Bioinformatics***

Initial bioinformatics approaches revealed 62 proteins with homology to known or predicted type-III effectors. These effectors encompass over 20 different families of proteins. Some of these families have well defined roles within host cells, such as the NleH/OspG family, where OspG has been shown to affect the ubiquitination of proteins associated with the I $\kappa$ B $\alpha$ /NF- $\kappa$ B pathway [361]. Other proteins contain domains that have been associated with mediating bacterial/human cell interactions. This includes the leucine rich repeat domain, and ankyrin repeat domain. The largest group of proteins in the list is those belonging to the NleG family, which contains 14 members, however analysis of the genome sequence of the genes encoding these homologues reveals that several of these copies are likely to be pseudogenes. Overall, analysis of the predicted effector list using the sequences of homologous proteins suggests that around a quarter of the list are pseudogenes, where the nucleotide sequence has been disrupted by nonsense or frameshift mutations.

#### ***3.3.2. Gene cloning and transfer***

Attempts were made to PCR amplify all genes identified from the bioinformatics screen. All but nine PCRs produced products of the correct size (see Figure 9). The majority of failed PCR reactions could be made to succeed by altering the annealing temperature used during thermal cycling. To optimise the reaction gradient PCR was

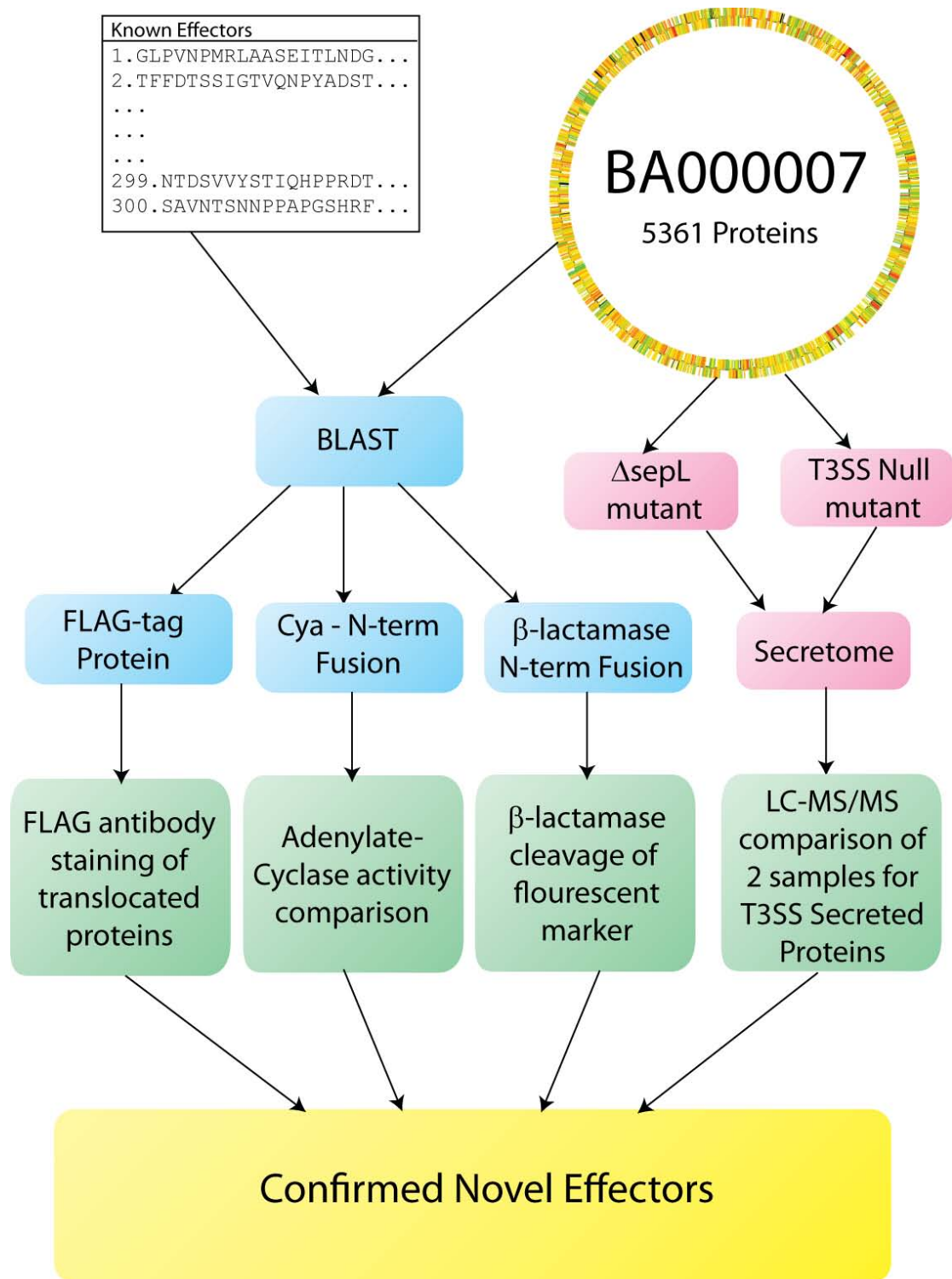


Figure 8. Experimental flow chart for determining novel effectors

used, with the annealing temperature varied between 41°C and 59°C across the heating block (results shown in gel in Figure 9). The remainder of failed PCRs were corrected by creating redesigned primers with more appropriate annealing temperatures.

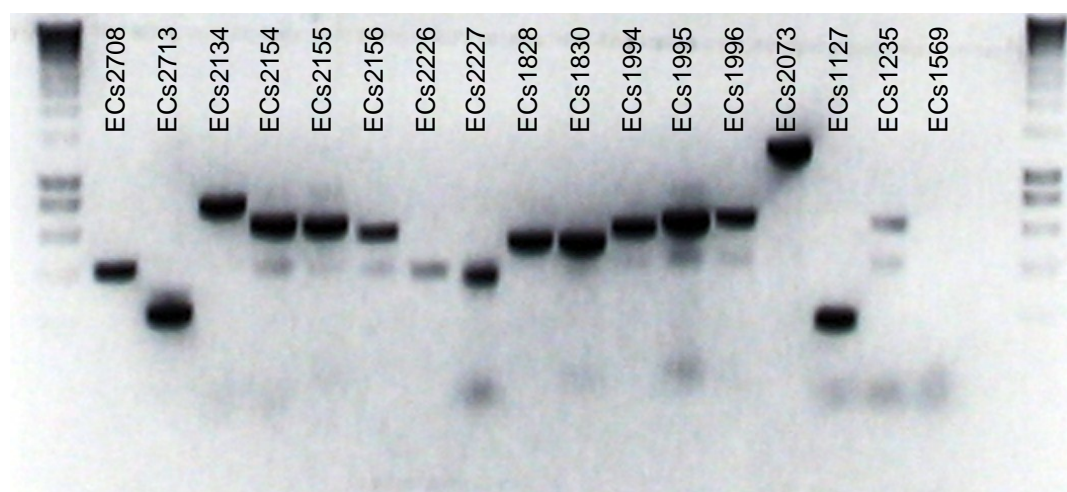
NotI digests and sequencing with M13 primers of transformed pENTR and pCX340gw plasmids revealed very few issues with the transformation process (digest gel shown in Figure 10). Transformed pCX340gw plasmids also went through the additional step of DNA sequencing using the M13 reverse primer to create complete coverage (where possible as determined by insert length) of the inserted DNA fragment (gel shown in Figure 10).

### ***3.3.3. Secretion and translocation assays***

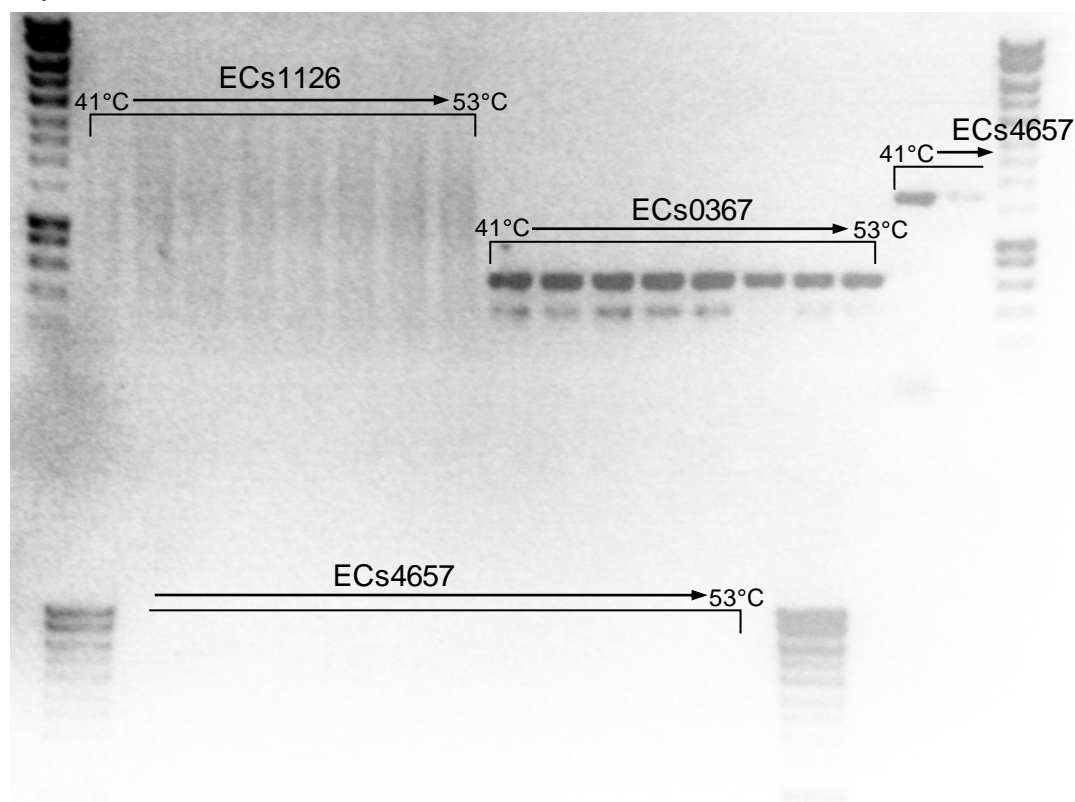
To confirm or reject the list of candidates obtained by homology, data from proteomics approaches was used. By comparing the culture supernatant from an EHEC mutant ( $\Delta sepL$ ) which constitutively secretes effectors, to a double mutant ( $\Delta sepL$ ,  $\Delta escR$ ) which contains a non-functional NF-T3SS, over 30 of the predicted effectors were shown to be secreted. Significantly, no additional candidates were found that were secreted but not identified by the homology screening already undertaken.

27 of the effector candidates were also shown to be translocated into host cells, by a combination of different assays: either through changes in cyclic AMP levels when the effector molecule was fused to adenylate cyclase CyaA; through visualisation of FLAG-tagged effectors using fluorescent antibodies; or using a  $\beta$ -lactamase reporter

A.



B.

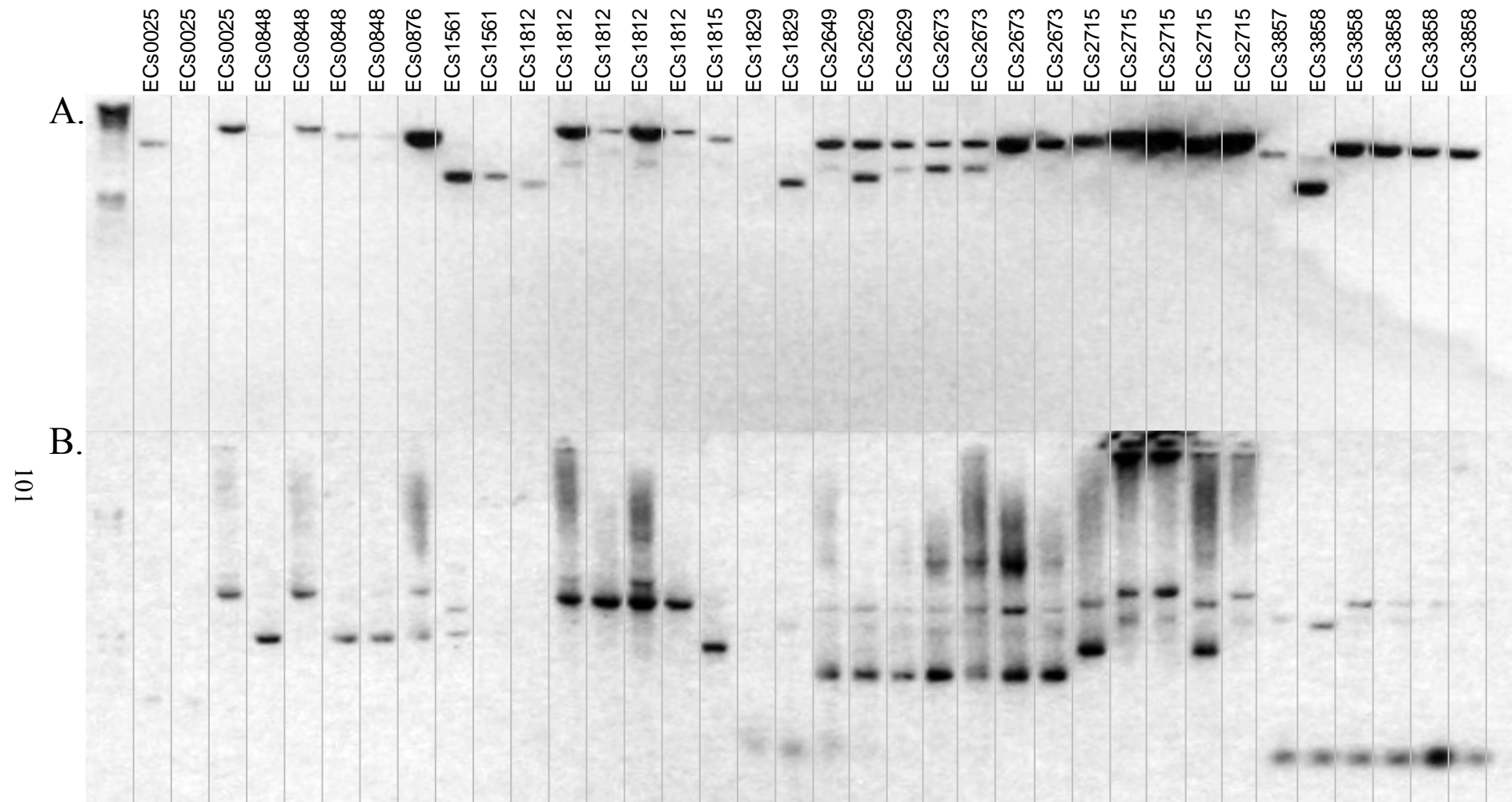


**Figure 9. PCR reaction products visualised on EtBr stained agarose gels**

**A.** Reaction products from PCR amplification of candidate effector genes. Ladder DNA is Hyperladder I (Bioline, Massachusetts, USA), Products are all present and of the correct size with the exception of ECs1569 (lane 18).

**B.** Reaction products from gradient PCR amplification of ECs1126, ECs0367 and ECs4657, eight lanes per gene. The first of the eight lanes for each gene amplification is from the coolest side of the gradient block (41°C) and the last of the eight from the hottest side of the gradient block. Whilst there was no product for ECs1126, ECs0367 demonstrates gradual reduction in an unwanted product as the temperature rises, whilst ECs4657 shows amplification only at low (<45°C) annealing temperatures





**Figure 10. NotI digests and M13 PCR amplification from pENTR plasmids visualised on EtBr stained agarose gels**

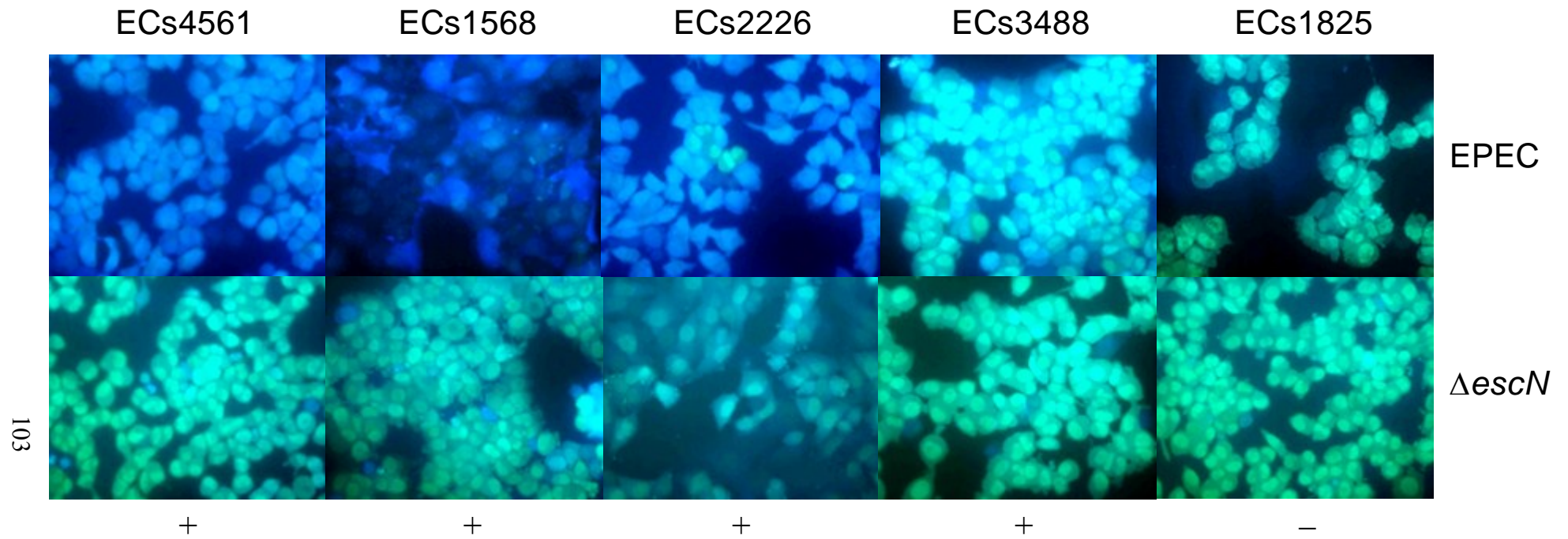
**A.** NotI digest of pENTR vector DNA containing candidate effector genes as listed above, correct band size is 2580bp + size of gene

**B.** PCR amplification of the same plasmids using M13 primers, correct band size is 271bp + size of gene

system (examples of  $\beta$ -lactamase reported results are shown in Figure 11), where the lactamase cleaves a fluorescent substrate (CCF2-AM) within the human cells leading to a change in wavelength of the emitted light. In total the proteomics and translocation assays showed 39 of the candidate effectors to be exported by the type-III secretion apparatus (Table 4).

With the exception of three genes (ECs0061, ECs0876 and ECs4653), all of the remaining confirmed effectors lie within prophage, or prophage like, islands on the chromosome. All 39 effectors lie in just fourteen separate loci. These exchangeable effector loci (EELs) comprise two pathogenicity islands (The LEE, and Sakai prophage like element 3, SpLE3), nine EELs within lambdoid prophages, two O-islands (loci present in *E. coli* O157:H7 but absent in *E. coli* K12), and one coli island (loci present in *E. coli* genomes but absent in related species such as *Salmonella enterica*), as determined by xBASE [362].

The lambda prophage encoded EELs share several distinctive characteristics. Firstly, they are all present within the same region of the prophage, located just downstream of the tail fibre genes, they always encode more than one effector gene, and stand out from the phage backbone in possessing extremely low GC content (Figure 12). Most interestingly however, is the fact that of the thirteen lambda prophages present on the Sakai chromosome, nine contain effector molecules. Of the remaining four prophages, two of them (Sp5 and Sp15) also contribute to the pathogenicity of *E. coli* by encoding shiga toxin genes. One prophage (Sp1) is interrupted by the insertion of another P4-like phage, and the remaining prophage (Sp8) also contains a passenger region downstream of the tail fibre genes, and contains a series of hypothetical proteins, of which several have domain matches to catalytic domains. Also, in the



**Figure 11. Results of  $\beta$ -lactamase assay for selected effector candidates**

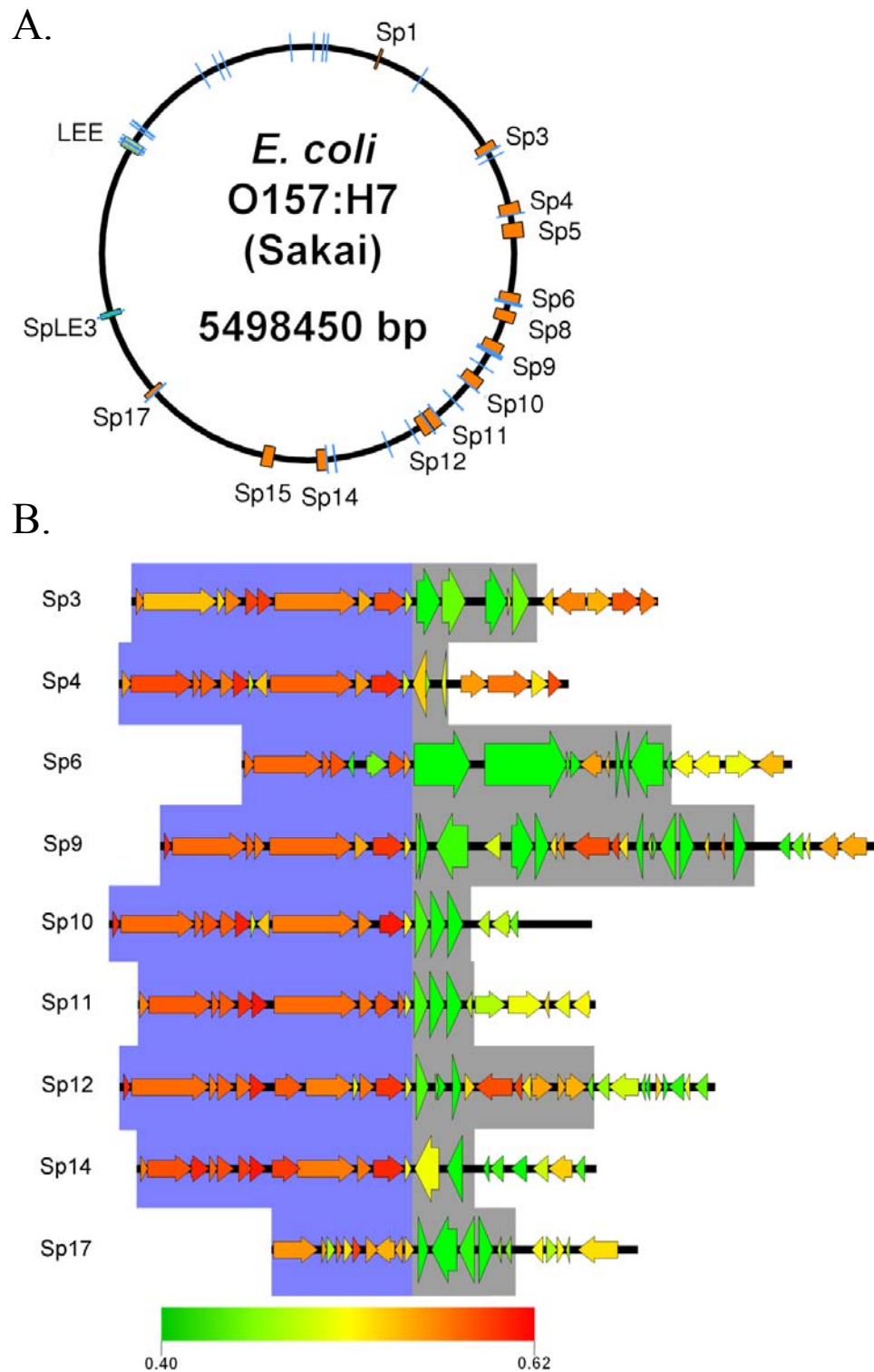
HeLa cells transfected with rabbit EPEC and a T3SS null mutant ( $\Delta escN$ ) visualised under light microscopy. Cells into which no tagged protein has been transfected show a green fluorescence produced by UV excitation of the uncleaved CCF2-AM molecule. Blue fluorescence indicates translocation of a  $\beta$ -lactamase tagged protein causing cleavage of the CCF2-AM molecule, changing the wavelength of light emitted from the molecule under UV excitation.

| Effector | Locus Tag | Family      | Locus  | Evidence |   |   |   |
|----------|-----------|-------------|--------|----------|---|---|---|
|          |           |             |        | S        | C | F | L |
| EspX1    | ECs0025   | PPR         | O-I 1  | ●        | ● | ● | ● |
| EspY1    | ECs0061   | SopD-N      | N C-I  | ●        | ● | ● | ○ |
| EspY2    | ECs0073   | SopD-N      | O-I 3  | ●        | ● | ● | ● |
| EspY3    | ECs0472   | SopD-N; PRR | C-I    | ●        | ● | ● | ● |
| NleB2-1  | ECs0846   | NleB        | Sp3    | ●        | ● | ● | ● |
| NleC     | ECs0847   | NleC        | Sp3    | ●        | ● | ● | ○ |
| NleH1-1  | ECs0848   | NleH        | Sp3    | ○        | ○ | ● | ● |
| NleD     | ECs0850   | NleD        | Sp3    | ●        | ● | ● | ○ |
| EspX2    | ECs0876   | PPR         | O-I 37 | ●        | ● | ● | ○ |
| EspF2-1  | ECs1126   | EspF        | Sp4    | ●        | ● | ● | ● |
| EspV     | ECs1127   | AvrA        | Sp4    | ●        | ● | ● | ● |
| EspX7    | ECs1560   | PPR; LRR    | Sp6    | ○        | ● | ● | ● |
| EspN     | ECs1561   | CNF         | Sp6    | ○        | ● | ● | ● |
| NleB2-2  | ECs1566   | NleB        | Sp6    | ●        | ● | ● | ● |
| EspO1-1  | ECs1567   | OspE        | Sp6    | ○        | ○ | ○ | ○ |
| EspK     | ECs1568   | LRR         | Sp6    | ○        | ● | ○ | ○ |
| NleG2-1  | ECs1810/1 | NleG        | Sp9    | ○        | ● | ● | ● |
| NleA     | ECs1812   | NleA        | Sp9    | ○        | ● | ● | ● |
| NleH1-2  | ECs1814   | NleH        | Sp9    | ○        | ○ | ○ | ○ |
| NleF     | ECs1815   | NleF        | Sp9    | ○        | ● | ● | ○ |
| EspO1-2  | ECs1821   | OspE        | Sp9    | ●        | ● | ● | ● |
| NleG     | ECs1824   | NleG        | Sp9    | ○        | ● | ○ | ● |
| EspM1    | ECs1825   | lpgB        | Sp9    | ○        | ● | ○ | ● |
| NleG9    | ECs1828   | NleG        | Sp9    | ●        | ● | ● | ● |
| NleG2-2  | ECs1994   | NleG        | Sp10   | ○        | ○ | ○ | ● |
| NleG6-1  | ECs1995   | NleG        | Sp10   | ○        | ○ | ○ | ● |
| NleG5-1  | ECs1996   | NleG        | Sp10   | ○        | ○ | ○ | ● |
| EspR1    | ECs2073   | LRR         | O-I 62 | ●        | ● | ● | ● |
| EspR2    | ECs2074/5 | LRR         | O-I 62 | ●        | ● | ● | ● |
| NleG5-2  | ECs2154   | NleG        | Sp11   | ○        | ● | ● | ● |
| NleG6-2  | ECs2155   | NleG        | Sp11   | ○        | ● | ● | ● |
| NleG2-3  | ECs2156   | NleG        | Sp11   | ○        | ● | ● | ● |
| NleG7    | ECs2226   | NleG        | Sp12   | ●        | ● | ● | ○ |
| NleG3    | ECs2227/8 | NleG        | Sp12   | ●        | ● | ● | ● |
| NleG2-4  | ECs2229   | NleG        | Sp12   | ●        | ● | ● | ● |
| EspL1    | ECs2427   | AR          | C-I    | ●        | ● | ● | ● |
| EspR3    | ECs2672   | LRR         | C-I    | ●        | ● | ● | ● |
| EspR4    | ECs2674   | LRR         | C-I    | ●        | ● | ● | ● |
| EspJ     | ECs2714   | EspJ        | Sp14   | ○        | ○ | ● | ● |
| TccP     | ECs2715   | EspF        | Sp14   | ○        | ● | ● | ● |
| EspM2    | ECs3485   | lpgB        | Sp17   | ○        | ● | ○ | ● |

|         |           |        |         |  |  |  |  |
|---------|-----------|--------|---------|--|--|--|--|
| NleG8-2 | ECs3486   | NleG   | Sp17    |  |  |  |  |
| EspW    | ECs3487   | HopW   | Sp17    |  |  |  |  |
| NleG6-3 | ECs3488   | NleG   | Sp17    |  |  |  |  |
| EspL2   | ECs3855   | AR     | SpLE3   |  |  |  |  |
| NleB1   | ECs3857   | NleB   | SpLE3   |  |  |  |  |
| NleE    | ECs3858   | NleE   | SpLE3   |  |  |  |  |
| EspF1   | ECs4550   | EspF   | LEE     |  |  |  |  |
| EspB    | ECs4554   | EspB   | LEE     |  |  |  |  |
| Tir     | ECs4561   | Tir    | LEE     |  |  |  |  |
| Map     | ECs4562   | lpgB   | LEE     |  |  |  |  |
| EspH    | ECs4564   | EspH   | LEE     |  |  |  |  |
| EspZ    | ECs4571   | EspZ   | LEE     |  |  |  |  |
| EspG    | ECs4590   | EspG   | LEE     |  |  |  |  |
| EspL3   | ECs4642/3 | AR     | O-I 152 |  |  |  |  |
| EspY4   | ECs4653   | SopD-N | O-I 153 |  |  |  |  |
| EspX3   | ECs4654/5 | PPR    | O-I 153 |  |  |  |  |
| EspY5   | ECs4657   | SopD-N | O-I 153 |  |  |  |  |
| EspL4   | ECs4935   | AR     | C-I     |  |  |  |  |
| EspX4   | ECs5021   | PPR    | C-I     |  |  |  |  |
| EspX5   | ECs5048   | PPR    | C-I     |  |  |  |  |
| EspX6   | ECs5295   | PPR    | O-I     |  |  |  |  |

**Table 4. T3SS effectors in the *E. coli* O157:H7 genome**

Families: LRR, leucine-rich repeats; AR, ankyrin repeats; PPR, pentapeptide repeats; SopD-N, SopD N-terminal domain. Location: Sp, Sakai prophage and prophage-like elements (highlighted in blue); LEE, Locus for enterocyte effacement, also known as SpLE4 (highlighted in orange); C-I, coli island (present in *E. coli* but not related species such as *S. enterica*); O-I, O-islands as determined by Perna *et al* [356]. Evidence: S, detected in secretome of  $\Delta$ sepL mutant; C, translocation of CyaA fusion detected; F, translocation of FLAG-tagged fusion detected; L, translocation of  $\beta$ -lactamase fusion detected. Rows highlighted in grey are predicted to be pseudogenes. A green button (and background) in an evidence column indicates a positive results. Similarly a red button and background indicates a negative result.



**Figure 12. Effector Locations and Phage effector loci in *E. coli* O157 Sakai**

A. *E. coli* chromosome with lambda prophage (orange), Prophage like element 3 and 4 (the LEE) (yellow), and effector locations (blue). B. Sakai prophage containing T3SS effectors. Gene colour represents GC content (see colour bar). Double height genes – effectors. Blue background – Phage backbone. Grey background – Effector passenger loci. Prophage aligned relative to tail fibre gene.

The detected candidate effectors are spaced throughout the genome, but those located within prophage are all found downstream of the prophage tail fibre gene. The effector candidates also have very low G+C content compared with the genome, and particularly compared to other genes within the prophage which have very high G+C content.

nine prophages which contain EELs, all but three of the 64 genes encoded in passenger compartments are putative or proven effectors, or are IS elements.

### **3.4. Discussion**

This study demonstrates that even in some of the better studied bacterial species there is still much to be learnt. The number of known type-III secretion effectors in *E. coli* numbered only five just a few years ago, a figure that has risen to over a dozen with recent studies [290, 360, 363-365]. This study shows that there are likely to be over three times this number of effectors present within the genome of the Sakai strain of *E. coli*. Whilst a function has yet to be determined for many of these new effectors, it suggests that the breadth of effectors used by the LEE type-III secretion system is far larger than was thought previously. It is also clear that prophage are a major source of effector genes, and the phage “meta-genome” has acted as significant agent in the evolution of pathogenicity in *E. coli*. It would seem then, that phages are an important source of natural variation in closely related *E. coli*, a situation conserved in other Enterobacteriaceae such as *Salmonella* [20, 42, 43].

#### **3.4.1. Recurrent domains and motifs**

The range of different proteins contained within the effectors identified within this study provides an intriguing insight into the breadth of measures utilised by EHEC to subvert the processes of host cells. Within the large number of effectors discovered in this survey there are a number of common motifs and domains that are shared amongst several proteins. There are, for example 14 proteins which are homologous to each other, and belong to the NleG group of proteins first identified by Deng *et al* [290]. There is a similar expansion in the number of NleG proteins in

Enteropathogenic *E. coli* (EPEC). This expansion in the number of NleG homologues is not mirrored in the other bacterium which contains an attaching/effacing T3SS, *Citrobacter rodentium*. There are also several repeat domains present amongst the proteins in this study, several of which are more commonly found in the eukaryotic domain of life. This includes domains such as the pentapeptide repeats, leucine rich repeats and ankryn repeats. In each case, there are examples of other bacterium which use these domains to interact with eukaryotic cells, some of which even secrete these proteins in a type-III dependant manner.

Pentapeptide repeats exist primarily in the prokaryotic domain, where they are mostly found in cyanobacteria. However, there are plenty of examples of this domain occurring within most bacterial phyla [366]. They occur in multiple repeats and have a motif of A(D/N)LXX [366]. The exact function of these repeats is unknown, although they have been predicted to have a targeting or structural function. Pentapeptide repeats are predicted to form a right handed beta-helical structure [366]. More recently this predicted structure has been implicated in fluoroquinolone resistance, by mimicking DNA, and hence disrupting the action of DNA gyrase [367]. What role pentapeptide repeats may play in affecting host cells is yet to be determined. All the identified proteins containing pentapeptide repeats within *E. coli* O157 Sakai show homology to the type-III secreted effectors SopA and PipB2 from *Salmonella enterica* serovar typhimurium. However in the case of both of these effectors, they appear to have a role in the vacuolar stage of invasion of host cells by salmonella [368, 369], a lifestyle which *E. coli* O157 does not undertake.

Leucine Rich Repeats (LRRs) fall into five different categories according to structural analysis [370], in general these repeats are 20-30 amino acids long and occur in



tandem. The repeat forms an  $\alpha$ -helix- $\beta$ -sheet secondary structure, and when present in number form a horseshoe type structure with the  $\beta$ -sheet on the inside of the horseshoe [371]. Searching using HMMER and the LRR\_1 model from PFAM reveals a total of six different proteins containing this domain within *E. coli* O157 Sakai, all of which were identified in the bioinformatics screen, although only two of these (ECs1560 and ECs1568) produced positive results in any experimental test. LRRs are responsible for diverse protein-protein interactions, and are utilised by several different bacteria to interact with host cells. Examples of this include *Listeria monocytogenes* protein Internalin B, which induces phagocytosis of the bacterium into host cells [372], and YopM from *Yersinia pestis*, which depletes Natural Killer cells *in vivo* [373].

Ankyrin Repeats are another form of protein-protein interaction domain, most commonly found in the eukaryotic domain of life. It is also suspected that the few known examples of ankyrin repeat domain proteins that occur in prokaryotes may be present as a result of horizontal gene transfer [374]. Four out of the five proteins predicted to have ankyrin repeats within the predicted set of effectors also contain a toxin\_15 domain. The toxin\_15 domain is best characterised in the ShET2 enterotoxin encoded by the *senA* gene, located on the invasion plasmid of *Shigella Flexneri* [375]. Within *Shigella* this protein is thought to be exported by the Mxi/Spa T3SS. Whilst the ShET2 protein is not recognised by PFAM or SMART [376] as having ankyrin repeats, BLAST identifies near full length homology between ShET2 and other proteins with ankyrin repeat domains, such as ECs2427 and ECs4935. There are multiple copies of the *senA* gene within the *Shigella* invasion plasmid where they are annotated as OspD, so it is not a great surprise to find multiple copies of a *senA* homologue encoded within the *E. coli* chromosome. The domain architecture of

toxin\_15 – ankyrin repeats is found in several other bacterial species such as *Yersinia* and *Ralstonia*.

Finally there are a group of proteins which show homology to the N-terminal region of SopD from *Salmonella*. All these proteins contain a WEX(I/M)xxFF motif which is found in several *Salmonella* effectors as well as effectors from *Edwardsiella* and *Sodalis* [377, 378]. Taken together with the information on distribution of effectors throughout the *E. coli* genome, these data demonstrate the extensive role of horizontal gene transfer in generating diversity and aiding spread of type-III secreted effectors, and in particular transfer by bacteriophage. It also demonstrates both the diversity of domains and activities undertaken by type-III effectors. The conserved motifs and domains, also suggest conserved methods of interaction between different T3SSs belonging to diverse bacteria and their target cells.

### **3.4.2. Pitfalls of screening and assays**

Bioinformatics tools have enabled scientists to assess genes and proteins at a rate that many *in vivo* and *in vitro* techniques are unable to match. The negative side to using such tools is that the analysis is several steps removed from the complexities of *in vivo* processes. At each step in-between are simplifications and assumptions which enable such mass analysis to be performed. However these simplifications and assumptions make such analyses less reliable.

In this regard genome wide bioinformatics analysis, can fulfil a variety of roles, but the results should be examined in the knowledge that conclusions established on homology alone are unlikely to be totally reliable. They can however provide an avenue to determine areas for further investigation, to generate hypotheses for future testing, or to narrow down a field of candidates where testing of all would not be

feasible.

In the case of this experiment, it can be seen that whole genome bioinformatic scans are useful in reducing a field of over 5000 candidate genes down to a list nearly two orders of magnitude smaller. Such a smaller list is amenable to more thorough experimental processes than a larger one, meaning that more definitive studies can be performed. The analysis of the secretome of the ‘always on’ mutant  $\Delta sepL$  provides a useful independent validation of the bioinformatics approach taken – in that no proteins were identified in the secretome which were not found in the bioinformatics study. It should be noted however, that by its very nature an always on T3SS may not export the same set of proteins as its wild-type version.

Similar issues lie with the various methods used to test for translocation of proteins via T3SSs. The addition of new domains to candidate proteins is liable to alter the overall shape and size of the protein in such a way as to prevent it being secreted via the T3SS apparatus. Fusing the reporter domain to only the N-terminal portion of the effector may also affect the ability for the T3SS or chaperones to detect any signal/binding sequence. In this regard the use of multiple translocation assays is of value. By threading together multiple lines of evidence: homology to other known effectors, secretion by a  $\Delta sepL$  mutant, translocation of *cyaA*/flag/ $\beta$ -lactamase chimera; the potential pitfalls of each approach becomes diminished. Of the thirty-nine proteins for which there is some form of experimental data to support the assertion of the protein being an effector, almost half have support from more than one method, and nearly a quarter are backed up with positive results from three or all four experimental procedures. When all these multiple sources of data are taken together they help to minimise the confounding effects found in each approach, and

also provide a sliding scale of certainty for the likelihood that any individual protein is a genuine T3SS effector.

### 3.5. Summary

The data presented in this chapter demonstrates that there are likely many more effectors present in the *E. coli* O157 Sakai genome than had been previously thought. The techniques used here also show that effectors are amenable to location through homology searching. Comparison with the independent proteomics data also demonstrate that homology searching for effectors can be both a sensitive and specific method for their location. However, the positive predictive value for the method is quite poor (63%), possibly due to the very diverse nature of effectors which leads to the detection of a larger number of false positive hits within genomes. As such homology searching is an ideal technique for determining candidate effectors for further examination, but not for conclusive proof of a protein's status as an effector, for which other techniques are best employed. The effectors located within *E. coli* demonstrate that there are often multiple copies of homologous effector proteins within a species, and that certain effectors are shared between multiple different bacteria. The distribution of effectors within the chromosome also highlights the degree to which horizontal gene transfer, and in particular transfer by bacteriophage, has to play in the distribution of effectors.

## CHAPTER 4 - SPECIFIC T3S SYSTEMS

### 4.1. Introduction

#### ***4.1.1. The Study of Model NF-T3SS systems***

Much of the research on T3SSs has been done on a very limited number of systems. Of the two and a half thousand papers found in PubMed using the search term “Type-III secretion” (search performed in April 2008), over 60% of the papers focus on the T3SSs from *Yersinia*, *Salmonella*, *Escherichia*, *Shigella* and *Pseudomonas*. In contrast some T3SSs have had practically no research directed towards them, such as the T3SS encoded on the megaplasmid of *Desulfovibrio vulgaris*, for which the same PubMed search finds only two articles: The paper describing the genome sequence of the bacterium, and a review article.

The body of work available on the better known systems is of great benefit to researchers in the field of type-III secretion, and work done on proteins conserved across the field of type-III secretion is of value to all. However as the work shown in Chapter 2 demonstrates, even amongst the conserved proteins there are subtle, but possibly important variations.

In this era where sequencing a bacterial genome is now a relatively simple process, and the number of bacterial genomes now sequenced continues to grow at an exponential rate (see Figure 16 in Chapter 5), the ability of the scientific community to analyse and annotate new genome sequence is now outstripped by its ability to generate new sequence data. As a result of this there are a large number of genomes which are annotated solely by transfer of the annotation from homologous genes in

other sequenced bacteria, and also many genomes which remain un-annotated. This process of semi-automated transfer of annotation from one genome to another on the whole would seem to work, there are however pitfalls to this approach. There are already several examples in the literature of errors in genome annotations, and studies have been done showing errors and their estimated rates throughout annotated genomes [379-383]. As a result of this, care needs to be taken in analysing data using such genome sequences, and one should be sceptical of any annotation which does not make sense in the context of the genomic locale and/or the bacterium.

In the case of Type-III secretion systems the reasonable level of sequence similarity shown between several conserved proteins means that even the most diverse members can often be identified using simple homology searches. Identification of novel T3SSs in newly sequenced genomes is of interest for several reasons: Firstly it provides insight into the diversity of T3SSs and the range of different ecological niches in which they have a role to play. Secondly they allow comparisons to be drawn between T3SS so that we can examine the core gene content of T3SSs and also see any unique characteristics which are confined to specific groups or even individual T3SSs, and in this way we can attempt a form of taxonomic classification of the systems, and compare this to other methods of classification such as those provided by sequence based phylogenetics.

#### ***4.1.2. Novel T3SSs found in diverse bacteria***

In an examination of the diversity of non-flagellar type-III secretion systems done by Foulter *et al* in 2002, they performed a phylogenetic analysis of conserved NF-T3SS genes which placed each T3SS under examination into five major groups: Ysc, Inv/Mxi/Spa, Esc-Ssa, Hrp1 and Hrp2 [297]. Since then there have been a large

number of additional genomes sequenced, of which several have T3SSs present (See Chapter 5). This includes T3SS systems encoded in bacteria which are unlikely to cluster with existing systems, and for which type-III secretion may perform a role unrelated to pathogenesis. These bacteria include the Chlamydiae (and Protochlamydia) [108, 301], *Rhizobium* [384], *Myxococcus xanthus* [385], *Verucomicrobium spinosum* [385], *Lawsonia intracellularis*, *Desulfovibrio vulgaris* [386] and *Hahella chejuensis* [387].

By examining these different systems we can begin we can see the differences between numerous different T3SSs now known, and analyse them to see if they contain similar conserved proteins to other systems, along with locating novel proteins involved in type-III secretion. Finally, the more diverse members of the T3SS family now being discovered also add information on the evolutionary history and genomic diversity of type-III secretion.

#### ***4.1.3. Annotation and analysis of novel T3SS systems***

Previous work by our group has already shown that there are certain proteins and domain features that are unique to the some of the groups shown in the work of Foulter *et al.* One such example of this is the SepL and YopN/TyeA families of proteins. YopN is homologous to the N-terminal region of SepL, whilst TyeA is homologous to the C-terminal of SepL. YopN and TyeA are encoded by adjacent genes, suggesting that YopN and TyeA arose as the result of a fission event from SepL (or *vice versa*). Within four of the five groups of T3SS described (Inv/Mxi/Spa, Esc-Ssa, Hrp1 and Hrp2) there is a homologue of SepL, but within the Ysc group, the YopN-TyeA combination is present.

Within the phylogenetic tree of the five groups of T3SSs mentioned above *Bordetella*

clusters in the same group as the Ysc systems, however it has a *sepL* type gene, rather than the YopN/TyeA case. There are other characteristics which are found only within the three sequenced *Bordetella* species, but not in the other members of the Ysc group of T3SSs. The major example of this is the ECF sigma factor type regulatory group of proteins which regulate the T3SSs of *Bordetella* species. The master regulator of the *Bordetella* type-III secretion system is the *bvgAS* locus [388]. This locus encodes BvgS a membrane bound sensor kinase which phosphorylates the other member of the locus BvgA, which then in turn alters expression of a wide range of virulence and colonisation factors around the *Bordetella* chromosome [389, 390]. BvgA controls the regulation of the NF-T3SS apparatus through a series of five genes located next to the *bsc* locus which encodes the NF-T3SS apparatus. These five *btr* genes: *btrS*, *btrU*, *btrX*, *btrW* and *btrV*, are all positively regulated by BvgA [391]. The protein products of these genes are homologous and act in a similar manner to the ECF sigma factor proteins from *Bacillus subtilis* [391, 392]. BtrS acts as the sigma factor, and is inhibited by the anti-sigma factor BtrW. The serine/threonine phosphatase BtrU activates BtrV, which in its activated state releases BtrS by binding BtrW [391, 392]. This ECF sigma factor type regulation system, which is traditionally only found in Gram-positive bacteria, is also found in *Chlamydia*, another bacterium with a T3SS [391].

#### **4.1.4. Aims**

It seems likely that the quality of genome annotation declines as the rate of DNA sequencing increases and errors in annotation continue to be propagated into newly deposited genome sequences. This chapter sets out to examine the quality of the annotation of newly identified NF-T3SSs, and to see what bioinformatics tools can do



to assess and improve the annotation quality. The genomes chosen in order to assess this were those where initial analysis suggested they would cluster within the Ysc/Bordetella group of NF-T3SSs. These genomes were chosen for several reasons: Firstly the Ysc group of NF-T3SSs is one of the better understood groups, meaning that there is a greater chance of making an inference from any comparison between new T3SSs and the existing systems; and also because there are already differences known between the Ysc and Bordetella groups, such as the regulatory mechanisms mentioned above.

This thorough analysis of these novel NF-T3SSs allows both comparison of their complement of genes, and phylogenetic clustering versus other Ysc NF-T3SSs. In doing so the data obtained may be able to locate single evolutionary changes which occurred to produce the larger variations in genetic makeup seen for example between Yersinia and Bordetella NF-T3SSs. Such side by side analyses will also add weight to any arguments made about errors made in, and changes required to, genome annotation data.

## **4.2. Methods**

### ***4.2.1. T3SS Region Prediction***

Starting with the complete genome sequences of *Lawsonia intracellularis* and *Hahella chejuensis*, and the megaplasmid of *Desulfovibrio vulgaris* (Genbank accession numbers AM180252, CP000155 and AE017286 respectively), T3SS regions were identified through HMMER searches using a database of domains related to both flagellar and non-flagellar type-III secretion. The list was generated by searching PFAM release 20 for all pfamA domains containing the terms “type-III”,

“type III” or “flagell\*” within their description or comment data. The domains chosen are shown in Table 5. Searches were performed using the hmmsearch program and both the Pfam\_ls and Pfam\_fs versions of the domain models. The start and end coordinates of the gene encoding each hit were recorded where the hit had a score greater than the gathering threshold set for the domain which matched against the proteins. Regions of the genome were clustered together where hits occurred within 50Kb of each other, and NF-T3SS regions were found by looking for clusters of genes containing proteins with domains common to both flagellar and non-flagellar systems, but not flagellar specific domains. Non-flagellar clusters were extracted from the genome based on the coordinates of the first and last identified proteins plus 50Kb of backbone sequence either side of the region.

#### ***4.2.2. Gene identification***

##### **4.2.2.1. HMMER searches**

Each protein encoded within the genomes of interest were searched using the complete PFAM database (release 20) in its Pfam\_ls and Pfam\_fs modes, using the hmmpfam program in order to identify all potential domains. Search results were used to confirm assignment of T3SS regions, and to help define the ends of the T3SS locus/loci.

##### **4.2.2.2. BLAST Searches**

Each protein within a putative T3SS locus was also searched with BLAST version 2.2.15 on Linux, using its BLASTP and PSI-BLAST modes using a database of bacterial proteins compiled from the bacterial genomes directory at the NCBI (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>) downloaded in April 2007.

| <b>PFAM Id</b> | <b><i>T<sub>GA</sub></i></b> | <b>Description</b>  |
|----------------|------------------------------|---|
| ATP-synt_ab    | -37.5, 55                    | ATP synthase alpha/beta family, nucleotide-binding domain |
| ATP-synt_ab_N  | 17, 16.4                     | ATP synthase alpha/beta family, beta-barrel domain        |
| Bac_export_1   | 25, 13                       | Bacterial export proteins, family 1                       |
| Bac_export_2   | 25, 25                       | FlhB HrpN YscU SpaS Family                                |
| Bac_export_3   | 25, 25                       | Bacterial export proteins, family 3                       |
| CesT           | 9.2, 16.5                    | Tir chaperone protein (CesT)                              |
| Chaperone_III  | 25, 25                       | Type III secretion chaperone domain                       |
| DspF           | 25, 25                       | DspF/AvrF protein   |
| EspA           | 25, 25                       | EspA-like secreted protein                                |
| EspB           | -18.3 , 12.2                 | Enterobacterial EspB protein                              |
| EspF           | 23, 25                       | EspF protein  |
| EspG           | 25, 25                       | EspG protein  |
| FHIPEP         | -452, 25                     | FHIPEP family   |
| FlaA           | 25, 25                       | Flagellar filament outer layer protein FlaA               |
| FlaE           | 25, 18                       | Flagellar basal body protein FlaE                         |
| FlaF           | 25, 25                       | Flagellar protein FlaF                                    |
| FlaG           | 25, 25                       | FlaG protein  |
| Flagellin_C    | 21, 14.8                     | Bacterial flagellin C-terminus                            |
| Flagellin_IN   | 25, 25                       | Flagellin hook IN motif                                   |
| Flagellin_N    | -15, 16                      | Bacterial flagellin N-terminus                            |
| FlbD           | 25, 25                       | Flagellar protein (FlbD)                                  |
| FlbT           | 25, 25                       | Flagellar protein FlbT                                    |
| FleQ           | 25, 25                       | Flagellar regulatory protein FleQ                         |
| Flg_bb_rod     | 25.3, 20.6                   | Flagella basal body rod protein                           |
| Flg_hook       | -2.3, 19.6                   | Flagellar hook-length control protein                     |
| FlgD           | -4, 25                       | Flagellar hook capping protein                            |
| FlgH           | -40, 25                      | Flagellar L-ring protein                                  |
| FlgI           | -205, 25                     | Flagellar P-ring protein                                  |
| FlgM           | 12.9, 19.1                   | Anti-sigma-28 factor, FlgM                                |
| FlgN           | -20, 25                      | FlgN protein  |
| FlhC           | 25, 25                       | Flagellar transcriptional activator (FlhC)                |
| FlhD           | 25, 25                       | Flagellar transcriptional activator (FlhD)                |
| FlhE           | 25, 25                       | Flagellar protein FlhE                                    |
| FliD_C         | 25, 25                       | Flagellar hook-associated protein 2 C-terminus            |
| FliD_N         | 25, 25                       | Flagellar hook-associated protein 2 C-terminus            |
| FliE           | 25, 25                       | Flagellar hook-basal body complex protein FliE            |
| FliG_C         | -6, 25                       | FliG C-terminal domain                                    |
| FliH           | -43, 15.7                    | Flagellar assembly protein FliH                           |
| FliJ           | 25, 25                       | Flagellar FliJ protein                                    |
| FliL           | 1.5, 16.5                    | Flagellar basal body-associated protein FliL              |
| FliM           | -46, 15                      | Flagellar motor switch protein FliM                       |
| FliO           | 25, 25                       | Flagellar biosynthesis protein, FliO                      |
| FliS           | 25, 25                       | Flagellar protein FliS                                    |

|                |             |  |
|----------------|-------------|--|
| FliT           | 25, 25      | Flagellar protein FliT                                     |
| HAP3           | 25, 25      | Putative flagellar hook-associated protein 3 (HAP3)        |
| HOOK           | 25, 25      | HOOK protein   |
| HrpF           | 25, 25      | HrpF protein   |
| HrpJ           | -2, 18.2    | Hypersensitivity response secretion protein HrpJ           |
| InvH           | 25, 25      | InvH outer membrane lipoprotein                            |
| IpaD           | 25, 25      | Invasion plasmid antigen IpaD                              |
| LcrV           | 25, 25      | V antigen (LcrV) protein                                   |
| MotA_ExbB      | -30, 18     | MotA/TolQ/ExbB proton channel family                       |
| NolV           | 25, 30      | Nodulation protein NolV                                    |
| NolX           | 25, 25      | NolX protein   |
| SAF            | 21.4, 18.8  | SAF domain   |
| Secretin       | 25, 19.1    | Bacterial type II and III secretion system protein         |
| SepL_SsaL      | 25, 25      | SepL/SsaL protein  |
| SepZ           | 25, 25      | SepZ   |
| SLT            | -4.1, 16.1  | Transglycosylase SLT domain                                |
| SseC           | 25, 25      | Secretion system effector C (SseC) like family             |
| Tir_receptor_C | 25, 25      | Translocated intimin receptor (Tir) C-terminus             |
| Tir_receptor_M | 25, 25      | Translocated intimin receptor (Tir) intimin-binding domain |
| Tir_receptor_N | 25, 25      | Translocated intimin receptor (Tir) N-terminus             |
| TyeA           | 1.1, 25     | TyeA   |
| YcgR           | 25, 25      | YcgR protein   |
| YscJ_FliF      | -34, 25     | Secretory protein of YscJ/FliF family                      |
| YscJ_FliF_C    | -31.6, 16.8 | Flagellar M-ring protein C-terminal                        |
| YscK           | -89.2, 25   | YOP proteins translocation protein K (YscK)                |
| YscO           | -2.2, 15.8  | Type III secretion protein YscO                            |

**Table 5. List of PFAM domains used to search for T3SS loci**

T<sub>GA</sub>: Gathering threshold for the domain. Two values are gathering thresholds for Pfam\_ls and Pfam\_fs hits respectively. Only hits which scored greater than this value were considered to actually contain this domain.

The gathering threshold is part of the PFAM dataset, and is the scoring threshold for a candidate domain, above which the domain will be considered to be a true member of this domain family. These values are chosen (often empirically) to prevent different domains appearing in multiple PFAM models, and also based on the overall degree of similarity amongst diverse members of the domain family.

BLAST was run with the filter and composition based statistics off using the BLOSUM62 matrix. BLASTX searches were also performed using the DNA sequence of the extracted regions as the query. In all cases the default settings were used. PSI-BLAST searches were run using the same starting parameters as were used for BLASTP, to convergence or a maximum of ten iterations, whichever occurred first. Relaxed search criteria were used to maximise the number of hits found, particularly to distant homologues.

### ***4.2.3. Analysis and annotation***

The extracted regions were loaded into the Artemis program, and re-annotated based on the information gained from the searches performed on the DNA and protein sequences. Alignments were created using T-coffee using its Gotoh pairwise dynamic programming option, or using HMMalign where PFAM domains were available to align sequences against. Phylogenetic trees were produced by ClustalX using alignments produced as described above, and bootstrapped where appropriate using 1000 replicates, then drawn using the MEGA 4 package [340].

## **4.3. Results**

### ***4.3.1. Comparison and re-annotation of the *Lawsonia intracellularis* T3SS regions***

Analysis of the genome of *Lawsonia intracellularis* reveals the presence of two loci within the chromosome which contain proteins belonging to non-flagellar T3SS apparatus. At the 3' end of locus one, and the 5' end of locus two there are two regions which show a high degree of similarity at the DNA level, however whilst this region has no CDSs annotated within it in locus one, there are a series of CDSs within

the equivalent section of locus two (annotated with locus tags LI1145-LI1149). None of proteins encoded by the CDSs show homology to any other proteins in the Genbank database, however a search of a DNA database using these regions as query sequences reveals homology to 5s, 16s and 23s ribosomal rRNA genes. Searching within these ribosomal RNA regions using tRNAsCAN also reveals the presence of two tRNAs. There are also several other duplicate genes within these two loci, including *lscJ*, *lscL* and *lscQ*, as well as what appears to be a gene fragment at the end of locus one just prior to the rRNA region which is homologous to the N-terminal region of the protein encoded by LI1150 from locus two.

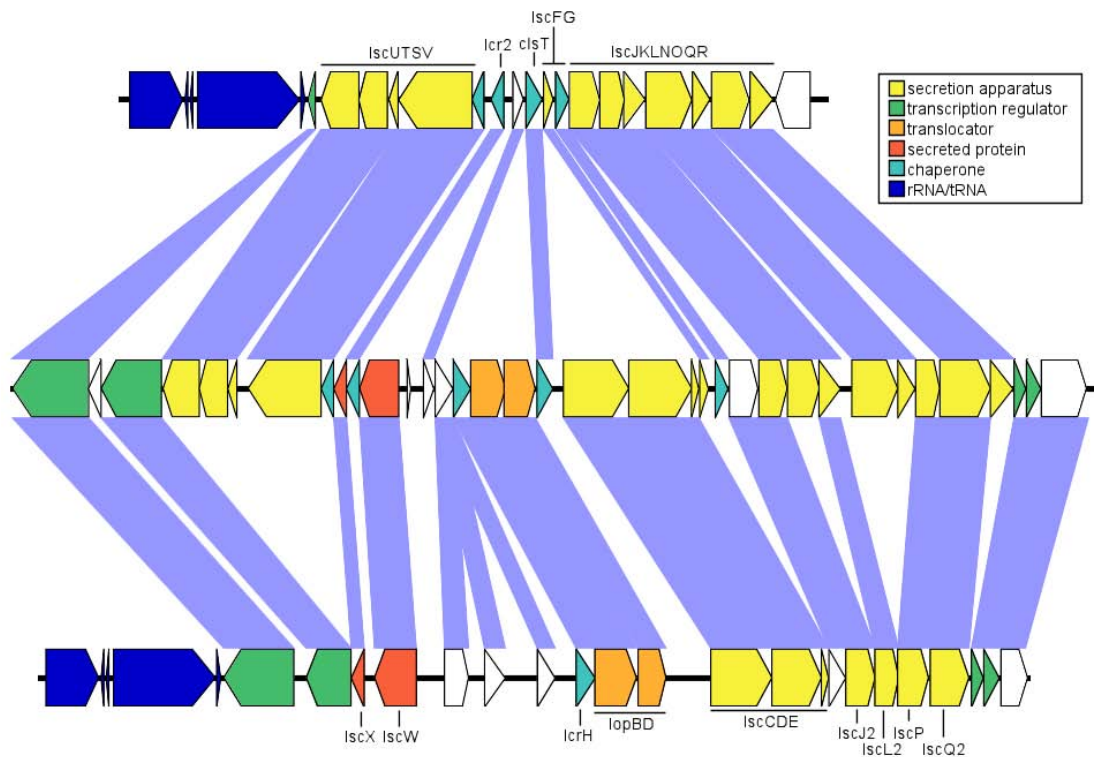
A more detailed analysis of the regions also reveals several annotation errors and missed genes within these regions. Following re-annotation, potential protein product descriptions and protein names were assigned to ten additional genes in region one, and twelve additional genes in region two (excluding the regions containing rRNA genes already mentioned). Furthermore, two additional genes (and one pseudogene) were identified in region one, and one additional gene was identified in region two. Comparison of these two regions to the T3SS locus encoded on the megaplasmid of *D. vulgaris* reveals that all bar one gene present in the *D. vulgaris* system has a homologous gene in one or both of the T3SS regions of *Lawsonia*. Suggesting that whilst neither region by itself is sufficient to encode a functional T3SS, together they contain all the genes required to produce a functional T3SS. A direct visual comparison of the two regions to the *D. vulgaris* system (see Figure 13) shows the discontinuous relationship between the two regions and NF-T3SS locus of *D. vulgaris*, and also the location of the several genes which are present in both regions: *lscJ*, *lscL* and *lscQ*. These duplicated genes actually show less similarity (at the translated protein level) to each other than they do to the equivalent gene in the T3SS

locus of *D. vulgaris*. The lists of genes in the two loci are listed in Table 6 and Table 7.

Interestingly, comparison of the *Lawsonia* and *Desulfovibrio* NF-T3SS loci also reveals several ORFs within the *Desulfovibrio* NF-T3SS locus which were not picked up as genes when the genome was annotated. This list of eight new genes produce a range of T3SS protein products including several apparatus proteins (SctO and SctK), four proteins homologous to NF-T3SS chaperones, and one gene which is conserved in *Lawsonia* but is not present in any other genomes. The final gene is conserved only in the two strains of *Desulfovibrio vulgaris* thus far sequenced (strains Hildenborough and DP4).

Both the T3SSs of *D. vulgaris* and *L. intracellularis* show the presence of an ECF sigma factor regulatory system, which is encoded at either end of the T3SS locus of *D. vulgaris*, and also at corresponding ends of the second T3SS locus of *Lawsonia*. DVUA0099 and LI1150 are homologous to the BtrU serine/threonine phosphatase protein from *Bordetella*, and contain the characteristic HAMP-PP2C domain architecture found in other members of the same regulatory family [391]. DVUA0101 and LI1151 are homologous to the ECF sigma factor protein BtrS. At the opposite end of the loci are the remaining two components of the regulatory system. DVUA0123 and LI1167 are homologous to the anti-anti sigma factor BtrV and finally DVUA0124 and LI1168 are homologous to the serine kinase anti-sigma factor BtrW.

There are also several additional interesting features in the two *Lawsonia* NF-T3SS loci including the presence of several T3SS chaperones: Members of the CesT (typified by CesT of *E. coli*, which chaperones the T3SS effector Tir), and TPR repeat chaperones. Despite the presence of these chaperones there are no apparent effectors



**Figure 13. Comparison of Lawsonia T3SS loci to Desulfovibrio vulgaris T3SS locus.**

Lawsonia locus one on top (and reverse complemented), locus two at bottom, Desulfovibrio vulgaris locus in middle. Genes coloured by type/function, blue bars between regions represent homology detected between genes.

The T3SS genes are split fairly evenly between the two loci, and synteny is well conserved between the Lawsonia clusters and Desulfovibrio. All but one gene in Desulfovibrio has at least one homologue in Lawsonia. This strong degree of conserved synteny between the desulfovibrio and lawsonia is in sharp contrast to the situation between Desulfovibrio and Bordetella where there is very little conserved synteny.



| <b>Locus</b>  | <b>Gene Name</b> | <b>Location</b> | <b>Function</b>                |
|---------------|------------------|-----------------|--------------------------------|
| LI0537        | IscR             | Inner Membrane  | T3S Apparatus                  |
| LI0538        | IscQ             | Inner Membrane  | T3S Apparatus                  |
| LI0539        | IscO             | Inner Membrane  | T3S Apparatus                  |
| LI0540        | IscN             | Cytoplasm/IM    | ATPase                         |
| LI0541        | IscL             | Inner Membrane  | T3S Apparatus                  |
| LI0542        | IscK             | Inner Membrane  | T3S Apparatus                  |
| LI0543        | IscJ             | Periplasm       | T3S Apparatus                  |
| LI0544        | IscG             | Cytoplasm       | Chaperone                      |
| LI0544A       | IscF             | Extracellular   | T3S Needle Protein             |
| LI0544B       | cltT             | Cytoplasm       | Chaperone                      |
| LI0545        | -                | ?               | Conserved hypothetical protein |
| LI0546        | Icr2             | Cytoplasm       | Chaperone                      |
| LI0547        | -                | Cytoplasm       | Chaperone                      |
| LI0548        | IscV             | Inner Membrane  | T3S Apparatus                  |
| LI0549        | IscS             | Inner Membrane  | T3S Apparatus                  |
| LI0550        | IscT             | Inner Membrane  | T3S Apparatus                  |
| LI0551        | IscU             | Inner Membrane  | T3S Apparatus                  |
| LI0551A       | -                | -               | Pseudogene*                    |
| LI5SA         | rrfA             | Cytoplasm       | 5S rRNA subunit                |
| LI23SA        | rrlA             | Cytoplasm       | 23S rRNA subunit               |
| LI_tRNA-Ala-1 | -                | Cytoplasm       | Alanine tRNA                   |
| LI_tRNA-Ile-1 | -                | Cytoplasm       | Isoleucine tRNA                |
| LI16SA        | rrsA             | Cytoplasm       | 16S rRNA subunit               |

\*Homologous to the N-terminal region of LI1150

**Table 6.** List of genes present in *Lawsonia* T3SS locus one

| <b>Locus</b>  | <b>Gene Name</b> | <b>Location</b>    | <b>Function</b>                 |
|---------------|------------------|--------------------|---------------------------------|
| LI16SB        | rrsB             | Cytoplasm          | 16S rRNA subunit                |
| LI_tRNA-Ile-2 | -                | Cytoplasm          | Isoleucine tRNA                 |
| LI_tRNA-Ala-2 | -                | Cytoplasm          | Alanine tRNA                    |
| LI23SB        | rrlB             | Cytoplasm          | 23S rRNA subunit                |
| LI5SB         | rrfB             | Cytoplasm          | 5S rRNA subunit                 |
| LI1150        | ltrU             | Cytoplasm          | Serine/threonine phosphatase    |
| LI1151        | ltrS             | Cytoplasm          | ECF Sigma Factor                |
| LI1152        | lscX             | Secreted Component | Putative T3SS secreted protein  |
| LI1153        | lscW             | Secreted Component | T3SS regulator/switch           |
| LI1154        | -                | ?                  | Conserved hypothetical protein* |
| LI1155        | -                | ?                  | Conserved hypothetical protein* |
| LI1156        | -                | ?                  | Conserved hypothetical protein* |
| LI1157        | lcrH             | Cytoplasm          | Chaperone                       |
| LI1158        | lopB             | Extracellular      | T3SS translocon                 |
| LI1159        | lopD             | Extracellular      | T3SS translocon                 |
| LI1160        | lscC             | Outer Membrane     | T3SS apparatus                  |
| LI1161        | lscD             | Inner Membrane     | T3SS apparatus                  |
| LI1161A       | lscE             | Inner Membrane     | T3SS apparatus                  |
| LI1162        | -                | ?                  | Conserved hypothetical protein  |
| LI1163        | lscJ2            | Periplasm          | T3SS apparatus                  |
| LI1164        | lscL2            | Inner Membrane     | T3SS apparatus                  |
| LI1165        | lscP             | Cytoplasm/IM       | T3SS needle length regulator    |
| LI1166        | lscQ2            | Inner Membrane     | T3SS apparatus                  |
| LI1167        | ltrV             | Cytoplasm          | Anti-Anti-Sigma factor          |
| LI1168        | ltrW             | Cytoplasm          | Anti-Sigma factor               |

\*Homologous to each other and homologous to DVUA0108

**Table 7. List of genes present in Lawsonia T3SS locus two**

within the loci, with one exception. Within locus two there is a homologue of YscX from *Yersinia*. YscX has been shown to be secreted by the *Yersinia* T3SS, and to be required for a fully functional T3SS [393-395]. There is however, no evidence in the literature to suggest a role for it as a ‘traditional’ effector. Outside of this one secreted protein, the only remaining candidate effector genes in the *Lawsonia* loci are homologous only to hypothetical genes in *Desulfovibrio*, and in some cases *Hahella*. Finally within *Lawsonia* locus two there are three genes adjacent to each other which are all homologues. These three genes are all also homologous to DVUA0108 from *D. vulgaris*, but to no other known proteins, and as such their role remains enigmatic.

#### **4.3.2. Comparison and re-annotation of the *Hahella chejuensis***

##### ***T3SS regions***

*Hahella chejuensis*, like *Lawsonia intracellularis* contains two T3SS loci. However in the case of *Hahella*, these two loci seem to contain a complete set of T3SS genes, which suggests that there are two complete and separate T3SSs present in the *Hahella* chromosome. Analysis of these two regions does not reveal any additional genes which are missing from the genome annotation of the two NF-T3SS regions, it does however show several small genes (~100 nucleotides in length) which show no homology to any proteins present in any other bacteria thus far genome sequenced, suggesting that they may not be actual coding sequences. There are also several genes in both regions which are annotated as hypothetical genes, but have clear homology to known T3SS genes. The list of genes in each locus are shown in Table 8 and Table 9.

As was the case for *Lawsonia*, there are several genes present which show homology to chaperones, or have chaperone characteristic domains (e.g. TPR repeats), but there seems to be a paucity genes which show homology to known effectors, particularly in

| <b>Locus Tag</b> | <b>Gene Name</b> | <b>Location</b>          | <b>Function</b>        |
|------------------|------------------|--------------------------|------------------------|
| HCH_03240        | HscC             | Outer Membrane           | T3SS apparatus         |
| HCH_03241        | -                | -                        | ?                      |
| HCH_03242        | HscD             | Inner Membrane           | T3SS apparatus         |
| HCH_03243        | -                | -                        | ?                      |
| HCH_03244        | HscF             | Extracellular            | T3SS Needle Protein    |
| HCH_03245        | -                | -                        | ?                      |
| HCH_03246        | HscI             | Inner Membrane           | T3SS apparatus         |
| HCH_03247        | HscJ             | Inner Membrane           | T3SS apparatus         |
| HCH_03248        | HscK             | Inner Membrane           | T3SS apparatus         |
| HCH_03249        | HscL             | Inner Membrane           | T3SS apparatus         |
| HCH_03251        | -                | -                        | Rhs family protein     |
| HCH_03252        | -                | -                        | ?                      |
| HCH_03253        | -                | -                        | Rhs family protein     |
| HCH_03254        | -                | -                        | ?                      |
| HCH_03255        | -                | -                        | ?                      |
| HCH_03256        | -                | -                        | ?                      |
| HCH_03257        | -                | -                        | ?                      |
| HCH_03258        | -                | -                        | ?                      |
| HCH_03259        | -                | -                        | Transposase            |
| HCH_03260        | HscU             | Inner Membrane           | T3SS apparatus         |
| HCH_03261        | HscT             | Inner Membrane           | T3SS apparatus         |
| HCH_03262        | HscS             | Inner Membrane           | T3SS apparatus         |
| HCH_03263        | HscR             | Inner Membrane           | T3SS apparatus         |
| HCH_03264        | HscQ             | Inner Membrane           | T3SS apparatus         |
| HCH_03266        | HscP             | Inner Membrane           | T3SS apparatus         |
| HCH_03267        | HscO             | Inner Membrane           | T3SS apparatus         |
| HCH_03268        | HscN             | Cytoplasm/Inner Membrane | T3SS ATPase            |
| HCH_03269        | -                | -                        | ABC transporter        |
| HCH_03270        | -                | -                        | Transposase            |
| HCH_03272        | -                | -                        | Mannose Isomerase      |
| HCH_03273        | HscV             | Inner Membrane           | T3SS apparatus         |
| HCH_03274        | -                | Cytoplasm                | Chaperone              |
| HCH_03275        | -                | -                        | ?                      |
| HCH_03276        | ShcN             | Cytoplasm                | Chaperone              |
| HCH_03277        | HopN             | Secreted Component       | T3SS regulator/switch  |
| HCH_03278        | HopD             | Extracellular            | T3SS translocon        |
| HCH_03279        | HopB             | Extracellular            | T3SS translocon        |
| HCH_03281        | ShcD             | Cytoplasm                | Chaperone              |
| HCH_03282        | -                | -                        | ?                      |
| HCH_03283        | -                | -                        | Conserved Hypothetical |
| HCH_03284        | -                | -                        | ?                      |
| HCH_03285        | -                | -                        | LuxR family regulator  |
| HCH_03286        | ChsT             | Cytoplasm                | Chaperone              |

**Table 8.** List of genes present in *Hahella* T3SS locus one

| <b>Locus Tag</b> | <b>Gene Name</b> | <b>Location</b>    | <b>Function</b>       |
|------------------|------------------|--------------------|-----------------------|
| HCH_05096        | HopD2            | Extracellular      | T3SS translocon       |
| HCH_05097        | HopB2            | Extracellular      | T3SS translocon       |
| HCH_05098        | ShcD2            | Cytoplasm          | Chaperone             |
| HCH_05099        | Hsp22            | Secreted Component | Secreted protein      |
| HCH_05100        | -                | -                  | ?                     |
| HCH_05101        | HscV2            | Intracellular      | T3SS apparatus        |
| HCH_05102        | HscY             | Cytoplasm          | Chaperone             |
| HCH_05103        | HscX             | Secreted Component | Secreted protein      |
| HCH_05104        | -                | -                  | ?                     |
| HCH_05105        | ShcN2            | Cytoplasm          | Chaperone             |
| HCH_05106        | HopN2            | Secreted Component | T3SS Regulator/Switch |
| HCH_05107        | HscN2            | Cytoplasm          | ATPase                |
| HCH_05108        | HscO2            | Intracellular      | T3SS apparatus        |
| HCH_05109        | -                | -                  | ?                     |
| HCH_05110        | HscQ2            | Intracellular      | T3SS apparatus        |
| HCH_05111        | HscR2            | Intracellular      | T3SS apparatus        |
| HCH_05112        | HscS2            | Intracellular      | T3SS apparatus        |
| HCH_05113        | HscT2            | Intracellular      | T3SS apparatus        |
| HCH_05114        | HscU2            | Intracellular      | T3SS apparatus        |
| HCH_05115        | -                | -                  | ?                     |
| HCH_05116        | -                | -                  | ?                     |
| HCH_05117        | -                | -                  | ?                     |
| HCH_05118        | -                | -                  | ?                     |
| HCH_05119        | HscL2            | Intracellular      | T3SS apparatus        |
| HCH_05121        | HscK2            | Intracellular      | T3SS apparatus        |
| HCH_05122        | HscJ2            | Intracellular      | T3SS apparatus        |
| HCH_05123        | HscI2            | Intracellular      | T3SS apparatus        |
| HCH_05124        | -                | -                  | ?                     |
| HCH_05126        | -                | -                  | ?                     |
| HCH_05127        | -                | -                  | ?                     |
| HCH_05128        | -                | -                  | ?                     |
| HCH_05129        | -                | -                  | ?                     |
| HCH_05130        | HscD2            | Intracellular      | T3SS apparatus        |
| HCH_05131        | HscC2            | Extracellular      | T3SS apparatus        |
| HCH_05132        | HscE             | Intracellular      | T3SS apparatus        |
| HCH_05133        | ChsT2            | Cytoplasm          | Chaperone             |
| HCH_05134        | -                | Cytoplasm          | Chaperone             |

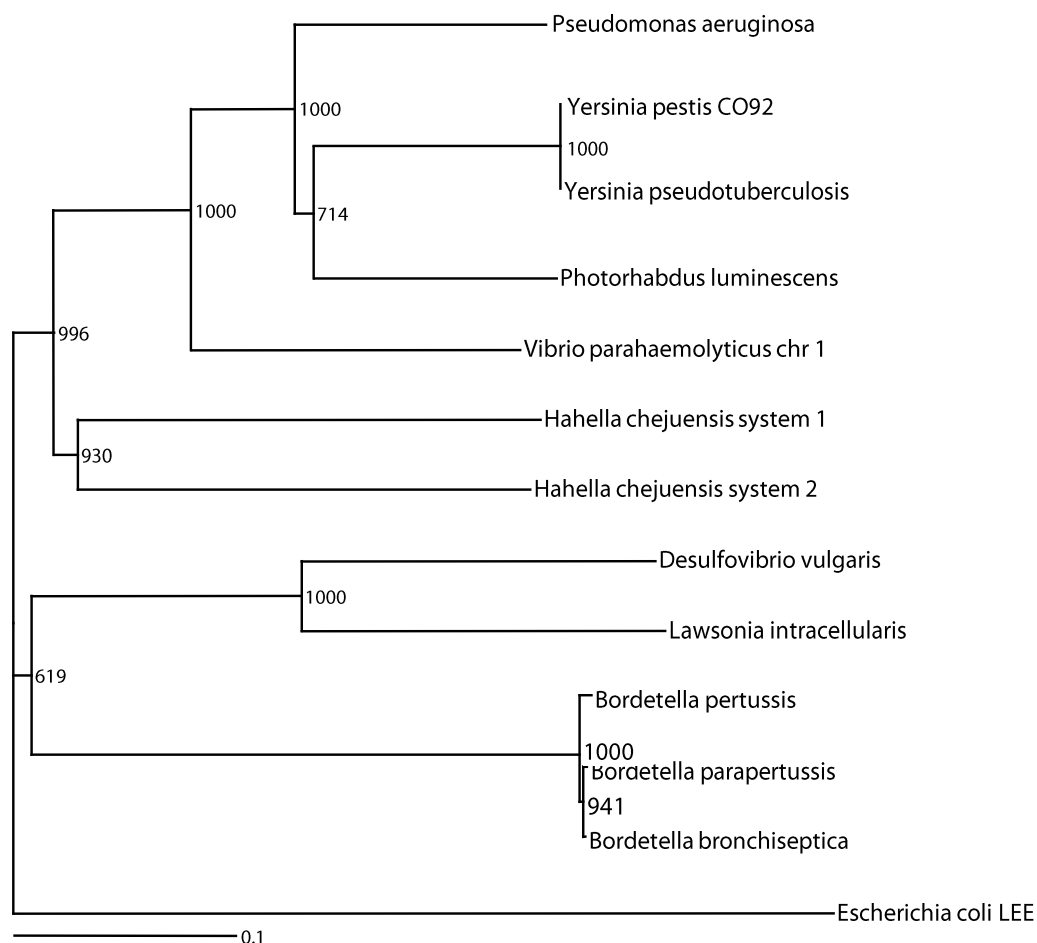
**Table 9.** List of genes present in *Hahella* T3SS locus two

locus one. Within locus two there are two genes which demonstrate homology to known T3SS effectors: HCH\_05103, which is homologous to YscX from *Yersinia*, a protein also found within the second T3SS locus of *Lawsonia*; and HCH\_05099 which is homologous to Bsp22, a T3SS secreted protein in *Bordetella pertussis* shown to be commonly detectable in clinical strains of the bacterium [396, 397].

There are also several genes missing from locus one and/or locus two, which in other systems are required for a functional secretion system. This includes the T3SS needle protein (SctF) which is absent from locus two. Conversely locus one appears to be missing the two chaperone proteins (SctE and SctG) required for SctF. These data suggest that in order for either system to function there must be some interplay between the two loci in order to form a functional system. Aside from the example of SctF above both systems appear to have a complete set of the core set of genes required to encode all the inner membrane apparatus, periplasmic spanning protein and outer membrane secretin.

Unlike other members of this NF-T3SS family neither locus encodes any genes that may be part of an ECF sigma regulatory system, meaning that regulation of the T3SSs in this bacterium is most likely achieved by other means. It does however also encode a SepL family protein in both loci like *Desulfovibrio*, *Lawsonia* and *Bordetella*, rather than as two separate genes (*yopN* and *tyeA*) as in *Yersinia*.

Finally it is interesting to examine these two systems side by side from an evolutionary perspective. Phylogenetic analysis places these two loci as each other's closest relative (see Figure 14), suggesting that these two loci most likely arose as a result of paralogy rather than separate horizontal transfer events. The appearance of paralogous genes within the chromosome has resulted in a great deal of divergence in



**Figure 14. Phylogenetic tree of Ysc type T3S systems**

Neighbour-joining tree drawn from a t-coffee alignment of conserved domains (Bac\_export\_1, Bac\_export\_2, Bac\_export\_3, FHIPEP, Secretin) from five proteins (SctT, SctU, SctS, SctV and SctC) belonging to Ysc T3SS systems. Tree is rooted on the LEE system from *E. coli*, and bootstrap values are calculated from 1000 replicates.

This tree demonstrates the separate subgroups of T3SSs belonging to the Ysc ‘supergroup’. The two systems of *Hahella* are close relatives to the traditional Ysc systems, whilst the *Desulfovibrio* and *Lawsonia* systems are more closely related to the *Bordetella* systems.

protein sequence for these two systems, resulting in the deep branch length observed in trees drawn based on the sequences of conserved proteins. A side by side comparison of the two clusters also reveals a large amount of gene rearrangements between the two loci, a situation which is likely explained by the large number of transposases located within the NF-T3SS locus one.

## 4.4. Discussion

### 4.4.1. Mis-annotation of T3SS Regions

The first thing which becomes clear from the analysis of the secretion systems present in these three bacteria is the quality of the annotation of the NF-T3SS loci. In all three cases there are genes which have been missed within the annotation, and numerous genes annotated simply as ‘hypothetical’, when clear homology exists with genes in other bacteria. This creates a series of problems for researchers looking to use the annotation data in any form of analysis. Whilst the problem of genes which do not have an annotation of product, or in some cases any form of gene name, can be solved simply by a quick examination of search results from programs such as BLAST or HMMER. More problematic, however, is the issue of genes entirely missing from the genome annotation. In such cases, researchers may end up making false assumptions about the functionality of certain systems and pathways within the bacterium based on the annotation, and the only way to correct this issue is to undertake a complete examination of the DNA sequence of the genome along with a re-calling of the coding sequences, in order to identify any genes which were not marked as such in the initial annotation of the genome.

In the example of the *Lawsonia* NF-T3SS loci, a researcher using the published



genome sequence and annotation would have assumed that there was no functional T3SS within the genome, owing to the lack of a gene encoding the NF-T3SS needle protein. However, the re-annotation shown here demonstrates that not only is this essential gene present but, there are several other genes present within the NF-T3SS loci which represent members of the NF-T3SS apparatus, and also as chaperones of NF-T3SS effectors. This problem of incomplete and partial annotation is a problem inherent in the growing rate of sequencing of bacterial genomes. Many of the institutes and groups now generating DNA sequence data are not necessarily experts in genome annotation. The ever reducing cost of sequencing bacterial genomes is likely to continue this trend, and in the process thousands, and maybe even tens of thousands of new genomes will be deposited in public databases over the next few years. Based on the evidence here it is likely that many if not most of these genomes will feature incomplete or inaccurate annotations. This leaves researchers with an important problem: Should genome sequences continue to be annotated by a small group of individuals (i.e. those directly associated with the sequencing project), producing what will most likely be an incomplete annotation; or should genomes be deposited as sequence data only, leaving the onus on other researchers to locate their genes of interest within the genome.

Each solution presents its own advantages and disadvantages. If the *status quo* remains then all annotations must be viewed with a sceptical eye, but none the less the annotation should give others a quick insight as to whether their gene/system of interest is present in the genome. The alternative solution removes any issues with the potential of false inference based on inaccurate data, but means that anyone interested in mining the genome space for information requires more than a passing knowledge of bioinformatics and genome annotation techniques. Whatever the solution, in the

future it would seem clear that when in doubt researchers should be sure of the annotation before making any conclusions based on it.

#### **4.4.2. Evolution through Paralogy**

One of the other interesting observations that can be made from observing the NF-T3SS loci present in *Lawsonia* and *Hahella* is the role that paralogy plays in their evolution. With the presence of multiple copies of a gene within a genome comes the possibility that one of the genes can evolve to fulfil a different function so long as at least one of the copies remains able to fulfil its original role. In this way many new mutations now become permissible without any noticeable (or at least not deleterious) change in phenotype.

The realities of this can be seen by examining the duplicated genes present within these two genomes in comparison with other genes from related NF-T3SS loci in other bacteria. In the example of the *Lawsonia* NF-T3SS loci there are a total of three genes, which at the translated protein level show less similarity to each other than they do to the equivalent genes from other bacteria. This situation is mirrored within *Hahella*, where genes present in two copies show a great deal of divergence in their translated protein sequences. Of the twenty-one genes present in duplicate over half show less similarity to each other than they do to genes present in bacteria such as *Lawsonia*, *Desulfovibrio*, *Vibrio*, *Yersinia* and *Photorhabdus*.

Despite the divergence of the two NF-T3SS loci present in *Hahella*, they are, phylogenetically speaking, their nearest relatives, suggesting a single inheritance and subsequent duplication, rather than two separate horizontal transfer events. However, the precise role that each of the T3SSs encoded within the *Hahella* chromosome plays has yet to be elucidated, and as such the effect that the divergence of the two clusters

has played remains unknown.

#### **4.4.3. Stepwise evolution and conserved regulation in T3SSs**

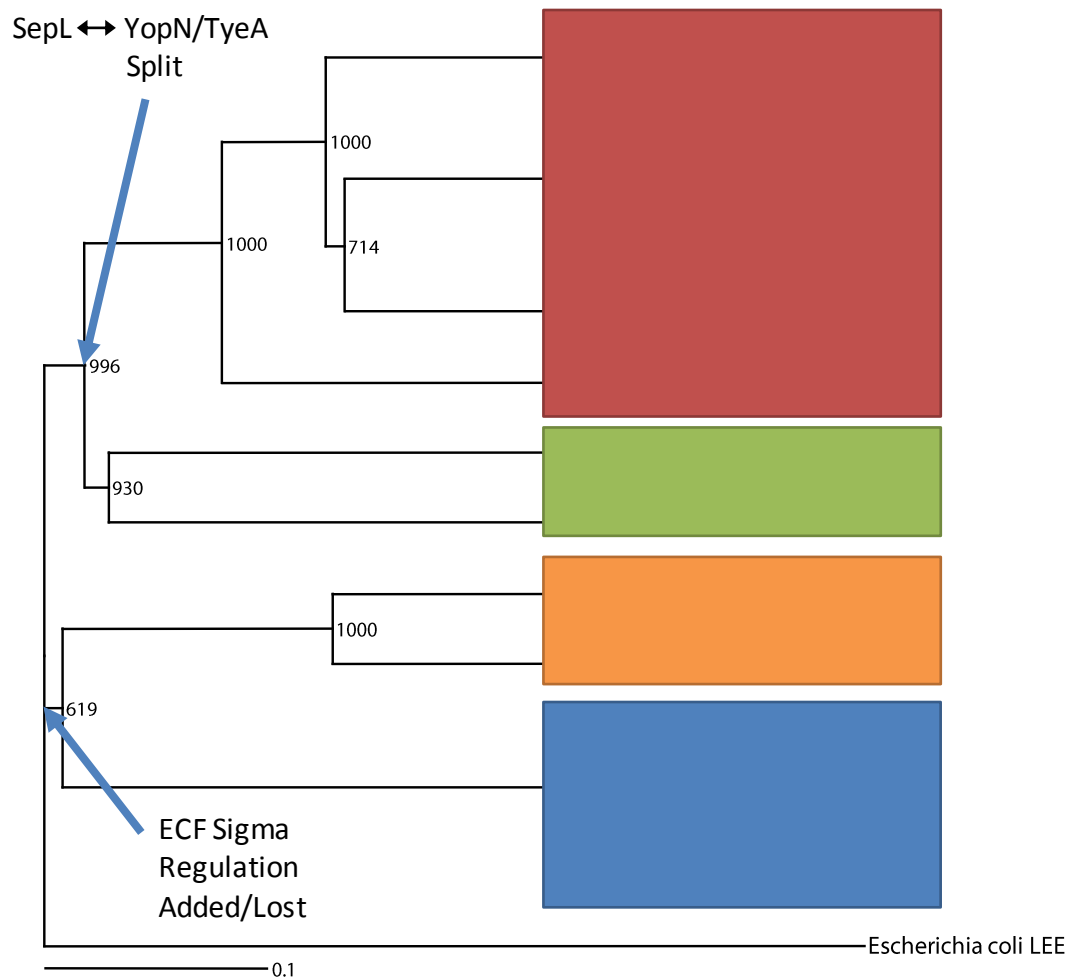
Examination of the NF-T3SSs examined in depth here shows the gradual evolution and changes that occur in similar NF-T3SS loci. When investigating these T3SSs in conjunction with other members of the Ysc NF-T3SS family (*Yersinia* and close relatives plus *Bordetella*), several key differences appear. The first of which is the phylogenetic relationship between the systems. There is clear splitting of the systems examined here into two groups: *Lawsonia* and *Desulfovibrio* which both cluster closest to the three NF-T3SS in *Bordetella* species, and the two *Hahella* T3SS which cluster closest to *Yersinia* species. This separate clustering suggests that there are actually at least two separate sub-groups of T3SSs within the ‘Ysc’ group as described by Foulter *et al.*

Not only does this analysis demonstrate the different classes or groups of NF-T3SSs revealed by phylogenetic analysis, but also the gradual gain/loss of genes within these clusters which results in the range of T3SSs we see in bacteria today. Also of interest is the distribution of these changes with reference to the ‘classical’ (i.e. sequence based) phylogenetic trees of T3SSs. Two of the key aspects of the systems analysed here are the presence/absence of SepL or YopN/TyeA homologues in systems which would appear to be the closest relatives of the ‘Ysc’ group of NF-T3SSs, and the regulation of these NF-T3SS systems by ECF sigma regulatory components. In both of these cases we see map the gene loss/gain events onto branching points on a phylogenetic tree (See Figure 15). Based on the evidence presented here it would seem that the ECF-sigma system was gained/lost following the point in time when the *Bordetella* (and *Desulfovibrio* & *Lawsonia*) systems diverged from the traditional Ysc

group systems (and *Hahella*). Similarly it is most likely the split of *sepL* into *yopN/tyeA*, or *vice versa*, occurred after the T3SSs of *Hahella* and *Yersinia* diverged. There are also more subtle gene additions/losses which have occurred amongst these T3SSs. One such example of this is the presence of a homologue of Bsp22, an effector protein first characterised in *Bordetella*, within the second T3SS locus of *Hahella* (but not found in *Lawsonia* or *Desulfovibrio*).

There are also several genes which we can see are ubiquitous to all the T3SSs systems within the Ysc/*Bordetella* super-group of T3SSs, but are not found in any other groups of T3SSs. This includes genes such as *yscX*, which encodes a secreted protein, *yscG* a gene encoding a chaperone, and *yscP*, which encodes the needle length regulator protein. *yscP* is a particularly interesting case, as whilst sequentially homologous proteins can only be found in Ysc group T3SSs, there are functionally analogous proteins found in the Inv-Mxi-Spa group of T3SSs (e.g. Spa32 and InvJ from *Shigella flexneri* and *Salmonella enterica* respectively). Since this group of proteins is so permissive of changes in primary sequence whilst retaining its function it is perhaps surprising to see that obvious sequence homology can be observed amongst systems within the same group.

Taken together the data obtained by comparing the genomes examined here with the other genomes belonging to the Ysc group of NF-T3SS offers an interesting view into the diversity in the gene complement of NF-T3SSs. Many of the genes within this group are well conserved throughout all the NF-T3SSs within the group. None the less there are some interesting differences outlined here, such as the different hypothetical mechanisms of regulation utilised by these systems. It is also interesting to note not only these differences but also the conserved features of the Ysc group



**Figure 15. Phylogenetic tree of Ysc type T3S systems, with groupings and evolutionary annotations**

The tree is drawn using the same method as Figure 14. Boxes around names of systems denote systems with similar gene complements. Arrows point to the most likely point at which the annotated changes took place.

This tree adds to the evidence suggesting different subgroups within the Ysc group, by demonstrating that there are not only sequence differences between conserved proteins within these subgroups, but also difference in their gene complement.

which are not found elsewhere. Such features are excellent case study in how different bacteria can evolve different methods to perform the same task.

## 4.5. Summary

Examination of the NF-T3SSs within the genomes inspected here reveals that the quality of the genome annotation in these bacteria is quite variable. In all the NF-T3SSs there were either genes missing or additional genes annotated which most probably do not exist. This in conjunction with the complete misannotation of the ribosomal RNA regions in *Lawsonia intracellularis* and the missing functional annotations of numerous genes despite clear homology data in all genomes suggests that comprehensive homology searching when used with caution can be used to transfer annotation details to a large number of genes, and also to determine coding sequences. This approach needs to be used with care however, particularly for multi-domain proteins or where the similarity is very low. The conserved genes amongst these and the other NF-T3SS loci with the Ysc groups allow for phylogenetic comparison of these systems, which shows the Ysc group of T3SSs splitting into several subgroups each with its own differences in gene complement, and also allow us to see points in evolutionary time when these changes such as the adoption/removal of regulation by ECF sigma systems amongst certain NF-T3SSs took place.

## CHAPTER 5 CONSERVED AND SPECIFIC FEATURES OF T3S SYSTEMS

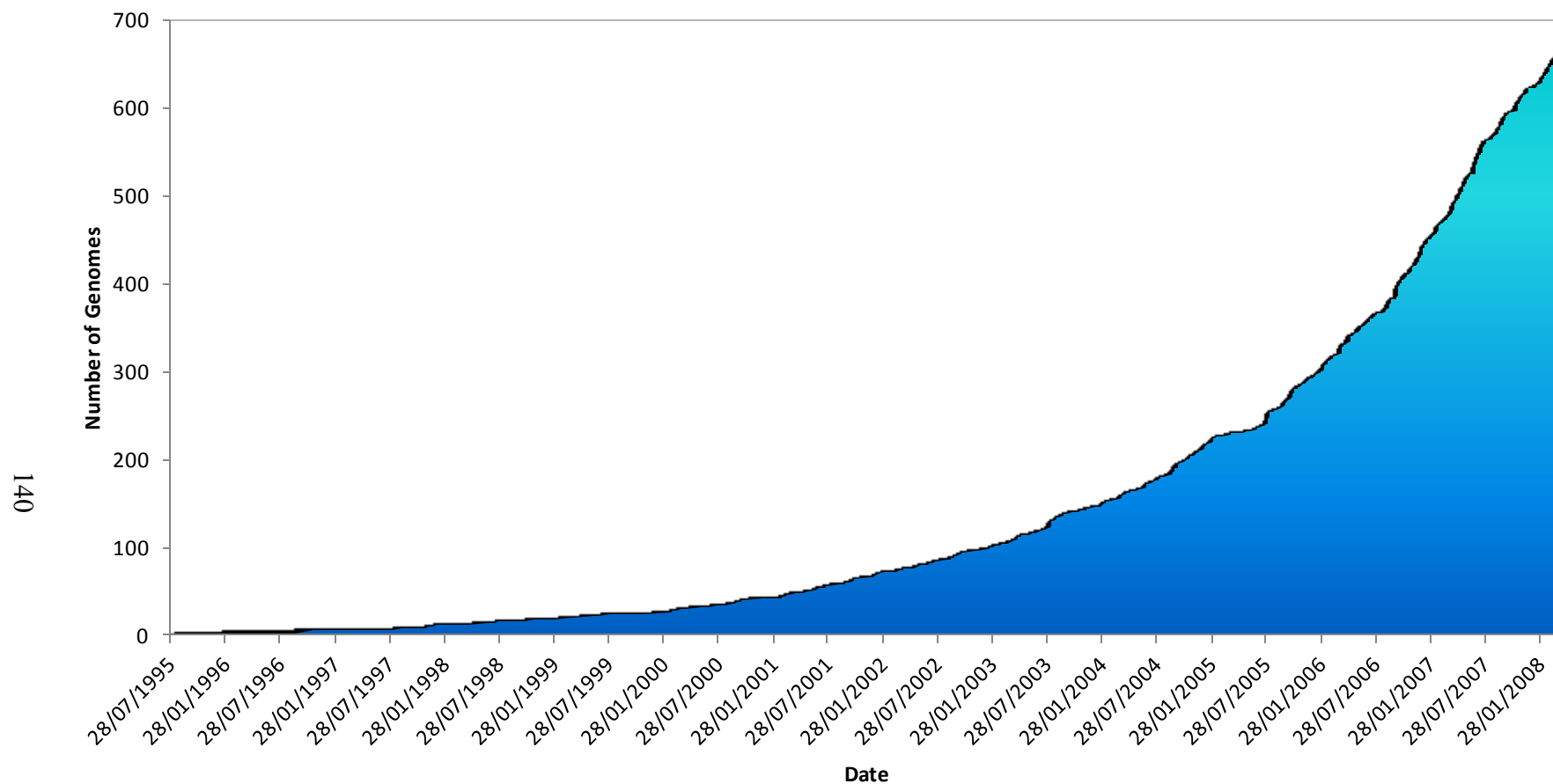
### 5.1. Introduction

#### *5.1.1. The breadth of type-III secretion*

One of the key problems which the prevalence of bacterial genome sequencing now presents the type-III secretion community with is the task of determining which of these genomes contain their secretion system of interest. Whilst there are a number of genes which are unique to the T3SSs apparatus, their presence in a genome do not necessarily denote that the system will or indeed could be functional. Secondary to this problem is the issue of determining whether such genes belong to flagellar or non-flagellar T3SSs.

Whilst in the past a great deal of time and effort was expended by those who sequenced the genome in annotating it, there are now however many genomes being deposited in databases solely as the shotgun reads of the genome without any attempt made to call open reading frames / coding sequences. As a result of this it is no longer possible for scientists working the field of type-III secretion to determine whether a bacterium contains a T3SS solely by examining the annotation of the genome. This problem is further compounded by the continued exponential growth in genome sequencing data (see Figure 16), making any sort of manual attempt to locate T3SSs a rapidly unfeasible task.

Previous attempts to define the breadth of NF-T3SSs have shown them to be located solely within the Proteobacterial and Chlamydial phyla of the Bacteria kingdom



**Figure 16. Graph showing expansion in bacterial genome data**

Graph shows number of genomes with a deposition date on or before that shown on the x-axis. Note the roughly exponential shape to the data, demonstrating the ever increasing number of genomes which have been and will be added to the public databases.



[196, 297, 385]. Their distribution within these phyla is by no means universal, nor uniform. There are for example no members of the epsilonproteobacteria with a non-flagellar T3SS (the same bacteria do have numerous examples of flagellar T3SSs). However, all other classes within the proteobacterial phylum do contain examples of non-flagellar T3SSs. This includes *Rhizobia* in the alphaproteobacteria [398], *Bordetella* and *Burkholderia* in the betaproteobacteria [388, 399], and *Desulfovibrio* and *Lawsonia* in the deltaproteobacterial class [386]. The vast majority of classical T3SSs are members of the gammaproteobacteria class. This includes the well-studied T3SSs of *Yersinia*, *Salmonella*, *Shigella* and *E. coli* [352, 400].

Within the Chlamydiae phylum there have been far fewer genomes sequenced than for Proteobacteria, however, of the nearly dozen or so genomes available there is evidence that most, if not all, have a non-flagellar T3SS. This group of NF-T3SSs also shows several distinct characteristics, such as the splitting of the NF-T3SS genes into multiple distinct loci. These systems also show a good degree of similarity in protein sequence and genomic locale. This is in stark contrast to the T3SSs present in Proteobacteria, where the closest phylogenetic relatives are found in diverse bacteria. For example the Ysc group of NF-T3SSs discussed in Chapter 4, consists of members of three separate classes of Proteobacteria (alpha-, beta- and gammaproteobacteria). Many of the T3SSs present in Proteobacteria also show evidence of horizontal gene transfer such as aberrant GC or codon usage compared to the genomic backbone [306], an attribute not shared by Chlamydial NF-T3SSs

All of these aspects place several obstacles in the path of those looking to determine those bacteria likely to possess a NF-T3SS. Whilst searches could be limited to just the two phyla where NF-T3SSs have already been found, one should be wary of the

fact that absence of evidence does not imply evidence of absence. As such any attempt to determine the breadth of non-flagellar type-III secretion amongst bacteria must begin by looking at the entire kingdom, not just a subset. This situation is even truer for flagellar systems, where many more bacterial phyla have already been shown to contain F-T3SSs. Fortunately however, there are also characteristics of T3SSs which can aid *in silico* searching of genome sequences. Firstly, they are for the most part encoded on one locus which will often stand out from the genome owing the biases in the nucleotides/codons it possesses. Even in the case of the Chlamydiales where the genes which encode the T3SS apparatus are found in multiple loci, there is still a tendency for the genes to remain together in a series of only 3-5 loci spaced around the chromosome [401]. The proteins which these genes encode also show a good degree of similarity and as such are amenable to location by homology searching techniques. In order to do this proteins must be selected which show both a good degree of similarity amongst all T3SSs otherwise standard homology searching methods will not work. Secondly, and also probably more obviously, the protein being used must be conserved amongst all T3SSs.

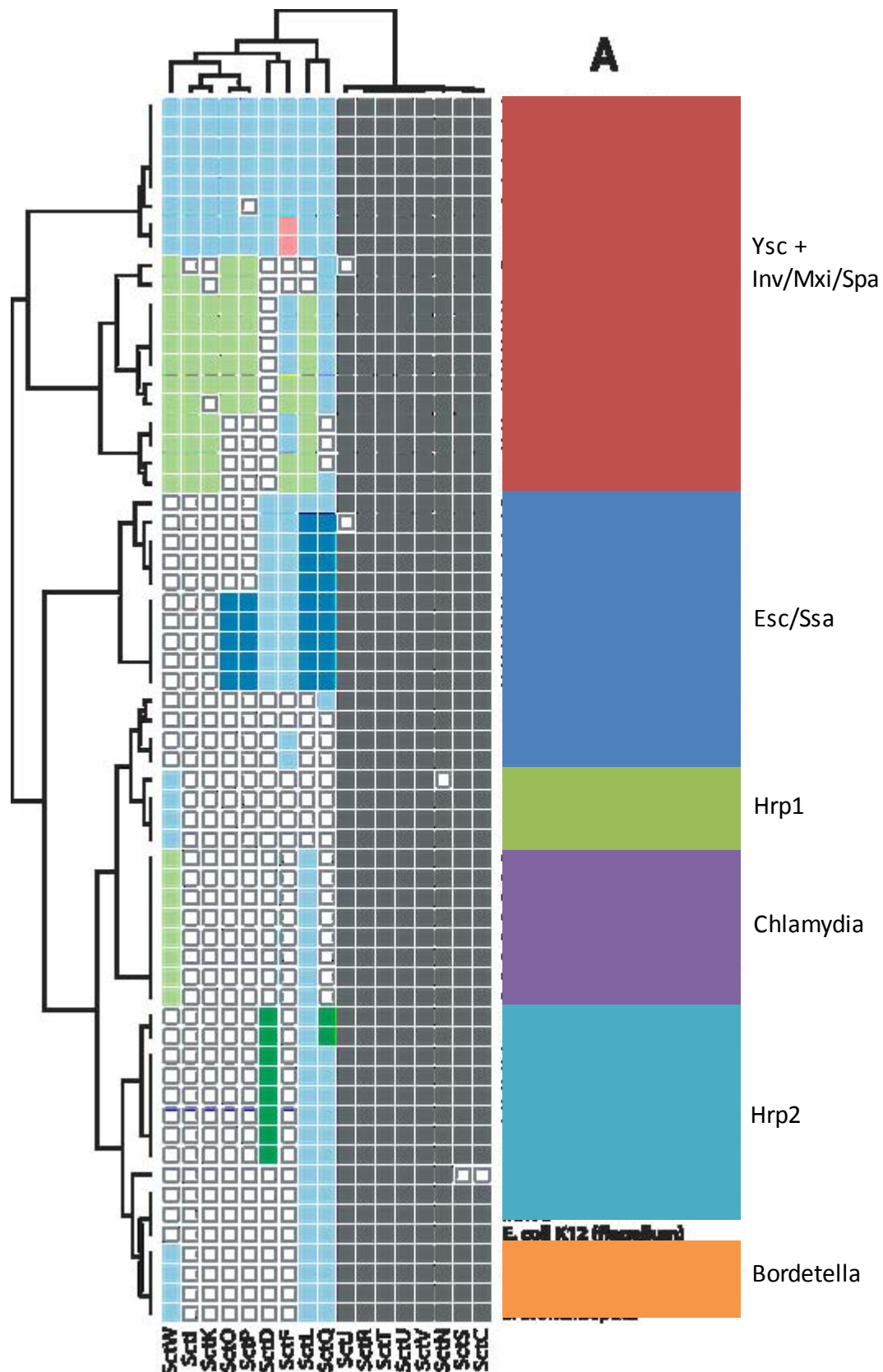
### ***5.1.2. Diverse gene complements in T3SSs***

Related to the problem of determining the diversity of T3SSs, is the task of determining what the complete gene complement of a T3SSs is. This job is fortunately aided by some of the characteristics mentioned above, that being the innate bias found in the DNA which encodes many T3SSs compared with the genomic backbone, and the sequence similarity seen amongst member proteins of T3SSs. By using these facts to locate and determine the boundaries of T3SSs, the complete T3SS locus (or loci) can be located in a bacterium's DNA, and in the

process the complete gene complement of the T3SS can be obtained. By looking at this set of gene products amongst all known T3SSs it should become possible to determine the conserved and unique sets of proteins belonging to these systems.

A more complex approach is to start with no *a priori* assumptions about gene loci or protein function and to attempt to cluster all proteins into families based on homology data alone. Such an approach was undertaken in a study by Medini *et al* [307], using an algorithm they term Overlap, which generates protein homology networks (PHNs) based on homology found through reciprocal BLAST searches. This approach generates a series of densely connected graphs where each graph represents a group of proteins with a conserved function. This approach is obviously computationally complex ( $O(n^2)$ ), and requires a large amount of CPU time in order to calculate all reciprocal BLAST pairs (in their study Medini *et al* used a database of ~750,000 proteins, which at 30 seconds per search, the approximate time for one BLAST search on a single 3.00GHz Intel processor with an equivalently sized database, would require 6250 CPU hours).

By extracting PHNs which contain proteins annotated as belonging to the T3SS apparatus they were able to survey both for the presence/absence of T3SSs in bacteria, but also to examine patterns of conservation of various T3SS components. As part of this study they found a series of proteins which showed conservation in blocks of T3SSs but were not universal to all T3SSs (See Figure 17). Such examples include several proteins which were already known to be not universal to all T3SSs such as the needle length regulation protein SctP [264, 266], and the outer membrane lipoprotein SctW (See Chapter 2). In fact it is interesting to note that the distribution of SctW proteins almost exactly mirrors the distribution of C-terminal regions in SctC



**Figure 17. T3SS families identified from Protein Homology Networks**

One column per protein family, empty squares represent absence, different colours represent different PHNs. Groupings of genomes on right based on nomenclature/data in [297]. Adapted from [307]. Empty squares: no member of this PHN in this bacterium, other colours: Member found. Different colours in same column: multiple distantly related PHNs in same superfamily.

(secretin) proteins (See Figure 6), with SctW being present in all T3SS groups apart from Esc/Ssa and Hrp2. As was seen in Chapter 4 the distribution of different proteins shows good correlation with sequence based phylogenetic trees, and with the NF-T3SS groupings overlain on Figure 17 patterns of gene presence/absence can be seen in the different groups. However, the downside of this approach is the lack of a human supervised element to the procedure, which may lead to under or over prediction of gene complements. For example EscD, an inner membrane component of the NF-T3SS in *E. coli* should appear in the SctD column, but mysteriously is listed as 'absent' based on the PHN data. Similarly five of the six proteins shown as absent in *Desulfovibrio vulgaris* are present in the data presented in Chapter 4.

Related to the issue of T3SS diversity is the problem of defining the minimal NF-T3SS, that being the smallest set of genes required to produce a fully functional secretion system. This may seem at first to be a trivial issue: simply survey all T3SSs and locate those genes present in all systems. This simplistic option however does not take into account the evidence that there are certain proteins required for secretion in some NF-T3SSs which are apparently absent in other systems. One such example of this is LcrV from *Yersinia* sp. In its absence the translocon proteins YopB and YopD are unable to assemble correctly to form a functional 'tip' to the needle of the T3SSs [274]. Whilst a  $\Delta lcrV$  mutant is capable of forming a functioning apparatus which can export proteins, it is unable to translocate these proteins into host cells, a function which requires correct assembly of YopB and YopD [402]. LcrV is found only in members of the Ysc group of T3SSs, but YopB and YopD homologues are found in a wide range of NF-T3SSs, which obviously are able to form a functioning translocon in the absence of LcrV.

Thus we must find a system to categorise member proteins of T3SSs which encapsulates this issue. Proteins which are common to all T3SSs are almost certain to be essential for secretion; those conserved only in groups of T3SSs are more problematic. These proteins may or may not be required for secretion, and homology searching alone is ill-equipped to provide the information needed to make the distinction.

### ***5.1.3. Mapping diversity to evolution***

Understanding the diversity shown amongst different T3SSs is key to our comprehension of how these systems function. Owing to the complex interplay between proteins within T3SSs it is hard to tease apart assembly of the secretion system and the individual function of proteins within it. It is this essential issue which had led to it being labelled an ‘irreducibly complex’ system by those in the intelligent design community [403-405]. The argument of irreducible complexity posits that systems that are composed of multiple proteins which interact and contribute to the function of the system, and where removal of any one of those proteins leads to the system to stop functioning, could not have evolved naturally [403].

However, this position runs counter to several lines of evidence: Firstly there are several proteins (for example the ATPase and Secretin components) within T3SSs which are found in multiple other cellular components, demonstrating the ability for individual proteins to function in a multitude of roles rather than just one ‘closed’ system which arose as a finished product. Secondly is modularity of T3SSs and the variation in their gene complement. The differences between T3SSs and their varying dependence on different proteins demonstrates that a stepwise process has been functioning in the development of the T3SSs found in different systems.

Sequence based phylogenetic techniques have been used in the past to determine different groups of T3SSs. One such example of this placed T3SSs into five separate groups: Ysc, Inv-Mxi-Spa, Esa-Ssa, Hrp1 and Hrp2 [297]. Each of these five groups can be seen in phylogenetic trees drawn using several different T3SS proteins, and interestingly can also be seen when using alignment free techniques to cluster T3SSs (See Figure 17). This data provides strong evidence for changes in T3SS gene complements during evolutionary events. For example we can postulate that changes occurred following the divergence of the Ysc/Inv-Mxi-Spa from other T3SSs which led to the recruitment of SctI to these systems. By mapping these progressive changes in gene gain/loss onto phylogenetic data we can begin to understand the changes which occurred and the order in which they occurred and hence begin to understand the gradual evolution of T3SSs (and also in the process add more evidence against T3SSs being ‘irreducibly complex’).

#### **5.1.4. Aims**

By locating the complete complement of T3SSs in bacteria by looking for proteins which are unique to these systems, the methods used in this chapter aim to demonstrate the breadth of bacteria which contain a T3SS. Of all the bacterial phyla for which there is a genome sequence so far only two seem to have a NF-T3SSs, and so it is unlikely that NF-T3SSs will be located in other phyla. Choosing proteins or domains which are common to all T3SSs should allow for location of all T3SS loci in sequenced genomes. By focusing attention on these loci, homology searching techniques can be employed without the requirement for massive amounts of computing resources, and genes can be identified which are present in sufficiently few loci that they may not have been located using a technique such as the PHN mapping

system mentioned above. This data set should be able to provide enough information about the different proteins encoded within these loci to make determinations about their degree of conservation amongst all known T3SSs. The expectation would be that proteins fall broadly into three categories: proteins conserved amongst all T3SSs, proteins conserved amongst multiple families of T3SSs, and proteins unique to individual or closely related T3SSs only. This information on patterns of conservation may also be used to compare the presence/absence of certain components to phylogenetic data, where once again patterns should emerge which demonstrate potential stepwise evolutionary changes which led to these differences amongst different T3SSs.

## **5.2. Methods**

### ***5.2.1. Locating T3SSs in completed genomes***

The complete set of bacterial genomes was downloaded from the NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria>) in June 2007. The proteins from each of these bacteria were concatenated together into a single FASTA format file, and then purged of redundancy at 100%. BLAST searches were performed using known T3SSs genes from *Yersinia pestis* and *E. coli* as query sequences and the non-redundant sequences generated above as the database with the BLASTP algorithm and the default parameters unless otherwise specified. HMMER searches were performed by the `hmmsearch` program on the same database as was used for the BLAST searches, using domain models related to type-III secretion (see Table 5), both `Pfam_ls` and `Pfam_fs` domain models, and their respective default alignment modes (global and local respectively) were used for this search. Relaxed criteria were chosen as any overprediction/false-positives could be corrected for at later search stages.



### ***5.2.2. Defining T3SS regions within genomes***

Proteins with hits to T3SSs domains were automatically clustered together based on their genomic locale. By default when two proteins lay within 50Kb of each other within the genome they were joined together and the coordinates of a T3SSs locus were defined as the start of the first identified T3SS gene in the cluster plus 25Kb of upstream sequence to the end of the last identified T3SS gene in the cluster plus 25Kb of downstream sequence. The protein products of all the genes in a cluster were then extracted and placed into a database of putative type-III secretion proteins

### ***5.2.3. Generating networks of related proteins***

The database of putative type-III secretion proteins generated in 5.2.2 was used to perform a complete reciprocal PSI-BLAST search (i.e. every protein sequence in the search database was used as a query sequence), with each PSI-BLAST search being run to convergence or a maximum of ten iterations. Data from these BLAST searches (i.e. iteration one of the PSI-BLAST search) was then used to define networks of homologous proteins. Proteins were considered to be part of the same network when BLAST reported a hit between the two proteins with an e-value less than 0.001. More stringent search criteria were chosen as the result set were to be analysed automatically. As any errors could not easily be corrected manually, and could be magnified by the subsequent networking process, a reduction in the false positive rate was desired. Networks were recursively grown from a starting protein based on the e-value cut-off criterion, until no new proteins could be added to the network. Networks were then drawn using the neato algorithm (an implementation of the Kamada-Kawai algorithm [406]), which comes as part of the GraphViz software suite [407]. Neato draws graphs based on a “spring” model, where nodes are pulled together based on

the weights of edges joining them, as such the weight associated can be thought of as the strength of a hypothetical spring joining the two nodes together. Graphs were exported in the scalable vector graphic (SVG) format. Networks were considered to be overlapping if they could be joined together using data from PSI-BLAST searches. The requirement for joining two networks being that there was a reciprocal hit between proteins from the two different networks with an e-value less than 0.001 in both directions.

#### ***5.2.4. Defining the conserved and specific sets of T3SS***

Phylogenetic trees were drawn using alignments generated from six separate conserved domains: Bac\_export\_1, Bac\_export\_2, Bac\_export\_3, FHIPEP, Secretin and YscJ\_FliF. For each protein where a domain hit was found, the region was extracted and all examples of the domain were then aligned against each other using T-Coffee with default parameters. Where domain hits were found in a T3SS locus for all six domains, the alignments were concatenated to produce a single alignment file. This alignment was then fed into ClustalW in order to produce a neighbour-joining phylogenetic tree. Each leaf of the tree represents a single T3SS locus, and network data (i.e. presence/absence of a protein belonging to the network in question, within that locus) were mapped onto the tree using a custom application written in Perl.

### **5.3. Results**

#### ***5.3.1. Finding T3SS Loci***

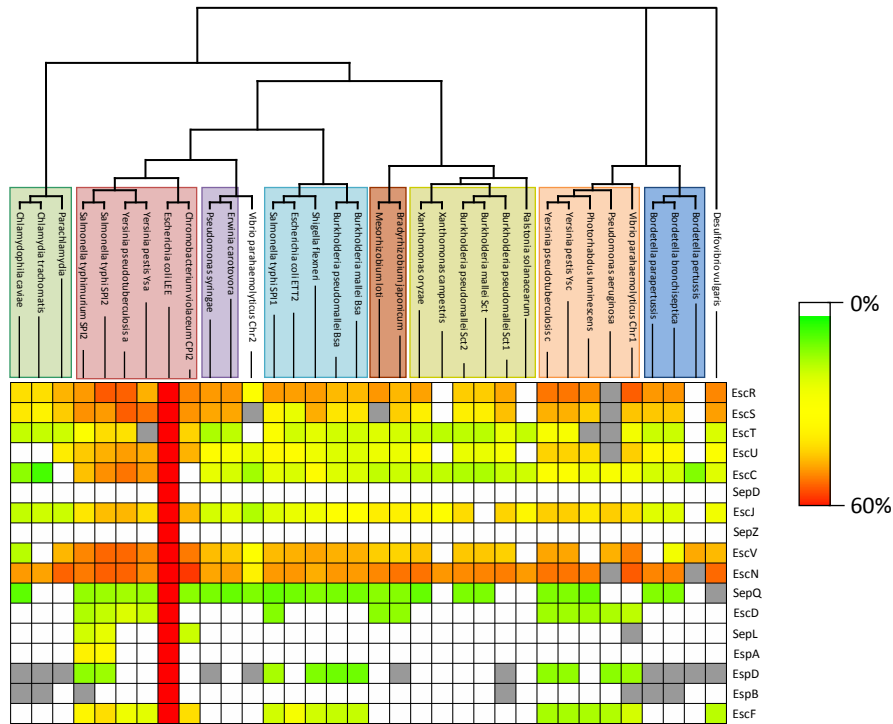
The T3SS genes encoded in the locus for enterocyte effacement (LEE) from *E. coli* O157:H7, and the plasmid encoded system from *Yersinia pestis* were extracted in order to test their ability to find proteins in other T3SSs. Filtering the results to just a

representative set of NF-T3SS containing genomes, PSI-BLASTS were used to locate homologous proteins between different systems. The results of these searches (See Figure 18) demonstrate the ability for the PSI-BLAST searches to find well conserved proteins (e.g. SctCNRSTUV). However, there are several systems where proteins should have been found, but were not (grey squares in Figure 18). In addition to this problem, there is also the issue of which PSI-BLAST settings to use. By altering the settings for filtering and compositional based statistics not only is the scoring affected, but there are also certain proteins where homology is only found when certain combinations of parameters are used (See Figure 19). SepQ and EspD are two such examples of this, where filtering masks out sufficiently large regions of the protein that no homology can be found with other proteins by BLAST.

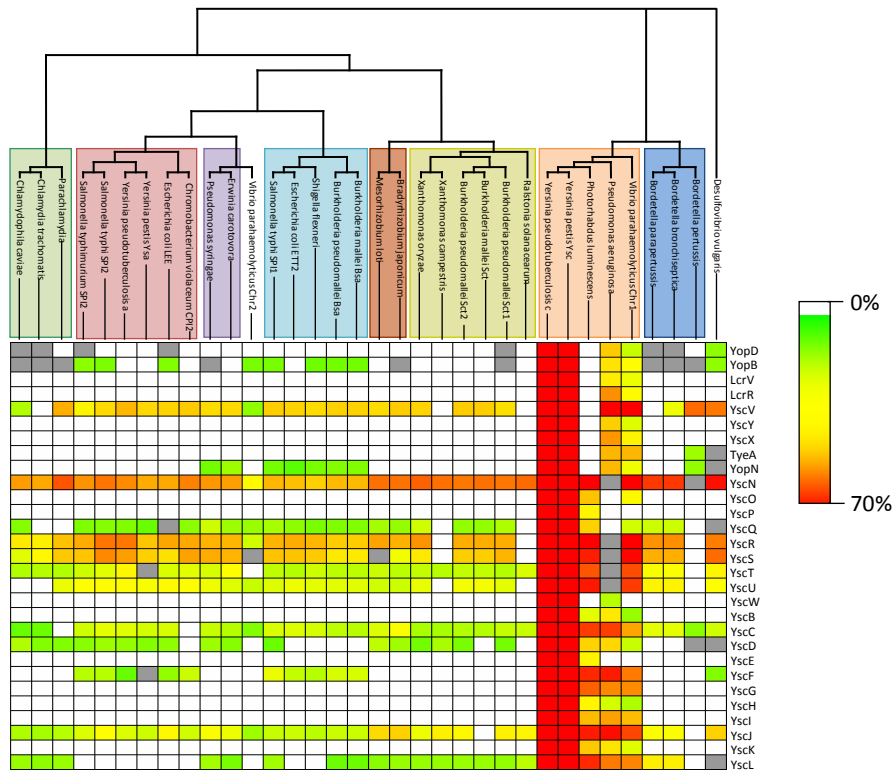
Searches using HMMER revealed a wide range in the number of hits returned by different T3SS related models. Some domains only find very few proteins, suggesting that they are not detecting the full set of related proteins, or the protein is not found in all type-III secretion systems. Conversely, there are also several domains which find considerably more proteins than might be expected. In these cases an examination of the results revealed that the domain model was also finding proteins unrelated to type-III secretion (for example MotA domain, which also finds ExsA and TonB domain proteins). As such, for the purposes of locating T3SS loci domains were chosen which would most likely be found in most or all T3SS, but not in other systems. Similarly domains were also chosen which could also be used to determine whether the system was flagellar or non-flagellar in nature (See Table 10).

Once the HMMER searches were completed, the data was saved into a database and a web based graphical user interface was produced in order to allow for manual

A.

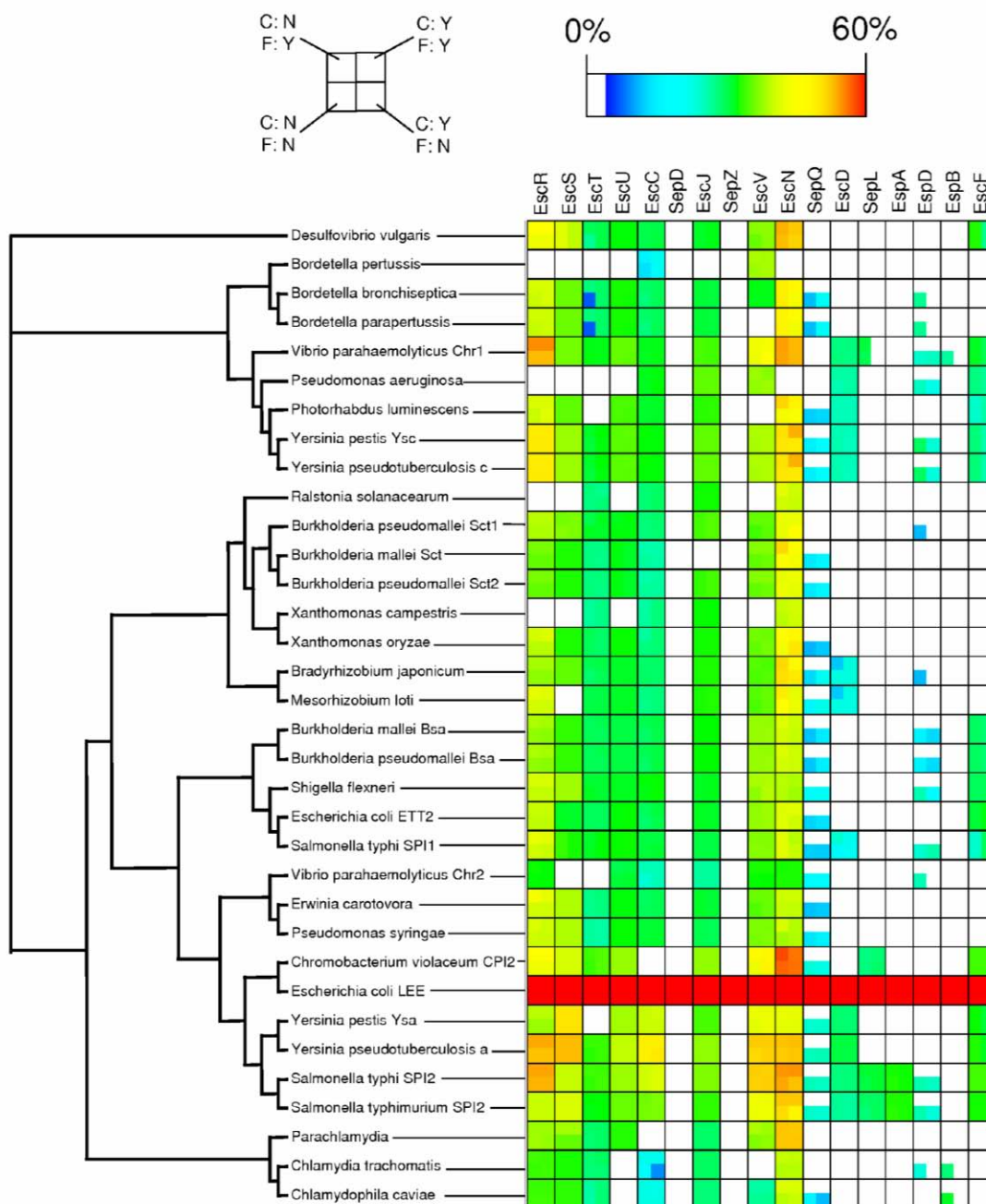


B.



**Figure 18. Heat maps of T3SS genes from *E. coli* and *Yersinia pestis***

Searches performed using Enterohaemorrhagic *E. coli*. (LEE System) (A) and *Yersinia pestis* (plasmid system) (B) genes as query sequences. Colouration based on degree of identity as reported by PSI-BLAST. White=No homology, Grey=Gene present but not detectable by PSI-BLAST. Phylogenetic tree drawn from an alignment of the ATPase protein (SctN). Note the rough conservation of patterns of conservation amongst related systems (highlighted in the same coloured block in the tree), and also the number of proteins not detected through PSI-BLAST which can be found using different starting points for PSI-BLAST search, or by analysis of gene synteny.



**Figure 19. Heat Map showing comparison of LEE system using 4 different BLAST settings**  
Each box is arranged into four smaller squares. Key: Top left – compositional based statistics (CBS) Off, low complexity filter (Filter) On; top right: CBS On, Filter On; bottom left: CBS Off, Filter Off, bottom right: CBS On, Filter Off.

Whilst the score for many proteins does not change dependant on the BLAST settings used, there are several exceptions such as SepQ and EspD where homologous are only found when filtering is turned off. Similarly for EspB, homologues are only found when both filtering and compositional based statistics are turned off. These results demonstrate the sensitivity of BLAST results to changes in starting parameters.

| <b>Domain</b> | <b>FS Hits</b> | <b>LS Hits</b> | <b>F/NF</b>   |
|---------------|----------------|----------------|---------------|
| Bac_export_1  | 374            | 358            | Both          |
| Bac_export_2  | 508            | 364            | Both          |
| Bac_export_3  | 362            | 365            | Both          |
| FHIPEP        | 409            | 371            | Both          |
| FlaA          | 23             | 17             |               |
| FlaE          | 313            | 263            | Flagellar     |
| FlaF          | 41             | 40             |               |
| FlaG          | 109            | 100            |               |
| FlbD          | 50             | 36             |               |
| FlbT          | 120            | 40             |               |
| FleQ          | 241            | 73             |               |
| Flg_bb_rod    | 1477           | 1462           |               |
| Flg_hook      | 248            | 244            | Flagellar     |
| FlgD          | 291            | 253            | Flagellar     |
| FlgH          | 230            | 212            | Flagellar     |
| FlgI          | 260            | 214            | Flagellar     |
| FlgN          | 114            | 116            |               |
| FliC          | 76             | 75             |               |
| FliD          | 84             | 73             |               |
| FliE          | 48             | 31             |               |
| FliD_C        | 345            | 234            | Flagellar     |
| FliD_N        | 272            | 227            |               |
| FliE          | 250            | 252            | Flagellar     |
| FliG_C        | 319            | 267            | Flagellar     |
| FliH          | 178            | 201            |               |
| FliJ          | 80             | 96             |               |
| FliL          | 467            | 274            |               |
| FliM          | 221            | 224            | Flagellar     |
| FliO          | 147            | 138            |               |
| FliS          | 230            | 219            | Flagellar     |
| FliT          | 42             | 38             |               |
| MotA_ExbB     | 990            | 1107           |               |
| Secretin      | 749            | 730            | Non-Flagellar |
| YscJ_FliF     | 309            | 151            | Both          |
| YscJ_FliF_C   | 147            | 135            |               |

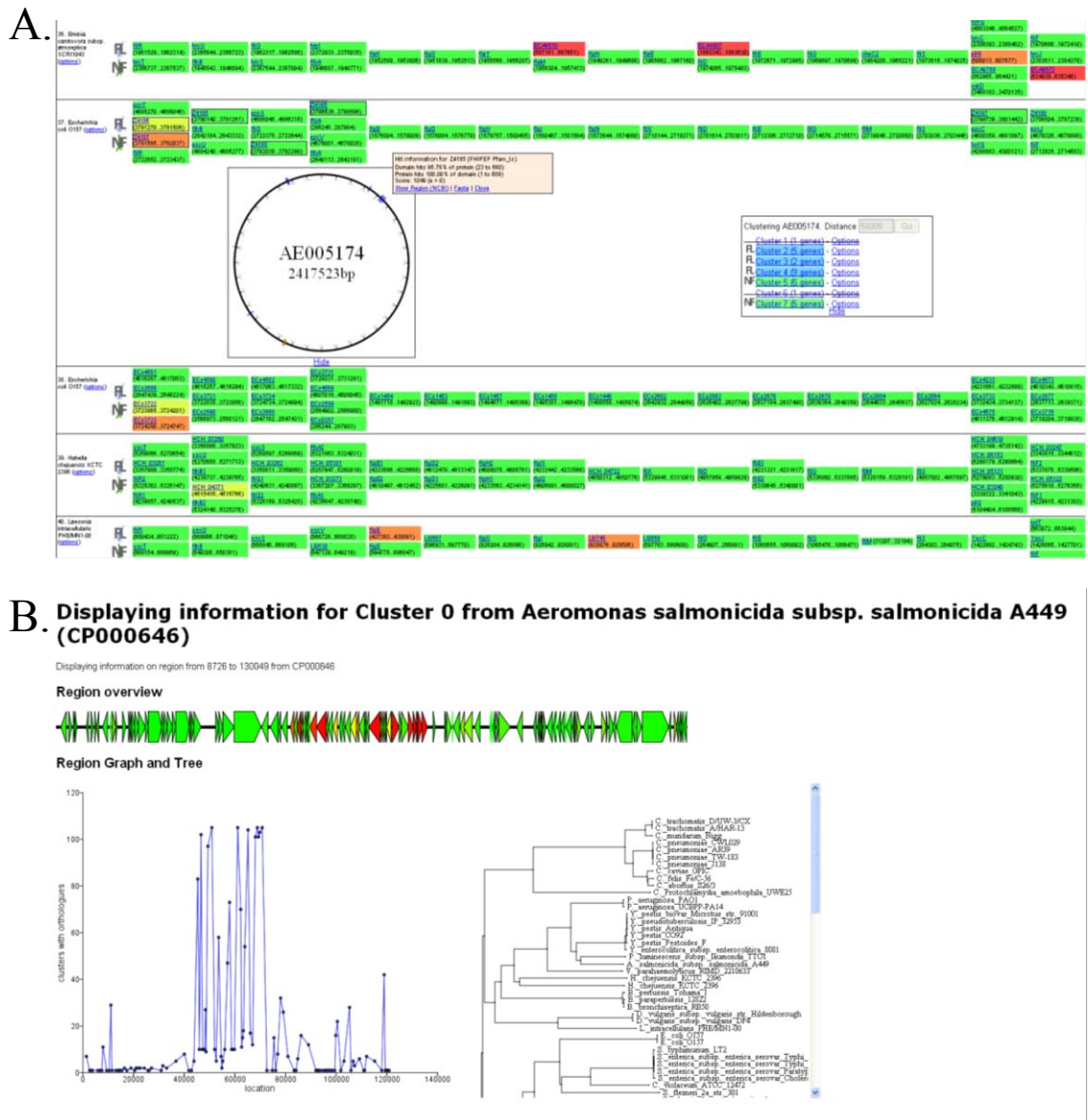
**Table 10. Pfam Domains and the number of hits found in bacterial genomes**

FS Hits: Number of hits to pfam fs domains (local alignment method). LS Hits: Number of hits to pfam\_ls domains (semi-global alignment method). F/NF: ability of the domain to find flagellar, non-flagellar or both types of loci. Colour key: Red: Domain not chosen for further use because number of hits was too small, orange: Domain not chosen because number of hits was too big, green: Ignored as FliD\_C model finds more of the same proteins, White: Domain used for T3SS locus finding.

curation of the data. The user interface groups HMMER hits together for each genome and allows for the genome to be categorised as containing flagellar or non-flagellar T3SS(s), both or neither. Similarly it allows for clusters of T3SSs genes to be visualised within the genome and decisions made as to whether they are clusters of flagellar or non-flagellar genes and also to determine the start and the end of clusters based on conservation and genomic data such as GC content. Several screenshots of the application can be seen in Figure 20.

### ***5.3.2. The distribution of flagellar and non-flagellar type-III secretion systems in sequenced bacterial genomes***

The set of bacterial genomes were searched using the domains listed in Table 10. There are a total of 445 different bacteria present in the database of genomes used, containing a total of 872 distinct chromosomes and plasmids. HMMER searches found hits in a total of 398 different bacteria (443 out of the 872 chromosomes/plasmids). After manual curation to remove genomes where the only hits were to the secretin domain model, or where all the domain hits were below the PFAM defined gathering threshold, there were a total of 239 different bacteria with a flagellar or non-flagellar system present. Of these 239, 228 contain at least one flagellar system and 73 contain at least one non-flagellar system (See Appendix 1 for the complete list of all bacteria containing a T3SS). The vast majority of non-flagellar T3SSs are contained within one locus. There are 106 non-flagellar T3SSs contained in a total of 130 loci. Those systems found in multiple loci are found in thirteen different bacteria: All eleven members of the Chlamydiae phylum, *Lawsonia intracellularis* and *Myxococcus Xanthus*. There are also 27 bacteria with more than one non-flagellar T3SS present (22 with two systems and five with three systems).

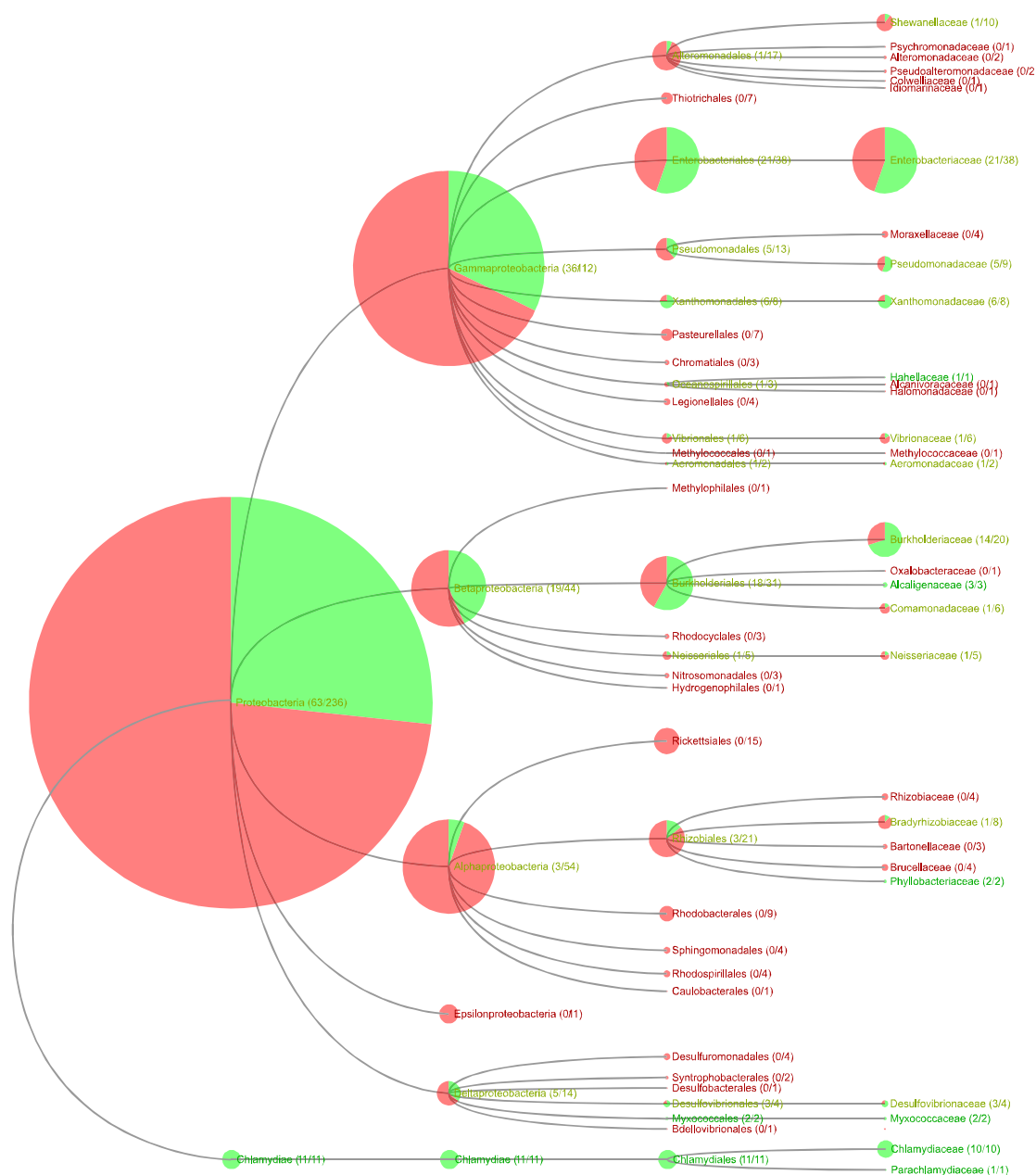


**Figure 20. Screen shots from T3SS finder web program**

A: Main screen, showing domain hits for different genomes. Background colour represents quality of domain hit. Also shown is the clustering information for the *E. coli* O157:H7 chromosome.

B: Graph and Schematic gene image showing degree of conservation of various proteins within the non-flagellar T3SS cluster from *Aeromonas salmonicida* (green: no conservation in other systems, red: conserved in all systems)





**Figure 21. Taxonomic tree showing distribution of non-flagellar T3SSs**

Tree data obtained from NCBI taxonomy database (<http://www.ncbi.nlm.nih.gov/Taxonomy/>). Root of the tree is 'Bacteria' (Kingdom), and the subsequent nodes shown are (from left to right): Phylum, class, order, family. No daughter nodes are shown for nodes with no non-flagellar T3SSs. Pie charts represent the number of bacteria within the given taxon (size of the chart), the proportion of bacteria within the taxon with a non-flagellar T3SS (green area of the chart), and the proportion of the bacteria within the taxon without a non-flagellar T3SS (red area of the chart). The colour of the taxonomic label is representative of the number of non-flagellar T3SSs (green: 100% presence, red: 0% presence, yellow: partial presence).

The two phyla shown are the only two with non-flagellar systems present. Non-flagellar systems are present in all classes of Proteobacteria apart from Epsilonproteobacteria. Non-flagellar system can also be found in all 11 members of the Chlamydiae phylum thus far sequenced.



The situation is more complicated for the flagellar T3SSs, where within the 228 different bacteria with a flagellar T3SS there are a total of 677 flagellar loci as detected by automated analysis, but at most only 255 flagellar T3SS based on numbers of proteins found. Around 40% (100 out of 255) of the flagellar T3SSs are encoded on only one locus; of the remainder there are several which have numerous flagellar loci. This is particularly true of members of the epsilonproteobacteria where it is not common to see more than ten loci containing flagellar genes. Spirochaetes also seem to have numerous flagellar gene loci, but not quite to the same degree as that seen in the epsilonproteobacteria.

Non-flagellar T3SSs are found in only two bacterial phyla: Proteobacteria and Chlamydiae (See Figure 21). Figure 21 also shows some of the taxons where non-flagellar T3SSs are commonly found. The first one of these to jump out is unsurprisingly the enterobacteriaceae, where 55% contain a non-flagellar T3SS, compared to just over 25% for all Proteobacteria. This data also confirms the lack of any non-flagellar T3SSs in the epsilonproteobacteria. Finally, all members of the Chlamydiae have a non-flagellar T3SS. Amongst the bacteria with a flagellar T3SS there are representatives of ten different phyla (see Figure 22): Actinobacteria, Aquificae, Bacteroidetes, Chlamydiae, Acidobacteria, Firmicutes, Planctomycetes, Proteobacteria, Spirochaetes and Thermotogae. The percentage of flagellar T3SS per phylum also appears to be much higher, for example over 75% of Proteobacteria have a flagellar T3SS versus just 25% for non-flagellar T3SSs in the same phylum.

PSI-BLASTS using proteins from both flagellar and non-flagellar loci produced in excess of 63 million homology pairs from distinct iterations. Trying to coerce this data into a manageable form or trying to produce homology networks from this data

proved to be unfeasible given the available computational resources, and so the remainder of the results in this chapter focus on non-flagellar systems only, which produce a much more modest 7.3 million homology pairs.

### **5.3.3. Phylogenetic groups of T3SSs**

The concatenated alignment of the six conserved domains produces a tree containing 96 out of the total of 106 T3SS loci (Figure 23). The remaining ten represent cases where not all six domains can be located within the genome. To prevent having to lose information by using fewer domains in the alignment, instead the most likely location for those other systems within the tree from all six domains was found by alternative means. Locations were determined locating these systems in trees produced using combinations of fewer domains (so long as all those domains were present in the system in question). Most of the absent systems fit where one would expect.

The absent *Yersinia* systems cluster with the other Yersinial Esc/Ssa systems, the absent *Burkholderia* systems cluster with the other Hrp2 group *Burkholderia* systems, and the absent *Shigella* systems cluster with the system from *S. flexneri*. The system from *Pseudomonas syringae* pv. tomato clusters with other Hrp1 systems, and the third system (by order in the genome) from *Sodalis glossinidius* clusters with the first system from the same bacterium, and the Inv/Mxi/Spa systems from *Chromobacterium violaceum* and *Salmonella enterica*. The final missing system not shown in Figure 23 is from chromosome 2 in *Vibrio parahaemolyticus*, which tends to cluster in different locations (either amongst the Hrp2 or Esc/Ssa groups) depending on the domain(s) used to draw the tree.

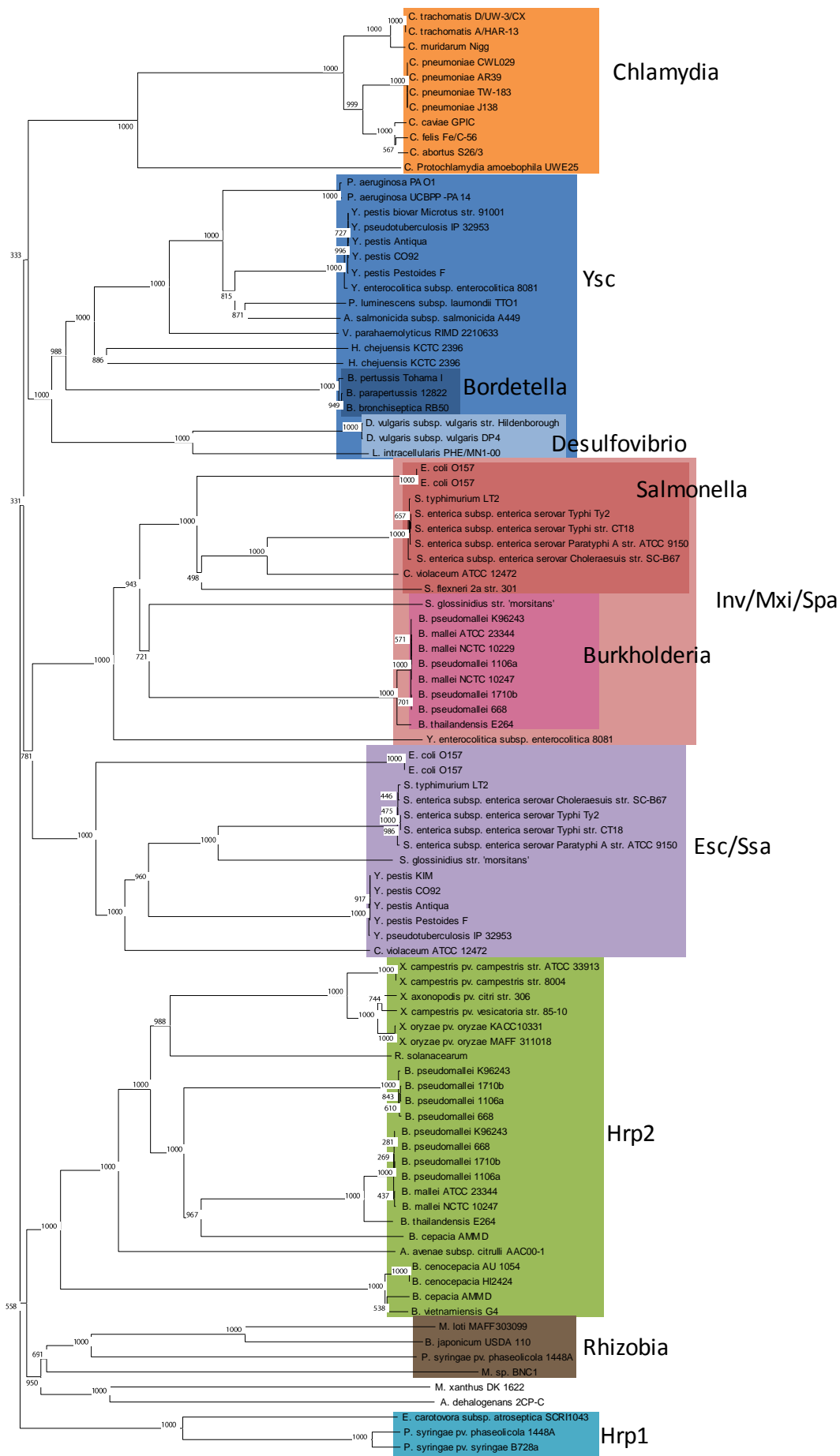


Figure 23. Phylogenetic tree of non-flagellar T3SSs

**Figure 23 (continued)**

Phylogenetic tree of non-flagellar type-III secretion systems. Tree drawn from an alignment of six separate domains (Bac\_export\_1, Bac\_export\_2, Bac\_export\_3, Secretin, FHIPEP, and YscJ\_FliF) using clustalw. This tree demonstrates a branching order which supports the five separate groups suggested by Foulter *et al* [297], as well as the presence of several novel groups including Chlamydia and Rhizobia. It also demonstrates the presence of sub groups within these main seven groups, such as the Bordetella and Desulfovibrio groups which are members of the Ysc group, and the splitting into the Inv-Mxi-Spa group into the Salmonella and Burkholderia groups

The phylogenetic tree also confirms the tree seen in Figure 14, in that it also places the *Hahella* systems next to the other Ysc systems, with the *Bordetella* systems in a close group with each other, and a third final group containing the T3SSs from *Desulfovibrio* and *Lawsonia*. The tree also shows several additional groups of T3SSs beyond those identified by Foulter *et al*: The Chlamydia group, containing all the T3SSs from *Chlamydia* and *Protochlamydia*; and the Rhizobia group, containing T3SSs from *Rhizobia*, *Bradyrhizobium* and *Mesorhizobium*. There are also sufficient numbers of systems within the tree to be able to begin to see more subgroups within each ‘group’ shown in the figure. For example the Hrp2 family systems from *Burkholderia* form their own separate group from the other Hrp2 systems from bacteria such as *Ralstonia* and *Xanthomonas*. A similar situation also exists for the Inv/Mxi/Spa systems, where the *Burkholderia* systems cluster separately from the other members of the group. Finally there is the potential for another group of systems to be present within the tree if there were more systems to support it: The Myxococcaceae group. At present there are only a couple of genomes from this taxonomic family available for sequence analysis, but the two present in this study: *Myxococcus xanthus* and *Anaeromyxobacter dehalogenans* do seem to cluster together in phylogenetic trees and their orthologous proteins are mutual best hits in BLAST searches.

#### **5.3.4. ‘Essential’ gene families**

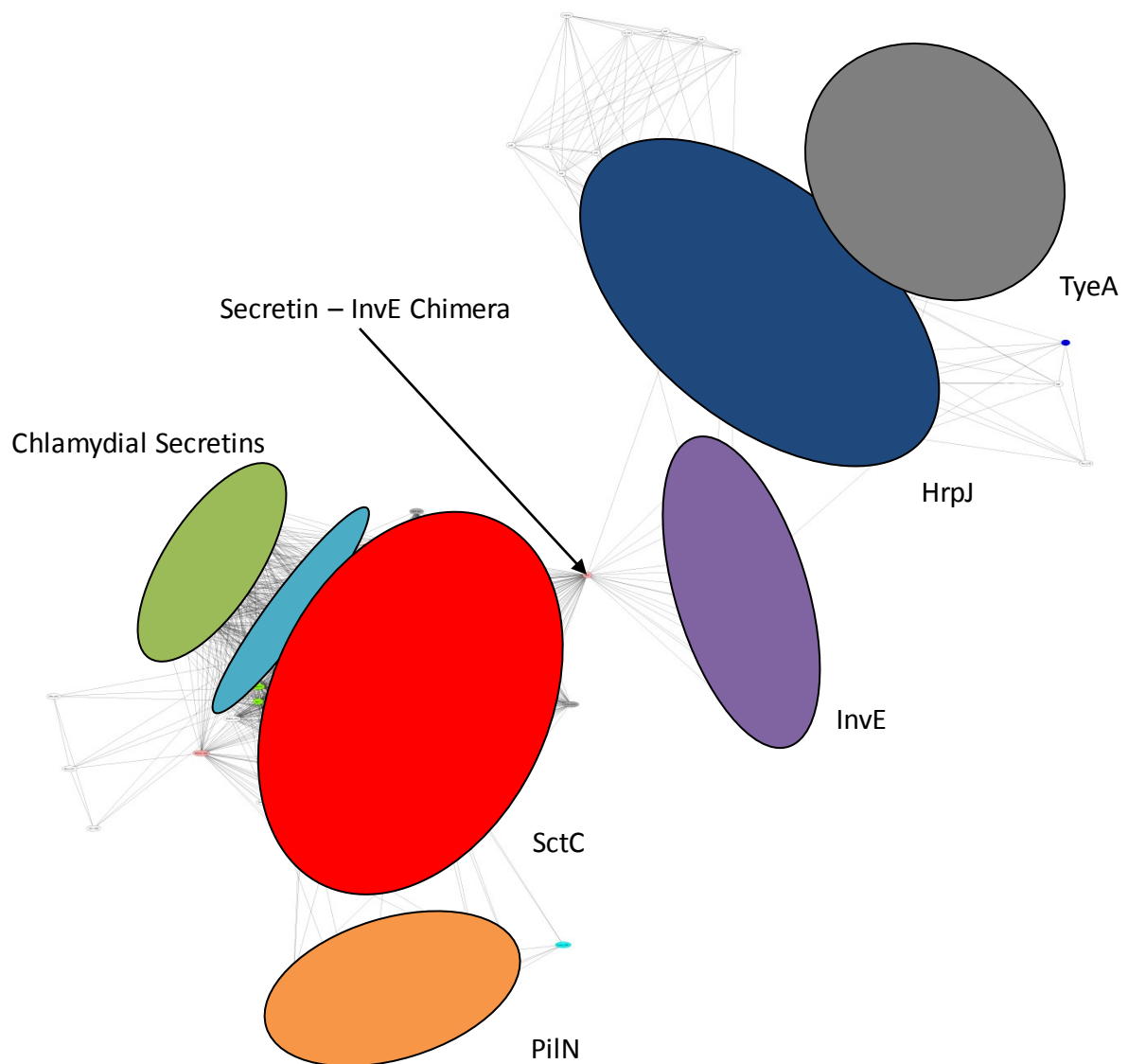
Generation of protein homology networks produced a total of 685 networks, of which fewer than 100 have more than ten members. This includes several networks which include more than 140 members, which is more than would be expected if there were only one copy of a given family of proteins per NF-T3SS locus. The two largest

examples have 326 and 167 members respectively, and closer examination reveals that these networks actually contain several sub networks connected by a single protein. In the case of the network with 167 members the two sub-networks contain members of the secretin protein family and the HrpJ/InvE family (See Figure 24). The protein which ties these two networks together is BsaO from the Inv-Mxi-Spa system of *Burkholderia pseudomallei* 1710b. The situation is similar for the largest network where the relevant sub-network (containing FHA domain proteins) is contained within a larger network which also contains BON and  $\sigma 54$  proteins present because of a chimeric protein from *Myxococcus xanthus*.

By flagging these misleading proteins during network generation the software then produces a total eleven networks where the size of the network suggests that the proteins present are conserved amongst most if not all the systems under examination. Of these eleven networks, ten contain proteins with a clear role in type-III secretion (See Table 11). Several of these networks are still too large to contain only one copy of the protein per NF-T3SS. The most likely cause of this is the generous boundaries set for the amount of sequence to be included in a T3SS locus. For example in the network in Figure 24 there are two highlighted groups of proteins which do not belong to T3SSs. These proteins in question either belong to phage systems which are highlighted in blue, or type-II secretion systems which are highlighted in orange.

All ten networks are shown mapped against NF-T3SS systems in Figure 25, which clearly shows the near complete conservation of nine out of the ten protein networks. The remaining network, that belonging to SctD, shows conservation amongst most systems, but initially appeared to be absent within systems belonging to the Hrp2 group of NF-T3SSs. In order to check whether this was correct the complete





**Figure 24. Homology Network showing Secretin and HrpJ/InvE proteins**  
 Proteins are coloured according to their domain architecture, translucent ellipses denote clusters of similar proteins. Note the connection of the two different groups of proteins by the single chimeric protein in the centre of the graph

| <i><b>Network Name</b></i> | <i><b>Network Size</b></i> | <i><b>T3SS Protein</b></i> |
|----------------------------|----------------------------|----------------------------|
| Secretin                   | 133                        | SctC                       |
| SepQ/SpoA                  | 123                        | SctQ                       |
| FHIPEP                     | 119                        | SctV                       |
| ATPase                     | 113                        | SctN                       |
| Bac_export_1               | 107                        | SctT                       |
| FliP                       | 106                        | SctR                       |
| Bac_export_2               | 106                        | SctU                       |
| YscJ_FliF                  | 102                        | SctJ                       |
| Bac_export_3               | 101                        | SctS                       |
| FHA                        | 78                         | SctD                       |

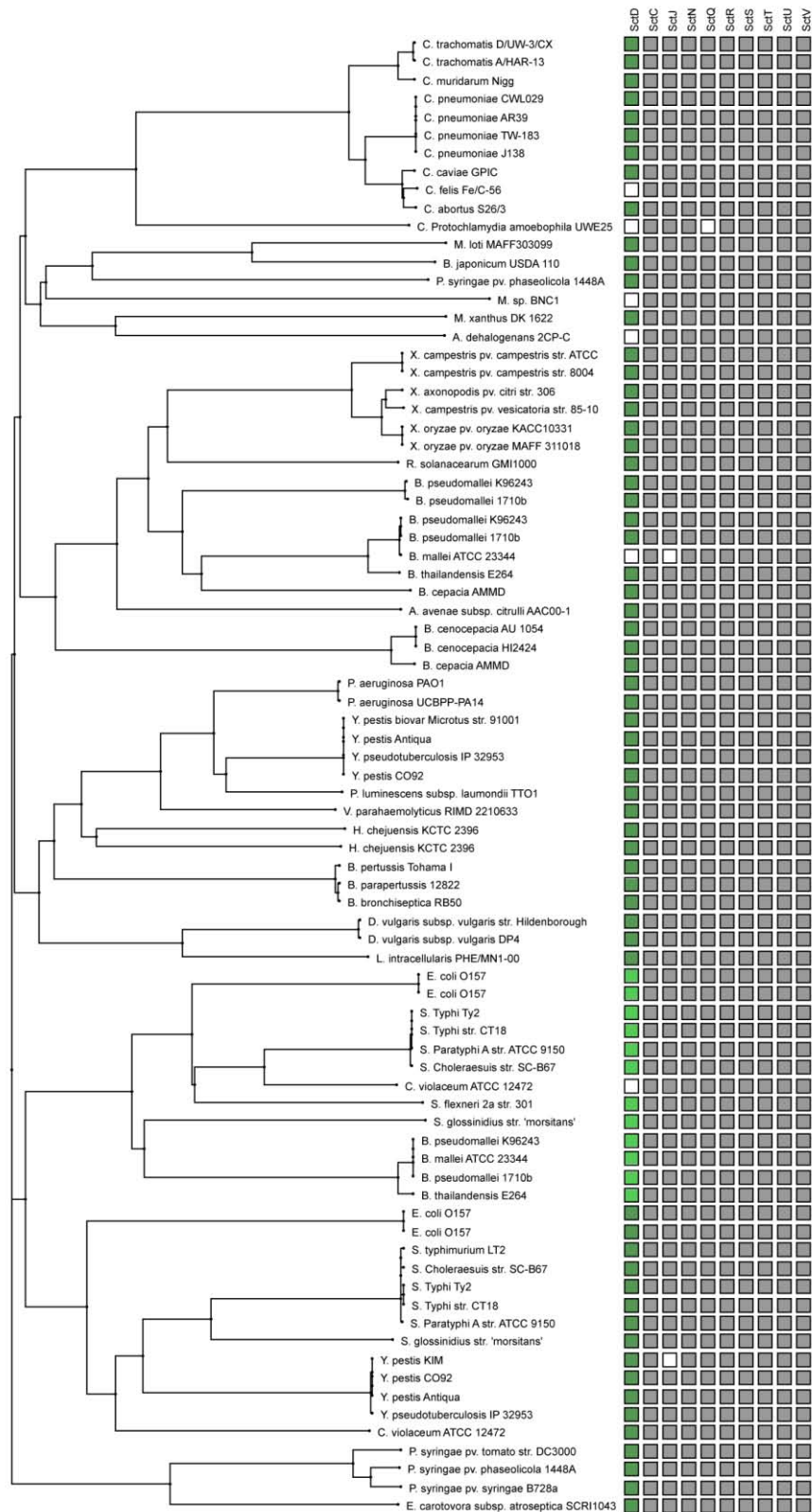
**Table 11. Largest protein networks containing T3SS proteins**

These protein networks contain the proteins conserved throughout all non-flagellar T3SSs

reciprocal PSI-BLAST data set was interrogated in order to locate any additional networks containing proteins which were homologous to proteins within the major SctD network, but where the homology could only be found using iterative BLAST searches. Such a network was indeed identified and the proteins located within this second SctD network are shown in a paler shade of green in Figure 25. Of the remaining missing proteins, several can also be found in some cases using the method of locating PSI-BLAST hits between proteins in the locus of interest and the network from which the locus appears to be absent.

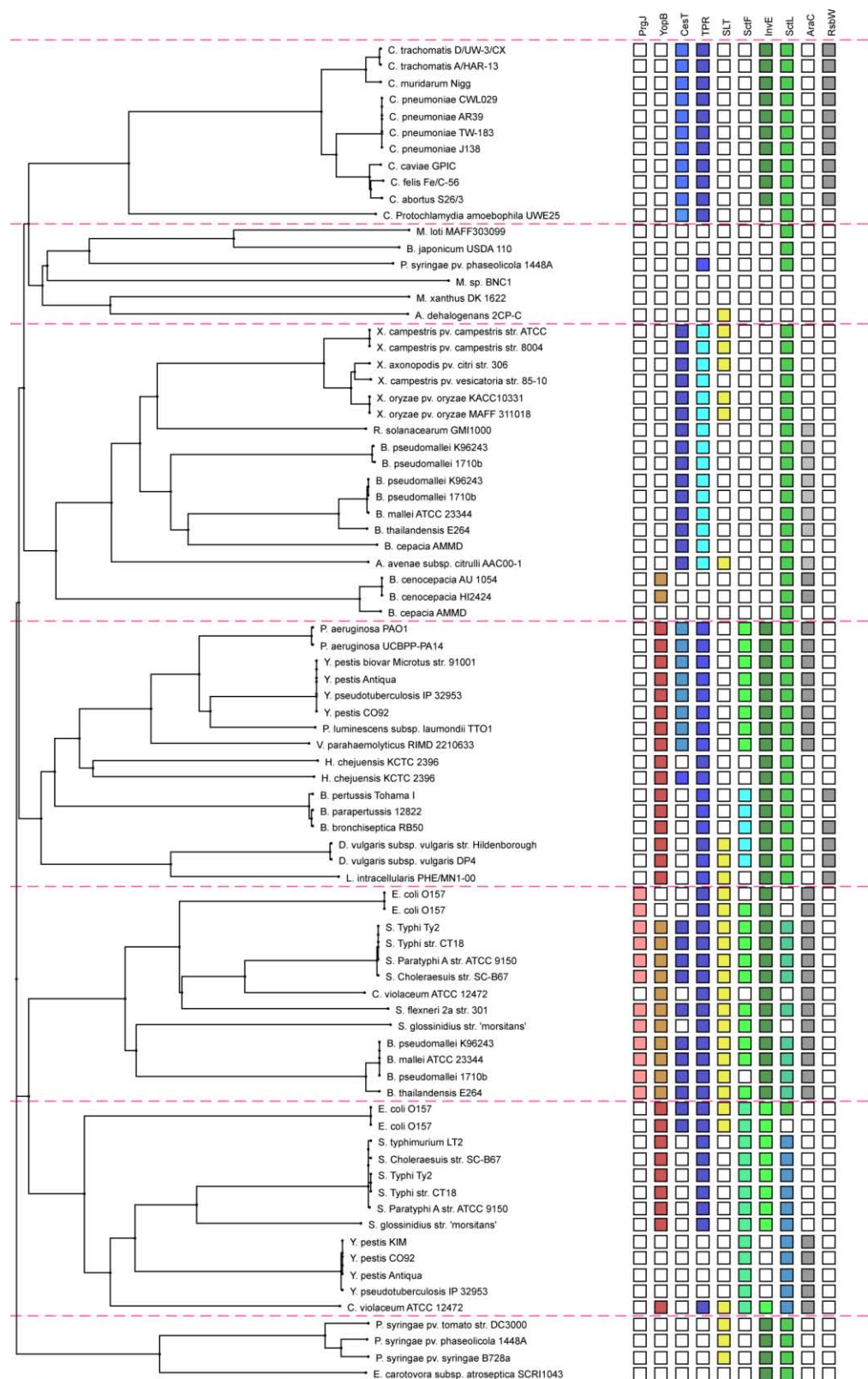
### ***5.3.5. Partially conserved gene families***

Beyond the ten protein networks conserved throughout non-flagellar type-III secretion there are then several networks which have numerous members but the size of the network precludes the network from being ubiquitous to all systems. Some of these networks span multiple groups of T3SSs, whilst others are isolated to single groups or subgroups. As can be seen from Figure 26, there are examples of chaperones, regulators and proteins from the apparatus and translocon which show only partial conservation amongst all the T3SSs under examination. For example the AraC family of regulators is only found in four different groups of T3SSs: The Hrp2, Ysc, Inv-Mxi-Spa and Esc/Ssa groups. One of the other examples in the figure below is that of SctF, the needle protein. Through use of PSI-BLAST data it is possible to stitch together several networks all of which contain needle proteins. However, even once these networks have been joined all together it would seem that there are no examples of this family of proteins within the Hrp1 or Hrp2 groups of T3SSs. This observation correlates excellently with the presence of the Hrp pillus within these groups suggesting that the pillus may fulfil the role of producing the complete needle,



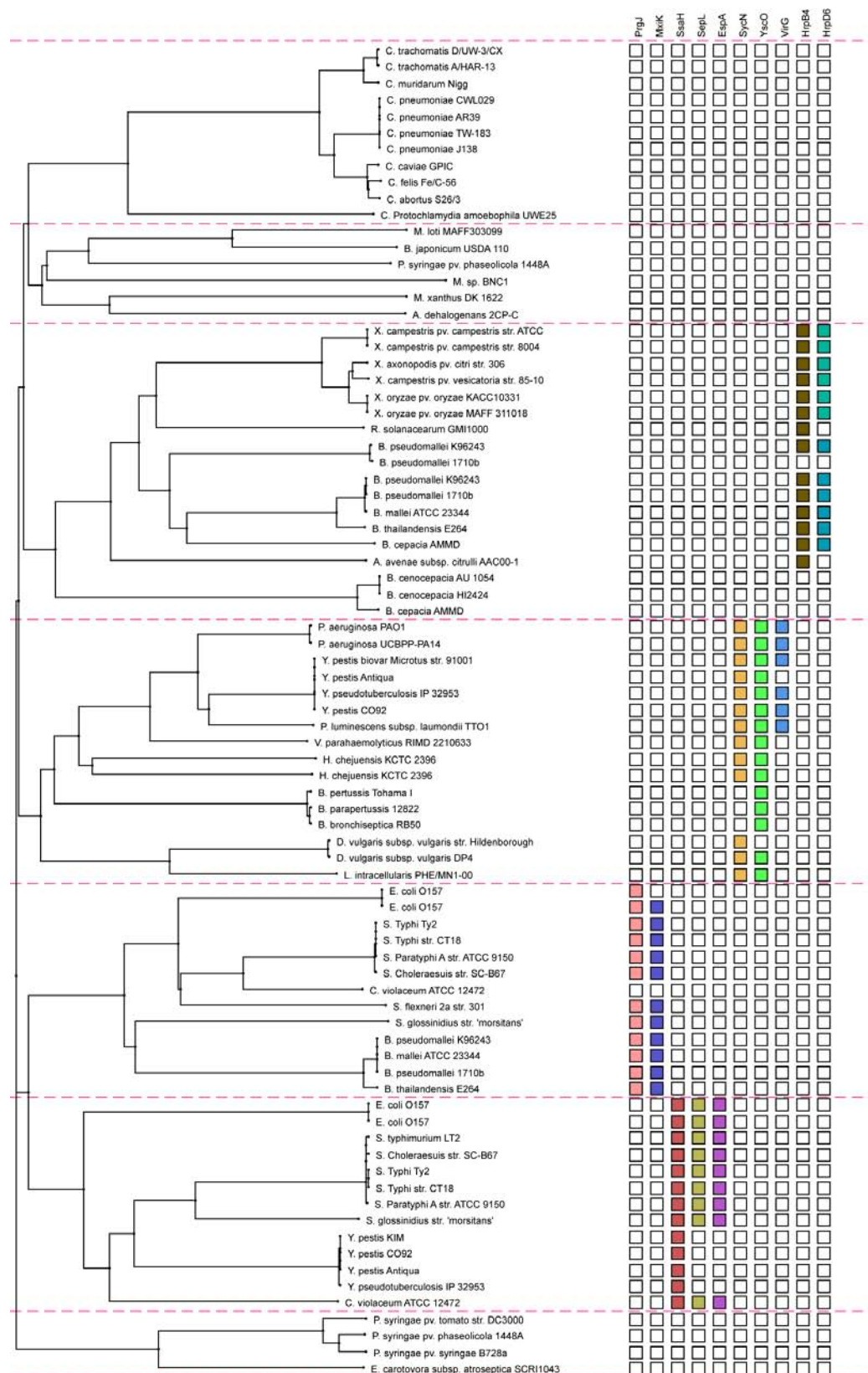
**Figure 25. Phylogenetic tree of T3SSs with highly conserved networks overlaid**

Phylogenetic tree drawn as per Figure 23, but without the YscJ\_FliF domain alignment. Boxes to the right represent presence (coloured box) or absence (empty box) of the protein in the system in question. The ten protein families listed here represent the set of proteins which are completely conserved amongst all non-flagellar T3SSs. The SctD proteins actually belong to two separate networks which can be linked by PSI-BLAST results. The shade of green shows which network each protein belongs to. Where all the proteins in a column belong to one family the column is coloured grey.



**Figure 26. Phylogenetic tree of T3SSs with well conserved networks overlaid**

Figure drawn as per Figure 25. White indicates protein absence. Different colours in the same column indicate multiple networks joined together using data available from PSI-BLAST searches alone. The proteins shown here are common to multiple but not all NF-T3SSs. These proteins include regulators, chaperones and structural components. Red dotted lines separate the tree into related groups of NF-T3SSs and protein conservation/absence is well correlated within group members



**Figure 27. Phylogenetic tree of T3SSs with partially conserved networks overlaid**

Figure drawn as per Figure 25. White indicates protein absence. Different colours in the same column indicate multiple networks joined together using data available from PSI-BLAST searches alone. These groups of proteins are a selection of proteins which are found only in one group of NF-T3SSs, and demonstrate the wide range of characteristics unique to particular groups or sub-groups of NF-T3SSs

thus making a ‘traditional’ needle protein unnecessary. Several of the groups of chaperones (CesT and TPR) are also very well conserved, being present in all but the Hrp1 and Rhizobial groups. Beyond the dozen or so examples of protein families which are conserved across multiple groups there are then also many examples of proteins which are conserved among just one group. This class of protein families numbers at least fifty and contains examples of all types of proteins involved in type-III secretion including effectors. Several examples of the protein families conserved in just one group of NF-T3SSs are shown in Figure 27.

### **5.3.6. ‘Absent’ proteins within the result set**

Within Figure 25, Figure 26 and Figure 27 there are numerous empty squares. Many of these squares (particularly for partially conserved networks) are expected. However there are several examples of empty squares where one would expect to find a hit. There are many reasons why this may occur and the examples that follow highlight some of the issues and causes for these absences.

Firstly within Figure 25, there are the two missing proteins within *Candidatus Protochlamydia amoebophila* UAE25: SctD and SctQ. A careful analysis of the whole genome (Accession BX908798) reveals another T3SS locus within the genome which was not identified by the initial screen with PFAM domains. This region includes several chaperone like proteins in addition to *pc1391* a gene which encodes an SctD family protein, and *pc1400* which encoded an SctQ family protein. In a similar manner *Chlamydia felis* Fe/C-56 contains a FHA domain protein similar to that found in other Chlamydia family T3SS loci, but which is encoded in a locus not found using the method described above. These omissions from the result set reveal the first issue with the methods used in this analysis. Where T3SSs loci are



spread around a chromosome or plasmid then loci become harder to detect using a subset of domain representative of T3SSs, and thus loci can be missed.

The second example of a ‘missing’ protein is SctJ from *Yersinia pestis* KIM (Chromosomally encoded Esc/Ssa system). A BLASTP search of the translated genome using YscJ from *Yersinia pestis* CO92 reveals no significant hits. However a TBLASTN search of the whole chromosome reveals a region of DNA which encompasses the pseudogene *y0521* and gene *y0522*. An alignment of the DNA from this region with the corresponding region from *Y. pestis* CO92 reveals 100% at the DNA level with the exception of a single cytosine residue inserted into the sequence from *Y. Pestis* KIM. The age of the DNA sequence, and fact that this insertion exists in a run of cytosine residues suggests that this may be a DNA sequencing artefact rather than reality. This highlights the second issue with the method used in this analysis, reliance on third party data and annotation. In order for homologues to be identified by BLAST in this analysis there must be a protein record in nr. An absence of a protein will, by necessity, mean no match being found by BLAST and other approaches; however, in this case that does not tell the full story.

There are also several examples of proteins families which are only found in one or two families but other evidence they may be more prevalent. One such example of this is the SepL family, which has been shown to be related to MxiC [408]. However the SepL network (shown in Figure 27) contains only proteins belonging to Esc/Ssa family T3SSs. An investigation of the PSI-BLAST data for SepL reveals several connections between the SepL network and other networks, many of which are uninformative; however it does identify multiple connections with the YopN network. Graphing of the connections between these two networks (Figure 28) shows the very





strong PSI-BLAST evidence for connecting the protein homology networks together. The network graph also shows the relationship between the different sub-families within the YopN network. Hrp1 proteins (HrpJ) are quite closely related to TyeA proteins, but not to other YopN proteins. TyeA/Hrp1 proteins are tied to the Ysc YopN proteins by the chimeric YopN-TyeA proteins from *Hahella*, *Desulfovibrio*, *Lawsonia* and *Bordetella* (see section 4.4.3 for more information on these proteins). These chimeric proteins also serve to tie together the other two groups of YopN proteins: Those from Chlamydia (which show a degree of similarity to each other that is far higher than for other sub-groups), and from Inv-Mxi-Spa T3SSs, which includes MxiC. This example serves to identify the third issue with this analysis, that BLAST by itself is often not enough to identify a full family of proteins, but straight acceptance of PSI-BLAST results in addition cannot be relied upon. In this case including all networks joined by PSI-BLAST to the SepL network would have resulted in a graph with an additional several hundred proteins including sigma-54 and two component regulators, helicases and HSP60 proteins. Thus manual interpolation of PSI-BLAST data is required to fully appreciate the full size of various protein families. This issue is discussed further in section 5.4.3.

The fourth omission actually illustrates two separate problems which the analysis method. In Figure 26, in the column for SctL proteins (part of the FliH family) there appears to be a hit in the LEE T3SS in *E. coli* O157:H7 EDL933 but not in *E. coli* O157:H7 str Sakai. This intuitively sounds wrong, as the two genomes are nearly identical at the nucleotide level. Investigation of the hit in EDL933 reveals this shows very low levels of similarity to only a small region of the SctL protein Psyr\_1197 from *Pseudomonas syringae*, and is in fact annotated as a transposase. Looking at its genomic locale also shows it to be in a prophage island just upstream of the LEE, as

might be expected if the annotation is correct. This fourth problem with the analysis this time highlights the issue of selecting appropriate criteria for accepting/rejecting BLAST hits, which in turn will affect the number of false positive hits. A more stringent e-value cutoff ( $< 0.0001$  rather than  $< 0.001$ ), or requiring a certain degree of coverage of the query protein (e.g.  $> 50\%$ ) by the BLAST hit, would have resulted in this hit being ignored.

Thus the absence of the SctL proteins from both *E. coli* O157:H7 strains in this analysis now needs explanation, as previous work has shown this protein exists within the LEE [219]. A search of all networks for Z5136 / ECs4584 (the SctL family proteins from *E. coli* O157:H7 EDL933 and *E. coli* O157:H7 str Sakai respectively) finds a single, network which contains just these two proteins alone. Since this network exists, why was it not included in the SctL column along with the three networks already identified by PSI-BLAST linkage? The four SctL family networks now identified are:

- The major family consisting of proteins from Ysc, Hrp1, Hrp2, Chlamydial and Rhizobial T3SSs
- A small family consisting of proteins from Inv-Mxi-Spa group
- A small family consisting of proteins from the Esc/Ssa group (apart from the LEE T3SS system)
- A two member network for proteins from the LEE

The first three networks were joined together based on PSI-BLAST hits from the major network to the two smaller networks. No PSI-BLAST evidence supports joining the major network to the LEE network. There are also no PSI-BLAST hits from the LEE system to any other networks. However, proteins from the Esc/Ssa

network do produce PSI-BLAST hits against the two LEE SctL proteins. Thus all four networks can be linked together. This fifth issue is actually closely related to the third one highlighted above regarding the integration of PSI-BLAST data. By the nature of the algorithm PSI-BLAST searches are sensitive to the query protein used to start the search. Thus homology detected by PSI-BLAST between protein A and protein B may only be detected when protein A is used as the query sequence, but not when protein B is used. Whilst the network approach used here does much to ameliorate this situation, there will still be cases where a relationship between two networks as identified by PSI-BLAST is unidirectional, and so relationships between networks may be missed dependant on which network is used as the starting point.

## **5.4. Discussion**

### ***5.4.1. Complexity in T3SS loci***

Whilst the data above ably demonstrates that there are only ten proteins totally conserved amongst all non-flagellar T3SSs, any T3SS must contain many more specific components in order to function. The list of ten contains only core structural components of the apparatus. It has no needle components, with which to extend the apparatus from the cell's surface, no translocon proteins with which to create a hole in the host cell's membrane, no effectors to channel through the system and no chaperone to target those effectors to the apparatus in the first place.

It is the diversity in the complements of these other proteins which allow T3SSs to function in diverse environments and with many different host organisms in order to produce a multitude of different outcomes. The evidence presented above demonstrates the differences in gene complement which allow different T3SSs to

function in different ecological niches. For example there are several interesting changes which have occurred in the Hrp groups of T3SSs which are related to their requirement to penetrate plant cell walls. The most obvious of these are the hrp pillus proteins which are responsible for producing the long extension on the end of the apparatus, beyond this there are other changes such as the apparent absence of a needle protein such as those found in the other groups of T3SSs. The normal translocon proteins are also absent from the Hrp groups of T3SSs, and instead they produce pores in host cell membranes through the use of the HrpZ protein [409].

This is just one of many examples of proteins which are localised to one group of T3SS, or even to just one species of T3SS. Manual examination of the smaller networks (those with  $< 10$  members) reveals that once the network size drops below six the networks almost always contain proteins belonging to just one species. Secondly, the average T3SS locus contains around 50 genes, meaning that after we take into account the ten conserved genes, and another roughly ten genes which are conserved amongst multiple groups of T3SSs, the remaining 30 (i.e. over half) will be genes unique to an individual group of T3SSs, unique to the species, or even unique to the bacterium in question.

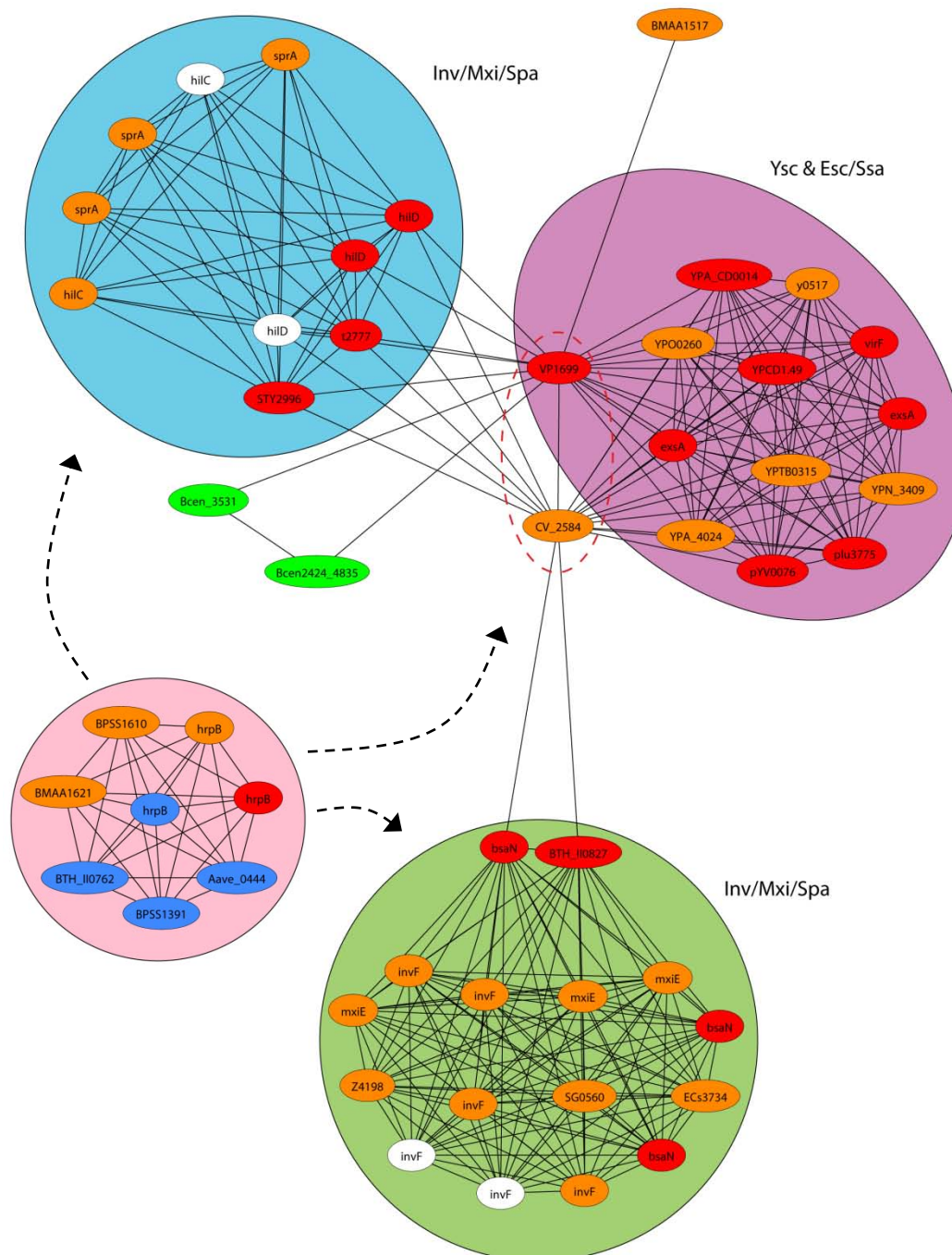
#### ***5.4.2. Mapping the change in gene complement to sequence phylogeny***

The figures presented above demonstrate the strong linkage between well-established methods of determining evolutionary distance with the gene complement of T3SSs. This provides excellent evidence for step-wise changes in T3SSs occurring over an evolutionary timescale, and as was shown in the analysis Ysc type systems in Chapter 4, it is possible to infer where such changes occurred by locating branch points in the

trex which split T3SSs into two groups: Those which have the gene, and those which do not. The level of overlap between the two measures of system similarity is practically complete. An examination of all networks related to type-III secretion (certain networks containing proteins such as transposases and phage components were excluded) with a size of six or greater revealed that almost every network could be resolved to a group or groups of T3SSs.

However, this interesting overlap between protein sequence and gene complement does not necessarily follow the precise patterns one might expect when examining individual network members. One such example of this is the homology network(s) containing members of the AraC regulator family. The two networks which contain this family of proteins can be connected through homologies found by PSI-BLAST, which forms connections between all eight members of the smaller network and 29 of the 43 members of the larger network. The smaller network forms a tight cluster with every node connecting to every other node within the graph, and comprises members of the Hrp2 family of T3SSs (these proteins are shown in light grey in the AraC column of Figure 26). Within the larger cluster the network is split into three separate regions which are joined together by just two proteins: CV\_2584 from *Chromobacterium violaceum* and VP1699 from *Vibrio parahaemolyticus* (see Figure 29). The members of this larger network are shown in dark grey within the AraC column of Figure 26, and include T3SSs from three separate groups: Ysc, Inv-Mxi-Spa and Esc/Ssa. Thus on the basis that there are three groups of T3SSs present in this network and three clusters within the network graph one might assume that each cluster contains proteins from one T3SS group alone. However this is not the case. As can be seen in Figure 29, two of the clusters contain members of the Inv-Mxi-Spa group, and the third group containing members of the Ysc group and the





**Figure 29. Graphical representation of AraC protein homology network**

Each node represents a protein, with the colour of the node representing domain structure (red: HTH\_AraC-HTH\_AraC, orange: HTH\_AraC, green: HTH\_AraC-AraC\_N, blue: TPR-HTH\_AraC). Each edge represents homology between two protein (BLAST e-value  $< 0.001$ ), length of edges are inversely proportional to  $-\log(\text{e-value})$ . Proteins contained within the pink circle represent a separate network of proteins which can only be connected to the main network through PSI-BLAST searches. These networks show one of the issues with automated clustering, should there be just one network containing all AraC members, or four separate networks of proteins produced by removing the weakly connected nodes.

Yersinia members of the Esc/Ssa group. This is an interesting result, but at least in the case of the separation of the two Inv-Mxi-Spa groups, not entirely surprising. The first group (contained within the blue circle in the figure) are members of the HilC/HilD family, which are responsible for regulating HilA, which in turn regulates transcription of the complete T3SS [279, 410]. The HilA family of regulators is confined to *Salmonella*, and the HilC/HilD sub group of the AraC protein network also only contains proteins from *Salmonella*. The other group of AraC proteins belonging to Inv-Mxi-Spa T3SSs are the InvF family of proteins which is ubiquitous to the group. In *Salmonella* InvF is downstream of HilA in the regulatory cascade which controls T3SS gene transcription [411]. It is however, the third group (outlined in purple) which confounds the assumption that the distribution of proteins should follow the phylogenetic groupings as the proteins which cluster together are from two separate T3SSs groups.

This network highlights one of the core issues of the methodology used in this study. As mentioned above the three networks are only joined together by just two proteins. In particular the link between the InvF group and the other two groups within the network is particularly weak compared to the strength of the link between the HilC/D group and the Ysc & Esc/Ssa group, thus maybe the InvF group should be separated into a separate network. Conversely there is also the matter of the AraC proteins from Hrp2 T3SSs which are not connected to the main network by BLAST searches, but show excellent connectivity to it through PSI-BLAST, in which case there is an equally strong case for their being just one network containing four clusters.

#### **5.4.3. Issues with automated locus and protein family finding**

The key issue in using automated approaches in order to find T3SS loci and assign



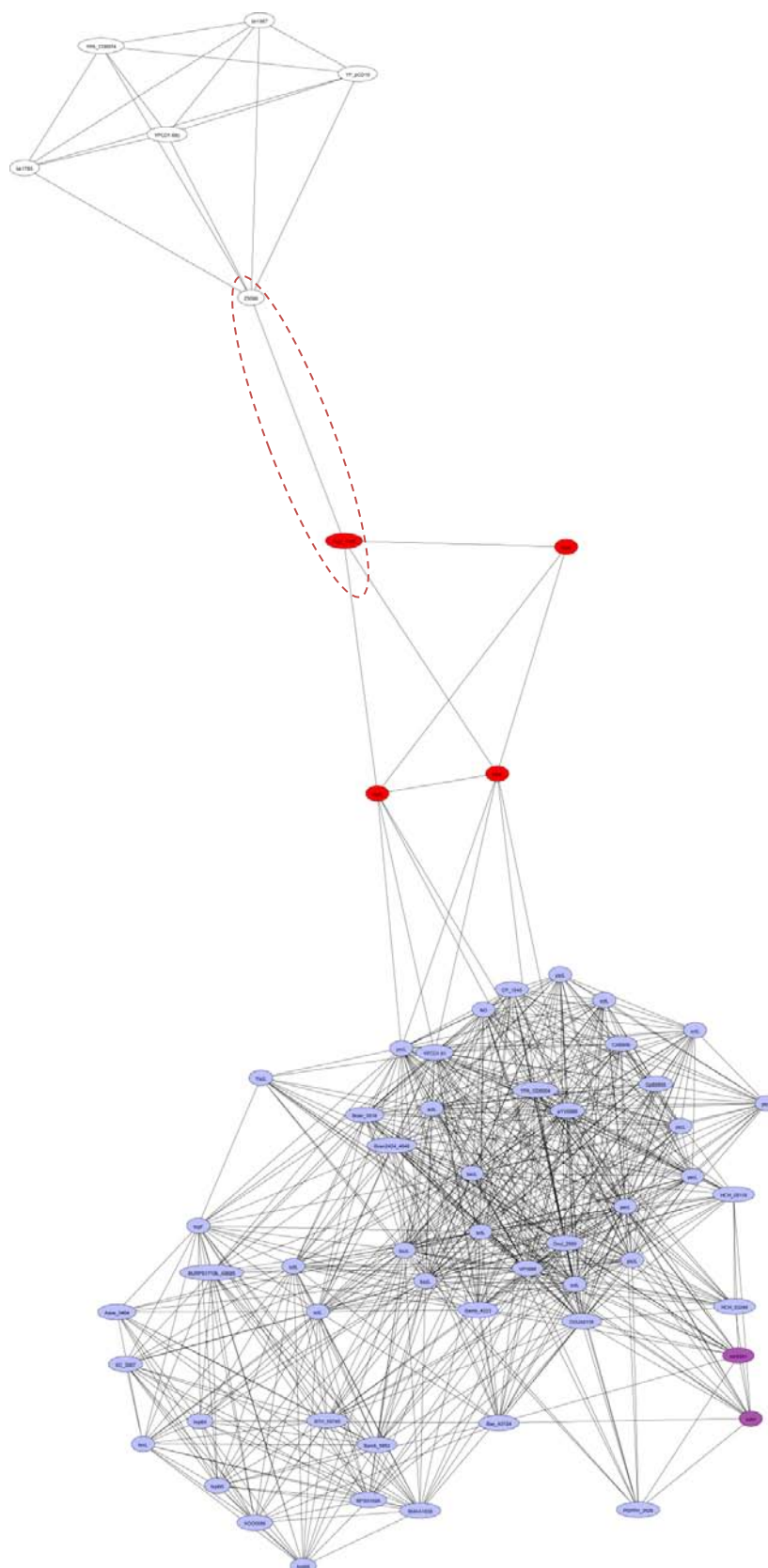
protein families is the rate at which errors are made, in particular false-positive and false-negative errors. In the case of finding T3SS loci the use of highly conserved proteins, and the fact that the vast majority of T3SSs are encoded in just one locus means that there were very few false positive hits, and most of these could be easily discarded as they were in isolated locations (i.e. did not cluster with other hits). Also, the relatively small number of clusters found by this approach meant that manual supervision of the clustering process was an amenable solution to improving data quality. This included manually redefining the start and end boundaries of the clusters for two purposes: Firstly to reduce the size of the search space for the subsequent BLAST searches, and secondly to minimise the possibility of proteins not associated with T3SS being clustered with T3SS proteins.

Given the numbers of proteins involved we have reached a stage where manual intervention and supervision of the process of assigning homology becomes impractical, and as such we must assess whether BLAST/PSI-BLAST with or without homology networking provides sufficiently accurate data. As was seen in previous studies, BLAST based homology networks can under-report homologous proteins, leading to the difference between the numbers of SctD proteins found in this study versus the study by Medini *et al.* Similarly the PSI-BLAST searches attempted here also missed homologous proteins dependant on what protein was used as the starting point of the search. The advantage of using networking approaches mean that any bias caused by the choice of starting protein is removed, as at some point every protein is used as the starting point. The down side of this approach is that it can magnify the effect of proteins which produce irrelevant hits. Once such example of this is the chimeric protein which joins together the Secretin and InvE/HrpJ protein families. In this case a simple analysis of the protein causing the join reveals the cause of the

problem. There are other examples of this effect can be seen in the AraC network above and, more prominently, in the SctL/HrpE network shown in Figure 30.

In both of these cases there are proteins or groups of proteins which show a low degree of connectivity compared to other members of the network. This is particularly true for the connection between the SctL and HrpE members of the network, and the Transposase\_25 proteins from plasmids present in *Yersinia*. In this case the fact that the two groups of proteins have known different roles and the relative weakness of the homology between Psyr\_1197 and Z5098 (highlighted in the dotted red oval) compared to Psyr\_1197 and the three other HrpE proteins, suggests that the link should be removed and the network partitioned into two separate ones.

In all the examples above manual intervention is required in order to tease apart or join together networks based on closer examination of the available homology data. With the exception of a couple of instances, BLAST on the whole seems to underpredict the size of homologous protein networks. Within the larger networks of proteins (those present in more than one group of T3SSs), attempts to link together networks using PSI-BLAST data often reveals additional relevant networks. For example the apparatus proteins SctD, SctL, SctF are all split into multiple networks, as are the CesT and TPR type chaperones. Given this information, one might assume that using PSI-BLAST data to perform unsupervised clustering of BLAST networks would be of some value. PSI-BLAST adds connections between 178 of the 685 networks; however it also produces a single network which contains a total of 102 BLAST networks and 1822 proteins. Obviously this result is incorrect, and represents a significant problem with utilising PSI-BLAST data, in that once irrelevant hits are included in the position specific scoring matrix the chance of the final result set



**Figure 30. Graphical representation of SctL/HrpE protein homology network**

Graph drawn as per Figure 29. Colour of nodes represents protein type (Blue: SctL, Purple: NoIV, Red: HrpE, White: Transposase\_25). This network is a clear example of where a weak BLAST hit joins two separate networks

containing only relevant protein hits diminishes dramatically. In this case however, the collection of 102 networks joined by PSI-BLAST acts to funnel all the unreliable PSI-BLAST data together, and leaves a small set of informative hits. The vast number of these hits join only two or three networks together, to produce a number of ‘super’ networks which encompass related proteins from one group of T3SSs. Examples of this include several proteins from the Esc/Ssa group such as SsaE, SsaF and SsaM where the two separate networks contain proteins from *Salmonella* and *E. coli* respectively. Similarly within the Ysc group there are examples of separate networks for members of the different subgroups: Ysc, *Desulfovibrio* and *Bordetella*, which can be joined together by PSI-BLAST information. Such protein groups include YscK, YscX and YscO.

There is however, data within the PSI-BLAST hits to support the joining of nine separate networks together. These networks consist of two groups of proteins (SctI and SctF) found in three different groups of T3SSs (Ysc, Inv-Mxi-Spa and Esc/Ssa). The PSI-BLAST data joins together the five SctI networks and the four SctF networks, and crucially, the two SctF proteins from *Desulfovibrio vulgaris* also provide good statistical backing (e-values between  $1 \times 10^{-10}$  and  $1 \times 10^{-15}$ ), albeit after 6 PSI-BLAST iterations, for joining the two groups together, validating our statement in an earlier paper that PrgI and PrgJ were homologous [219] (See Figure 31 and Figure 32).

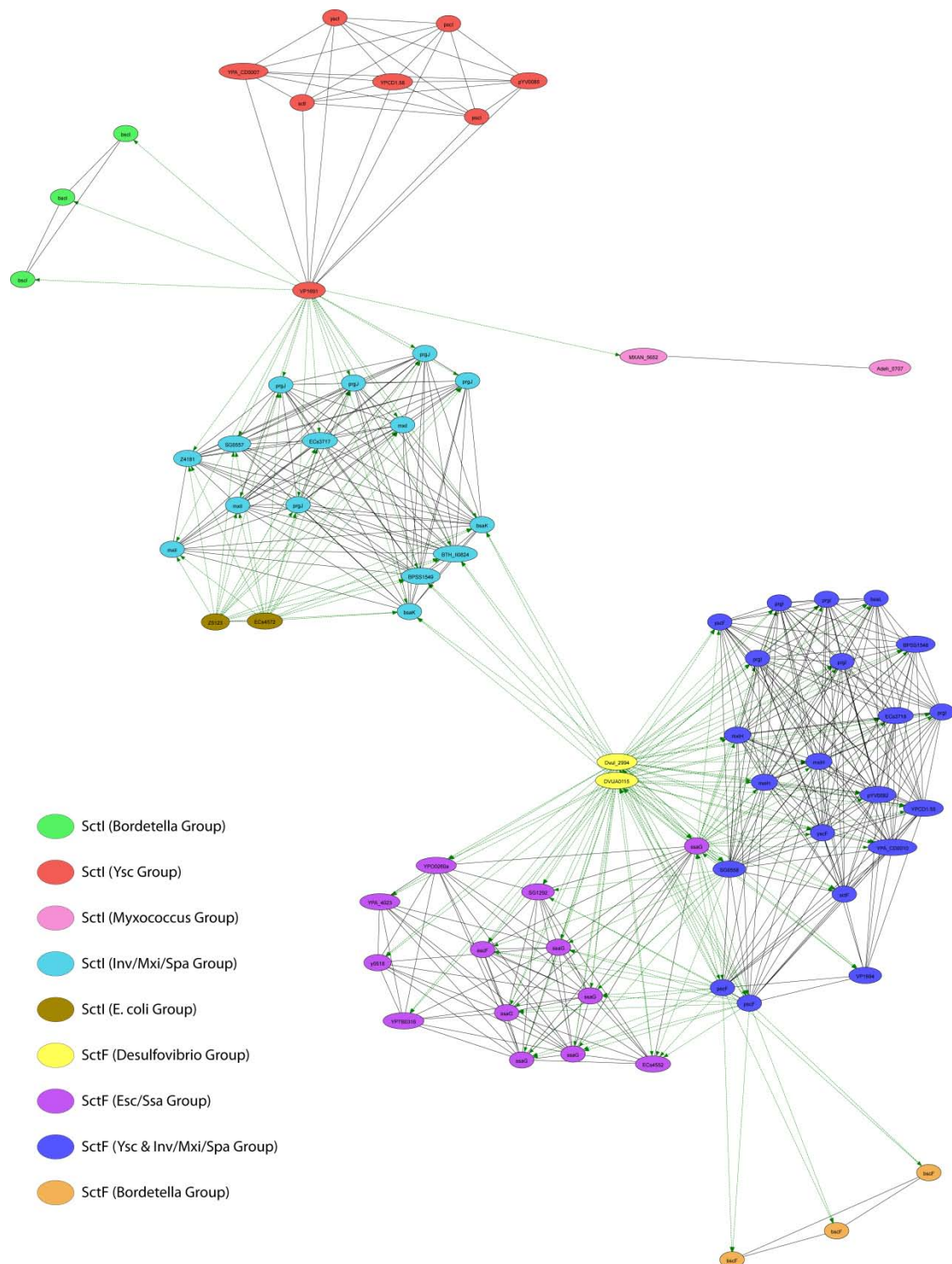
In either case it becomes clear that automated homology approaches used in this and other studies are not without their pitfalls. BLAST tends to underpredict the number of homologous proteins, while PSI-BLAST tends to overpredict. Although in some cases pruning of nodes with a low number of connections may help matters, it would

not for example detect the Secretin-InvE chimera protein shown in Figure 24. The ability for these types of tools to allow bulk analysis of proteins in volumes which would be impossible by manual means alone needs to be measured against the caution one must apply when accepting their results verbatim. These tools also have nothing to add when homology searches show only hits to other unknown or hypothetical proteins, or when no homology could be found at all, as is the case for over 15% of the proteins predicted to form part of a T3SS in this study. In these cases we are still entirely reliant on lab based and other techniques in order to predict their potential role. In the absence of such evidence we are limited to speculation on their role and relevance based solely evolutionary inference.

#### ***5.4.4. Non-detection of known proteins***

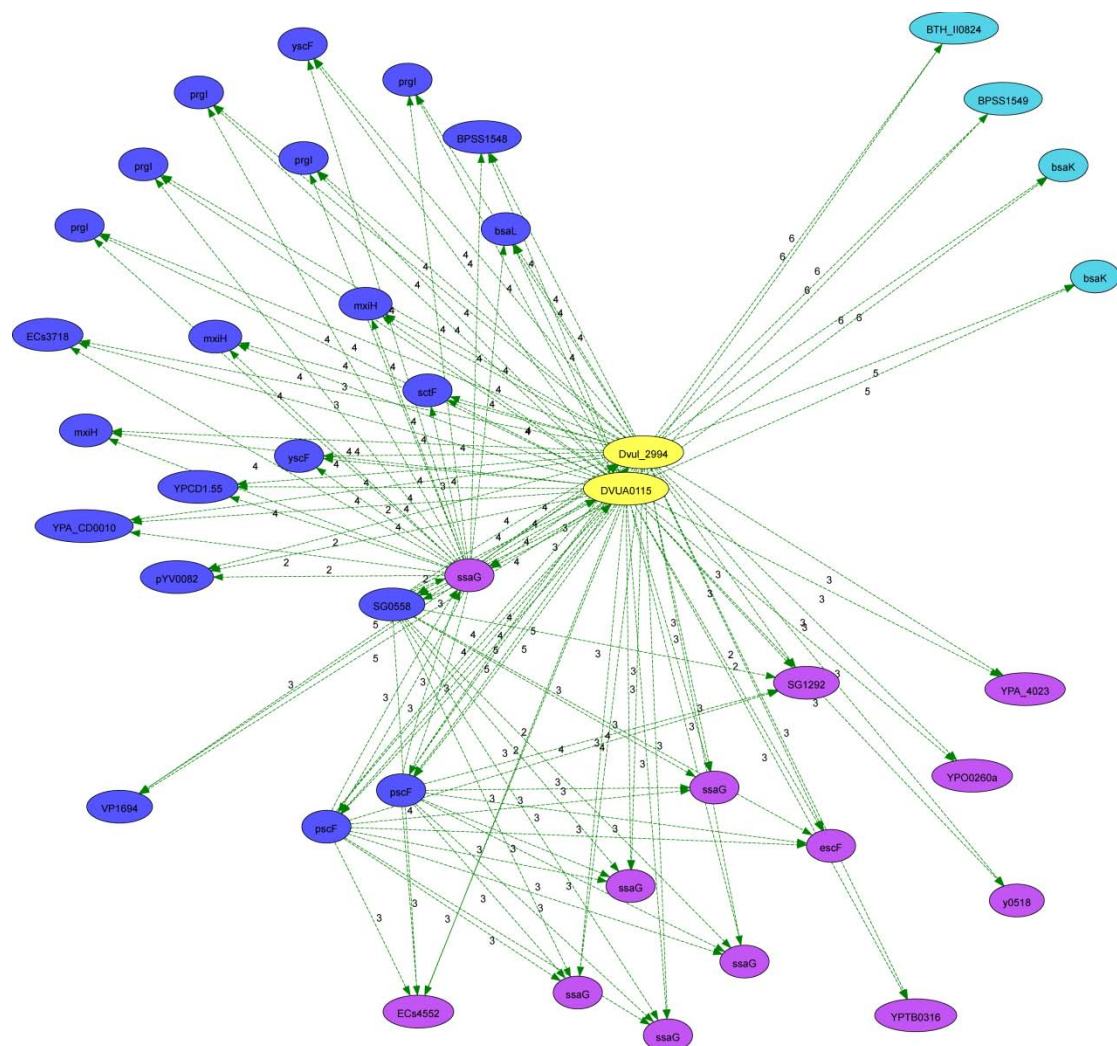
Section 5.3.6 above lists several examples of proteins which would have been expected to be detected within this bioinformatics analysis, but for various reasons were not. Each of these examples provides an insight into the problems that can be encountered when implementing the search methodology used in this analysis. Starting with the first step of locus determination highlights the breadth of problems which this form of analysis can be applied to, and circumstances where it may not be appropriate for T3SSs.

In order to make the calculation of an all-versus-all BLAST data set given the available computational resources it was necessary to reduce the number of proteins down from the approaching 2 million proteins which represent all proteins from all sequenced bacteria – and a potential search space of  $2 \times 10^{12}$  pairwise protein comparisons, to a number several orders of magnitude smaller.



**Figure 31. Graph of SctF and SctI networks joined by PSI-BLAST**

Nodes represent proteins, and solid black edges represent BLAST homology (edge length inversely proportional e-value). Dashed green lines represent PSI-BLAST homology between networks, arrow heads represent direction of search (arrow start: query protein, arrow head: hit protein). Colour of the node represents the original (BLAST) network to which the protein belonged. Note how PSI-BLAST hits are key to being able to join proteins which show little homology to each other.



**Figure 32. Graph of SctF networks showing PSI-BLAST homology to *D. vulgaris* network**  
 Drawn from the same data as Figure 31. Edges representing BLAST homology have been removed but nodes are still clustered based on this data. Only proteins showing homology by PSI-BLAST to Dvul\_2994 or DVUA0115 are shown, and the numbers on the dashed green edges represent the first PSI-BLAST iteration in which homology between the two proteins was found

This necessity meant that large portions of the whole protein database had to be excluded. With only a few exceptions T3SS loci are only encoded on a single locus within the genome. Thus, a process which excluded proteins encoded outside of these loci is a simple method to reduce the protein search space. Such a shortcut however means that situations where the T3SS is not encoded in a single locus present a problem for the analysis.

This was particularly a problem for the Chlamydial T3SSs where the genes encoding the system are found in multiple loci. There are several approaches that can be taken to solve this problem. Either the whole genome can be considered to be the T3SS locus, or the genome can be split into several small loci, each containing a subset of the complete T3SS gene complement. Choosing the former option would result in the production of clusters containing irrelevant proteins that are conserved throughout Chlamydiae, choosing the latter increases the likelihood that T3SS genes will be absent from the latter stages of the analysis because their location was not annotated as part of a T3SS locus.

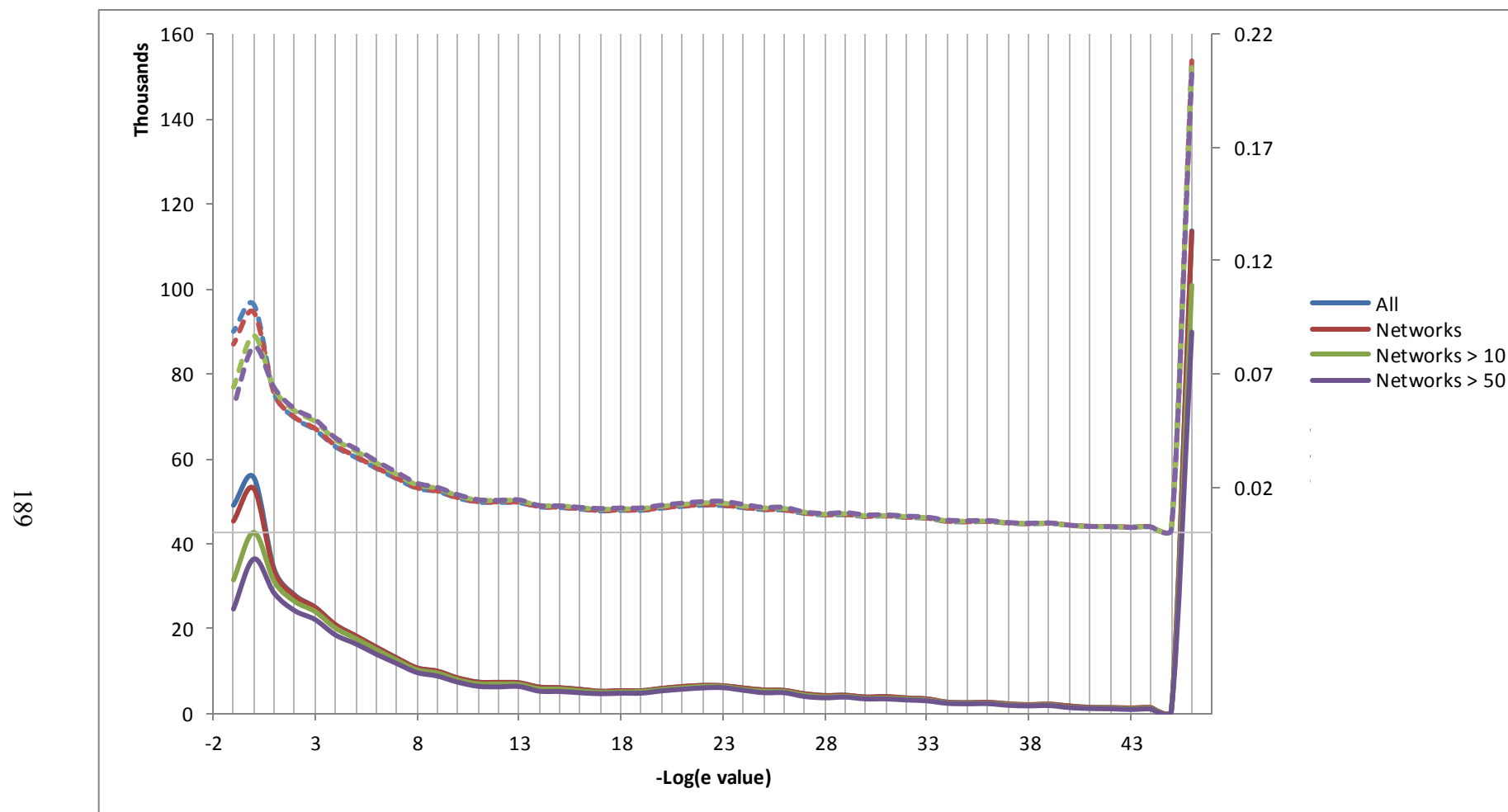
Whilst the latter approach was taken to simplify later analysis, locus determination within the Chlamydial genomes presented several additional challenges, as some of the normal clues as to the start and end of cluster lie are not present in these genomes. Firstly their T3SS genes do not show compositional bias compared to the chromosomal backbone that Proteobacterial T3SS do, and looking for degree of conservation of genes by BLAST is also not as beneficial compared to other genomes, as it is common to see co-linearity conserved around the T3SS loci, again unlike proteobacterial T3SSs. In defence of choosing the locus approach there were numerous networks in the final analysis with ten members. These networks contain



the equivalent proteins from the ten different chlamydial genomes under examination in this analysis. In the vast majority of cases contained proteins unrelated to type-III secretion (based on examination of the annotation of members of these networks). Thus it is likely that including all proteins found in all chlamydial species would have resulted in a massive expansion in the number of ten and eleven (for cases where an equivalent protein was also present in *Protochlamydia*) member networks, making meaningful examination of any networks of this size impossible.

Following the creation of T3SS loci the next stage was the automated all-versus-all homology search using BLAST and PSI-BLAST. As has been mentioned elsewhere in this analysis there is a constant trade-off between false positive and false negative rates when using choosing not just whether to use an iterated BLAST, but also whether to employ filtering, compositional based statistics and what e-value cutoff to use.

In particular, the decision on the e-value cutoff at which BLAST hits were included in the network analysis has a large effect on the final make up, in terms of numbers and size, of the networks. There are several examples of separate networks which were only later associated with one another by PSI-BLAST results, that would have been placed into a single network by BLAST results alone (one such example of this is the YopN/TyeA/SepL/MxiC family) had a larger cutoff been used. Altering the e-value cutoff in order to ensure that this was the case though would have a dramatic effect on the size of the networks produced. Figure 33 shows the composition of different subsets of the BLAST result set based on the membership of networks of different sizes. When only larger networks are considered then there is a decrease in number (and relative proportion) of hits present in these networks with less significant scores



**Figure 33. Number of BLAST hits by e-value.**

All: All BLAST hits. Networks: BLAST hits that form part of a Network. Networks >10: BLAST hits that form part of a network with more than 10 members. Networks >50: BLAST hits which form part of a network with more than 50 members. Dotted lines: Equivalent data plotted as a proportion of total hits (on secondary y axis)

( $e\text{-value} \geq 10^{-3}$ ), suggesting that less significant hits are more likely to have a role in the formation of smaller networks. On the basis that smaller networks are more likely to contain proteins unrelated to type-III secretion, or consist only of hypothetical proteins, then a decision to use a larger  $e\text{-value}$  cutoff would be likely just produce more small networks rather than expand the membership of larger networks.

Conversely, even at the selected  $e\text{-value}$  cutoff of  $10^{-3}$  there were false-positive BLAST results which confounded analysis of the results (e.g. LEE family SctL proteins). As there were approximately 10000 proteins in the search database, BLAST would have compared a total of 50 million protein pairings. Using this data an  $e\text{-value}$  cutoff can be determined in order to select an appropriate false positive rate. Thus, for example, if the desire was to have less than one false-positive result in the entire search output, then an  $e\text{-value}$  cutoff of  $\leq 2 \times 10^{-8}$  should be chosen. The choice of a larger cutoff value means that there will be a larger number of false positive results (approximately 5000, based on purely theoretical considerations), but as can be seen from the graph in Figure 33, even at  $e\text{-values}$  several orders of magnitude larger than  $10^{-8}$  there are very few hits that are not members of large networks.

Given the nature of the networking and PSI-BLAST portions of this analysis an  $e\text{-value}$  cutoff smaller than  $10^{-3}$  could have been chosen without a dramatic effect on the final result as many hits above this value subsequently appeared in PSI-BLAST results with an  $e\text{-value}$  several orders of magnitude smaller than the cutoff – thus the end result would be networks of similar size and structure, but with a larger shift to greater reliance on PSI-BLAST results to perform the networking analysis. Whilst it is hard to ascertain the precise effect of such a change due the combinatorial effect of the change over multiple PSI-BLAST iterations, the likelihood is that stricter

tolerances for inclusion in a PSSM would have meant that more informative connections being made between BLAST networks using PSI-BLAST, rather than the formation of a 102 network 'supergroup'.

The decisions on cutoffs and tolerances are important factors in interpreting the results of the analysis as a whole. The use of an e-value cutoff at  $10^{-3}$  strikes a balance between sensitivity and specificity, and thus there will be a proportion of both false positive and false negative results in the search output. Being able to repeat the analysis using a smaller cutoff would present an interesting comparison, as any positive hit could be more definitively relied on by the virtue of there being fewer false-positives in the result set. The benefits of using a larger cutoff are less clear, as whilst the false negative rate will fall as a result, overall analysis of the data set will become more difficult owing to the increased size of the result set and the exponential effect of the cutoff change on PSI-BLAST results.

Beyond choices of e-values there are also alternative approaches which could have been taken to improve the quality of the analysis. Construction of the PSSM was done based on the results of the first iteration of hits from an initial BLAST search. However, given that we already have sets of related proteins from which to form informative PSSMs in the sets of BLAST networked proteins, analysis of the PSSMs or equivalent HMMs for use in HMMER would be able to show similarity between networks based on consensus sampling and homology searching using these models. Similarly, inter- and intra-network analysis could be strengthened by assessing node connectivity. If a node (protein) within an individual network is well connected to multiple other members of the network, then its membership of the network is well supported. However, nodes which are subject to much lower levels of connectivity

(particularly if this is in comparison to the average level of connectivity across the network) should be subject to more rigorous examination, or exclusion from the network. Such methods would result in the correct action being taken with regards to the network shown in Figure 30.

The breadth of the protein set under examination here almost certainly precludes any meaningful analysis via *in vitro* or *in vivo* techniques of the data set as a whole. However, the individual networks themselves do present an interesting data set when looked at in isolation. Where the set of proteins numbers in the tens, as is the case for most networks, then more in depth *in silico* analyses are possible. For example, secondary structure can be determined for groups of interest to examine conservation amongst members of the same network. Similarly other features such as transmembrane domains can be looked for. This would particularly be of interest for T3SS apparatus proteins which lie in the inner or outer membrane. Information from these types of analyses can add additional lines of evidence to support or refute assertions about common functionality amongst the group. Similarly if structures are available for members of the group then homology modelling can be employed to attempt to align members of the group to a known 3D structure, with similar aims as secondary structure prediction.

Finally, it is worth examining the overall value of this automated approach in the light of the knowledge and intervention required to use it, and the information it provides. The first issue worth noting is one already looked at in section 1.1.6 – the cases where conserved sequence does not imply conserved function (and *vice versa*). Within type III secretion there are several examples of proteins (most notably class IB chaperones, and FHA domain proteins), where presence or absence of sequence

homology data is insufficient to draw conclusions on the presence, absence or function of such proteins. In the case of FHA domain proteins for example, an additional search for serine-threonine protein kinases/phosphatases would also be required to inform any discussion on the protein's activities and functions.

Beyond the fact that any sequence homology search is using sequence similarity/identity as a proxy measure for structural and functional conservation, the next issue surrounds one of the major topics covered in this section – the selection of appropriate search criteria. Whilst BLAST can be run with minimal intervention or understanding on the side of the user, when automated and multi-step processes are involved small alterations in starting parameters can have a strong effect on the results of the search, and this effect cannot be corrected as easily by manual intervention. Getting the most from BLAST and other related bioinformatics tools requires not only an understanding of the underlying biology, but also of the statistics underpinning their algorithms. As Jones and Swindells have pointed out “Many biologists design their own experiments with exquisite care yet still assume that the results from programs with more than 20 adjustable parameters are 100% reliable” [412]. Thus the tooling developed during this analysis could not be applied to another data set with the expectation that it would return appropriate results without alterations in the way the tools are configured and parameters used.

Similarly one must also consider precisely what the aim of the analysis is, and thus what the result set should represent. In this case parameters were chosen which balanced between sensitivity and specificity. However, if the intention was to filter a data set for further analysis (such as was the case in Chapter 3), then a larger e-value would be required to ensure near 100% sensitivity. Conversely if the aim is to

produce a definitive data set of only true positives then a smaller e-value would be required to produce near 100% specificity.

Finally, an understanding of the biology under examination, and a significant amount of time is still required to make best use of the data presented. As has been shown here, individual analysis of networking results is required to be sure of the data presented. Without knowledge of the system(s) under analysis then determining false positive links within networks becomes more difficult (e.g. Figure 30), as does decisions regarding inclusion of PSI-BLAST data in the analysis (e.g. Figure 29). Further, whilst the analysis methods used are capable of distilling down massive amounts of data into a more comprehensible level, examining each network in turn still requires a substantial amount of time. There are nearly 700 networks, plus over 150 links between networks generated from PSI-BLAST data. In order to get the most from the data generated each one of these networks and links needs to be examined, a process which would take several weeks of constant work even if only a few minutes were spent on each.

## **5.5. Summary**

The data presented in this chapter demonstrates the distribution of T3SSs amongst bacteria. Whilst there are only two phyla which contain an NF-T3SS, between flagellar and non-flagellar systems there are a total of ten phyla which contain a T3SS (from a total of 15 phyla with a genome sequence). This includes several phyla where only one or two genomes have been sequenced, and so there is still a chance that a T3SS may be found within these phyla. The approach of searching for hits to T3SS proteins and joining the into loci works very well for locating and calculating the number of NF-T3SSs since the vast majority of these systems (especially in

Proteobacteria) are contained in one locus. This assumption does not work as well for F-T3SSs where it is more common to see the system broken into multiple loci, particularly in epsilon-proteobacteria and Spirochetes. Within non-flagellar T3S loci there are a total of ten genes which are conserved throughout. By using a selection of these a well-supported phylogenetic tree can be constructed which confirms the different groups of T3SSs identified elsewhere, and also identified several additional groups and subgroups. The networks created by joining together BLAST hits show that there are a number of protein which exist within solely within one or several of these groups of subgroups, demonstrating that there are clear differences between the gene complement of different T3SS loci, and that the time these changes occurred can be placed at certain points in evolutionary space.

As with work presented in earlier chapters, the data presented here shows the need for careful decisions to be made when starting homology searches. Such searches are very unlikely to ever be 100% sensitive and specific, thus trade-offs must be made to ensure that wither all the results returned will be correct (in which case some true homologues will be missed), or that the result set will contain all homologues (but will also contain a number of false negatives). That being said, the data presented here show how homology searching can be used to powerfully examine numbers of genes and bacteria that simply would not be possible using *in vitro* techniques, and in doing so give novel insights into these genes and bacteria. Gaining such insights is not without problems however: curation of appropriate loci takes time, as does the analysis of the end product of the automated tools; the processes used within the analysis make it sensitive to the starting parameters used; and the tools do not present an opportunity to gain insight without significant knowledge of the system under analysis. As has been shown in Chapter 4, over-reliance on *in silico* tools can result in inaccurate data being presented into the public domain.



## **CHAPTER 6 - DISCUSSION**

### **6.1. The current view of type-III secretion**

Type-III secretion systems are complex organelles. The number of proteins they use in order to produce a functional secretion system is substantially larger than that required for a functional type-I or type-V secretion system, but closer to being on a par with the number of proteins found in type-II and type-IV secretion systems. What is it about these systems which means they require in the order of ten proteins or more to achieve what can be done by just three, or even one protein? One might initially be tempted to make an argument based on role and environment, but this falls flat in light of the evidence that the evolutionarily closest non-flagellar T3SSs are found in diverse bacteria performing different functions.

NF-T3SSs function in a wide variety of important situations. They are responsible for mediating the interaction between bacteria and a wide range of cell types in humans and other animals as well as being responsible for both pathogenesis and symbiosis with plant cells. The differences between NF-T3SSs responsible for interaction with plants and those responsible for interaction with animals are interesting in that it is one of the only key examples of a clear link between differences in the structural gene complement of a T3SS and its target. The evolutionary groups of T3SSs shown in this study also break down into groups containing animal T3SSs or plant T3SSs, but not both.

Beyond the differences between plant and animal NF-T3SSs, the differences between NF-T3SSs and the effect these have become less clear. For example the Esc/Ssa group

of T3SSs contain proteins homologous to EspA. EspA filaments extend from the end of the T3SS needle apparatus (formed by EspF in *E. coli*), and the translocation apparatus sits on the end of the EspA filament. As the filament is an order of magnitude longer than the needles typically formed by EspF like proteins (60 vs 600nm) [413-415], it should come as no surprise that regulation of the needle length such as is found in the T3SSs of *Yersinia* [413], is not seen in these systems. However, the role of EspA and why it is a conserved part of Esc/Ssa T3SSs is unclear, since Inv-Mxi-Spa T3SSs are able to infect similar cell types without such an extracellular appendage.

The myriad of forms seen today in different non-flagellar T3SSs suggests a complex series of events of gene gain and gene loss in order to produce them. There are for example numerous examples of paralogy and horizontal transfer to be found across the spectrum of T3SSs. The vast majority of non-flagellar T3SSs show evidence of horizontal gene transfer of the entire apparatus into the host genome. There are also examples of the separate inheritance of effector genes through their transfer by bacteriophage, a phenomenon already seen *Salmonella*[20], and now also well demonstrated within *E. coli* (as seen in Chapter 3). The inheritance of multiple core components of non-flagellar T3SSs from other systems such as ATP-synthases and Type-II secretion systems also displays the complex interplay between these systems and suggests a series of ancient events which resulted in paralogues of genes present in these systems being inherited by the ancestor of T3SSs.

The extensive presence of markers of horizontal gene transfer amongst Proteobacterial non-flagellar T3SSs demonstrate the role this mode of transfer has had on shaping the diversity of species containing T3SSs. As a result of this however, it

becomes very hard if not impossible to determine the bacterium which first contained the ancestral T3SS. The universal presence of NF-T3SSs in bacteria belonging to the Chlamydia phylum, along with the absence of any evidence of horizontal gene transfer may suggest that the ancestral bacterium belonged to this phylum.

Regardless of their ancestry, it is clear that there have been a series of important evolutionary events which have shaped the different groups of T3SSs seen today. The data presented in Chapter 5 demonstrates the diversity of proteins seen between different NF-T3SSs, and the presence of numerous proteins within NF-T3SS loci which appear to be conserved just amongst single groups or even sub-groups of systems. Many of these loss/gain events can also be mapped to specific points in the phylogenetic tree giving us clues as to when these events may have happened. These semi-conserved proteins fulfil various roles from altering the structural aspects of the apparatus, such as is the case for EspA and HrpZ, to changing the regulatory mechanisms used by the system such as the ECF-sigma system seen in *Bordetella*, *Desulfovibrio* and *Lawsonia*.

## **6.2. Discovering new T3SSs**

The data shown here confirm and enhance much of what was already known or suspected in the field of type-III secretion. Horizontal gene transfer is a key evolutionary aspect of NF-T3SSs and their diversity. The incongruence between standard phylogenetic trees such as those drawn using ribosomal RNA sequences and those drawn with T3SS proteins give clear evidence of the large number of horizontal gene transfer events required to rationalise the two trees. There are a large numbers of effectors in *E. coli* in locations outside of the NF-T3SS locus, in common with numerous bacteria such as *Salmonella* [20], *Pseudomonas* [288], and *Ralstonia* [289].

The presence of some of these effectors in genomic islands which show evidence of horizontal gene transfer also demonstrates that this method of gene transfer is also important in the transfer of effectors independent of the T3SS apparatus.

Given this information it is perhaps surprising that NF-T3SSs have only been found in two different phyla. Is this the true breadth of NF-T3SSs or will additional systems be discovered in additional phyla? This question and the ability to answer it also relate back to some of the issues mentioned in discussions earlier in this thesis. Namely: What role should the sequencers of the genome play in also annotating the genome, and what method should be used in order to locate T3SSs? Given the data presented in Chapter 4 and Chapter 5, there is a strong argument to be made for no annotation to be made by those who sequenced the genome. Instead the genome should be deposited purely as just its DNA sequence data, and for annotation to be added in a purely automated fashion. By doing this and allowing transfer of annotation from a limited number of highly curated genomes, hopefully errors in genome annotation will be minimised, but those who are unable to perform their own annotation will still be able to glean information from new genome data.

As was shown in Chapter 5 the role that homology searching techniques can have in determining whether homology exists between two genes or proteins must be carefully considered. No tool is currently able to produce predictions with no false negatives or false positives, and so one must instead determine an acceptable specificity and sensitivity level for the tool(s) used. For example BLAST is very specific in that it is unlikely to report a significant hit between two non-homologous proteins. However, it does lose some sensitivity by doing such. PSI-BLAST is the converse example of this in that it is more sensitive, but achieves this by being less

specific. Users of genome data which has been annotated automatically should be aware of these issues and thus treat this data with caution.

Tools such as the locus finder and networking tool presented in Chapter 5 can help improve the results of tools such as BLAST and HMMER by limiting the search space to limited sets of genes and proteins. This approach is obviously limited to families of genes (and their protein products) which are known to occupy definable loci within the genome. Similarly networking homology search results needs to be done with care, since it can magnify the effects of single proteins on the search result, and also can produce confusing results for multi-domain proteins.

### **6.3. The role of bioinformatics in T3SS research**

Given the availability of high quality DNA sequence, and the tools to analyse it, bioinformatics presents an appealing option for those wishing to survey many aspects of biology. Bioinformatic tools present a simple opportunity for those interested in a particular bacterium to find about the range of systems and pathways that it may have. From the other side of the picture it also allows those interested in a particular system or pathway to survey the number and range of bacteria which possess it.

Both of these aspects of bioinformatics have been used in this study in an attempt to learn more about the role of specific T3SS proteins, and the range of bacteria which contain T3SS proteins in general. The good degree of similarity shown between structural proteins in T3SSs allows for easy identification of T3SS containing bacteria, and to make educated estimations about T3SS loci and the genes they contain. Assigning annotations to individual genes within these loci is a more problematic process, and can require a much more labour intensive process to ensure

data quality is maintained.

Tools such as the locus finder described here make it easy for new T3SSs to be found given a genome with coding sequences called within it. Similarly this approach can be used for other systems found in single loci (see Appendix 2). It also allow for predictions to be made as to the function of novel genes, where homology data support it. These tools do not however abrogate the need for many other forms of scientific enquiry in order to enhance our knowledge of T3SSs. Large numbers of the genes identified as being part of a T3SS locus are annotated as hypothetical, and no predictions can be made as to their function. In these cases one must resort to techniques such as those show in Chapter 3, which allow for experimental techniques to validate predictions based on bioinformatics.

Similarly Chapter 5 demonstrates the role which bioinformatics tools can play in identifying whole T3SSs, defining their gene complement and characterising their evolutionary relationships to other systems. The approach used in this chapter relied heavily on hidden Markov models as a generic starting point for identifying T3SSs, and BLAST/PSI-BLAST for determining homology between proteins in different T3SS loci. HMMs and BLAST present an ideal opportunity for mining large data sets as they are quick to run and scale well as the size of the data to be analysed increases. However, these tools also have their disadvantages, HMMer searches for example require a pre-calculated model based on a defined alignment of related proteins, whilst BLAST has a wide number of tuneable parameters which will have mixed effects dependent on the query sequenced provided to the program.

These shortcomings result in some of the errors highlighted in section 5.3.6 and 5.4.4 regarding the sensitivity and specificity of the BLAST/PSI-BLAST result set.

Selection of an appropriate result set from a BLAST search requires careful determination of the search criteria used for each search. For example in the case of low-complexity filtering, when the filter is off common low complexity proteins can easily become included in BLAST results (particularly PSI-BLAST results), however, with the filter on some protein families (e.g. YopD/EspB) will produce incomplete result sets. As mentioned previously, with the search criteria used in Chapter 5, BLAST will on the whole tend to under-predict the size of homologous proteins families, whilst PSI-BLAST will tend to over predict. There are changes which could have been made to the inclusion/exclusion criteria such as smaller e-values, but the only effect this would have is to alter the balance between sensitivity and specificity. Thus whilst this approach provides a computationally easy method to assay T3SSs, there is still a large amount of manual work required to tease out the maximum amount of information from the results.

Given the limitations in the analytical method used in Chapter 5, and the breadth of bioinformatics tools available, it would be possible to improve the accuracy of the final result by the combination of multiple tools together. In this way a form of protein identification funnel could be formed. The first two steps of the funnel would be the steps already used, the finding of relevant loci, and the linking of homologous proteins together using PSI-BLAST. There are then several other tools which could then be used to further refine the networks. For example creation of HMMer models based on each BLAST subnetwork in a PSI-BLAST network would then be used to identify outliers or problematic proteins such as the multi-domain protein shown in Figure 24. More accurate homology searching tools such as FASTA, or even a direct Smith-Waterman comparison would also be possible on these smaller datasets, providing more accurate statistics and reducing false-positive hits.

Alternatively, there are several methods which could be used to reduce the issues over-prediction encountered when running PSI-BLAST. For example, there are alternative methods to initiate a PSI-BLAST search that a set of BLAST results. Other homology searching tools such as dynamic-programming alignment algorithms and multiple alignment methods can be used to prepare the position specific matrix. In these cases the non-directionality of PSI-BLAST hits will be minimised, as will the likelihood of PSI-BLAST pulling in false-positive hits.

Additional tools can also be used as independent validators of the BLAST data set. Domain databases such as PFAM/SMART/INTERPRO can be used to define domain architectures for proteins within a homology network to see if they are all related or if there are subfamilies within the network, or proteins which do not belong in the group. Similarly, structural modelling and prediction tools can also be used to examine members of homology networks in order to see if they retain a conserved secondary structure, or if homology modelling tools can successfully map the protein sequence onto the known structure of a protein in the same homology network. This structural information along with detailed alignment information when assessed against domain, motif and other information about key residues (e.g. binding sites) can be put together to produce a much fuller picture which allows for detailed comparison in the light of multiple data sources. In this regard such an analysis would in many regards be similar to the approach used in Chapter 3 where multiple independent experimental techniques were used to validate a candidate protein as an T3SS effector or not. However in this case the combination of sequence similarity searching, domain finding, structural prediction and external annotation serves not only to provide multiple sources to confirm an assignment of homology but also to add deeper predictions about function.



## REFERENCES

1. Ryle, A.P., et al., *The disulphide bonds of insulin*. Biochem J, 1955. **60**(4): p. 541-56.
2. Brown, H., F. Sanger, and R. Kitai, *The structure of pig and sheep insulins*. Biochem J, 1955. **60**(4): p. 556-65.
3. Zuckerkandl, E., R.T. Jones, and L. Pauling, *A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis*. Proc Natl Acad Sci U S A, 1960. **46**(10): p. 1349-60.
4. Konigsberg, W., G. Guidotti, and R.J. Hill, *The amino acid sequence of the alpha chain of human hemoglobin*. J Biol Chem, 1961. **236**: p. PC55-PC56.
5. Cantor, C.R. and T.H. Jukes, *The repetition of homologous sequences in the polypeptide chains of certain cytochromes and globins*. Proc Natl Acad Sci U S A, 1966. **56**(1): p. 177-84.
6. Margoliash, E., *Primary Structure and Evolution of Cytochrome C*. Proc Natl Acad Sci U S A, 1963. **50**: p. 672-9.
7. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
8. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd*. Science, 1995. **269**(5223): p. 496-512.
9. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
10. Bernal, A., U. Ear, and N. Kyrpides, *Genomes OnLine Database (GOLD): a monitor of genome projects world-wide*. Nucleic Acids Res, 2001. **29**(1): p. 126-7.
11. Health, N.I.f., *Genbank Overview*, National Institute for Health.
12. Li, W.-H., *Molecular evolution*. 1997, Sunderland, Massachusetts: Sinauer Associates. XV, 487.
13. Kimura, M., *Evolutionary rate at the molecular level*. Nature, 1968. **217**(5129): p. 624-6.
14. King, J.L. and T.H. Jukes, *Non-Darwinian evolution*. Science, 1969. **164**(881): p. 788-98.
15. Mazodier, P. and J. Davies, *Gene transfer between distantly related bacteria*. Annu Rev Genet, 1991. **25**: p. 147-71.
16. Chen, I., P.J. Christie, and D. Dubnau, *The ins and outs of DNA transfer in bacteria*. Science, 2005. **310**(5753): p. 1456-60.
17. Cascales, E. and P.J. Christie, *The versatile bacterial type IV secretion systems*. Nat Rev Microbiol, 2003. **1**(2): p. 137-49.
18. Schroder, G. and E. Lanka, *The mating pair formation system of conjugative plasmids-A versatile secretion machinery for transfer of proteins and DNA*. Plasmid, 2005. **54**(1): p. 1-25.
19. Llosa, M. and F. de la Cruz, *Bacterial conjugation: a potential tool for genomic engineering*. Res Microbiol, 2005. **156**(1): p. 1-6.
20. Brussow, H., C. Canchaya, and W.D. Hardt, *Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion*. Microbiol Mol Biol Rev, 2004. **68**(3): p. 560-602, table of contents.
21. Juhala, R.J., et al., *Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages*. J Mol Biol, 2000. **299**(1): p. 27-51.

22. Hendrix, R.W., et al., *The origins and ongoing evolution of viruses*. Trends Microbiol, 2000. **8**(11): p. 504-8.
23. Davison, J., *Genetic exchange between bacteria in the environment*. Plasmid, 1999. **42**(2): p. 73-91.
24. Rawlings, D.E. and E. Tietze, *Comparative biology of IncQ and IncQ-like plasmids*. Microbiol Mol Biol Rev, 2001. **65**(4): p. 481-96, table of contents.
25. Nesbo, C.L., M. Dlutek, and W.F. Doolittle, *Recombination in Thermotoga: implications for species concepts and biogeography*. Genetics, 2006. **172**(2): p. 759-69.
26. Salanoubat, M., et al., *Genome sequence of the plant pathogen Ralstonia solanacearum*. Nature, 2002. **415**(6871): p. 497-502.
27. Wachino, J., et al., *Horizontal transfer of blaCMY-bearing plasmids among clinical Escherichia coli and Klebsiella pneumoniae isolates and emergence of cefepime-hydrolyzing CMY-19*. Antimicrob Agents Chemother, 2006. **50**(2): p. 534-41.
28. Wu, T.L., et al., *CMY-2 beta-lactamase-carrying community-acquired urinary tract Escherichia coli: genetic correlation with Salmonella enterica serotypes Choleraesuis and Typhimurium*. Int J Antimicrob Agents, 2007. **29**(4): p. 410-6.
29. Parkhill, J., et al., *Genome sequence of Yersinia pestis, the causative agent of plague*. Nature, 2001. **413**(6855): p. 523-7.
30. Wei, J., et al., *Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T*. Infect Immun, 2003. **71**(5): p. 2775-86.
31. Thomson, N.R., et al., *The Complete Genome Sequence and Comparative Genome Analysis of the High Pathogenicity Yersinia enterocolitica Strain 8081*. PLoS Genet, 2006. **2**(12): p. e206.
32. Wommack, K.E. and R.R. Colwell, *Virioplankton: viruses in aquatic ecosystems*. Microbiol Mol Biol Rev, 2000. **64**(1): p. 69-114.
33. University, A.N., *Press Release: Star Count: ANU Astronomer Makes Best Yet*. 2003.
34. Pedulla, M.L., et al., *Origins of highly mosaic mycobacteriophage genomes*. Cell, 2003. **113**(2): p. 171-82.
35. Canchaya, C., et al., *Prophage genomics*. Microbiol Mol Biol Rev, 2003. **67**(2): p. 238-76, table of contents.
36. Waldor, M.K. and J.J. Mekalanos, *Lysogenic conversion by a filamentous phage encoding cholera toxin*. Science, 1996. **272**(5270): p. 1910-4.
37. O'Brien, A.D., et al., *Shiga-like toxin-converting phages from Escherichia coli strains that cause hemorrhagic colitis or infantile diarrhea*. Science, 1984. **226**(4675): p. 694-6.
38. Freeman, V.J., *Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae*. J Bacteriol, 1951. **61**(6): p. 675-88.
39. Wright, A., *Mechanism of conversion of the salmonella O antigen by bacteriophage epsilon 34*. J Bacteriol, 1971. **105**(3): p. 927-36.
40. Mirolid, S., et al., *Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic Salmonella typhimurium strain*. Proc Natl Acad Sci U S A, 1999. **96**(17): p. 9845-50.
41. Figueroa-Bossi, N. and L. Bossi, *Inducible prophages contribute to Salmonella virulence in mice*. Mol Microbiol, 1999. **33**(1): p. 167-76.
42. Figueroa-Bossi, N., et al., *Variable assortment of prophages provides a*

- transferable repertoire of pathogenic determinants in Salmonella*. Mol Microbiol, 2001. **39**(2): p. 260-71.
43. Thomson, N., et al., *The role of prophage-like elements in the diversity of Salmonella enterica serovars*. J Mol Biol, 2004. **339**(2): p. 279-300.
  44. Owen, R., *On the Archetype and Homologies of the Vertebrate Skeleton*. 1848, London: Murray.
  45. Zuckerkandl, E. and L. Pauling, eds. *Evolutionary divergence and convergence of proteins*. Evolving Genes and Proteins, ed. V. Bryson and H. Vogel. 1965, Academic: New York.
  46. Fitch, W.M., *Distinguishing homologous from analogous proteins*. Syst Zool, 1970. **19**(2): p. 99-113.
  47. Koonin, E.V., *Orthologs, paralog, and evolutionary genomics*. Annu Rev Genet, 2005. **39**: p. 309-38.
  48. Bellman, R.E., *Dynamic programming*. 1957, Princeton - N.J.: Princeton University Press. 339.
  49. Dreyfus, S.E. and A.M. Law, *The Art and theory of dynamic programming*. 1977, New York a.o.: Academic Press. XIII, 284.
  50. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. J Mol Biol, 1970. **48**(3): p. 443-53.
  51. Gotoh, O., *An improved algorithm for matching biological sequences*. J Mol Biol, 1982. **162**(3): p. 705-8.
  52. Durbin, R., *Biological sequence analysis probabilistic models of proteins and nucleic acids*. 9th print. ed. 2004, Cambridge: Cambridge University Press. 356.
  53. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences*. J Mol Biol, 1981. **147**(1): p. 195-7.
  54. Karlin, S. and G. Ghandour, *Comparative statistics for DNA and protein sequences: multiple sequence analysis*. Proc Natl Acad Sci U S A, 1985. **82**(18): p. 6186-90.
  55. Karlin, S. and G. Ghandour, *Multiple-alphabet amino acid sequence comparisons of the immunoglobulin kappa-chain constant domain*. Proc Natl Acad Sci U S A, 1985. **82**(24): p. 8597-601.
  56. Dayhoff, M.O. and National Biomedical Research Foundation., *Atlas of protein sequence and structure*, National Biomedical Research Foundation.: Silver Spring, Md., p. v.
  57. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks*. Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
  58. Altschul, S.F., et al., *Basic local alignment search tool*. J Mol Biol, 1990. **215**(3): p. 403-10.
  59. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
  60. Karlin, S. and S.F. Altschul, *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes*. Proc Natl Acad Sci U S A, 1990. **87**(6): p. 2264-8.
  61. Gumbel, E.J., *Statistics of extremes*. 1958, New York: Columbia University Press. 375.
  62. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison*. Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2444-8.

63. Pearson, W.R., *Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.* Genomics, 1991. **11**(3): p. 635-50.
64. Pearson, W.R., *Comparison of methods for searching protein sequence databases.* Protein Sci, 1995. **4**(6): p. 1145-60.
65. Shpaer, E.G., et al., *Sensitivity and selectivity in protein similarity searches: a comparison of Smith-Waterman in hardware to BLAST and FASTA.* Genomics, 1996. **38**(2): p. 179-91.
66. Salamov, A.A., et al., *Combining sensitive database searches with multiple intermediates to detect distant homologues.* Protein Eng, 1999. **12**(2): p. 95-100.
67. Taylor, W.R., *Identification of protein sequence homology by consensus template alignment.* J Mol Biol, 1986. **188**(2): p. 233-58.
68. Dodd, I.B. and J.B. Egan, *Systematic method for the detection of potential lambda Cro-like DNA-binding regions in proteins.* J Mol Biol, 1987. **194**(3): p. 557-64.
69. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-8.
70. Patthy, L., *Detecting homology of distantly related proteins with consensus sequences.* J Mol Biol, 1987. **198**(4): p. 567-77.
71. Henikoff, S. and J.G. Henikoff, *Embedding strategies for effective use of information from multiple sequence alignments.* Protein Sci, 1997. **6**(3): p. 698-705.
72. Bork, P., C. Sander, and A. Valencia, *An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins.* Proc Natl Acad Sci U S A, 1992. **89**(16): p. 7290-4.
73. Tatusov, R.L., S.F. Altschul, and E.V. Koonin, *Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks.* Proc Natl Acad Sci U S A, 1994. **91**(25): p. 12091-5.
74. Yi, T.M. and E.S. Lander, *Recognition of related proteins by iterative template refinement (ITR).* Protein Sci, 1994. **3**(8): p. 1315-28.
75. Schultz, J., et al., *SAM as a protein interaction domain involved in developmental regulation.* Protein Sci, 1997. **6**(1): p. 249-53.
76. Henikoff, S. and J.G. Henikoff, *Position-based sequence weights.* J Mol Biol, 1994. **243**(4): p. 574-8.
77. Nielsen, H. and A. Krogh, *Prediction of signal peptides and signal anchors by a hidden Markov model.* in *Proc Int Conf Intell Syst Mol Biol.* 1998. California: AAAI Press.
78. Gerstein, M., E.L. Sonnhammer, and C. Chothia, *Volume changes in protein evolution.* J Mol Biol, 1994. **236**(4): p. 1067-78.
79. Fraenkel, A.S., *Complexity of protein folding.* Bull Math Biol, 1993. **55**(6): p. 1199-210.
80. Lei, H. and Y. Duan, *Ab initio folding of albumin binding domain from all-atom molecular dynamics simulation.* J Phys Chem B, 2007. **111**(19): p. 5458-63.
81. Dill, K.A., et al., *The protein folding problem.* Annu Rev Biophys, 2008. **37**: p. 289-316.
82. Shirts, M. and V.S. Pande, *COMPUTING: Screen Savers of the World Unite!* Science, 2000. **290**(5498): p. 1903-1904.

83. <http://folding.stanford.edu/>. Available from: <http://folding.stanford.edu/>.
84. Cohen, B.I., S.R. Presnell, and F.E. Cohen, *Origins of structural diversity within sequentially identical hexapeptides*. Protein Sci, 1993. **2**(12): p. 2134-45.
85. Kabsch, W. and C. Sander, *On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations*. Proc Natl Acad Sci U S A, 1984. **81**(4): p. 1075-8.
86. Sternberg, M.J. and S.A. Islam, *Local protein sequence similarity does not imply a structural relationship*. Protein Eng, 1990. **4**(2): p. 125-31.
87. Punta, M. and Y. Ofran, *The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function*. PLoS Comput Biol, 2008. **4**(10): p. e1000160.
88. Tian, W. and J. Skolnick, *How well is enzyme function conserved as a function of pairwise sequence identity?* J Mol Biol, 2003. **333**(4): p. 863-82.
89. Yang, A.S. and B. Honig, *An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence*. J Mol Biol, 2000. **301**(3): p. 679-89.
90. Gibrat, J.F., T. Madej, and S.H. Bryant, *Surprising similarities in structure comparison*. Curr Opin Struct Biol, 1996. **6**(3): p. 377-85.
91. Yang, A.S. and B. Honig, *An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments*. J Mol Biol, 2000. **301**(3): p. 691-711.
92. Stahl, N. and G.D. Yancopoulos, *The alphas, betas, and kinases of cytokine receptor complexes*. Cell, 1993. **74**(4): p. 587-90.
93. Rozwarski, D.A., et al., *Structural comparisons among the short-chain helical cytokines*. Structure, 1994. **2**(3): p. 159-73.
94. Anfinsen, C.B., *Principles that govern the folding of protein chains*. Science, 1973. **181**(96): p. 223-30.
95. Pan, K.M., et al., *Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins*. Proc Natl Acad Sci U S A, 1993. **90**(23): p. 10962-6.
96. Stahl, N., et al., *Structural studies of the scrapie prion protein using mass spectrometry and amino acid sequencing*. Biochemistry, 1993. **32**(8): p. 1991-2002.
97. Prusiner, S.B., *Prions*. Proc Natl Acad Sci U S A, 1998. **95**(23): p. 13363-83.
98. Alexander, P.A., et al., *A minimal sequence code for switching protein structure and function*. Proc Natl Acad Sci U S A, 2009. **106**(50): p. 21149-54.
99. Abrahams, J.P., et al., *Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria*. Nature, 1994. **370**(6491): p. 621-8.
100. Milgrom, Y.M., L.L. Ehler, and P.D. Boyer, *ATP binding at noncatalytic sites of soluble chloroplast F1-ATPase is required for expression of the enzyme activity*. J Biol Chem, 1990. **265**(31): p. 18725-8.
101. Jault, J.M. and W.S. Allison, *Slow binding of ATP to noncatalytic nucleotide binding sites which accelerates catalysis is responsible for apparent negative cooperativity exhibited by the bovine mitochondrial F1-ATPase*. J Biol Chem, 1993. **268**(3): p. 1558-66.
102. Goward, C.R. and D.J. Nicholls, *Malate dehydrogenase: a model for structure, evolution, and catalysis*. Protein Sci, 1994. **3**(10): p. 1883-8.

103. Nicholls, D.J., et al., *The importance of arginine 102 for the substrate specificity of Escherichia coli malate dehydrogenase*. Biochem Biophys Res Commun, 1992. **189**(2): p. 1057-62.
104. Manning, G., et al., *The protein kinase complement of the human genome*. Science, 2002. **298**(5600): p. 1912-34.
105. Anamika, K., N. Garnier, and N. Srinivasan, *Functional diversity of human protein kinase splice variants marks significant expansion of human kinome*. BMC Genomics, 2009. **10**: p. 622.
106. Sadowski, M.I. and D.T. Jones, *The sequence-structure relationship and protein function prediction*. Curr Opin Struct Biol, 2009. **19**(3): p. 357-62.
107. Pallen, M., R. Chaudhuri, and A. Khan, *Bacterial FHA domains: neglected players in the phospho-threonine signalling game?* Trends Microbiol, 2002. **10**(12): p. 556-63.
108. Horn, M., et al., *Illuminating the evolutionary history of chlamydiae*. Science, 2004. **304**(5671): p. 728-30.
109. Bork, P., L. Holm, and C. Sander, *The immunoglobulin fold. Structural classification, sequence patterns and common core*. J Mol Biol, 1994. **242**(4): p. 309-20.
110. Halaby, D.M., A. Poupon, and J. Mornon, *The immunoglobulin fold family: sequence analysis and 3D structure comparisons*. Protein Eng, 1999. **12**(7): p. 563-71.
111. Lilic, M., M. Vujanac, and C.E. Stebbins, *A common structural motif in the binding of virulence factors to bacterial secretion chaperones*. Mol Cell, 2006. **21**(5): p. 653-64.
112. Reis, K.J., E.M. Ayoub, and M.D. Boyle, *Streptococcal Fc receptors. II. Comparison of the reactivity of a receptor from a group C streptococcus with staphylococcal protein A*. J Immunol, 1984. **132**(6): p. 3098-102.
113. Akesson, P., et al., *Protein H--a novel IgG binding bacterial protein*. Mol Immunol, 1990. **27**(6): p. 523-31.
114. Frick, I.M., et al., *Convergent evolution among immunoglobulin G-binding bacterial proteins*. Proc Natl Acad Sci U S A, 1992. **89**(18): p. 8532-6.
115. Kikuchi, T., *Analysis of 3D structural differences in the IgG-binding domains based on the interresidue average-distance statistics*. Amino Acids, 2008. **35**(3): p. 541-9.
116. He, Y., et al., *Solution NMR structures of IgG binding domains with artificially evolved high levels of sequence identity but different folds*. Biochemistry, 2005. **44**(43): p. 14055-61.
117. Wistow, G. and J. Piatigorsky, *Recruitment of enzymes as lens structural proteins*. Science, 1987. **236**(4808): p. 1554-6.
118. Fernald, R.D., *Casting a genetic light on the evolution of eyes*. Science, 2006. **313**(5795): p. 1914-8.
119. Taylor, J.S. and J. Raes, *Duplication and divergence: the evolution of new genes and old ideas*. Annu Rev Genet, 2004. **38**: p. 615-43.
120. Fernald, R.D., *Evolution of eyes*. Curr Opin Neurobiol, 2000. **10**(4): p. 444-50.
121. Soundararajan, V., et al., *Atomic interaction networks in the core of protein domains and their native folds*. PLoS One, 2010. **5**(2): p. e9391.
122. Simpson, J.A. and E.S.C. Weiner, *The Oxford English dictionary*. Second ed. 1989, Oxford: Clarendon Press. 20 vols.
123. Sneath, P.H.A. and R.R. Sokal, *Numerical taxonomy the principles and practice of numerical classification*. 1973, San Francisco: Freeman. XV, 573.

124. Cavalli-Sforza, L.L. and A.W. Edwards, *Phylogenetic analysis. Models and estimation procedures*. Am J Hum Genet, 1967. **19**(3 Pt 1): p. 233-57.
125. Rzhetsky, A. and M. Nei, *Theoretical foundation of the minimum-evolution method of phylogenetic inference*. Mol Biol Evol, 1993. **10**(5): p. 1073-95.
126. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees*. Mol Biol Evol, 1987. **4**(4): p. 406-25.
127. Qi, J., B. Wang, and B.I. Hao, *Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach*. J Mol Evol, 2004. **58**(1): p. 1-11.
128. Henz, S.R., et al., *Whole-genome prokaryotic phylogeny*. Bioinformatics, 2005. **21**(10): p. 2329-35.
129. Snel, B., M.A. Huynen, and B.E. Dutilh, *Genome trees and the nature of genome evolution*. Annu Rev Microbiol, 2005. **59**: p. 191-209.
130. Henderson, I.R., et al., *Type V protein secretion pathway: the autotransporter story*. Microbiol Mol Biol Rev, 2004. **68**(4): p. 692-744.
131. Greenfield, J.J. and S. High, *The Sec61 complex is located in both the ER and the ER-Golgi intermediate compartment*. J Cell Sci, 1999. **112** ( Pt 10): p. 1477-86.
132. Van den Berg, B., et al., *X-ray structure of a protein-conducting channel*. Nature, 2004. **427**(6969): p. 36-44.
133. Rusch, S.L. and D.A. Kendall, *Oligomeric states of the SecA and SecYEG core components of the bacterial Sec translocon*. Biochim Biophys Acta, 2007. **1768**(1): p. 5-12.
134. Ullers, R.S., et al., *SecB is a bona fide generalized chaperone in Escherichia coli*. Proc Natl Acad Sci U S A, 2004. **101**(20): p. 7583-8.
135. Randall, L.L., et al., *Asymmetric binding between SecA and SecB two symmetric proteins: implications for function in export*. J Mol Biol, 2005. **348**(2): p. 479-89.
136. Walter, P., I. Ibrahim, and G. Blobel, *Translocation of proteins across the endoplasmic reticulum. I. Signal recognition protein (SRP) binds to in-vitro-assembled polysomes synthesizing secretory protein*. J Cell Biol, 1981. **91**(2 Pt 1): p. 545-50.
137. Valent, Q.A., et al., *The Escherichia coli SRP and SecB targeting pathways converge at the translocon*. Embo J, 1998. **17**(9): p. 2504-12.
138. Angelini, S., S. Deitermann, and H.G. Koch, *FtsY, the bacterial signal-recognition particle receptor, interacts functionally and physically with the SecYEG translocon*. EMBO Rep, 2005. **6**(5): p. 476-81.
139. Muller, M., *Twin-arginine-specific protein export in Escherichia coli*. Res Microbiol, 2005. **156**(2): p. 131-6.
140. Lee, P.A., D. Tullman-Ereck, and G. Georgiou, *The bacterial twin-arginine translocation pathway*. Annu Rev Microbiol, 2006. **60**: p. 373-95.
141. Jongbloed, J.D., et al., *TatC is a specificity determinant for protein secretion via the twin-arginine translocation pathway*. J Biol Chem, 2000. **275**(52): p. 41350-7.
142. De Leeuw, E., et al., *Membrane interactions and self-association of the TatA and TatB components of the twin-arginine translocation pathway*. FEBS Lett, 2001. **506**(2): p. 143-8.
143. Ray, N., et al., *Location and mobility of twin arginine translocase subunits in the Escherichia coli plasma membrane*. J Biol Chem, 2005. **280**(18): p. 17961-8.

144. Alami, M., et al., *Differential interactions between a twin-arginine signal peptide and its translocase in Escherichia coli*. Mol Cell, 2003. **12**(4): p. 937-46.
145. Emanuelsson, O., et al., *Locating proteins in the cell using TargetP, SignalP and related tools*. Nat Protoc, 2007. **2**(4): p. 953-71.
146. Valent, Q.A., et al., *Early events in preprotein recognition in E. coli: interaction of SRP and trigger factor with nascent polypeptides*. Embo J, 1995. **14**(22): p. 5494-505.
147. Cristobal, S., et al., *Competition between Sec- and TAT-dependent protein translocation in Escherichia coli*. Embo J, 1999. **18**(11): p. 2982-90.
148. Delepelaire, P., *Type I secretion in gram-negative bacteria*. Biochim Biophys Acta, 2004. **1694**(1-3): p. 149-61.
149. Dawson, R.J. and K.P. Locher, *Structure of a bacterial multidrug ABC transporter*. Nature, 2006. **443**(7108): p. 180-5.
150. Murakami, S., et al., *Crystal structure of bacterial multidrug efflux transporter AcrB*. Nature, 2002. **419**(6907): p. 587-93.
151. Binet, R. and C. Wandersman, *Protein secretion by hybrid bacterial ABC-transporters: specific functions of the membrane ATPase and the membrane fusion protein*. Embo J, 1995. **14**(10): p. 2298-306.
152. Thanabalu, T., et al., *Substrate-induced assembly of a contiguous channel for protein export from E.coli: reversible bridging of an inner-membrane translocase to an outer membrane exit pore*. Embo J, 1998. **17**(22): p. 6487-96.
153. Sharff, A., et al., *The role of the TolC family in protein transport and multidrug efflux. From stereochemical certainty to mechanistic hypothesis*. Eur J Biochem, 2001. **268**(19): p. 5011-26.
154. Betts, H.J., et al., *Bacterial Secretion Systems*, in *Bacterial-epithelial cell cross-talk molecular mechanisms in pathogenesis*, B.A. McCormick, Editor. 2006, Cambridge University Press: Cambridge. p. 439.
155. Touze, T., et al., *Interactions underlying assembly of the Escherichia coli AcrAB-TolC multidrug efflux system*. Mol Microbiol, 2004. **53**(2): p. 697-706.
156. Koronakis, V., et al., *Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export*. Nature, 2000. **405**(6789): p. 914-9.
157. Letoffe, S., J.M. Ghigo, and C. Wandersman, *Iron acquisition from heme and hemoglobin by a Serratia marcescens extracellular protein*. Proc Natl Acad Sci U S A, 1994. **91**(21): p. 9876-80.
158. Hinsä, S.M., et al., *Transition from reversible to irreversible attachment during biofilm formation by Pseudomonas fluorescens WCS365 requires an ABC transporter and a large secreted protein*. Mol Microbiol, 2003. **49**(4): p. 905-18.
159. Guzzo, J., et al., *The secretion genes of Pseudomonas aeruginosa alkaline protease are functionally related to those of Erwinia chrysanthemi proteases and Escherichia coli alpha-haemolysin*. Mol Microbiol, 1991. **5**(2): p. 447-53.
160. Letoffe, S., P. Delepelaire, and C. Wandersman, *Cloning and expression in Escherichia coli of the Serratia marcescens metalloprotease gene: secretion of the protease from E. coli in the presence of the Erwinia chrysanthemi protease secretion functions*. J Bacteriol, 1991. **173**(7): p. 2160-6.
161. Duong, F., A. Lazdunski, and M. Murgier, *Protein secretion by heterologous bacterial ABC-transporters: the C-terminus secretion signal of the secreted*



- protein confers high recognition specificity.* Mol Microbiol, 1996. **21**(3): p. 459-70.
162. Filloux, A., *The underlying mechanisms of type II protein secretion.* Biochim Biophys Acta, 2004. **1694**(1-3): p. 163-79.
  163. Voulhoux, R., et al., *Involvement of the twin-arginine translocation system in protein secretion via the type II pathway.* Embo J, 2001. **20**(23): p. 6735-41.
  164. Sandkvist, M., *Biology of type II secretion.* Mol Microbiol, 2001. **40**(2): p. 271-83.
  165. Sandkvist, M., et al., *Interaction between the autokinase EpsE and EpsL in the cytoplasmic membrane is required for extracellular secretion in Vibrio cholerae.* Embo J, 1995. **14**(8): p. 1664-73.
  166. Camberg, J.L. and M. Sandkvist, *Molecular analysis of the Vibrio cholerae type II secretion ATPase EpsE.* J Bacteriol, 2005. **187**(1): p. 249-56.
  167. Camberg, J.L., et al., *Synergistic stimulation of EpsE ATP hydrolysis by EpsL and acidic phospholipids.* Embo J, 2007. **26**(1): p. 19-27.
  168. Sandkvist, M., et al., *Direct interaction of the EpsL and EpsM proteins of the general secretion apparatus in Vibrio cholerae.* J Bacteriol, 1999. **181**(10): p. 3129-35.
  169. Sandkvist, M., et al., *Two regions of EpsL involved in species-specific protein-protein interactions with EpsE and EpsM of the general secretion pathway in Vibrio cholerae.* J Bacteriol, 2000. **182**(3): p. 742-8.
  170. Py, B., L. Loiseau, and F. Barras, *An inner membrane platform in the type II secretion machinery of Gram-negative bacteria.* EMBO Rep, 2001. **2**(3): p. 244-8.
  171. Nouwen, N., et al., *Secretin PulD: association with pilot PulS, structure, and ion-conducting channel formation.* Proc Natl Acad Sci U S A, 1999. **96**(14): p. 8173-7.
  172. Daefler, S., et al., *The C-terminal domain of the secretin PulD contains the binding site for its cognate chaperone, PulS, and confers PulS dependence on pIVf1 function.* Mol Microbiol, 1997. **24**(3): p. 465-75.
  173. Shevchik, V.E. and G. Condemine, *Functional characterization of the Erwinia chrysanthemi OutS protein, an element of a type II secretion system.* Microbiology, 1998. **144** ( Pt 11): p. 3219-28.
  174. Christie, P.J., *Agrobacterium tumefaciens T-complex transport apparatus: a paradigm for a new family of multifunctional transporters in eubacteria.* J Bacteriol, 1997. **179**(10): p. 3085-94.
  175. Christie, P.J., *Type IV secretion: the Agrobacterium VirB/D4 and related conjugation systems.* Biochim Biophys Acta, 2004. **1694**(1-3): p. 219-34.
  176. Shirasu, K., et al., *An inner-membrane-associated virulence protein essential for T-DNA transfer from Agrobacterium tumefaciens to plants exhibits ATPase activity and similarities to conjugative transfer genes.* Mol Microbiol, 1994. **11**(3): p. 581-8.
  177. Dang, T.A. and P.J. Christie, *The VirB4 ATPase of Agrobacterium tumefaciens is a cytoplasmic membrane protein exposed at the periplasmic surface.* J Bacteriol, 1997. **179**(2): p. 453-62.
  178. Stephens, K.M., C. Roush, and E. Nester, *Agrobacterium tumefaciens VirB11 protein requires a consensus nucleotide-binding site for function in virulence.* J Bacteriol, 1995. **177**(1): p. 27-36.
  179. Das, A. and Y.H. Xie, *The Agrobacterium T-DNA transport pore proteins VirB8, VirB9, and VirB10 interact with one another.* J Bacteriol, 2000. **182**(3):

- p. 758-63.
180. Ward, D.V., et al., *Peptide linkage mapping of the Agrobacterium tumefaciens vir-encoded type IV secretion system reveals protein subassemblies*. Proc Natl Acad Sci U S A, 2002. **99**(17): p. 11493-500.
  181. Lai, E.M., et al., *Genetic and environmental factors affecting T-pilin export and T-pilus biogenesis in relation to flagellation of Agrobacterium tumefaciens*. J Bacteriol, 2000. **182**(13): p. 3705-16.
  182. Yuan, Q., et al., *Identification of the VirB4-VirB8-VirB5-VirB2 pilus assembly sequence of type IV secretion systems*. J Biol Chem, 2005. **280**(28): p. 26349-59.
  183. Spudich, G.M., et al., *Intermolecular disulfide bonds stabilize VirB7 homodimers and VirB7/VirB9 heterodimers during biogenesis of the Agrobacterium tumefaciens T-complex transport apparatus*. Proc Natl Acad Sci U S A, 1996. **93**(15): p. 7512-7.
  184. Baron, C., Y.R. Thorstenson, and P.C. Zambryski, *The lipoprotein VirB7 interacts with VirB9 in the membranes of Agrobacterium tumefaciens*. J Bacteriol, 1997. **179**(4): p. 1211-8.
  185. Fernandez, D., et al., *The Agrobacterium tumefaciens VirB7 lipoprotein is required for stabilization of VirB proteins during assembly of the T-complex transport apparatus*. J Bacteriol, 1996. **178**(11): p. 3168-76.
  186. Llosa, M., et al., *The N- and C-terminal portions of the Agrobacterium VirB1 protein independently enhance tumorigenesis*. J Bacteriol, 2000. **182**(12): p. 3437-45.
  187. Zupan, J., et al., *VirB1\* promotes T-pilus formation in the vir-type IV secretion system of Agrobacterium tumefaciens*. J Bacteriol, 2007.
  188. Henderson, I.R., F. Navarro-Garcia, and J.P. Nataro, *The great escape: structure and function of the autotransporter proteins*. Trends Microbiol, 1998. **6**(9): p. 370-8.
  189. Henderson, I.R., R. Cappello, and J.P. Nataro, *Autotransporter proteins, evolution and redefining protein secretion*. Trends Microbiol, 2000. **8**(12): p. 529-32.
  190. Hoiczky, E., et al., *Structure and sequence analysis of Yersinia YadA and Moraxella UspAs reveal a novel class of adhesins*. Embo J, 2000. **19**(22): p. 5989-99.
  191. Guedin, S., et al., *Novel topological features of FhaC, the outer membrane transporter involved in the secretion of the Bordetella pertussis filamentous hemagglutinin*. J Biol Chem, 2000. **275**(39): p. 30202-10.
  192. Jacob-Dubuisson, F., C. Loch, and R. Antoine, *Two-partner secretion in Gram-negative bacteria: a thrifty, specific pathway for large virulence proteins*. Mol Microbiol, 2001. **40**(2): p. 306-13.
  193. Schonherr, R., et al., *Amino acid replacements in the Serratia marcescens haemolysin ShIA define sites involved in activation and secretion*. Mol Microbiol, 1993. **9**(6): p. 1229-37.
  194. Jacob-Dubuisson, F., et al., *Lack of functional complementation between Bordetella pertussis filamentous hemagglutinin and Proteus mirabilis HpmA hemolysin secretion machineries*. J Bacteriol, 1997. **179**(3): p. 775-83.
  195. Tamm, A., et al., *Hydrophobic domains affect the collagen-binding specificity and surface polymerization as well as the virulence potential of the YadA protein of Yersinia enterocolitica*. Mol Microbiol, 1993. **10**(5): p. 995-1011.
  196. Hueck, C.J., *Type III protein secretion systems in bacterial pathogens of*

- animals and plants*. Microbiol Mol Biol Rev, 1998. **62**(2): p. 379-433.
197. Marlovits, T.C., et al., *Structural insights into the assembly of the type III secretion needle complex*. Science, 2004. **306**(5698): p. 1040-2.
  198. Burghout, P., et al., *Structure and electrophysiological properties of the YscC secretin from the type III secretion system of Yersinia enterocolitica*. J Bacteriol, 2004. **186**(14): p. 4645-54.
  199. Lario, P.I., et al., *Structure and biochemical analysis of a secretin pilot protein*. Embo J, 2005. **24**(6): p. 1111-21.
  200. Yip, C.K., et al., *Structural characterization of the molecular platform for type III secretion system assembly*. Nature, 2005. **435**(7042): p. 702-7.
  201. Phan, J., B.P. Austin, and D.S. Waugh, *Crystal structure of the Yersinia type III secretion protein YscE*. Protein Sci, 2005. **14**(10): p. 2759-63.
  202. Slepentin, A., L.M. de la Maza, and E.M. Peterson, *Interaction between components of the type III secretion system of Chlamydiaceae*. J Bacteriol, 2005. **187**(2): p. 473-9.
  203. Creasey, E.A., et al., *Yeast two-hybrid system survey of interactions between LEE-encoded proteins of enteropathogenic Escherichia coli*. Microbiology, 2003. **149**(Pt 8): p. 2093-106.
  204. Jackson, M.W. and G.V. Plano, *Interactions between type III secretion apparatus components from Yersinia pestis detected using the yeast two-hybrid system*. FEMS Microbiol Lett, 2000. **186**(1): p. 85-90.
  205. Morita-Ishihara, T., et al., *Shigella Spa33 is an essential C-ring component of type III secretion machinery*. J Biol Chem, 2006. **281**(1): p. 599-607.
  206. Woestyn, S., et al., *YscN, the putative energizer of the Yersinia Yop secretion machinery*. J Bacteriol, 1994. **176**(6): p. 1561-9.
  207. Akeda, Y. and J.E. Galan, *Genetic analysis of the Salmonella enterica type III secretion-associated ATPase InvC defines discrete functional domains*. J Bacteriol, 2004. **186**(8): p. 2402-12.
  208. Vogler, A.P., et al., *Salmonella typhimurium mutants defective in flagellar filament regrowth and sequence similarity of Flil to F0F1, vacuolar, and archaeobacterial ATPase subunits*. J Bacteriol, 1991. **173**(11): p. 3564-72.
  209. Dreyfus, G., et al., *Genetic and biochemical analysis of Salmonella typhimurium Flil, a flagellar protein related to the catalytic subunit of the F0F1 ATPase and to virulence proteins of mammalian and plant pathogens*. J Bacteriol, 1993. **175**(10): p. 3131-8.
  210. Eichelberg, K., C.C. Ginocchio, and J.E. Galan, *Molecular and functional characterization of the Salmonella typhimurium invasion genes invB and invC: homology of InvC to the F0F1 ATPase family of proteins*. J Bacteriol, 1994. **176**(15): p. 4501-10.
  211. Gruber, G., et al., *Structure-function relationships of A-, F- and V-ATPases*. J Exp Biol, 2001. **204**(Pt 15): p. 2597-605.
  212. Blocker, A., K. Komoriya, and S. Aizawa, *Type III secretion systems and bacterial flagella: insights into their function from structural similarities*. Proc Natl Acad Sci U S A, 2003. **100**(6): p. 3027-30.
  213. Pozidis, C., et al., *Type III protein translocase: HrcN is a peripheral ATPase that is activated by oligomerization*. J Biol Chem, 2003. **278**(28): p. 25816-24.
  214. Minamino, T., et al., *Oligomerization of the bacterial flagellar ATPase Flil is controlled by its extreme N-terminal region*. J Mol Biol, 2006. **360**(2): p. 510-9.
  215. Claret, L., et al., *Oligomerization and activation of the Flil ATPase central to*

- bacterial flagellum assembly*. Mol Microbiol, 2003. **48**(5): p. 1349-55.
216. Zarivach, R., et al., *Structural analysis of a prototypical ATPase from the type III secretion system*. Nat Struct Mol Biol, 2007. **14**(2): p. 131-7.
  217. Muller, S.A., et al., *Double hexameric ring assembly of the type III protein translocase ATPase HrcN*. Mol Microbiol, 2006. **61**(1): p. 119-25.
  218. Sorg, J.A., B. Blaylock, and O. Schneewind, *Secretion signal recognition by YscN, the Yersinia type III secretion ATPase*. Proc Natl Acad Sci U S A, 2006. **103**(44): p. 16490-5.
  219. Pallen, M.J., S.A. Beatson, and C.M. Bailey, *Bioinformatics analysis of the locus for enterocyte effacement provides novel insights into type-III secretion*. BMC Microbiol, 2005. **5**(1): p. 9.
  220. Auvray, F., et al., *Intrinsic membrane targeting of the flagellar export ATPase FliI: interaction with acidic phospholipids and FliH*. J Mol Biol, 2002. **318**(4): p. 941-50.
  221. Minamino, T. and R.M. MacNab, *FliH, a soluble component of the type III flagellar export apparatus of Salmonella, forms a complex with FliI and inhibits its ATPase activity*. Mol Microbiol, 2000. **37**(6): p. 1494-503.
  222. Blaylock, B., et al., *Characterization of the Yersinia enterocolitica type III secretion ATPase YscN and its regulator, YscL*. J Bacteriol, 2006. **188**(10): p. 3525-34.
  223. Jouihri, N., et al., *MxiK and MxiN interact with the Spa47 ATPase and are required for transit of the needle components MxiH and MxiI, but not of Ipa proteins, through the type III secretion apparatus of Shigella flexneri*. Mol Microbiol, 2003. **49**(3): p. 755-67.
  224. Zhao, R., et al., *FliN is a major structural protein of the C-ring in the Salmonella typhimurium flagellar basal body*. J Mol Biol, 1996. **261**(2): p. 195-208.
  225. Gonzalez-Pedrajo, B., et al., *Interactions between C ring proteins and export apparatus components: a possible mechanism for facilitating type III protein export*. Mol Microbiol, 2006. **60**(4): p. 984-98.
  226. McMurry, J.L., J.W. Murphy, and B. Gonzalez-Pedrajo, *The FliN-FliH interaction mediates localization of flagellar export ATPase FliI to the C ring complex*. Biochemistry, 2006. **45**(39): p. 11790-8.
  227. Aizawa, S.I., *Bacterial flagella and type III secretion systems*. FEMS Microbiol Lett, 2001. **202**(2): p. 157-64.
  228. McMurry, J.L., et al., *Analysis of the cytoplasmic domains of Salmonella FlhA and interactions with components of the flagellar export machinery*. J Bacteriol, 2004. **186**(22): p. 7586-92.
  229. Minamino, T. and R.M. Macnab, *Components of the Salmonella flagellar export apparatus and classification of export substrates*. J Bacteriol, 1999. **181**(5): p. 1388-94.
  230. Ohnishi, K., et al., *The FliO, FliP, FliQ, and FliR proteins of Salmonella typhimurium: putative components for flagellar assembly*. J Bacteriol, 1997. **179**(19): p. 6092-9.
  231. Malakooti, J., B. Ely, and P. Matsumura, *Molecular characterization, nucleotide sequence, and expression of the fliO, fliP, fliQ, and fliR genes of Escherichia coli*. J Bacteriol, 1994. **176**(1): p. 189-97.
  232. Fan, F., et al., *The FliP and FliR proteins of Salmonella typhimurium, putative components of the type III flagellar export apparatus, are located in the flagellar basal body*. Mol Microbiol, 1997. **26**(5): p. 1035-46.

233. Fraser, G.M., et al, *Substrate specificity of type III flagellar protein export in Salmonella is controlled by subdomain interactions in FlhB*. Mol Microbiol, 2003. **48**(4): p. 1043-57.
234. Minamino, T. and R.M. Macnab, *Domain structure of Salmonella FlhB, a flagellar export component responsible for substrate specificity switching*. J Bacteriol, 2000. **182**(17): p. 4906-14.
235. Edqvist, P.J., et al, *YscP and YscU regulate substrate specificity of the Yersinia type III secretion system*. J Bacteriol, 2003. **185**(7): p. 2259-66.
236. Sorg, I., et al, *YscU recognizes translocators as export substrates of the Yersinia injectisome*. Embo J, 2007. **26**(12): p. 3015-24.
237. Ogino, T., et al, *Assembly of the type III secretion apparatus of enteropathogenic Escherichia coli*. J Bacteriol, 2006. **188**(8): p. 2801-11.
238. Durocher, D. and S.P. Jackson, *The FHA domain*. FEBS Lett, 2002. **513**(1): p. 58-66.
239. Mougous, J.D., et al, *Threonine phosphorylation post-translationally regulates protein secretion in Pseudomonas aeruginosa*. Nat Cell Biol, 2007. **9**(7): p. 797-803.
240. Plano, G.V., S.S. Barve, and S.C. Straley, *LcrD, a membrane-bound regulator of the Yersinia pestis low-calcium response*. J Bacteriol, 1991. **173**(22): p. 7293-303.
241. Ginocchio, C.C. and J.E. Galan, *Functional conservation among members of the Salmonella typhimurium InvA family of proteins*. Infect Immun, 1995. **63**(2): p. 729-32.
242. Crepin, V.F., et al, *Structural and functional studies of the enteropathogenic Escherichia coli type III needle complex protein EscJ*. Mol Microbiol, 2005. **55**(6): p. 1658-70.
243. Ueno, T., K. Oosawa, and S. Aizawa, *Domain structures of the MS ring component protein (FliF) of the flagellar basal body of Salmonella typhimurium*. J Mol Biol, 1994. **236**(2): p. 546-55.
244. Grunenfelder, B., S. Gehrig, and U. Jenal, *Role of the cytoplasmic C terminus of the FliF motor protein in flagellar assembly and rotation*. J Bacteriol, 2003. **185**(5): p. 1624-33.
245. Francis, N.R., et al, *Localization of the Salmonella typhimurium flagellar switch protein FliG to the cytoplasmic M-ring face of the basal body*. Proc Natl Acad Sci U S A, 1992. **89**(14): p. 6304-8.
246. Genin, S. and C.A. Boucher, *A superfamily of proteins involved in different secretion pathways in gram-negative bacteria: modular structure and specificity of the N-terminal domain*. Mol Gen Genet, 1994. **243**(1): p. 112-8.
247. Collins, R.F., et al, *Structure of the Neisseria meningitidis outer membrane PilQ secretin complex at 12 Å resolution*. J Biol Chem, 2004. **279**(38): p. 39750-6.
248. Collins, R.F., et al, *Analysis of the PilQ secretin from Neisseria meningitidis by transmission electron microscopy reveals a dodecameric quaternary structure*. J Bacteriol, 2001. **183**(13): p. 3825-32.
249. Burghout, P., et al, *Role of the pilot protein YscW in the biogenesis of the YscC secretin in Yersinia enterocolitica*. J Bacteriol, 2004. **186**(16): p. 5366-75.
250. Schuch, R. and A.T. Maurelli, *MxiM and MxiJ, base elements of the Mxi-Spa type III secretion system of Shigella, interact with and stabilize the MxiD secretin in the cell envelope*. J Bacteriol, 2001. **183**(24): p. 6991-8.

251. Daefler, S. and M. Russel, *The Salmonella typhimurium InvH protein is an outer membrane lipoprotein required for the proper localization of InvG*. Mol Microbiol, 1998. **28**(6): p. 1367-80.
252. Hoiczyk, E. and G. Blobel, *Polymerization of a single protein of the pathogen Yersinia enterocolitica into needles punctures eukaryotic cells*. Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4669-74.
253. Torruellas, J., et al., *The Yersinia pestis type III secretion needle plays a role in the regulation of Yop secretion*. Mol Microbiol, 2005. **57**(6): p. 1719-33.
254. Kenjale, R., et al., *The needle component of the type III secretion of Shigella regulates the activity of the secretion apparatus*. J Biol Chem, 2005. **280**(52): p. 42929-37.
255. Heesemann, J., B. Algermissen, and R. Laufs, *Genetically manipulated virulence of Yersinia enterocolitica*. Infect Immun, 1984. **46**(1): p. 105-10.
256. Schubot, F.D., et al., *Three-dimensional structure of a macromolecular assembly that regulates type III secretion in Yersinia pestis*. J Mol Biol, 2005. **346**(4): p. 1147-61.
257. Cordes, F.S., et al., *Helical structure of the needle of the type III secretion system of Shigella flexneri*. J Biol Chem, 2003. **278**(19): p. 17103-7.
258. Deane, J.E., et al., *Molecular model of a type III secretion system needle: Implications for host-cell sensing*. Proc Natl Acad Sci U S A, 2006. **103**(33): p. 12529-33.
259. Knutton, S., et al., *A novel EspA-associated surface organelle of enteropathogenic Escherichia coli involved in protein translocation into epithelial cells*. Embo J, 1998. **17**(8): p. 2166-76.
260. Roine, E., et al., *Hrp pilus: an hrp-dependent bacterial surface appendage produced by Pseudomonas syringae pv. tomato DC3000*. Proc Natl Acad Sci U S A, 1997. **94**(7): p. 3459-64.
261. Hu, W., et al., *Immunogold labeling of Hrp pili of Pseudomonas syringae pv. tomato assembled in minimal medium and in planta*. Mol Plant Microbe Interact, 2001. **14**(2): p. 234-41.
262. Magdalena, J., et al., *Spa32 regulates a switch in substrate specificity of the type III secretion of Shigella flexneri from needle components to Ipa proteins*. J Bacteriol, 2002. **184**(13): p. 3433-41.
263. Kubori, T., et al., *Molecular characterization and assembly of the needle complex of the Salmonella typhimurium type III protein secretion system*. Proc Natl Acad Sci U S A, 2000. **97**(18): p. 10225-30.
264. Journet, L., et al., *The needle length of bacterial injectisomes is determined by a molecular ruler*. Science, 2003. **302**(5651): p. 1757-60.
265. Makishima, S., et al., *Length of the flagellar hook and the capacity of the type III export apparatus*. Science, 2001. **291**(5512): p. 2411-3.
266. Tamano, K., et al., *Shigella Spa32 is an essential secretory protein for functional type III secretion machinery and uniformity of its needle length*. J Bacteriol, 2002. **184**(5): p. 1244-52.
267. Russmann, H., et al., *Molecular and functional analysis of the type III secretion signal of the Salmonella enterica InvJ protein*. Mol Microbiol, 2002. **46**(3): p. 769-79.
268. Shibata, S., et al., *FliK regulates flagellar hook length as an internal ruler*. Mol Microbiol, 2007. **64**(5): p. 1404-15.
269. Minamino, T., et al., *Domain organization and function of Salmonella FliK, a flagellar hook-length control protein*. J Mol Biol, 2004. **341**(2): p. 491-502.

270. Agrain, C., et al., *Characterization of a Type III secretion substrate specificity switch (T3S4) domain in YscP from Yersinia enterocolitica*. Mol Microbiol, 2005. **56**(1): p. 54-67.
271. Lafont, F. and F.G. van der Goot, *Oiling the key hole*. Mol Microbiol, 2005. **56**(3): p. 575-7.
272. Hayward, R.D., et al., *Cholesterol binding by the bacterial type III translocon is essential for virulence effector delivery into mammalian cells*. Mol Microbiol, 2005. **56**(3): p. 590-603.
273. Pettersson, J., et al., *The V-antigen of Yersinia is surface exposed before target cell contact and involved in virulence protein translocation*. Mol Microbiol, 1999. **32**(5): p. 961-76.
274. Sarker, M.R., et al., *The Yersinia Yop virulon: LcrV is required for extrusion of the translocators YopB and YopD*. J Bacteriol, 1998. **180**(5): p. 1207-14.
275. Derewenda, U., et al., *The structure of Yersinia pestis V-antigen, an essential virulence factor and mediator of immunity against plague*. Structure, 2004. **12**(2): p. 301-6.
276. Kummer, L.W., et al., *Antibodies and cytokines independently protect against pneumonic plague*. Vaccine, 2008. **26**(52): p. 6901-7.
277. Francis, M.S., H. Wolf-Watz, and A. Forsberg, *Regulation of type III secretion systems*. Curr Opin Microbiol, 2002. **5**(2): p. 166-72.
278. Pegues, D.A., et al., *PhoP/PhoQ transcriptional repression of Salmonella typhimurium invasion genes: evidence for a role in protein secretion*. Mol Microbiol, 1995. **17**(1): p. 169-81.
279. Akbar, S., et al., *AraC/XylS family members, HilD and HilC, directly activate virulence gene expression independently of HilA in Salmonella typhimurium*. Mol Microbiol, 2003. **47**(3): p. 715-28.
280. Schechter, L.M., et al., *The small nucleoid-binding proteins H-NS, HU, and Fis affect hilA expression in Salmonella enterica serovar Typhimurium*. Infect Immun, 2003. **71**(9): p. 5432-5.
281. Darwin, K.H. and V.L. Miller, *Type III secretion chaperone-dependent regulation: activation of virulence genes by SicA and InvF in Salmonella typhimurium*. Embo J, 2001. **20**(8): p. 1850-62.
282. Page, A.L. and C. Parsot, *Chaperones of the type III secretion pathway: jacks of all trades*. Mol Microbiol, 2002. **46**(1): p. 1-11.
283. Parsot, C., C. Hamiaux, and A.L. Page, *The various and varying roles of specific chaperones in type III secretion systems*. Curr Opin Microbiol, 2003. **6**(1): p. 7-14.
284. Wattiau, P., et al., *Individual chaperones required for Yop secretion by Yersinia*. Proc Natl Acad Sci U S A, 1994. **91**(22): p. 10493-7.
285. Neyt, C. and G.R. Cornelis, *Role of SycD, the chaperone of the Yersinia Yop translocators YopB and YopD*. Mol Microbiol, 1999. **31**(1): p. 143-56.
286. Trulzsch, K., et al., *Analysis of chaperone-dependent Yop secretion/translocation and effector function using a mini-virulence plasmid of Yersinia enterocolitica*. Int J Med Microbiol, 2003. **293**(2-3): p. 167-77.
287. Wilharm, G., et al., *Yersinia enterocolitica type III secretion: evidence for the ability to transport proteins that are folded prior to secretion*. BMC Microbiol, 2004. **4**(1): p. 27.
288. Guttman, D.S., et al., *A functional screen for the type III (Hrp) secretome of the plant pathogen Pseudomonas syringae*. Science, 2002. **295**(5560): p. 1722-6.

289. Mukaihara, T., et al., *Genetic screening of Hrp type III-related pathogenicity genes controlled by the HrpB transcriptional activator in Ralstonia solanacearum*. Mol Microbiol, 2004. **54**(4): p. 863-75.
290. Deng, W., et al., *Dissecting virulence: systematic and functional analyses of a pathogenicity island*. Proc Natl Acad Sci U S A, 2004. **101**(10): p. 3597-602.
291. Gruenheid, S., et al., *Identification and characterization of NleA, a non-LEE-encoded type III translocated virulence factor of enterohaemorrhagic Escherichia coli O157:H7*. Mol Microbiol, 2004. **51**(5): p. 1233-49.
292. Charpentier, X. and E. Oswald, *Identification of the secretion and translocation domain of the enteropathogenic and enterohemorrhagic Escherichia coli effector Cif, using TEM-1 beta-lactamase as a new fluorescence-based reporter*. J Bacteriol, 2004. **186**(16): p. 5486-95.
293. Anderson, D.M. and O. Schneewind, *A mRNA signal for the type III secretion of Yop proteins by Yersinia enterocolitica*. Science, 1997. **278**(5340): p. 1140-3.
294. Sory, M.P., et al., *Identification of the YopE and YopH domains required for secretion and internalization into the cytosol of macrophages, using the cyaA gene fusion approach*. Proc Natl Acad Sci U S A, 1995. **92**(26): p. 11998-2002.
295. Lloyd, S.A., et al., *Yersinia YopE is targeted for type III secretion by N-terminal, not mRNA, signals*. Mol Microbiol, 2001. **39**(2): p. 520-31.
296. Lloyd, S.A., et al., *Targeting exported substrates to the Yersinia TTSS: different functions for different signals?* Trends Microbiol, 2001. **9**(8): p. 367-71.
297. Foulter, B., et al., *Characterization of the ysa pathogenicity locus in the chromosome of Yersinia enterocolitica and phylogeny analysis of type III secretion systems*. J Mol Evol, 2002. **55**(1): p. 37-51.
298. Stephens, R.S., et al., *Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis*. Science, 1998. **282**(5389): p. 754-9.
299. Read, T.D., et al., *Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39*. Nucleic Acids Res, 2000. **28**(6): p. 1397-406.
300. Hsia, R.C., et al., *Type III secretion genes identify a putative virulence locus of Chlamydia*. Mol Microbiol, 1997. **25**(2): p. 351-9.
301. Kalman, S., et al., *Comparative genomes of Chlamydia pneumoniae and C. trachomatis*. Nat Genet, 1999. **21**(4): p. 385-9.
302. Pallen, M.J., C.W. Penn, and R.R. Chaudhuri, *Bacterial flagellar diversity in the post-genomic era*. Trends Microbiol, 2005. **13**(4): p. 143-9.
303. Ren, C.P., et al., *The ETT2 gene cluster, encoding a second type III secretion system from Escherichia coli, is present in the majority of strains but has undergone widespread mutational attrition*. J Bacteriol, 2004. **186**(11): p. 3547-60.
304. Betts, H.J., R.R. Chaudhuri, and M.J. Pallen, *An analysis of type-III secretion gene clusters in Chromobacterium violaceum*. Trends Microbiol, 2004. **12**(11): p. 476-82.
305. Mecsas, J.J. and E.J. Strauss, *Molecular mechanisms of bacterial virulence: type III secretion and pathogenicity islands*. Emerg Infect Dis, 1996. **2**(4): p. 270-88.
306. Gophna, U., E.Z. Ron, and D. Graur, *Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events*. Gene, 2003. **312**: p. 151-63.



307. Medini, D., A. Covacci, and C. Donati, *Protein homology network families reveal step-wise diversification of Type III and Type IV secretion systems*. PLoS Comput Biol, 2006. **2**(12): p. e173.
308. Blocker, A.J., et al., *What's the point of the type III secretion system needle?* Proc Natl Acad Sci U S A, 2008. **105**(18): p. 6507-13.
309. Cordes, F.S., et al., *Helical packing of needles from functionally altered Shigella type III secretion systems*. J Mol Biol, 2005. **354**(2): p. 206-11.
310. Moraes, T.F., T. Spreter, and N.C. Strynadka, *Piecing together the type III injectisome of bacterial pathogens*. Curr Opin Struct Biol, 2008. **18**(2): p. 258-66.
311. Gogarten, J.P., et al., *Evolution and isoforms of V-ATPase subunits*. J Exp Biol, 1992. **172**: p. 137-47.
312. Nouwen, N., et al., *Domain structure of secretin PulD revealed by limited proteolysis and electron microscopy*. Embo J, 2000. **19**(10): p. 2229-36.
313. Guilvout, I., et al., *Genetic dissection of the outer membrane secretin PulD: are there distinct domains for multimerization and secretion specificity?* J Bacteriol, 1999. **181**(23): p. 7212-20.
314. Daefler, S., M. Russel, and P. Model, *Module swaps between related translocator proteins pIV(f1), pIV(IKe) and PulD: identification of a specificity domain*. J Mol Biol, 1997. **266**(5): p. 978-92.
315. Walker, J.E., et al., *Structural aspects of proton-pumping ATPases*. Philos Trans R Soc Lond B Biol Sci, 1990. **326**(1236): p. 367-78.
316. Boyer, P.D., *The binding change mechanism for ATP synthase--some probabilities and possibilities*. Biochim Biophys Acta, 1993. **1140**(3): p. 215-50.
317. Masaike, T., et al., *Rotation of F(1)-ATPase and the hinge residues of the beta subunit*. J Exp Biol, 2000. **203**(Pt 1): p. 1-8.
318. Stock, D., A.G. Leslie, and J.E. Walker, *Molecular architecture of the rotary motor in ATP synthase*. Science, 1999. **286**(5445): p. 1700-5.
319. Sambongi, Y., et al., *Mechanical rotation of the c subunit oligomer in ATP synthase (F0F1): direct observation*. Science, 1999. **286**(5445): p. 1722-4.
320. Kaim, G., U. Matthey, and P. Dimroth, *Mode of interaction of the single a subunit with the multimeric c subunits during the translocation of the coupling ions by F1F0 ATPases*. Embo J, 1998. **17**(3): p. 688-95.
321. Cruz, J.A., C.A. Radkowski, and R.E. McCarty, *Functional Consequences of Deletions of the N Terminus of the [epsilon] Subunit of the Chloroplast ATP Synthase*. Plant Physiol, 1997. **113**(4): p. 1185-1192.
322. Cipriano, D.J. and S.D. Dunn, *The role of the epsilon subunit in the Escherichia coli ATP synthase. The C-terminal domain is required for efficient energy coupling*. J Biol Chem, 2006. **281**(1): p. 501-7.
323. Feniouk, B.A., T. Suzuki, and M. Yoshida, *The role of subunit epsilon in the catalysis and regulation of F0F1-ATP synthase*. Biochim Biophys Acta, 2006. **1757**(5-6): p. 326-38.
324. Smith, J.B. and P.C. Sternweis, *Purification of membrane attachment and inhibitory subunits of the proton translocating adenosine triphosphatase from Escherichia coli*. Biochemistry, 1977. **16**(2): p. 306-11.
325. Laget, P.P. and J.B. Smith, *Inhibitory properties of endogenous subunit epsilon in the Escherichia coli F1 ATPase*. Arch Biochem Biophys, 1979. **197**(1): p. 83-9.
326. Dunn, S.D. and R.G. Tozer, *Activation and inhibition of the Escherichia coli*

- F1-ATPase by monoclonal antibodies which recognize the epsilon subunit.* Arch Biochem Biophys, 1987. **253**(1): p. 73-80.
327. Lotscher, H.R., C. deJong, and R.A. Capaldi, *Interconversion of high and low adenosinetriphosphatase activity forms of Escherichia coli F1 by the detergent lauryldimethylamine oxide.* Biochemistry, 1984. **23**(18): p. 4140-3.
  328. Ketchum, C.J. and R.K. Nakamoto, *A mutation in the Escherichia coli F0F1-ATP synthase rotor, gammaE208K, perturbs conformational coupling between transport and catalysis.* J Biol Chem, 1998. **273**(35): p. 22292-7.
  329. Peskova, Y.B. and R.K. Nakamoto, *Catalytic control and coupling efficiency of the Escherichia coli FoF1 ATP synthase: influence of the Fo sector and epsilon subunit on the catalytic transition state.* Biochemistry, 2000. **39**(38): p. 11830-6.
  330. Suzuki, T., et al., *F0F1-ATPase/synthase is geared to the synthesis mode by conformational rearrangement of epsilon subunit in response to proton motive force and ADP/ATP balance.* J Biol Chem, 2003. **278**(47): p. 46840-6.
  331. Tsunoda, S.P., et al., *Large conformational changes of the epsilon subunit in the bacterial F1F0 ATP synthase provide a ratchet action to regulate this rotary motor enzyme.* Proc Natl Acad Sci U S A, 2001. **98**(12): p. 6560-4.
  332. Lu, M., et al., *The amino-terminal domain of the E subunit of vacuolar H(+)-ATPase (V-ATPase) interacts with the H subunit and is required for V-ATPase function.* J Biol Chem, 2002. **277**(41): p. 38409-15.
  333. Fethiere, J., et al., *Building the stator of the yeast vacuolar-ATPase: specific interaction between subunits E and G.* J Biol Chem, 2004. **279**(39): p. 40670-6.
  334. Ohira, M., et al., *The E and G subunits of the yeast V-ATPase interact tightly and are both present at more than one copy per V1 complex.* J Biol Chem, 2006. **281**(32): p. 22752-60.
  335. Aris, J.P. and R.D. Simoni, *Cross-linking and labeling of the Escherichia coli F1F0-ATP synthase reveal a compact hydrophilic portion of F0 close to an F1 catalytic subunit.* J Biol Chem, 1983. **258**(23): p. 14599-609.
  336. Sawada, K., et al., *Interaction of the delta and b subunits contributes to F1 and F0 interaction in the Escherichia coli F1F0-ATPase.* J Biol Chem, 1997. **272**(48): p. 30047-53.
  337. Lane, M.C., P.W. O'Toole, and S.A. Moore, *Molecular basis of the interaction between the flagellar export proteins FliI and FliH from Helicobacter pylori.* J Biol Chem, 2006. **281**(1): p. 508-17.
  338. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2007.
  339. <http://hmmer.wustl.edu>. Available from: <http://hmmer.wustl.edu>.
  340. Tamura, K., et al., *MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.* Mol Biol Evol, 2007. **24**(8): p. 1596-9.
  341. Suhre, K. and J.M. Claverie, *FusionDB: a database for in-depth analysis of prokaryotic gene fusion events.* Nucleic Acids Res, 2004. **32**(Database issue): p. D273-6.
  342. Dunn, S.D., et al., *The b subunit of Escherichia coli ATP synthase.* J Bioenerg Biomembr, 2000. **32**(4): p. 347-55.
  343. Gonzalez-Pedrajo, B., et al., *Molecular dissection of Salmonella FliH, a regulator of the ATPase FliI and the type III flagellar protein export pathway.* Mol Microbiol, 2002. **45**(4): p. 967-82.
  344. Dmitriev, O., et al., *Structure of the membrane domain of subunit b of the*

- Escherichia coli F0F1 ATP synthase*. J Biol Chem, 1999. **274**(22): p. 15598-604.
345. Paul, K., J.G. Harmon, and D.F. Blair, *Mutational analysis of the flagellar rotor protein FliN: identification of surfaces important for flagellar assembly and switching*. J Bacteriol, 2006. **188**(14): p. 5240-8.
  346. Fraser, G.M., et al., *Interactions of FliJ with the Salmonella type III flagellar export apparatus*. J Bacteriol, 2003. **185**(18): p. 5546-54.
  347. Minamino, T., et al., *The ATPase FliI can interact with the type III flagellar protein export apparatus in the absence of its regulator, FliH*. J Bacteriol, 2003. **185**(13): p. 3983-8.
  348. Rossier, O., G. Van den Ackerveken, and U. Bonas, *HrpB2 and HrpF from Xanthomonas are type III-secreted proteins and essential for pathogenicity and recognition by the host plant*. Mol Microbiol, 2000. **38**(4): p. 828-38.
  349. Hardt, W.D., H. Urlaub, and J.E. Galan, *A substrate of the centisome 63 type III protein secretion system of Salmonella typhimurium is encoded by a cryptic bacteriophage*. Proc Natl Acad Sci U S A, 1998. **95**(5): p. 2574-9.
  350. Miold, S., et al., *Salmonella host cell invasion emerged by acquisition of a mosaic of separate genetic elements, including Salmonella pathogenicity island 1 (SPI1), SPI5, and sopE2*. J Bacteriol, 2001. **183**(7): p. 2348-58.
  351. Mudgett, M.B. and B.J. Staskawicz, *Characterization of the Pseudomonas syringae pv. tomato AvrRpt2 protein: demonstration of secretion and processing during bacterial pathogenesis*. Mol Microbiol, 1999. **32**(5): p. 927-41.
  352. Jarvis, K.G., et al., *Enteropathogenic Escherichia coli contains a putative type III secretion system necessary for the export of proteins involved in attaching and effacing lesion formation*. Proc Natl Acad Sci U S A, 1995. **92**(17): p. 7996-8000.
  353. Jarvis, K.G. and J.B. Kaper, *Secretion of extracellular proteins by enterohemorrhagic Escherichia coli via a putative type III secretion system*. Infect Immun, 1996. **64**(11): p. 4826-9.
  354. McDaniel, T.K., et al., *A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens*. Proc Natl Acad Sci U S A, 1995. **92**(5): p. 1664-8.
  355. Hayashi, T., et al., *Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12*. DNA Res, 2001. **8**(1): p. 11-22.
  356. Perna, N.T., et al., *Genome sequence of enterohaemorrhagic Escherichia coli O157:H7*. Nature, 2001. **409**(6819): p. 529-33.
  357. Datsenko, K.A. and B.L. Wanner, *One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products*. Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6640-5.
  358. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers*. Methods Mol Biol, 2000. **132**: p. 365-86.
  359. Nougayrede, J.P., et al., *The long-term cytoskeletal rearrangement induced by rabbit enteropathogenic Escherichia coli is Esp dependent but intimin independent*. Mol Microbiol, 1999. **31**(1): p. 19-30.
  360. Marches, O., et al., *Enteropathogenic and enterohaemorrhagic Escherichia coli deliver a novel effector called Cif, which blocks cell cycle G2/M transition*. Mol Microbiol, 2003. **50**(5): p. 1553-67.
  361. Kim, D.W., et al., *The Shigella flexneri effector OspG interferes with innate*

- immune responses by targeting ubiquitin-conjugating enzymes*. Proc Natl Acad Sci U S A, 2005. **102**(39): p. 14046-51.
362. Chaudhuri, R.R., et al., *xBASE2: a comprehensive resource for comparative bacterial genomics*. Nucleic Acids Res, 2008. **36**(Database issue): p. D543-6.
  363. Mundy, R., et al., *Identification of a novel Citrobacter rodentium type III secreted protein, EspI, and roles of this and other secreted proteins in infection*. Infect Immun, 2004. **72**(4): p. 2288-302.
  364. Garmendia, J., et al., *TccP is an enterohaemorrhagic Escherichia coli O157:H7 type III effector protein that couples Tir to the actin-cytoskeleton*. Cell Microbiol, 2004. **6**(12): p. 1167-83.
  365. Dahan, S., et al., *EspJ is a prophage-carried type III effector protein of attaching and effacing pathogens that modulates infection dynamics*. Infect Immun, 2005. **73**(2): p. 679-86.
  366. Bateman, A., A.G. Murzin, and S.A. Teichmann, *Structure and distribution of pentapeptide repeats in bacteria*. Protein Sci, 1998. **7**(6): p. 1477-80.
  367. Hegde, S.S., et al., *A fluoroquinolone resistance protein from Mycobacterium tuberculosis that mimics DNA*. Science, 2005. **308**(5727): p. 1480-3.
  368. Henry, T., et al., *The Salmonella effector protein PipB2 is a linker for kinesin-I*. Proc Natl Acad Sci U S A, 2006. **103**(36): p. 13497-502.
  369. Zhang, Y., et al., *Recognition and ubiquitination of Salmonella type III effector SopA by a ubiquitin E3 ligase, HsRMA1*. J Biol Chem, 2005. **280**(46): p. 38682-8.
  370. Kajava, A.V. and B. Kobe, *Assessment of the ability to model proteins with leucine-rich repeats in light of the latest structural information*. Protein Sci, 2002. **11**(5): p. 1082-90.
  371. Kobe, B. and J. Deisenhofer, *Proteins with leucine-rich repeats*. Curr Opin Struct Biol, 1995. **5**(3): p. 409-16.
  372. Kobe, B. and A.V. Kajava, *The leucine-rich repeat as a protein recognition motif*. Curr Opin Struct Biol, 2001. **11**(6): p. 725-32.
  373. Kerschen, E.J., et al., *The plague virulence protein YopM targets the innate immune response by causing a global depletion of NK cells*. Infect Immun, 2004. **72**(8): p. 4589-602.
  374. Bork, P., *Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally?* Proteins, 1993. **17**(4): p. 363-74.
  375. Nataro, J.P., et al., *Identification and cloning of a novel plasmid-encoded enterotoxin of enteroinvasive Escherichia coli and Shigella strains*. Infect Immun, 1995. **63**(12): p. 4721-8.
  376. Letunic, I., et al., *SMART 4.0: towards genomic data integration*. Nucleic Acids Res, 2004. **32**(Database issue): p. D142-4.
  377. Miao, E.A. and S.I. Miller, *A conserved amino acid sequence directing intracellular type III secretion by Salmonella typhimurium*. Proc Natl Acad Sci U S A, 2000. **97**(13): p. 7539-44.
  378. Brumell, J.H., et al., *SopD2 is a novel type III secreted effector of Salmonella typhimurium that targets late endocytic compartments upon delivery into host cells*. Traffic, 2003. **4**(1): p. 36-48.
  379. Pallen, M.J., *The ESAT-6/WXG100 superfamily -- and a new Gram-positive secretion system?* Trends Microbiol, 2002. **10**(5): p. 209-12.
  380. Devos, D. and A. Valencia, *Intrinsic errors in genome annotation*. Trends Genet, 2001. **17**(8): p. 429-31.

381. Bork, P. and E.V. Koonin, *Predicting functions from protein sequences--where are the bottlenecks?* Nat Genet, 1998. **18**(4): p. 313-8.
382. Muller, A., R.M. MacCallum, and M.J. Sternberg, *Benchmarking PSI-BLAST in genome annotation.* J Mol Biol, 1999. **293**(5): p. 1257-71.
383. Valencia, A., *Automatic annotation of protein function.* Curr Opin Struct Biol, 2005. **15**(3): p. 267-74.
384. Freiberg, C., et al., *Molecular basis of symbiosis between Rhizobium and legumes.* Nature, 1997. **387**(6631): p. 394-401.
385. Pallen, M.J., S.A. Beatson, and C.M. Bailey, *Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective.* FEMS Microbiol Rev, 2005. **29**(2): p. 201-29.
386. Heidelberg, J.F., et al., *The genome sequence of the anaerobic, sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough.* Nat Biotechnol, 2004. **22**(5): p. 554-9.
387. Jeong, H., et al., *Genomic blueprint of Hahella chejuensis, a marine microbe producing an algicidal agent.* Nucleic Acids Res, 2005. **33**(22): p. 7066-73.
388. Yuk, M.H., E.T. Harvill, and J.F. Miller, *The BvgAS virulence control system regulates type III secretion in Bordetella bronchiseptica.* Mol Microbiol, 1998. **28**(5): p. 945-59.
389. Bock, A. and R. Gross, *The BvgAS two-component system of Bordetella spp.: a versatile modulator of virulence gene expression.* Int J Med Microbiol, 2001. **291**(2): p. 119-30.
390. Uhl, M.A. and J.F. Miller, *Autophosphorylation and phosphotransfer in the Bordetella pertussis BvgAS signal transduction cascade.* Proc Natl Acad Sci U S A, 1994. **91**(3): p. 1163-7.
391. Mattoo, S., et al., *Regulation of type III secretion in Bordetella.* Mol Microbiol, 2004. **52**(4): p. 1201-14.
392. Kozak, N.A., et al., *Interactions between partner switcher orthologs BtrW and BtrV regulate type III secretion in Bordetella.* J Bacteriol, 2005. **187**(16): p. 5665-76.
393. Iriarte, M. and G.R. Cornelis, *Identification of SycN, YscX, and YscY, three new elements of the Yersinia yop virulon.* J Bacteriol, 1999. **181**(2): p. 675-80.
394. Day, J.B. and G.V. Plano, *The Yersinia pestis YscY protein directly binds YscX, a secreted component of the type III secretion machinery.* J Bacteriol, 2000. **182**(7): p. 1834-43.
395. Brome, J.E., et al., *Mapping of a YscY binding domain within the LcrH chaperone that is required for regulation of Yersinia type III secretion.* J Bacteriol, 2005. **187**(22): p. 7738-52.
396. Yuk, M.H., et al., *Modulation of host immune responses, induction of apoptosis and inhibition of NF-kappaB activation by the Bordetella type III secretion system.* Mol Microbiol, 2000. **35**(5): p. 991-1004.
397. Fennelly, N.K., et al., *Bordetella pertussis expresses a functional type III secretion system that subverts protective innate and adaptive immune responses.* Infect Immun, 2008. **76**(3): p. 1257-66.
398. Viprey, V., et al., *Symbiotic implications of type III protein secretion machinery in Rhizobium.* Mol Microbiol, 1998. **28**(6): p. 1381-9.
399. Winstanley, C., B.A. Hales, and C.A. Hart, *Evidence for the presence in Burkholderia pseudomallei of a type III secretion system-associated gene cluster.* J Med Microbiol, 1999. **48**(7): p. 649-56.
400. Rosqvist, R., et al., *Functional conservation of the secretion and translocation*

- machinery for virulence proteins of yersiniae, salmonellae and shigellae.* Embo J, 1995. **14**(17): p. 4187-95.
401. Peters, J., et al., *Type III secretion a la Chlamydia.* Trends Microbiol, 2007. **15**(6): p. 241-51.
  402. Neyt, C. and G.R. Cornelis, *Insertion of a Yop translocation pore into the macrophage plasma membrane by Yersinia enterocolitica: requirement for translocators YopB and YopD, but not LcrG.* Mol Microbiol, 1999. **33**(5): p. 971-81.
  403. Behe, M.J., *Darwin's black box : the biochemical challenge to evolution.* 1996, New York ; London: The Free Press. xii, 307p.
  404. Behe, M.J., *The Challenge of Irreducible Complexity.* Natural History, 2002. **111**.
  405. Dembski, W.A., *No free lunch : why specified complexity cannot be purchased without intelligence.* 2002, Lanham, Md. ; Oxford: Rowman & Littlefield. 336p.
  406. Kamada, T. and S. Kawai, *An algorithm for drawing general undirected graphs.* 1989, Elsevier North-Holland, Inc. p. 7-15.
  407. GraphViz. Available from: <http://www.graphviz.org/>.
  408. Deane, J.E., et al., *Structures of the Shigella flexneri type 3 secretion system protein MxiC reveal conformational variability amongst homologues.* J Mol Biol, 2008. **377**(4): p. 985-92.
  409. Lee, J., et al., *HrpZ(PspH) from the plant pathogen Pseudomonas syringae pv. phaseolicola binds to lipid bilayers and forms an ion-conducting pore in vitro.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 289-94.
  410. Schechter, L.M., S.M. Damrauer, and C.A. Lee, *Two AraC/XylS family members can independently counteract the effect of repressing sequences upstream of the hilA promoter.* Mol Microbiol, 1999. **32**(3): p. 629-42.
  411. Eichelberg, K. and J.E. Galan, *Differential regulation of Salmonella typhimurium type III secreted proteins by pathogenicity island 1 (SPI-1)-encoded transcriptional activators InvF and hilA.* Infect Immun, 1999. **67**(8): p. 4099-105.
  412. Jones, D.T. and M.B. Swindells, *Getting the most from PSI-BLAST.* Trends Biochem Sci, 2002. **27**(3): p. 161-4.
  413. Mota, L.J., et al., *Bacterial injectisomes: needle length does matter.* Science, 2005. **307**(5713): p. 1278.
  414. Sekiya, K., et al., *Supramolecular structure of the enteropathogenic Escherichia coli type III secretion system and its direct interaction with the EspA-sheath-like structure.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11638-43.
  415. Tamano, K., et al., *Supramolecular structure of the Shigella type III secretion machinery: the needle part is changeable in length and essential for delivery of effectors.* Embo J, 2000. **19**(15): p. 3876-87.

## Appendix 1 – Complete list of all genomes containing a T3SS

| <i>Bacterium</i>                                    | <i>Genome</i>           | <i>Accession</i>     | <i>Type</i> |
|---|-------------------------|----------------------|-------------|
| Acidobacteria bacterium Ellin345                    | Chromosome              | CP000360             | Flagellar   |
| Acidothermus cellulolyticus 11B                     | Chromosome              | CP000481             | Flagellar   |
| Acidovorax avenae subsp. citrulli AAC00-1           | Chromosome              | CP000512             | Multiple    |
| Acidovorax sp. JS42                                 | Chromosome              | CP000539             | Flagellar   |
| Aeromonas hydrophila subsp. hydrophila ATCC 7966    | Chromosome              | CP000462             | Flagellar   |
| Aeromonas salmonicida subsp. salmonicida A449       | Chromosome<br>Plasmid 5 | CP000644<br>CP000646 | Multiple    |
| Agrobacterium tumefaciens str. C58                  | Chromosome              | AE007869             | Flagellar   |
| Alkalilimnicola ehrlichei MLHE-1                    | Chromosome              | CP000453             | Flagellar   |
| Anaeromyxobacter dehalogenans 2CP-C                 | Chromosome              | CP000251             | Multiple    |
| Aquifex aeolicus VF5                                | Chromosome              | AE000657             | Flagellar   |
| Azoarcus sp. BH72                                   | Chromosome              | AM406670             | Flagellar   |
| Azoarcus sp. EbN1                                   | Chromosome              | CR555306             | Flagellar   |
| Bacillus anthracis str. 'Ames Ancestor'             | Chromosome              | AE017334             | Flagellar   |
| Bacillus anthracis str. Ames                        | Chromosome              | AE016879             | Flagellar   |
| Bacillus anthracis str. Sterne                      | Chromosome              | AE017225             | Flagellar   |
| Bacillus cereus ATCC 10987                          | Chromosome              | AE017194             | Flagellar   |
| Bacillus cereus ATCC 14579                          | Chromosome              | AE016877             | Flagellar   |
| Bacillus cereus E33L                                | Chromosome              | CP000001             | Flagellar   |
| Bacillus clausii KSM-K16                            | Chromosome              | AP006627             | Flagellar   |
| Bacillus halodurans C-125                           | Chromosome              | BA000004             | Flagellar   |
| Bacillus licheniformis ATCC 14580                   | Chromosome              | AE017333             | Flagellar   |
| Bacillus subtilis subsp. subtilis str. 168          | Chromosome              | AL009126             | Flagellar   |
| Bacillus thuringiensis serovar konkukian str. 97-27 | Chromosome              | AE017355             | Flagellar   |
| Bacillus thuringiensis str. Al Hakam                | Chromosome              | CP000485             | Flagellar   |
| Bartonella bacilliformis KC583                      | Chromosome              | CP000524             | Flagellar   |
| Bdellovibrio bacteriovorus HD100                    | Chromosome              | BX842601             | Flagellar   |
| Bordetella bronchiseptica RB50                      | Chromosome              | BX470250             | Multiple    |
| Bordetella parapertussis 12822                      | Chromosome              | BX470249             | Multiple    |
| Bordetella pertussis Tohama I                       | Chromosome              | BX470248             | Multiple    |
| Borrelia afzelii PKo                                | Chromosome              | CP000395             | Flagellar   |
| Borrelia burgdorferi B31                            | Chromosome              | AE000783             | Flagellar   |
| Borrelia garinii PBi                                | Chromosome linear       | CP000013             | Flagellar   |
| Bradyrhizobium japonicum USDA 110                   | Chromosome              | BA000040             | Multiple    |
| Brucella abortus biovar 1 str. 9-941                | Chromosome 2            | AE017224             | Flagellar   |
| Brucella melitensis 16M                             | Chromosome 2            | AE008918             | Flagellar   |
| Brucella melitensis biovar Abortus 2308             | Chromosome 2            | AM040265             | Flagellar   |

|   |              |          |               |
|---|--------------|----------|---------------|
| Brucella suis 1330                                | Chromosome 2 | AE014292 | Flagellar     |
| Buchnera aphidicola str. APS (Acyrtosiphon pisum) | Chromosome   | BA000003 | Flagellar     |
| Buchnera aphidicola str. Bp (Baizongia pistaciae) | Chromosome   | AE016826 | Flagellar     |
| Buchnera aphidicola str. Cc (Cinara cedri)        | Chromosome   | CP000263 | Flagellar     |
| Buchnera aphidicola str. Sg (Schizaphis graminum) | Chromosome   | AE013218 | Flagellar     |
| Burkholderia cenocepacia AU 1054                  | Chromosome 1 | CP000378 | Multiple      |
|   | Chromosome 2 | CP000379 |               |
| Burkholderia cenocepacia HI2424                   | Chromosome 1 | CP000458 | Multiple      |
|   | Chromosome 2 | CP000459 |               |
| Burkholderia cepacia AMMD                         | Chromosome 1 | CP000440 | Multiple      |
|   | Chromosome 2 | CP000441 |               |
|   | Chromosome 3 | CP000442 |               |
| Burkholderia mallei ATCC 23344                    | Chromosome 1 | CP000010 | Multiple      |
|   | Chromosome 2 | CP000011 |               |
| Burkholderia mallei NCTC 10229                    | Chromosome 1 | CP000545 | Multiple      |
|   | Chromosome 2 | CP000546 |               |
| Burkholderia mallei NCTC 10247                    | Chromosome 1 | CP000547 | Multiple      |
|   | Chromosome 2 | CP000548 |               |
| Burkholderia mallei SAVP1                         | Chromosome 2 | CP000526 | Flagellar     |
| Burkholderia pseudomallei 1106a                   | Chromosome 1 | CP000572 | Multiple      |
|   | Chromosome 2 | CP000573 |               |
| Burkholderia pseudomallei 1710b                   | Chromosome 1 | CP000124 | Multiple      |
|   | Chromosome 2 | CP000125 |               |
| Burkholderia pseudomallei 668                     | Chromosome 1 | CP000570 | Multiple      |
|   | Chromosome 2 | CP000571 |               |
| Burkholderia pseudomallei K96243                  | Chromosome 1 | BX571965 | Multiple      |
|   | Chromosome 2 | BX571966 |               |
| Burkholderia sp. 383                              | Chromosome 1 | CP000151 | Flagellar     |
| Burkholderia thailandensis E264                   | Chromosome 1 | CP000085 | Multiple      |
|   | Chromosome 2 | CP000086 |               |
| Burkholderia vietnamiensis G4                     | Chromosome 1 | CP000614 | Multiple      |
|   | Chromosome 2 | CP000615 |               |
| Burkholderia xenovorans LB400                     | Chromosome 1 | CP000270 | Multiple      |
| Campylobacter fetus subsp. fetus 82-40            | Chromosome   | CP000487 | Flagellar     |
| Campylobacter jejuni RM1221                       | Chromosome   | CP000025 | Flagellar     |
| Campylobacter jejuni subsp. jejuni 81-176         | Chromosome   | CP000538 | Flagellar     |
| Campylobacter jejuni subsp. jejuni NCTC 11168     | Chromosome   | AL111168 | Flagellar     |
| Candidatus Protochlamydia amoebophila UWE25       | Chromosome   | BX908798 | Non-Flagellar |
| Carboxydotherrmus hydrogenoformans Z-2901         | Chromosome   | CP000141 | Flagellar     |
| Caulobacter crescentus CB15                       | Chromosome   | AE005673 | Flagellar     |
| Chlamydia muridarum Nigg                          | Chromosome   | AE002160 | Non-Flagellar |
| Chlamydia trachomatis A/HAR-13                    | Chromosome   | CP000051 | Non-Flagellar |



|   |                 |          |               |
|---|-----------------|----------|---------------|
| <i>Chlamydia trachomatis</i> D/UW-3/CX                                  | Chromosome      | AE001273 | Non-Flagellar |
| <i>Chlamydomonas abortus</i> S26/3                                      | Chromosome      | CR848038 | Non-Flagellar |
| <i>Chlamydomonas caviae</i> GPIC  | Chromosome      | AE015925 | Non-Flagellar |
| <i>Chlamydomonas felis</i> Fe/C-56                                      | Chromosome      | AP006861 | Non-Flagellar |
| <i>Chlamydomonas pneumoniae</i> AR39                                    | Chromosome      | AE002161 | Non-Flagellar |
| <i>Chlamydomonas pneumoniae</i> CWL029                                  | Chromosome      | AE001363 | Non-Flagellar |
| <i>Chlamydomonas pneumoniae</i> J138                                    | Chromosome      | BA000008 | Non-Flagellar |
| <i>Chlamydomonas pneumoniae</i> TW-183                                  | Chromosome      | AE009440 | Non-Flagellar |
| <i>Chromobacterium violaceum</i> ATCC 12472                             | Chromosome      | AE016825 | Multiple      |
| <i>Chromohalobacter salexigens</i> DSM 3043                             | Chromosome      | CP000285 | Flagellar     |
| <i>Clostridium acetobutylicum</i> ATCC 824                              | Chromosome      | AE001437 | Flagellar     |
| <i>Clostridium difficile</i> 630  | Chromosome      | AM180355 | Flagellar     |
| <i>Clostridium novyi</i> NT   | Chromosome      | CP000382 | Flagellar     |
| <i>Clostridium tetani</i> E88   | Chromosome      | AE015927 | Flagellar     |
| <i>Clostridium thermocellum</i> ATCC 27405                              | Chromosome      | CP000568 | Flagellar     |
| <i>Colwellia psychrerythraea</i> 34H                                    | Chromosome      | CP000083 | Flagellar     |
| <i>Dechloromonas aromatica</i> RCB                                      | Chromosome      | CP000089 | Flagellar     |
| <i>Desulfotobacterium hafniense</i> Y51                                 | Chromosome      | AP008230 | Flagellar     |
| <i>Desulfotalea psychrophila</i> LSv54                                  | Chromosome      | CR522870 | Flagellar     |
| <i>Desulfotomaculum reducens</i> MI-1                                   | Chromosome      | CP000612 | Flagellar     |
| <i>Desulfovibrio desulfuricans</i> G20                                  | Chromosome      | CP000112 | Flagellar     |
| <i>Desulfovibrio vulgaris</i> subsp. <i>vulgaris</i> DP4                | Chromosome      | CP000527 | Multiple      |
|   | Plasmid pDVUL01 | CP000528 |               |
| <i>Desulfovibrio vulgaris</i> subsp. <i>vulgaris</i> str. Hildenborough | Chromosome      | AE017285 | Multiple      |
|   | Plasmid pDV     | AE017286 |               |
| <i>Erwinia carotovora</i> subsp. <i>atroseptica</i> SCRI1043            | Chromosome      | BX950851 | Multiple      |
| <i>Escherichia coli</i> 536   | Chromosome      | CP000247 | Flagellar     |
| <i>Escherichia coli</i> APEC O1   | Chromosome      | CP000468 | Flagellar     |
| <i>Escherichia coli</i> CFT073  | Chromosome      | AE014075 | Flagellar     |
| <i>Escherichia coli</i> K12   | Chromosome      | U00096   | Flagellar     |
| <i>Escherichia coli</i> O157 (EDL933)                                   | Chromosome      | AE005174 | Multiple      |
| <i>Escherichia coli</i> O157 (Sakai)                                    | Chromosome      | BA000007 | Multiple      |
| <i>Escherichia coli</i> UTI89   | Chromosome      | CP000243 | Flagellar     |
| <i>Escherichia coli</i> W3110   | Chromosome      | AP009048 | Flagellar     |
| <i>Geobacillus kaustophilus</i> HTA426                                  | Chromosome      | BA000043 | Flagellar     |
| <i>Geobacillus thermodenitrificans</i> NG80-2                           | Chromosome      | CP000557 | Flagellar     |
| <i>Geobacter metallireducens</i> GS-15                                  | Chromosome      | CP000148 | Flagellar     |
| <i>Geobacter sulfurreducens</i> PCA                                     | Chromosome      | AE017180 | Flagellar     |
| <i>Gluconobacter oxydans</i> 621H                                       | Chromosome      | CP000009 | Flagellar     |
| <i>Hahella chejuensis</i> KCTC 2396                                     | Chromosome      | CP000155 | Multiple      |
| <i>Halorhodospira halophila</i> SL1                                     | Chromosome      | CP000544 | Flagellar     |
| <i>Helicobacter acinonychis</i> str. Sheeba                             | Chromosome      | AM260522 | Flagellar     |

|   |              |          |           |
|---|--------------|----------|-----------|
| <i>Helicobacter hepaticus</i> ATCC 51449                                    | Chromosome   | AE017125 | Flagellar |
| <i>Helicobacter pylori</i> 26695  | Chromosome   | AE000511 | Flagellar |
| <i>Helicobacter pylori</i> HPAG1  | Chromosome   | CP000241 | Flagellar |
| <i>Helicobacter pylori</i> J99  | Chromosome   | AE001439 | Flagellar |
| <i>Herminiimonas arsenicoxydans</i>   | Chromosome   | CU207211 | Flagellar |
| <i>Hyphomonas neptunium</i> ATCC 15444                                      | Chromosome   | CP000158 | Flagellar |
| <i>Idiomarina loihiensis</i> L2TR   | Chromosome   | AE017340 | Flagellar |
| <i>Jannaschia</i> sp. CCS1  | Chromosome   | CP000264 | Flagellar |
| <i>Lawsonia intracellularis</i> PHE/MN1-00                                  | Chromosome   | AM180252 | Multiple  |
| <i>Legionella pneumophila</i> str. Lens                                     | Chromosome   | CR628337 | Flagellar |
| <i>Legionella pneumophila</i> str. Paris                                    | Chromosome   | CR628336 | Flagellar |
| <i>Legionella pneumophila</i> subsp. <i>pneumophila</i> str. Philadelphia 1 | Chromosome   | AE017354 | Flagellar |
| <i>Leptospira borgpetersenii</i> serovar Hardjo-bovis JB197                 | Chromosome 1 | CP000350 | Flagellar |
| <i>Leptospira borgpetersenii</i> serovar Hardjo-bovis L550                  | Chromosome 1 | CP000348 | Flagellar |
| <i>Leptospira interrogans</i> serovar Copenhageni str. Fiocruz L1-130       | Chromosome 1 | AE016823 | Flagellar |
| <i>Leptospira interrogans</i> serovar Lai str. 56601                        | Chromosome 1 | AE010300 | Flagellar |
| <i>Listeria innocua</i> Clip11262   | Chromosome   | AL592022 | Flagellar |
| <i>Listeria monocytogenes</i> EGD-e   | Chromosome   | AL591824 | Flagellar |
| <i>Listeria monocytogenes</i> str. 4b F2365                                 | Chromosome   | AE017262 | Flagellar |
| <i>Listeria welshimeri</i> serovar 6b str. SLCC5334                         | Chromosome   | AM263198 | Flagellar |
| <i>Magnetococcus</i> sp. MC-1   | Chromosome   | CP000471 | Flagellar |
| <i>Magnetospirillum magneticum</i> AMB-1                                    | Chromosome   | AP007255 | Flagellar |
| <i>Maricaulis maris</i> MCS10   | Chromosome   | CP000449 | Flagellar |
| <i>Marinobacter aquaeolei</i> VT8   | Chromosome   | CP000514 | Flagellar |
| <i>Mesorhizobium loti</i> MAFF303099  | Chromosome   | BA000012 | Multiple  |
| <i>Mesorhizobium</i> sp. BNC1   | Chromosome   | CP000390 | Multiple  |
| <i>Methylobium petroleiphilum</i> PM1                                       | Chromosome   | CP000555 | Flagellar |
| <i>Methylobacillus flagellatus</i> KT                                       | Chromosome   | CP000284 | Flagellar |
| <i>Moorella thermoacetica</i> ATCC 39073                                    | Chromosome   | CP000232 | Flagellar |
| <i>Myxococcus xanthus</i> DK 1622   | Chromosome   | CP000113 | Multiple  |
| <i>Nitrobacter hamburgensis</i> X14   | Chromosome   | CP000319 | Flagellar |
| <i>Nitrobacter winogradskyi</i> Nb-255                                      | Chromosome   | CP000115 | Flagellar |
| <i>Nitrosococcus oceani</i> ATCC 19707                                      | Chromosome   | CP000127 | Flagellar |
| <i>Nitrosomonas europaea</i> ATCC 19718                                     | Chromosome   | AL954747 | Flagellar |
| <i>Nitrosomonas eutropha</i> C91  | Chromosome   | CP000450 | Flagellar |
| <i>Nitrospira multiformis</i> ATCC 25196                                    | Chromosome 1 | CP000103 | Flagellar |
| <i>Nocardioides</i> sp. JS614   | Chromosome   | CP000509 | Flagellar |
| <i>Oceanobacillus iheyensis</i> HTE831                                      | Chromosome   | BA000028 | Flagellar |

|   |                        |          |           |
|---|------------------------|----------|-----------|
| Paracoccus denitrificans PD1222                                       | Chromosome 1           | CP000489 | Flagellar |
| Pelobacter carbinolicus DSM 2380                                      | Chromosome             | CP000142 | Flagellar |
| Pelobacter propionicus DSM 2379                                       | Chromosome             | CP000482 | Flagellar |
| Photobacterium profundum SS9  | Chromosome 1           | CR354531 | Flagellar |
| Photorhabdus luminescens subsp. laumondii TTO1                        | Chromosome             | BX470251 | Multiple  |
| Pseudoalteromonas atlantica T6c                                       | Chromosome             | CP000388 | Flagellar |
| Pseudoalteromonas haloplanktis TAC125                                 | Chromosome 1           | CR954246 | Flagellar |
| Pseudomonas   | Chromosome             | CT573326 | Flagellar |
| Pseudomonas aeruginosa PAO1   | Chromosome             | AE004091 | Multiple  |
| Pseudomonas aeruginosa UCBPP-PA14                                     | Chromosome             | CP000438 | Multiple  |
| Pseudomonas fluorescens Pf-5  | Chromosome             | CP000076 | Flagellar |
| Pseudomonas fluorescens PfO-1   | Chromosome             | CP000094 | Flagellar |
| Pseudomonas putida KT2440   | Chromosome             | AE015451 | Flagellar |
| Pseudomonas syringae pv. phaseolicola 1448A                           | Chromosome             | CP000058 | Multiple  |
| Pseudomonas syringae pv. syringae B728a                               | Chromosome             | CP000075 | Multiple  |
| Pseudomonas syringae pv. tomato str. DC3000                           | Chromosome             | AE016853 | Multiple  |
| Psychromonas ingrahamii 37  | Chromosome             | CP000510 | Flagellar |
| Ralstonia eutropha H16  | Chromosome 2           | AM260480 | Flagellar |
| Ralstonia eutropha JMP134   | Chromosome 2           | CP000091 | Flagellar |
| Ralstonia metallidurans CH34  | Plasmid<br>megaplasmid | CP000353 | Flagellar |
| Ralstonia solanacearum  | Chromosome             | AL646053 | Multiple  |
| Rhizobium etli CFN 42   | Chromosome             | CP000133 | Flagellar |
| Rhizobium leguminosarum bv. viciae 3841                               | Chromosome             | AM236080 | Flagellar |
| Rhodobacter sphaeroides 2.4.1   | Chromosome 1           | CP000143 | Flagellar |
| Rhodobacter sphaeroides ATCC 17029                                    | Chromosome 1           | CP000577 | Flagellar |
| Rhodoferrax ferrireducens T118  | Chromosome             | CP000267 | Flagellar |
| Rhodopirellula baltica SH 1   | Chromosome             | BX119912 | Flagellar |
| Rhodopseudomonas palustris BisA53                                     | Chromosome             | CP000463 | Flagellar |
| Rhodopseudomonas palustris BisB18                                     | Chromosome             | CP000301 | Flagellar |
| Rhodopseudomonas palustris BisB5                                      | Chromosome             | CP000283 | Flagellar |
| Rhodopseudomonas palustris CGA009                                     | Chromosome             | BX571963 | Flagellar |
| Rhodopseudomonas palustris HaA2                                       | Chromosome             | CP000250 | Flagellar |
| Rhodospirillum rubrum ATCC 11170                                      | Chromosome             | CP000230 | Flagellar |
| Roseobacter denitrificans OCh 114                                     | Chromosome             | CP000362 | Flagellar |
| Saccharophagus degradans 2-40   | Chromosome             | CP000282 | Flagellar |
| Salinibacter ruber DSM 13855  | Chromosome             | CP000159 | Flagellar |
| Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67  | Chromosome             | AE017220 | Multiple  |
| Salmonella enterica subsp. enterica serovar Paratyphi Astr. ATCC 9150 | Chromosome             | CP000026 | Multiple  |

|   |                                |                      |           |
|---|--------------------------------|----------------------|-----------|
| Salmonella enterica subsp. enterica serovar Typhi str. CT18 | Chromosome                     | AL513382             | Multiple  |
| Salmonella enterica subsp. enterica serovar Typhi Ty2       | Chromosome                     | AE014613             | Multiple  |
| Salmonella typhimurium LT2                                  | Chromosome                     | AE006468             | Multiple  |
| Shewanella amazonensis SB2B                                 | Chromosome                     | CP000507             | Flagellar |
| Shewanella baltica OS155                                    | Chromosome                     | CP000563             | Flagellar |
| Shewanella denitrificans OS217                              | Chromosome                     | CP000302             | Flagellar |
| Shewanella frigidimarina NCIMB 400                          | Chromosome                     | CP000447             | Flagellar |
| Shewanella loihica PV-4                                     | Chromosome                     | CP000606             | Flagellar |
| Shewanella oneidensis MR-1                                  | Chromosome                     | AE014299             | Flagellar |
| Shewanella sp. ANA-3  | Chromosome 1                   | CP000469             | Flagellar |
| Shewanella sp. MR-4   | Chromosome                     | CP000446             | Flagellar |
| Shewanella sp. MR-7   | Chromosome                     | CP000444             | Flagellar |
| Shewanella sp. W3-18-1                                      | Chromosome                     | CP000503             | Flagellar |
| Shigella boydii Sb227                                       | Chromosome                     | CP000036             | Flagellar |
| Shigella dysenteriae Sd197                                  | Chromosome<br>Plasmid pSD1_197 | CP000034<br>CP000035 | Multiple  |
| Shigella flexneri 2a str. 2457T                             | Chromosome                     | AE014073             | Flagellar |
| Shigella flexneri 2a str. 301                               | Chromosome<br>Plasmid pCP301   | AE005674<br>AF386526 | Multiple  |
| Shigella flexneri 5 str. 8401                               | Chromosome                     | CP000266             | Flagellar |
| Shigella sonnei Ss046                                       | Chromosome<br>Plasmid pSS_046  | CP000038<br>CP000039 | Multiple  |
| Silicibacter pomeroyi DSS-3                                 | Chromosome                     | CP000031             | Flagellar |
| Silicibacter sp. TM1040                                     | Chromosome                     | CP000377             | Flagellar |
| Sinorhizobium meliloti 1021                                 | Chromosome                     | AL591688             | Flagellar |
| Sodalis glossinidius str. 'morsitans'                       | Chromosome                     | AP008232             | Multiple  |
| Solibacter usitatus Ellin6076                               | Chromosome                     | CP000473             | Flagellar |
| Sphingopyxis alaskensis RB2256                              | Chromosome                     | CP000356             | Flagellar |
| Symbiobacterium thermophilum IAM 14863                      | Chromosome                     | AP006840             | Flagellar |
| Syntrophomonas wolfei subsp. wolfei str. Goettingen         | Chromosome                     | CP000448             | Flagellar |
| Syntrophus aciditrophicus SB                                | Chromosome                     | CP000252             | Flagellar |
| Thermoanaerobacter tengcongensis MB4                        | Chromosome                     | AE008691             | Flagellar |
| Thermotoga maritima MSB8                                    | Chromosome                     | AE000512             | Flagellar |
| Thiobacillus denitrificans ATCC 25259                       | Chromosome                     | CP000116             | Flagellar |
| Thiomicrospira crunogena XCL-2                              | Chromosome                     | CP000109             | Flagellar |
| Thiomicrospira denitrificans ATCC 33889                     | Chromosome                     | CP000153             | Flagellar |
| Treponema denticola ATCC 35405                              | Chromosome                     | AE017226             | Flagellar |
| Treponema pallidum subsp. pallidum str. Nichols             | Chromosome                     | AE000520             | Flagellar |
| Verminephrobacter eiseniae EF01-2                           | Chromosome                     | CP000542             | Flagellar |
| Vibrio cholerae O1 biovar eltor str.                        | Chromosome 1                   | AE003852             | Flagellar |

|   |                  |          |           |
|---|------------------|----------|-----------|
| N16961  |                  |          |           |
| <i>Vibrio fischeri</i> ES114  | Chromosome 1     | CP000020 | Flagellar |
| <i>Vibrio parahaemolyticus</i> RIMD 2210633                                   | Chromosome 1     | BA000031 | Multiple  |
|   | Chromosome 2     | BA000032 |           |
| <i>Vibrio vulnificus</i> CMCP6  | Chromosome 1     | AE016795 | Flagellar |
| <i>Vibrio vulnificus</i> YJ016  | Chromosome 1     | BA000037 | Flagellar |
| <i>Wigglesworthia glossinidia</i> endosymbiont of <i>Glossina brevipalpis</i> | Chromosome       | BA000021 | Flagellar |
| <i>Wolinella succinogenes</i> DSM 1740  | Chromosome       | BX571656 | Flagellar |
| <i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306                       | Chromosome       | AE008923 | Multiple  |
| <i>Xanthomonas campestris</i> pv. <i>campestris</i> str. 8004                 | Chromosome       | CP000050 | Multiple  |
| <i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC 33913           | Chromosome       | AE008922 | Multiple  |
| <i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> str. 85-10               | Chromosome       | AM039952 | Multiple  |
| <i>Xanthomonas oryzae</i> pv. <i>oryzae</i> KACC10331                         | Chromosome       | AE013598 | Multiple  |
| <i>Xanthomonas oryzae</i> pv. <i>oryzae</i> MAFF 311018                       | Chromosome       | AP008229 | Multiple  |
| <i>Yersinia enterocolitica</i> subsp. <i>enterocolitica</i> 8081              | Chromosome       | AM286415 | Multiple  |
|   | Plasmid pYVe8081 | AM286416 |           |
| <i>Yersinia pestis</i> Antiqua  | Chromosome       | CP000308 | Multiple  |
|   | Plasmid pCD      | CP000311 |           |
| <i>Yersinia pestis</i> biovar <i>Microtus</i> str. 91001                      | Chromosome       | AE017042 | Multiple  |
|   | Plasmid pCD1     | AE017043 |           |
| <i>Yersinia pestis</i> CO92   | Chromosome       | AL590842 | Multiple  |
|   | Plasmid pCD1     | AL117189 |           |
| <i>Yersinia pestis</i> KIM  | Chromosome       | AE009952 | Multiple  |
| <i>Yersinia pestis</i> Nepal516   | Chromosome       | CP000305 | Multiple  |
| <i>Yersinia pestis</i> Pestoides F  | Chromosome       | CP000668 | Multiple  |
|   | Plasmid CD       | CP000669 |           |
| <i>Yersinia pseudotuberculosis</i> IP 32953                                   | Chromosome       | BX936398 | Multiple  |
|   | Plasmid pYV      | BX936399 |           |
| <i>Zymomonas mobilis</i> subsp. <i>mobilis</i> ZM4                            | Chromosome       | AE008692 | Flagellar |

**Supplementary Table 1. List of genomes with T3SS present**

Systems detected through HMMER searches using flagellar and non-flagellar specific domains. Rows coloured by type: Blue: Flagellar T3SS found only, Red: Non-Flagellar T3SS found only, Orange: Flagellar and Non-Flagellar T3SSs found

## **Appendix 2 – Locus finding approaches applied to type-VI secretion systems**

As part of work undertaken by our lab group on the field of type-VI secretion systems (T6SSs) a survey of the number of bacteria containing a T6SS was carried out using the same locus finder software as was used in the T3SS finder in Conserved and Specific features of T3S systems. In this survey a series of domains were chosen based on HMMER searches of the T6SS of *Vibrio cholerae*. The domains found within this T6SS which were then used within the locus finder software were: DUF1305, DUF770, DUF876, DUF877, DUF879, ImcF-related and ImpA-rel\_N (PFAM accessions: PF06996, PF05591, PF05936, PF05943, PF05947, PF06761 and PF06812 respectively). These domains were then clustered using the same 50Kb cut-off as was used for clustering T3SSs, and the results collated to produce a comprehensive list of bacteria containing type-VI secretion systems. This data was used as part of a published review on the field, and this review follows in this appendix.