AUTHOR:	**Joseph Richard John Healey**
DEGREE:	**Mathematical Biology and Biophysical Chemistry**
TITLE:	***Photorhabdus* Virulence Cassettes: understanding the structure, and genomic role, of a novel bacterial toxin secretion weapon**
DATE OF DEPOSIT:	**May, 2018**

I **agree** that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I **agree** that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

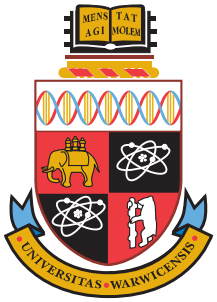
> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: .....

USER DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE	SIGNATURE	ADDRESS
.....	.....	.....
.....	.....	.....
.....	.....	.....
.....	.....	.....
.....	.....	.....



Photorhabdus Virulence Cassettes: understanding the structure, and genomic role, of a novel bacterial toxin secretion weapon

by

Joseph Richard John Healey

Doctoral Thesis
Submitted to the University of Warwick
for the degree of
Doctor of Philosophy

Supervisors: Dr. Nicholas R. Waterfield and Prof. Matthew I. Gibson

MOAC CDT in collaboration with:
Warwick Medical School and Warwick Chemistry Department

THE UNIVERSITY OF WARWICK

Contents

List of Figures

List of Tables

List of Equations

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The author previously submitted a Masters thesis, entitled "*Photorhabdus asymbiotica* as a Model Organism for Understanding Emerging Human Pathogens", for the degree of M.Sc. Mathematical Biology and Biophysical Chemistry, for a related project in 2014. While both pieces of work are the authors own, due to sharing commonality in research topic, there may be some superficial resemblance in introductory content and methods.

All work and data analysis was conducted by the author except in the cases outlined below.

Crystallography screening and diffraction testing was kindly performed with Dr. Avinash Punekar (University of Warwick).

List of data provided and/or analysis carried out by collaborators. Parts of this thesis have been published by the author: List of publications including submitted papers.

"I would like to describe a field, in which little has been done,
but in which an enormous amount can be done in principle.

This field is not quite the same as others in that it will not tell us much [...].
Furthermore, a point that is most important is that it would have an
enormous number of technical applications.

What I want to talk about is the problem of manipulating and
***controlling things on a small scale...**"*

— RICHARD P. FEYNMAN

The quotes used in the opening of this thesis and in the epigraphs of individual chapters have come from the transcript of Richard Feynman's discourse "There's plenty of room at the bottom", as reproduced in the book "Plenty of Room for Biology at the Bottom: An Introduction to Bionanotechnology" by E. Gazit and A. Mitraki

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Abbreviations

pnf	*Photorhabdus* Necrosis Factor
ATCC	American Type Culture Collection
F	Forward
LPS	lipopolysaccharide
Mbp	Mega-basepairs
NEB	New England Biolabs
PCR	Polymerase Chain Reaction
PVC	*Photorhabdus* Virulence Cassette
R	Reverse
RCF	Relative Centrifugal Force
T6SS	Type 6 Secretion System
TVISS	see T6SS
WMS	Warwick Medical School
bp	basepairs
gDNA	genomic DeoxyriboNucleic Acid
gp##	gene product ##
hcp	hemolysin co-regulated protein
kbp	kilo-basepairs
ppGpp	Guanosine pentaphosphate
qRT-PCR	quantitative Reverse-Transcriptase Polymerase Chain Reaction

Fill out full abbreviation list

Part I

Introduction & Methodology

Chapter 1

Introduction

*"So, there is **plenty** of room at the bottom!"*

Richard P. Feynman

The focus of this PhD is on an unusual protein secretion system and 'bacterial nano-scale weapon' known as the *Photorhabdus* Virulence Cassette (PVC), produced by members of the *Photorhabdus* genus. The study of the system, throughout this candidature, was primarily motivated in 2 ways. Firstly, the PVCs are of interest from a fundamental biology standpoint, given their uniqueness. A better understanding of the mechanistics of the system and its precise role in the environment could be immensely valuable to studies of virulence, microbial ecology, and structural biology, among others. Recent studies, which are discussed in detail in coming sections, are beginning to suggest a much more widespread and pervasive role for structures such as these, and therefore understanding as many of the naturally occurring variants as possible will be key.

Secondly, given the PVCs putative role as a targeted protein delivery mechanism, it stands to reason that there may be huge potential in the system for use as a first-of-its-kind bio-nanotechnological drug delivery mechanism. From the outset, the lab work conducted was intended to explore the potential for 'functionalisation' of the PVC system and its use as a biotechnological tool.

Before discussing the PVC system however, it is logical to discuss the incredibly unusual host bacterium from which it derives - *Photorhabdus*.

1.1 *Photorhabdus*

1.1.1 The *Photorhabdus* genus: the same but different

Photorhabdus describes a genus of extremely effective (primarily) insect pathogens. The prevailing literature, and even a visit to the current Wikipedia page, for *Photorhabdus*, shows that 3 species have been formally recognised within the clade - *P. luminescens*, *P. asymbiotica*, and *P. temperata*. More recently however, further species have begun to be defined, for instance, the new species *P. heterorhabditis*, has been proposed (Naidoo et al., 2015). This will likely continue, as further species/strains are isolated, and existing genomic annotation is corrected. *Photorhabdus* is, itself, only a relatively recently recognised clade, having been demarcated from the related *Xenorhabdus* in the 1990s (Saux et al., 1999; Boemare et al., 1993).

Even within the genus, a remarkable degree of diversity can be seen (and is an important recurring point in this thesis), reflected in a plethora of subspecies/strains that are recognised (Peat et al., 2010). In the case of *P. asymbiotica*, the presence of a unique plasmid (and in certain strains, more than 1 (Wilkinson et al., 2010)), and chromosomal differences with yet to be understood mechanisms, allow for infection of higher order organisms, including humans. However, this is not the case for all members of the *P. asymbiotica* clade, and there exist genotypically *P. asymbiotica* strains, which do not exhibit all the same phenotypic traits.

Upon first sequencing of the *P. luminescens* genome, 4,839 genes were predicted at a genome size of 5.69 Megabases (Duchaud et al., 2003); for *P. asymbiotica*, that number was 4,417, with a genome size of just over 5 Megabases. Despite this genome reduction, comparative genomics has shown that each species carries around a megabase of unique sequence (Wilkinson et al., 2009). Our own work (unpublished) has demonstrated that the core genome of the clade consists of some 673 chromosomal genes and, for the relevant strains, 19 plasmid genes, meaning that, a considerable amount of any given *Photorhabdus* genome is strain specific (and this number will no doubt shrink as more genomes are studied). Unsurprisingly, these stark genetic differences can manifest in substantial phenotypic differences. As mentioned, certain *P. asymbiotica* strains can infect

mammalian and human hosts, and in order to do this it must be capable of withstanding an adaptive immune system (which the normal insect hosts lack) (Lemaitre and Hoffmann, 2007), as well as the higher body temperatures of homeotherms. Insects are *poikilothermic*, meaning that their body temperatures vary considerably, in line with the environmental temperature. *P. luminescens* are unable to withstand temperatures much in excess of approximately 34 °C, whereas *P. asymbiotica* strains are viable up to roughly 38 °C. As is so often the case with biological systems however, there are exceptions to this 'rule'. Namely, European isolates which are genetically closest to *P. asymbiotica* strains, have been demonstrated to not be capable of human infection and thermotolerance, like their other *P. asymbiotica* counterparts from the USA and Australia (Peat et al., 2010; Mulley et al., 2015). Figure 1.1 below shows, in a schematic manner, the host and temperature restrictions of some exemplar strains from each species.

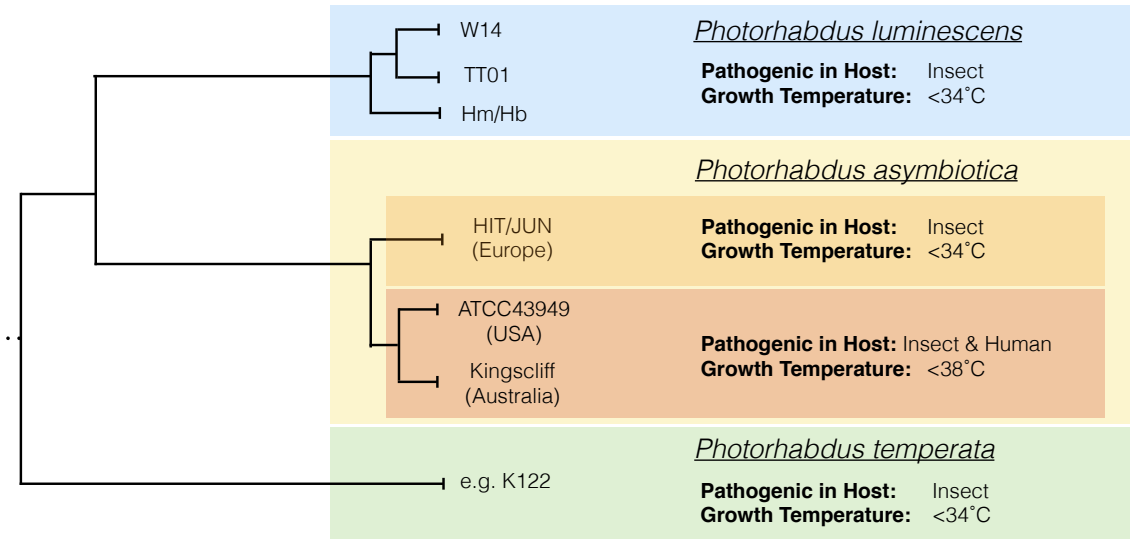


Figure 1.1 | A SCHEMATIC DIAGRAM DEPICTING THE SUBCLADES WITHIN THE *Photorhabdus* GENUS. Not drawn to scale. The schematic shows the host ranges and thermotolerance of the archetypical species within the *Photorhabdus* genus, with some exemplar strains. Adapted from (Waterfield et al., 2009), and reproduced from my own Masters thesis "*Photorhabdus asymbiotica* as a Model Organism for Understanding Emerging Human Pathogens".

These differences in genetic content and pathogenicity notwithstanding, all *Photorhabdus* strains are obligately associated with *Heterorhabditid* nematodes (see upcoming Section 1.1.3 on page 7 for a detailed explanation). Consequently, all strains must maintain the capability of associating with the host. These interactions are undoubtedly extremely complex, requiring all manner of genetic components as well as transcriptional and trans-

lational epigenetic regulation, and the molecular basis for this symbiosis is still yet to be fully understood, though some progress has been made. *Photorhabdus* is known to exhibit a kind of phase variance, and previous studies have demonstrated that secondary phase bacteria are unable to support symbiosis, being termed "symbiosis-deficient"; though *Photorhabdus* phase variance appears to differ somewhat from that of other bacteria in being uni-directional (Ffrench-Constant et al., 2003). In the process of bioconversion of an infected insect, late phase *Photorhabdus* have been shown to produce an array of secreted proteins such as proteases, toxins, and antimicrobials, to degrade the cadavers, and ward off non-*Photorhabdus* competition (be it from other microbes, or from scavenging insects) (Daborn et al., 2001; Baur et al., 1998).

Thus, every *Photorhabdus* strain studied to date maintains within its genome, all the symbiosis factors required for association with the nematode vector. This includes any 'standard' genetic determinants, but also any regulatory and epigenetic mechanisms. All the strains also maintain virulence factors and bioconversion enzymes required to cause lethal infection and biomass conversion of an insect prey. In the case of *P. asymbiotica*, they must do this in spite of a genome reduction (though with the gain of one or more plasmids), as well as harbour all the necessary genetic apparatuses to confer infectiousness in higher order homeotherms (including, but not limited to: thermotolerance, adaptive immune resistance/evasion, facultative intracellularity). This has given rise to the hypothesis that rather than maintain a repertoire of 'anti-insect' and 'anti-mammalian' virulence factors etc., that instead, the virulence factors it has are efficacious against cell types from both organisms (Waterfield et al., 2004).

1.1.2 A biological 'box of tricks'

There are a number of extremely interesting and unusual aspects of *Photorhabdus* cellular biochemistry and physiology that make it a fascinating study organism. A member of the *Enterobacteriaceae*, *Photorhabdus* is a motile, Gram negative rod shaped Gammaproteobacterium, which is partially what gives it its name: "rhabdus" from the Greek, "*rhábdos*" - "rod" or "wand". The former portion of its name is derived from perhaps its most striking characteristic: bioluminescence (Greek: "*phôs*" - light). *Photorhabdus* is still the only

known terrestrial bacterium that exhibits bioluminescence, and does so through possession of the full *luxCDABE* operon (Peat et al., 2010; Clarke and Joyce, 2008; Farmer et al., 1989; Gerrard et al., 2003). Why *Photorhabdus* has maintained the operon is still a mystery, but hypotheses include use as a signalling mechanism, similar to its marine counterparts, in symbiosis (signalling to the nematode that a cadaver is populated for instance), or possibly as a virulence mechanism to deal with oxygen free radicals and enhance survival - however there is sparse evidence for these theories, and valid counters to all of them (Waterfield et al., 2009). Nevertheless, it has lead to interesting anecdotes about a phenomenon observed during the American Civil War, known as "Angel's Glow", which has made its way in to some popular media including being covered in the well-known educational magazine "Mental Floss" (Durham, 2001; Soniak, 2012). The phenomenon observed that soldiers who were wounded in the conflict, had a greater average survival rate if their wounds glowed. The subsequent rationale for this is that their wounds may have been infected with *Photorhabdus*, which produces a myriad of antimicrobial compounds and toxins, killing off competition, including more virulent human pathogens that would have otherwise killed the individual.

The last point is another profoundly interesting and important aspect of *Photorhabdus* biology. At the time of sequencing, it was discovered that *Photorhabdus* has a greater proportion of its genome dedicated to secondary metabolite and toxin production than any other bacterium - including the model for secondary metabolite production, *Streptomyces* (6% vs. 3.8%) (Waterfield et al., 2009; Duchaud et al., 2003). Among these secondary metabolites, a stilbene compound has been previously identified, *3,5-Dihydroxy-4-isopropyl-trans-stilbene*, for which *Photorhabdus* is unique in being the only non-Plant organism known to produce it (Joyce et al., 2008). The compound itself has been shown to be a potent and broad range antimicrobial (Hu and Webster, 2000). Consequently, the burgeoning field of 'bio-prospecting' ('genome mining') (Shi and Bode, 2018), has begun to turn it's attention to *Photorhabdus* as a source of novel compounds - particularly important as we continue to try and combat the threat of antimicrobial resistance (Orozco et al., 2016), and helpful as *Photorhabdus* researchers, as it is leading to an increase in the number of available genomes and roles for many of the unknown genes. This will no doubt

continue to affirm *Photorhabdus*' place within the biotechnology world, complementing the exploitation which is already underway of the *lux* operon, and the organism itself for biopesticides - and, we hope, the PVCs in due course.

1.1.3 The life cycle of a pathogen and mutualist

Photorhabdus is an obligate pathogen and symbiont (Ffrench-Constant et al., 2003). Much of the research interest in the organism to date has been specifically because of this unusual life style. There are abundant examples of symbiotic microorganisms and pathogenic organisms, but very few where both lifestyles are found to be exhibited by a single organism. Trying to unravel the complex molecular basis for this is a huge task, making *Photorhabdus* an unusual and valuable emerging model.

Photorhabdus is a seemingly ubiquitous soil dwelling bacterium, having been isolated from all over the world, though most commonly near coastlines. However, it is not thought to survive exposed in the soil by itself. Instead, it is found in mutualistic symbiosis with entomopathogenic soil nematodes, specifically members of the *Heterorhabditidae*. In fact, such is the specificity of this mutualism, that different nematode species are known to harbour only particular bacterial species - the closely related *Xenorhabdus* are associated with *Steinernema* nematodes instead, for example (Chaston et al., 2011). The 'bacterium-nematode complex' has potent demonstrated lethality against members of the *Lepidoptera*, *Coleoptera*, *Hymenoptera*, and *Dictyoptera* (Naidoo et al., 2015), and has been used for many years now as a biopesticide (Waterfield et al., 2009). In the soil, the bacteria are associated with the so-called "Infective Juvenile" (IJ) (also known as a "Dauer juvenile") stage of the nematode host, which is free living and actively seeking insects to infect/parasitise.

As Figure 1.2 on the following page shows, the cycle begins with free-living Infective Juvenile nematodes in the soil. The IJs are associated with their *Photorhabdus* symbiotes, where the bacteria reside in the lumen of the nematode gut. Conversely, in the *Xenorhabdus-Stiernema* complex, the bacteria are relegated to quiescent growth in a specialised region of the intestine known as the 'receptacle'. IJs are a specialised alternative third developmental stage which are non-feeding, self-fertile hermaphrodites, with increased resilience to environmental stresses (by retaining an additional cuticle layer

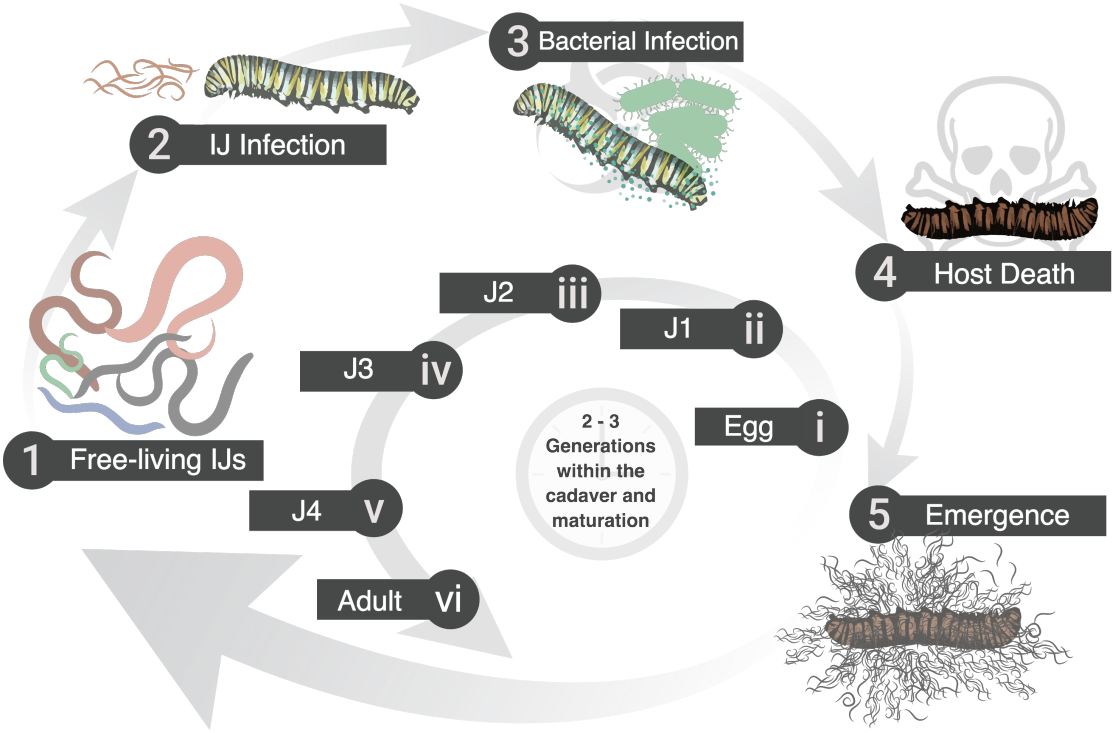


Figure 1.2 | The infection lifecycle of the Entomopathogenic Nematode complex.
1) The free living "Infective Juveniles" (IJs) in the soil seek out a new insect host to prey on. **2)** IJs ingress in to the insect prey either through natural openings or boring through the cuticle. **3)** *Photorhabdus* bacteria are regurgitated by the EPN in to the open blood system of the insect (hemocoel). **4.** The bacteria reproduce and express virulence factors to kill the prey in a matter of hours. **5)** Bio-conversion of the cadaver biomass in to additional bacteria provides a food source for the continued reproduction of nematodes. **i-vi**) During replication on the cadaver, juvenile nematodes reach maturation and conduct their sexual reproduction phase. Millions of next generation IJs then leave the cadaver, reassociating with the bacteria to find the next prey insect. Adapted from `http://www.giabr.gd.cn/kxcb/kpdt/201405/t20140516_234014.html`, and in turn (Ffrench-Constant et al., 2003)

(Ciche and Ensign, 2003)). IJs seek out prey insects within the soil to parasitise, and enter the organisms open circulatory system (or "hemocoel") via natural openings such as the spiracles (breathing tubes), mouth or anus. Alternatively, the nematodes bear a sharp tooth-like structure at the mouth which they can use to bore through the cuticle of the organism. Once inside, the nematodes regurgitate their bacterial 'payload', which is typically less than 200 individual cells, speaking to the potency of the entomopathogenic activity of *Photorhabdus*, which then employ sophisticated molecular tools to evade the immune response and establish an infection. The regurgitation and triggering of developmental processes for the nematode are induced by compounds within the insect hemolymph (blood) (Ciche and Ensign, 2003). Interestingly, the same paper by Ciche and colleagues showed that Grace's insect medium could not replicate this effect, suggesting

that it is only very specific compounds involved in the process which are not reproduced in artificial media, though these compounds could not be identified specifically. Over the course of the next ≈36 hours, the IJs developmental switch, brought about by the insect environment, triggers feeding behaviour. The bacteraemia is lethal to the host insect, due to rapid proliferation and production of many exoenzymes and virulence factors. The bacteria therefore digest and bioconvert the cadaver, and the nematodes feed on the new bacterial biomass. Some of the ingested bacteria adhere to the nematode intestine and invade the rectal gland cells, restoring the EPN complex. While growing and maturing on the insect cadaver, the nematodes can complete their maturation to adults, having been larval juveniles up to this point. Their development to adults also leads to a dioecious stage, rather than the hermaphroditic one seen in the IJs, meaning that the nematodes can undertake sexual reproduction. Once the cadaver is exhausted, the EPN complexes vacate the site and go off in search of fresh prey, repeating the cycle.

There are a number of theories suggesting bases for the control of symbiosis and the switch to pathogenicity (though a full review is beyond the scope and need of this thesis), including the use of the *lux* operon as mentioned earlier, and recent studies have shown that many of the secondary metabolites that *Photorhabdus* produces have roles in this mechanism. It has been observed that mutants in *relA* and *spoT*, which are both ppGpp alarmone synthases, become deficient in secondary metabolism and in symbiosis, but not in virulence (Bager et al., 2016). Mutants in the malate dehydrogenase enzyme (*mdh*), exhibited similar behaviour, with no effect on virulence, but becoming incapable of mutualism (and unable to produce light, pigments, and the previously mentioned stilbene compound; all of which are hallmarks of post-exponential phase secondary metabolism). *mdh* is a central enzyme in the Krebs' Cycle, implicating it in both central and secondary metabolism (Lango and Clarke, 2010). Similarly, mutants in *hfq*, a global post-transcriptional regulator that is widespread in bacteria demonstrated complete abolishment of all known secondary metabolite production, and a concomitant failure to associate with the nematode vector (Tobias et al., 2016). Large-scale lifestyle decisions, such as sporulation in *Bacillus*, a process thought to involve as much as half of the genome, may be analogous. Certainly, there are similarities in that both processes

require a functional Krebs' cycle (Stephens, 1998), and so it seems likely that a complex process such as symbiosis could also involve a significant proportion of the genome. It is probable, therefore, that any or all of the aforementioned theories are true, and that there are a vast number of pathways working together to fettle and control the symbiosis and pathogenicity process.

1.2 The *Photorhabdus* Virulence Cassette

Not much is known for certain when it comes to the *Photorhabdus* Virulence Cassettes, and even less has been published. To date, there has only been a single paper on the discovery and biology of the *bona fide* PVCs (Yang et al., 2006). However, an increasing number of papers have appeared, particularly in the last ≈5 years, which have attempted to understand how PVCs 'fit in', in a wider context, and have begun to speculate on the roles of various genes. While there is nothing wrong with this in principle, much of the biology is still lacking, and it is not always constructive to try and constrain a biological entity to fit within the criteria for different systems. With the rapid proliferation of structural data for analogous systems however, thanks in no small part to the advancements in cryo-electron microscopy for studying large macromolecular complexes, there has never been a better time to study fascinating structures such as these.

As these structures are complex, multipartite and still quite enigmatic, this section will serve as a 'guided tour' through the various components of the PVC structures, as well as draw analogies against other, better characterised, related structures as it was understood before this project began, to help the reader understand the chapters to come. Chapter 3 on page 95 will continue in this vein, in the context of what has been learnt since, with the advantages of more complete databases and a better biological understanding, to fill in some of the gaps.

1.2.1 Discovery of the PVCs

Upon sequencing of the first *Photorhabdus* genomes, the PVCs were identified as putative prophage regions the *P. luminescens* TT01 genome, in 4 tandem repeats which demonstrated unusual % GC content. When a cosmid library was constructed from the *P.*

asymbiotica ATCC43949 genome, clones harbouring the operon for a particular PVC with the so-called 'Pnf' (*Photorhabdus* necrosis factor) cognate effector demonstrated high levels of injectable toxicity against whole insects - killing them in as little as 15 minutes, and earning them their name ("Virulence Cassettes") (Yang et al., 2006; Waterfield et al., 2008). It became apparent from these early experiments that the PVCs represented a novel kind of toxin delivery and translocation mechanism, and similar patterns of toxicity could be identified in other cosmid clones which bore other PVC variants. The pnf effector of this particularly potent PVC, is homologous to the active site domain of cnf1 (Cytonecrosis Factor) of uropathogenic *Eschericia coli*, and works in the same way, by activating Rho GTPase proteins inside the target cell, which leads to cytoskeleton depolymerisation (Landraud et al., 2004; Buetow et al., 2001). Further inspection showed this cosmid to contain what we have now come to recognise as a PVC, with its associated effector, and several more cosmids within the library were identified which contained various other PVCs. It was observed that a number of the cosmid-borne PVC operons are defective in some way, suggesting that the obtained colonies were those where the cosmids had been inactivated in some way such that they could be tolerated. This potential self-toxicity is the subject of Chapter 6 on page 185.

The PVCs were able to be purified/enriched from the cosmid supernatants to identify the basis of the toxicity, using diethylaminoethyl-sepharose resins and upon imaging via electron microscope, sure enough, phage like structures were apparent in the samples. Subsequent immunogold staining using antibodies raised against the Pnf toxin showed that, when disrupted, the structures appeared to be loaded with the toxins. Figure 1.3 on the next page shows some of the original microscopy, reproduced from the Yang et al. (2006) study, as well as some more recent micrographs from the lab and an ongoing collaboration with the Max Plank Institute in Dortmund, showing cleaner samples. Preliminary data from the collaboration with the Max-Planck Institute resolving the atomistic structure of the PVCs is now beginning to vindicate the many assumptions about the PVC architecture, biology and function.

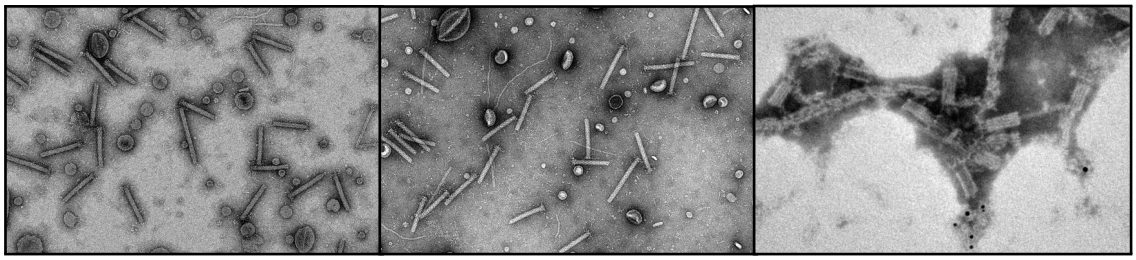


Figure 1.3 | A SELECTION OF ELECTRON MICROGRAPHS OF THE PVCs.
A selection of micrographs are shown here revealing their caudate structure and the resemblance to other contractile tail systems. The left and centre panels show more recent, but unpublished data from an ongoing collaboration with the Max Planck Institute at Dortmund, purified via Cesium Chloride buoyant density gradient ultracentrifugation. The right hand panel shows one of the earliest images of the PVCs ever obtained, via diethylaminoethyl sepharose resin elution (the white mass in the background). The black dots in the bottom-center of this panel correspond to immunogold antibody staining against the payload molecules released from the syringes.

1.2.2 *Photorhabdus* is a PVC Addict

A single PVC is a remarkable biological entity. However, *Photorhabdus* has chosen not to stop here. Within the genomes studied to date, there are as many as 6 distinct PVC operons, each with 1 or more associated toxin effectors, in any single genome. The fact that each one has a hallmark effector or effectors has been used since their discovery to delineate which PVC is under discussion (Yang et al., 2006) - with some exceptions. In several cases, the PVCs were simply named for their positions in the genome. Specifically, in *P. luminescens* TT01, the 4 tandem PVCs mentioned in the previous section were simply named "Unit1", "Unit2", "Unit3", and "Unit4". In *P. asymbiotica* genomes, there is also a distinct "Unit1", but confusingly, it is not most homologous to the "Unit1"s of *P. luminescens*.

With this in mind, this is an ideal moment to briefly explain the nomenclature that will be used throughout this thesis when referring to the PVCs. The PVC with the cognate pnf toxin that was mentioned in the previous section will be used as an example. In order to denote each PVC, the nomenclature "PVCpnf" will be used. Where a distinction between inter-strain variants of the same operon is required, this will be followed by the strain name itself, as in: "PVCpnf ATCC43949" or "PVCpnf Kingscliff". Furthermore, when a specific gene is under discussion, "PVCpnf" will be followed by the numerical identifier for that gene; so for the first protein of the *P. asymbiotica* ATCC43949 strain Pnf operon, the nomenclature will become: "PVCpnf1 ATCC43949". In cases where this thesis refers

to just a specific locus, across all operons, the terminology will simply be "PVC1" - i.e. the first locus, in all operons syntenically.

Figure 1.4 on the following page demonstrates the variants from the *P. luminescens* TT01 and *P. asymbiotica* genomes. The existence of phage-like contractile tails in a myriad of genomes has now been demonstrated, however, *Photorhabdus* appears to remain somewhat unique, in that no other organisms have been found, so far, which harbour so many forms of the same structure in a single genome. Even if one examines *Xenorhabdus* strains, which are as closely related as it is possible to be outside of the immediate *Photorhabdus* genus, it is only possible to identify single copies of PVC-like operons.

Naturally, this leads to questions about how and why *Photorhabdus* is able to maintain so many copies of operons which have a high degree of internal and inter-operon paralogy. Conventional wisdom would suggest that at least 1 of these extra operons might drift to the point of removal/pseudogenicity. For certain, there are substantial genetic differences between different PVCs, and drift has most likely played a part in this, but nevertheless they persist, which suggests the selective pressure is sufficient on each PVC to maintain them. There are a few speculative rationales that this could be the case. Firstly, it's possible that *Photorhabdus*' life cycle is so competitive that any additional toxin systems of a net benefit to the organism, despite their high metabolic cost, with each fulfilling sufficiently different roles. This would further mesh with the observation that *Photorhabdus* elaborates the largest repertoire of toxins known so far. An alternative idea however is that not all PVCs are evolved with a toxic effect in mind, and may have host modulatory effects - which will be covered in more detail in a later section. There is some preliminary evidence that different PVCs perform different roles, perhaps with toxicity to particular tissue/cell types, or responding to different environmental cues.

This almost paradoxical variability-yet-conservation is a recurring theme in this thesis, and is also useful for understanding any single one of the PVCs.

1.2.3 PVCs as contractile nanomachines

The initial electron microscopy, and homologies observed to R-type pyocins (Ge et al., 2015a), the Antifeeding prophage (Heymann et al., 2013), and other prophage sequences,

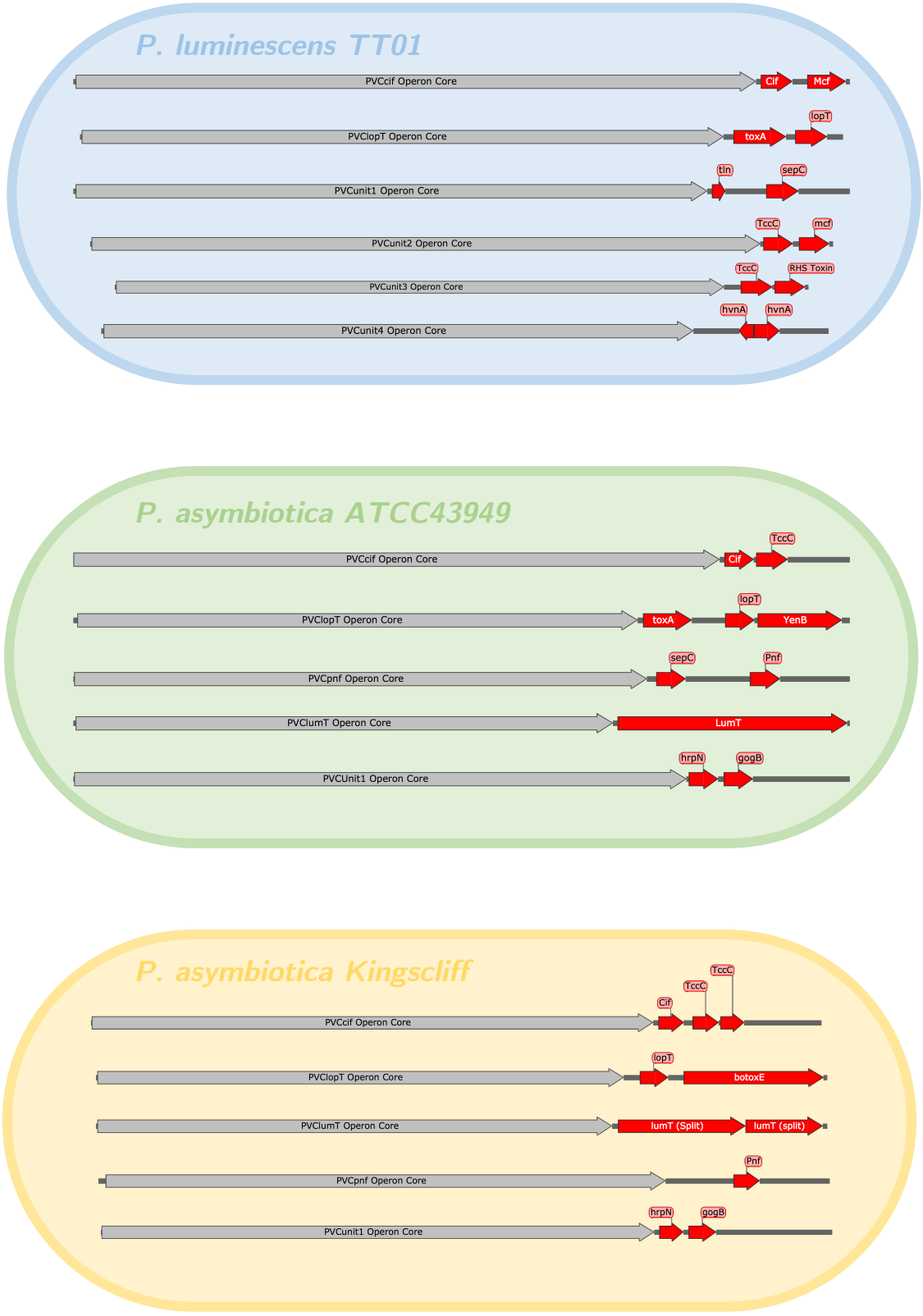


Figure 1.4 | The PVC variants found in 3 *Photorhabdus* genomes.

A schematic of the variant forms of PVCs found in 3 different genomes. The operon core is denoted in a single grey arrow, and the diverse effector proteins which distinguish the operons are identified in red. Each operon is to scale, though the scale between operons is not identical, though they are all approximately 23 - 25 kbp.

which have had their structures resolved or EM studies conducted, provided compelling evidence that the PVCs elaborated a similar caudate structure. Moreover, it has become increasingly apparent in recent years that the entire mechanism of 'contractile machines' is an evolutionarily conserved structure which appears time and time again in nature - and not just in the form of prophages, which is what many of these devices have been mistaken for to date (Kube and Wendler, 2015a; Sarris et al., 2014; Brackmann et al., 2017). In particular, in the Sarris et al. paper, contractile tail mechanisms of various forms have been demonstrated to be widespread with a remarkable diversity of functions, even in the Archaea. Much as the bacteriophage biosphere has become increasingly well understood to actively shape the bacterial biosphere, it is now becoming similarly apparent that phage-like structures will have had (and are having) a similarly decisive role in shaping ecosystems and evolution (Ffrench-Constant and Dowling, 2014). The following sections now detail the state of knowledge for the well studied cousins of PVCs; readers are also encouraged to look at the original Kube and Wendler, Taylor et al., and Sarris et al. papers for 2 excellent reviews from both a structural and genetic perspective, accordingly (Kube and Wendler, 2015a; Sarris et al., 2014; Taylor et al., 2018).

1.2.3.1 Of PVCs and Phage

Contractile tail nanomachines are typified by the bacteriophage order *Caudovirales* (from the Latin: "*cauda*" - "tail"), so it makes sense to start here. In particular, those of the *Myoviridae* family, to which the well known model T4 phage belongs (Ackermann, 1998). Phages have been studied for over 100 years now, after their original discovery at the beginning of the 20th Century by Félix d'Hérelle (D'Hérelle, 1917). d'Hérelle is also credited with conceptualising phage therapy, which is becoming increasingly relevant again with the rise of antibiotic resistance.

The tail structures of the T4 bacteriophage were the first structures ever to be resolved by electron microscopic (EM) density reconstruction as far back as 1975 (Amos and Klug, 1975); and with the recent explosion of Cryo-EM data, and the so called "Resolution Revolution" (Kühlbrandt, 2014), we have a clearer understanding of these elegant and staggeringly complex macromolecular machines than ever before. The tail tube, capsid,

and the intricate baseplate complex of T4 have now been solved to atomic or near atomic resolution (Aksyuk et al., 2009a; Kostyuchenko et al., 2003, 2005; Fokine et al., 2004, 2013; Rossmann et al., 2004; Taylor et al., 2016a; Lan et al., 2014), and is probably the single most well studied structural entity. Figure 1.5 shows some of the actual micrographs of T4 collected to date, and Figure 1.7 on page 21 shows a collection of the now resolved structures reproduced from the literature. Even at a glance, its quite apparent that these entities share similar origins and architectures.

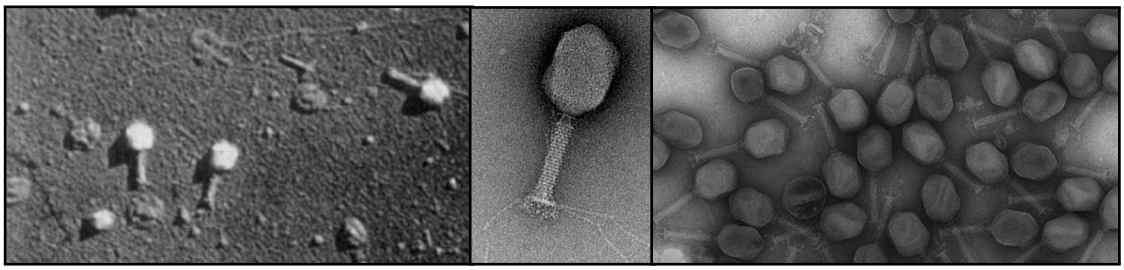


Figure 1.5 | Electron Micrographs of T4 bacteriophage virions.
The left panel shows an early T4 phage micrograph from 1958, reproduced from `https://www.molbio.unige.ch/eng/about/history`. The centre panel shows a close up of the T4 phage, revealing the tail fibres, baseplate, capsid and tube in detail, reproduced from Knott and Genoud (2013). The right hand panel shows a scaled up experiment, purifying virions for phage therapy, reproduced from Bourdin et al. (2014).

Despite their superficial resemblance in gross structure, PVCs appear considerably simpler than T4. As non-replicative entities, of course, the PVCs lack any machinery associated with this function (including the capsid packaging mechanism and replicative enzymes), but also differ quite substantially in structural proteins. PVC operons are typically around 25-30 kilobases in length, and usually encode approximately 25 proteins at most, whereas the T4 genome is nearly 170 kb, and encodes 289 proteins (Miller et al., 2003).

From the very earliest annotations of the *Photorhabdus* genomes, it was evident that there was shared homology for at least some of the genes in the operon. In particular, the inner and outer sheath proteins matched comparatively well, picking up annotations as gp19 proteins and major sheath proteins (gp18) respectively, whilst the majority of the operon remained as purely 'hypothetical proteins'. Figure 1.6 on page 20 shows how the many different subunits of the T4 sheath and baseplate complex together. The tail is comprised of 2 concentric hollow cylinders. The interior tube is comprised of stacked, helically offset, hexameric toroids of gp19. Similarly, the outer sheath of gp18 which

provides the force for contraction has a helical hexameric cylindrical shape. Figure 1.7A on page 21 shows the helical nature of both the inner and outer sheaths well - a single "protofilament" of the outer gp18 is depicted in its extended (green) and contracted state (orange). In the relaxed state, the offset is approximately 17.2°, and in the contracted state, this twist increases to 32.9° (Kube and Wendler, 2015a; Kostyuchenko et al., 2005; Leiman et al., 2004). The exact mechanism of contraction for contractile tail systems is thought to be highly conserved, despite often significant differences in structure and sequence between structural homologues, and will be discussed for all the upcoming systems in Section 1.2.4 on page 52.

There is evidence from heterologous expression of the analogous inner tubes of the Type 6 Secretion System (the Hcp1 protein) that the homohexameric toroids spontaneously self-assemble (Ballister et al., 2008), and that polymerisation begins from the gp27-gp5 spike tip complex (the so called "baseplate hub complex" (Lan et al., 2014)) (Kanamaru et al., 2002). The gp18 homohexamers then also polymerise around the growing interior tube. The tail tube length is controlled by 3 further proteins, which have been identified as gp29, a "tape-measure" protein, and a tube terminator/cap protein complex of gp15 and gp3. The tail tube tape measure protein was identified by elongation and truncation experiments, with the actual tube length varying in accordance with the expansion of shrinking of the tape measure protein (Abuladze et al., 1994), and proteins serving similar roles have been identified in other contractile tail systems (Katsura, 1987; Katsura and Hendrix, 1984; Isao, 1990).

Despite having a couple of identifiable baseplate-like proteins, the PVCs appear to have radically reduced baseplates overall, though it must serve almost exactly the same purpose and function. It was possible to spot some assorted similarities to the gp6 baseplate component proteins, though in the absence of a full PVC structure, its still unclear exactly where these proteins will fit, and their exact role in the final structure. The T4 baseplate complex is exceedingly intricate, comprising some 18 different protein types (including the baseplate spike/hub, and the tail fibres), and roughly 57 separate protein molecules (some of which are, themselves, made up from multiple chains). As can be seen in Figure 1.6 on page 20 from Leiman et al. (2004), 6 "wedge" complexes are

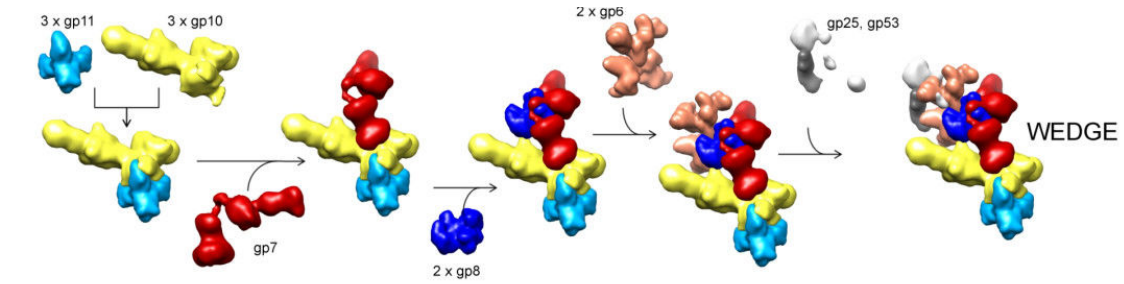
formed from a gp6-gp7-gp8-gp10-gp11-gp25-gp53 complex. Each of these 6 wedges then come together around the baseplate hub spike complex (gp27-gp5), itself comprised of 3 different proteins and at least 7 distinct chains. A further 12 proteins are added (6 each of gp9, and the tail fibres - gp12). Next, a heterodimeric toroid collar of gp48 and gp54 is then added to the apex of the dome shaped complex, similar to the keystone at the top of an arch. When scrutinising the T4 baseplate it perhaps makes sense that of all of the proteins for the PVCs to maintain detectable homology to, gp6 is the best hit, as it sits in close register to the collar and spike complex and is therefore a minimal component of the complex (depicted in light orange in Figure 1.6A on page 20).

Though the PVC and T4 baseplates are likely to be substantially different in structure, the baseplate hubs/spike complexes appear to share more in common. Existing annotation attributes VgrG protein homology to the spike (which is associated with the T6SS - see Section 1.2.3.4 on page 35), rather than gp27-gp5, though these are extremely similar structures - among the most structurally conserved and easily identifiable amongst all caudate apparatuses despite often having as little as 13% sequence identity (Veesler and Cambillau, 2011; Leiman et al., 2009; Basler, 2015a). The T4 tail spike complex retains an Oligosaccharide/Oligonucleotide binding domain ("OB-fold" - a 5-stranded anti-parallel β-barrel (Murzin, 1993)) and a lysozyme domain which appears to be lacking from the VgrG, instead being functionalised with assorted alternative enzymatic activities (Pukatzki et al., 2007; Kanamaru et al., 2002) - an extensive discussion of VgrG will be saved for Section 1.2.3.4 on page 35.

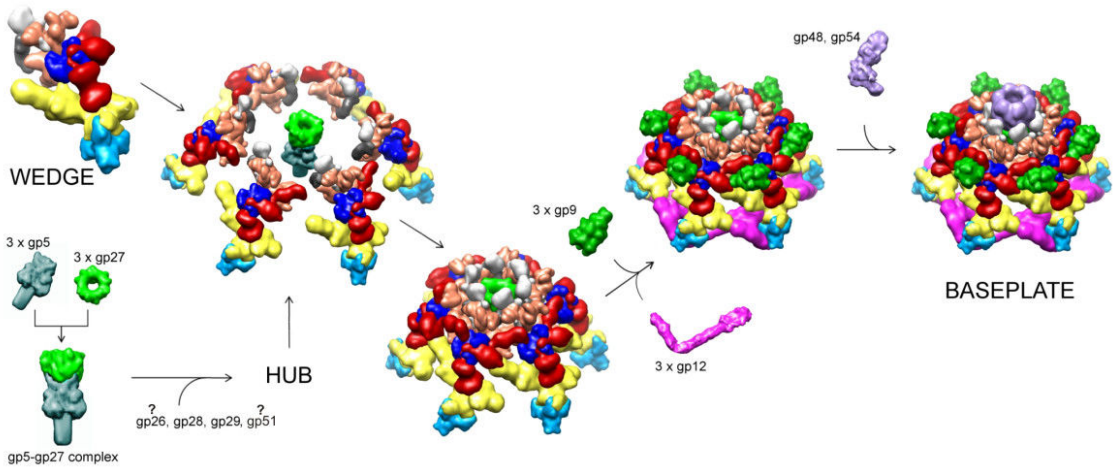
Finally, the PVCs are thought to contain putative tail-fibre like genes, proposed to be for cell binding in the same manner as T4. So far, there appears to be no evidence of both long and short fibres as is the case in T4 however (Bartual et al., 2010; Thomassen et al., 2003). Again, the PVCs seem to elaborate a much simpler version of these analogous structures. The long tail fibres of T4 are comprised of 4 proteins: gp37 and gp34 form the main trimeric body of the fibre, but are separated in to a "proximal" (thigh) and "distal" (shin) end by gp35 hinge (a so-called "Knee cap" which induces a kink in the structure allowing them to fold away when in unfavourable infection conditions). At the upper end of the 'shin' a trimer of gp36 is also present completing the knee joint (Leiman et al., 2010).

The long tail fibres are anchored in to the baseplate structure by 6 trimer complexes of the gp9 protein (Figure 1.6B on the following page). At the outermost edge of the dome, the 6 short tail fibres comprised of gp12 can be seen wreathing the edge in the folded state (in Figure 1.6B on the next page they can be seen pinkish-purple; in Figure 1.7D on page 21 the short fibres can be seen in their extended state). The short tail fibres are known to be capable of folding correctly without the need for additional chaperones (Leiman et al., 2010; Goldberg et al., 1997; Ali et al., 2003). For the PVCs there appears to only be a single tail-fibre like gene, referred to as PVC13 due to its general syntenic location, and is the focus of Chapter 5 on page 138, where a more detailed introduction to the tail fibre structure and proposed function can also be found.

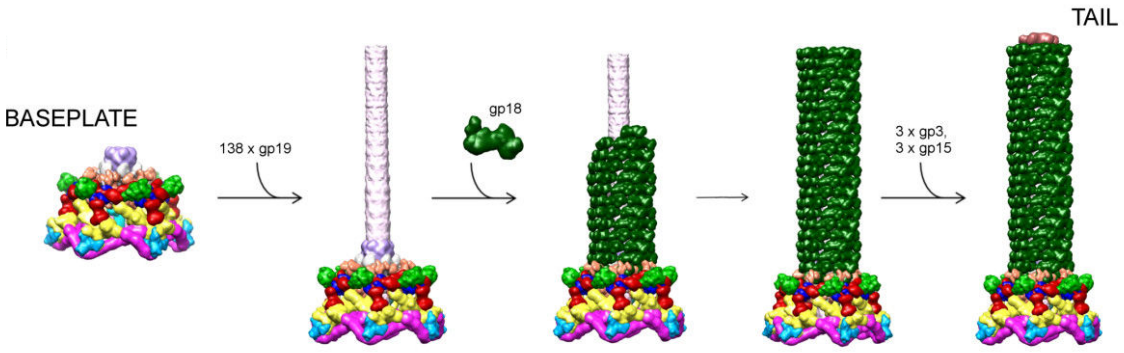
This review will not consider the assembly of the T4 capsid, as the PVCs do not contain analogous structure, but for an excellent all-round review of the full assembly of T4, see Yap and Rossmann (2014a) and Leiman et al. (2010).



(A)



(B)



(C)

Figure 1.6 | The structural components of T4 and their stoichiometric assembly.
(A) The formation of the "baseplate wedge" subunit, which is, itself comprised of 6 different proteins and which makes up the majority of the baseplate. **(B)** Shows the formation of the complete baseplate, where the spike baseplate hub complex and tail fibres are added. The overall baseplate is made up of 6 wedge complexes which are further complexed together, with the addition of tail fibres and a number of other baseplate proteins including the collar. **(C)** A depiction of the complex between the baseplate structure and the polymerisation of the tail tube. The collar interfaces with the interior tube, around which almost 150 copies of gp18 are helically polymerised before termination and capping. Adapted and reproduced from Leiman et al. (2004) and Yap and Rossmann (2014a).

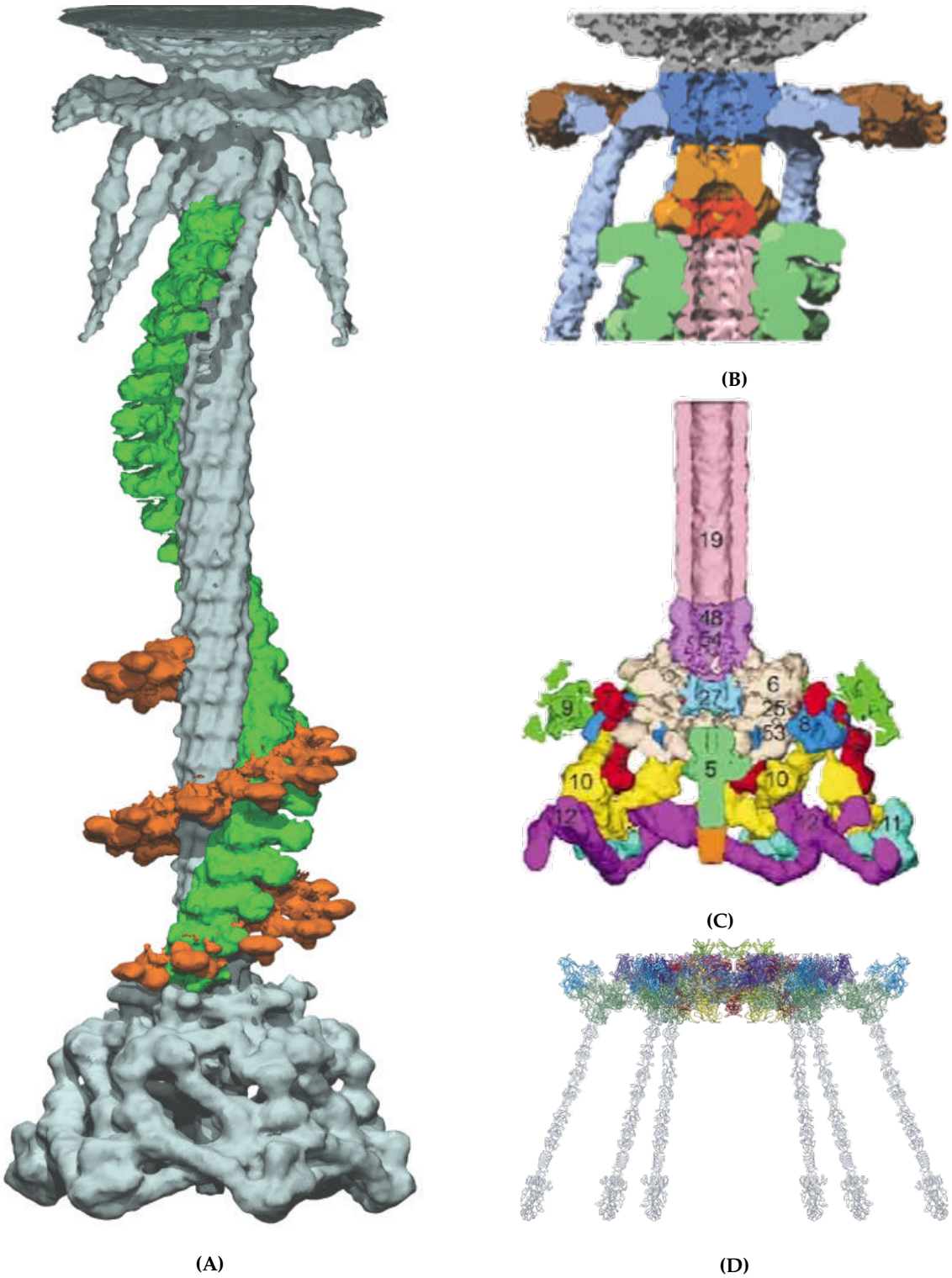


Figure 1.7 | A selection of the resolved structural components of Bacteriophage T4.
(A) The T4 EM density reproduced from Kostyuchenko et al. (2005), the helical outersheath protofilaments are shown in the extended (green) and contracted (orange) conformations. **(B)** The architecture of the 'neck'/'collar' region of the T4 phage, showing the top of the tube. Adapted and reproduced from Kostyuchenko *et al.* also. **(C)** The intricate baseplate architecture (shown as a slice-through), adapted and reproduced from Kostyuchenko et al. (2003). **(D)** The structure of the lower baseplate complex, showing the extended short tail fibres (they are 'retracted' in **A** and **C**) adapted from Taylor et al. (2016b). The structure of the genome-containing capsid has been omitted, as the similarity to the tail and baseplate is more relevant to this thesis.

1.2.3.2 Of PVCs and R-type Pyocins/Tailocins

The R-type pyocins, particularly those of *Pseudomonas aeruginosa*, have been among the longest studied caudate structures, if Myophages such as those just discussed are discounted, with papers describing their structure and activity as far back as 1965 (ichi Ishii et al., 1965), and they were discovered as early as 1954 (Jacob, 1954). Nevertheless, it took until 2015 for the structure of one of these tailocin structures to be resolved fully (Ge et al., 2015a) (see Figure 1.9 on page 25). A rapidly growing body of data on the specificity, activity and structure of these types of macromolecular complexes is appearing. 'Tailocins' have attracted much attention recently due to their potential use as an alternative to phage therapy. The prospect of utilising bacteriophages has made the public and some of the scientific community understandably nervous, due to their uncontrolled, rapid replication within bacteria, and the introduction of foreign DNA in to the body's microbiome. Tailocins have alleviated some of these concerns due to their highly specific bactericidal activity, similar to phage, but without containing any nucleic material and thus no replicative capacity, and they appear tractable for engineering (Scholl and Martin, 2008).

Tailocins are so called as they are comprised of bacteriophage tail tube, baseplate and fibre structures, without a capsid or head (ichi Ishii et al., 1965). Bacteria have co-opted these structures in to their genomes such that they can be used as highly specific antimicrobials against other, potentially closely related bacteria, providing considerably higher selective toxicity than is attainable through small molecule antimicrobials (Heo et al., 2007). This section specifically focuses on the "R-Type" pyocins, which are considered a subclass of bacteriocins (protein or peptide toxic molecules effective against other bacteria; colicins are another well known example). They take their name from the fact that they are 'Rod'-like phage tails, being demarcated from the F ('flexious') and S (soluble) type bacteriocins. They derive the name 'pyocin' from their discovery in *P. aeruginosa*, as mentioned, as it was renamed from *Pseudomonas pyocynia*. The F-type pyocins are also phage tail like structures, but the tails are not straight tail rods, instead being somewhat curved, and crucially, they are noncontractile, meaning they are more closely related to P2 and λ phages, than T-even *Caudoviriales* (Michel-Briand and Baysse, 2002; Nakayama

et al., 2000). The S-type bacteriocins are small, soluble antimicrobials, more reminiscent of small molecule compounds, and cannot be sedimented or visualised by EM, unlike the F and R types (Heo et al., 2007; Kageyama, 1975). As with the previous section on T4, Figure 1.8 shows a selection of R-type pyocin molecules as observed via EM. Hopefully the reader can already appreciate the similarities between the PVCs as shown in Figure 1.3 on page 12 and the pyocins in the figures below.

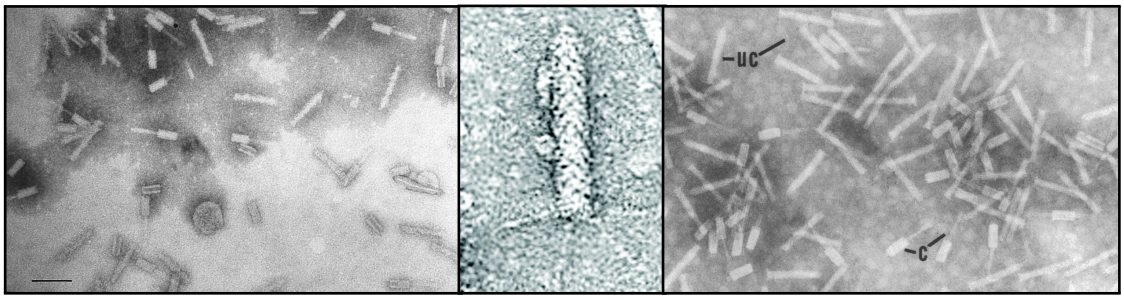


Figure 1.8 | Electron Micrographs of *Pseudomonas* R-type pyocin particles
The left panel shows a number of R-type pyocin molecules purified. The contracted nature of several of the particles reveals clearly the size difference between the inner and outer sheaths. Reproduced from Lee et al. (1999). The central panel shows a close up image of an individual pyocin particle, where the caudate structure and presence of at least 4 tail fibres is apparent (reproduced from Williams et al. (2008). The right hand panel shows R-type pyocin particles in a semi-purified form. The annotations on the image from the original document denote **UC** - uncontracted particles, and **C** - contracted particles. Reproduced from Morse et al. (1976).

R-type pyocins exert their antimicrobial activity in a similar fashion to bacteriophages, by first using their tail fibre proteins to bind with the lipopolysaccharides (LPS) of other Gram negative bacterial cells of closely related strains. The binding occurs strongly, which provides the necessary anchorage for the next step of toxicity - puncturing. The contractile system, as in the Myophages, drills the tail tube and spike in to the surface of the cell, creating a pore. Unlike the Myophages however, the pyocins contain no translocated material (DNA nor protein) and instead, simply cause a rapid and lethal depolarisation of the bacterial membrane (Uratani and Hoshino, 1984). The consensus, at least, is that no material is translocated, though some papers have shown single stranded nucleotide cargoes (Lee et al., 1999) - this may be an exception, rather than the rule though. Such is the efficacy and potency of this mechanism of killing, that the pyocins demonstrate 'single hit kinetics', meaning a single pyocin complex is sufficient to kill an individual cell (Ohkawa et al., 1973). Roughly 100-200 pyocins can be produced from a single host bacterium, with the first active complexes matured after as little as 45 minutes after

induction (Michel-Briand and Baysse, 2002; Shinomiya, 1972; Scholl and Martin, 2008).

The R-type pyocins represent a step further along the evolutionary path from phage to PVC-like systems, having undergone the relevant streamlining, by removal of the capsid genes and the replicative machinery associated. The pyocins are also a good example of the ubiquity of contractile tail systems in nature, underscoring their potentially pivotal role in the shaping of ecosystems, being elaborated by around 90% of *Pseudomonas* strains (Michel-Briand and Baysse, 2002), being widespread amongst Gram negatives (particularly among other *Enterobacteriaceae*) (Coetzee et al., 1968) and examples also being found in Gram positives such as *Listeria* (Zink et al., 1995) and *Staphylococcus* (Birmingham and Pattee, 1981; Scholl and Martin, 2008).

Even with this seeming ubiquity among various clades within bacteria, it's interesting to observe and speculate at this point, on the possible link between these microbial 'weapons' and their abundance in species of bacteria which are thought to have some marine origins. *Pseudomonas* has long been known to be associated with marine and generally aquatic environments, and this has been a long running hypothesis for the origins of *Photorhabdus* itself (two 'smoking guns' for this being that it has retained the *lux* operon, which is otherwise exclusive to marine organisms, and its frequent isolation near coastlines). It seems that the ability to produce caudate structure which can be deployed at a distance could have some extra utility in aquatic environments - and some further examples of innovative tailocin like structures are discussed in **??** 1.2.3.5.1 on page 47 and **??** 1.2.3.5.2 on page 48. Persson et al. (2009) have also made a similar observation, when they studied the prevalence of various pathogenicity islands in marine organisms from the Global Ocean Sampling dataset, including islands like the Antifeeding prophage (see Section 1.2.3.3 on page 29).

The seminal paper which finally resolved the intricacies of the the structure of the R-type pyocins was that of Ge et al. (2015a). Not only were they able to obtain high resolution EM maps of the structure, they managed to resolve, atomistically, both the pre- and post-contraction states. Figure 1.9 on the following page reproduces this data.

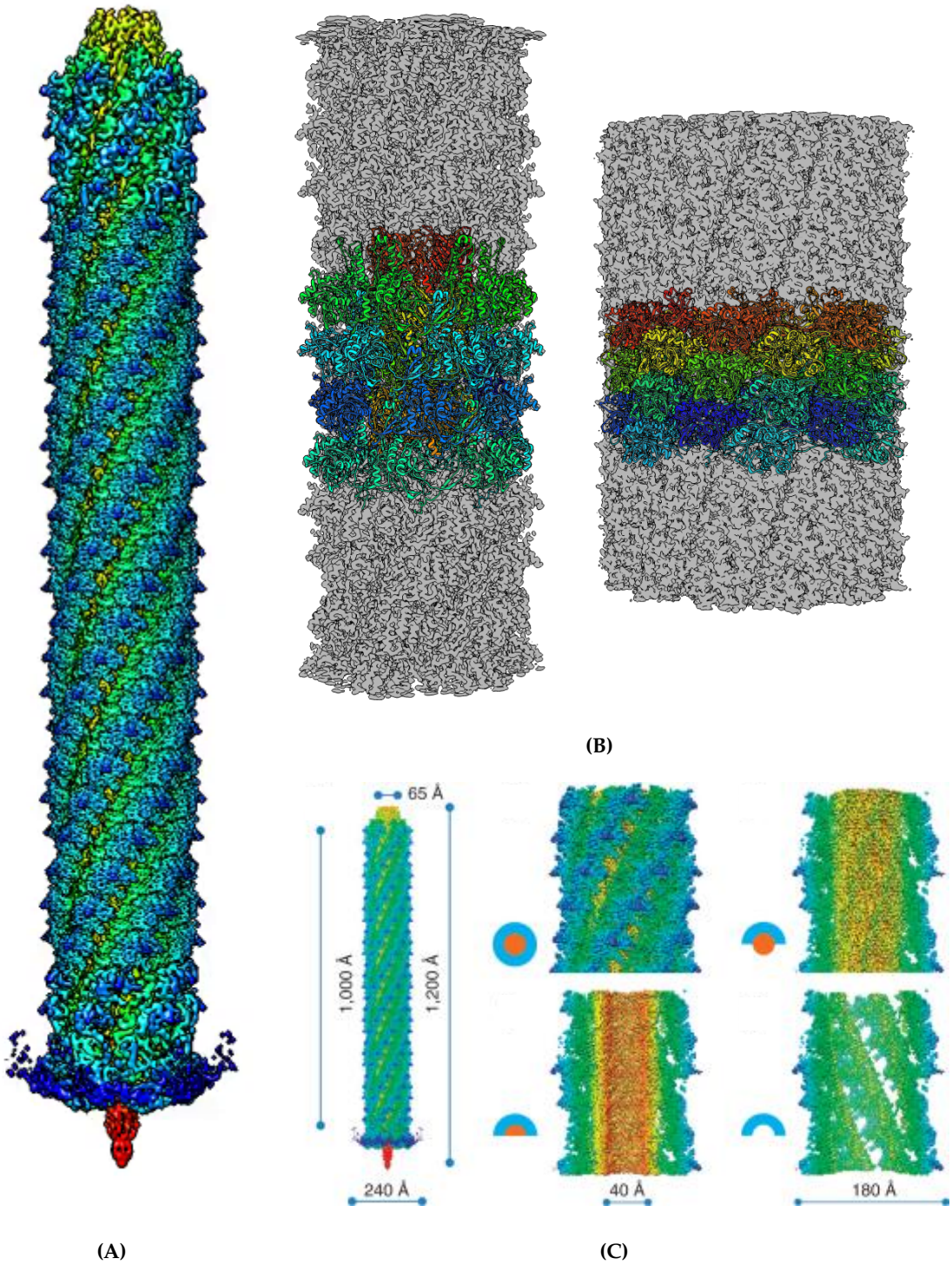


Figure 1.9 | The structures of the R-type pyocin from Ge et al. (2015a).
(A) The reconstructed pyocin EM density reproduced from the Supplementary Video 1 of the Ge et al. (2015a) paper as a snapshot. The structure is coloured according to its distance from the central axis (colder colours are further away from the centre). **(B)** Shows the extended and contracted sheath structures for the pyocin from EMDB-6270 (extended) and EMDB-6271 (contracted) with the fitted PDBs 3J9Q and 3J9R respectively. These figures were made using the published data, but reproduced independently using UCSF Chimera (Pettersen et al., 2004). **(C)** Shows sequential cut-aways of the sheaths with Ångstrom measurements of the inner and outer diameters and lengths. The coloured circles adjacent to each sub-panel are a key to which tube faces have been sliced through (orange = inner, blue = outer).

From Figure 1.9 on page 25 the helical nature of the outer sheath is quite apparent. Interestingly though, unlike the T4 phage structure, the inner sheath is not helically offset, instead being a direct stack of hexamers forming the equivalent of the gp19 toroids. The outer sheaths are helically offset relative to one another by 18.3°, with a right handed spiral, and are translated by 38.4° along the vertical helix axis (Ge et al., 2015a; Kube and Wendler, 2015a). Thus the hexameric helix is, in effect, more 'tightly wound', by having a greater deal of twist, and less vertical rise per unit versus the T4 sheath (in the extended configuration). The outer sheath differs further still as it is comprised of much simpler monomers. The molecular weight of the gp18 monomers is $\approx$ 71.3 kDa, whereas the equivalent outer sheath protein in the R-type pyocin is only 41.2 kDa. This can also be seen from the structures themselves, as the pyocin monomers seem to lack the protrusions that the T4 gp18 protomers have to quite the same degree, though there is still a noticeable ridge-trough-ridge architecture to the tube (Kube and Wendler, 2015a). From the atomic reconstruction, it was shown by Ge and colleagues that each protomer of the outer sheath interacts with the adjacent 2 protomers via extensions of the N and C termini of the individual monomers with the C-terminal reaching out to the monomer to the right, and the N-terminal to the left. Thus the outer sheath of the R-type pyocin actually more closely resembles a mesh, like a chain-link fence, encompassing the inner sheath, but with the ability to transduce a contraction force along its length (Ge et al., 2015a). The bottom left panel of Figure 1.9C on page 25 demonstrates this, as the outer sheath cutaway can be seen through completely from the interior. This interaction has also been observed in bacterial pili and Type 6 Secretion Systems, and previously referred to as the "β-augmentation mechanism" (Remaut and Waksman, 2006). Mutagenesis studies showed that these interwoven strands were essential for the contractile mechanism in the T6SS, though were not required for assembly, suggesting that hydrostatics are largely responsible (which is also consistent with the spontaneous self assembly of Hcp monomers seen in Ballister et al. (2008)) (Kudryashev et al., 2015; Clemens et al., 2015). The Ge et al. paper also made the observation that there is no structural interaction between subunits of the outer sheath beyond the terminal extensions. The primary interactions in the outer sheath are actually along individual helical protofilaments - i.e. one subunit interacts with the

subunits above-right, and below-left of it.

The inner sheaths rings display complementary surface charge, further suggesting that they are self-assembled in a hydrostatically driven manner. In effect, each disk could be considered a bar magnet, with an electrostatic 'north and south pole' (really an electrostatic dipole), ensuring they assemble correctly in a head-to-tail fashion (Ge et al., 2015a). The inner sheath monomers of the pyocin consists of 2 anti-parallel β-sheets, which are orthogonal to one another in strand direction by approximately 90°. It was noted that they form a similar structure to the well known 'jelly roll' or 'cupin' fold where 2 sets of 4 β-strands are opposed to one another (Richardson, 1981; Dunwell et al., 2004), but are actually thought to be unrelated, despite this domain being highly conserved in other viral and (to a lesser degree) cellular protein sequences. The 6 monomers combine to form one of the largest β-barrel structures to be resolved yet with 24 β-strands forming the inner circumference of the lumen (Ge et al., 2015a).

Ge and colleagues have recapitulated an often seen homology modelling approach for the core lumen of the pyocin, and compared its surface charge, and smooth bore to the orthologues from phage λ (Pell et al., 2009), the T6SS (Jobichen et al., 2010), and phage PS1. They observed that the inner lumen of the R-type pyocins are primarily negatively charged, consistent with its putative role in depolarisation of cells by de-protonating the cell interior. Conversely, the inner sheaths of phage are typically positively charged to assist in the conveyance of negatively charged DNA. As the electrostatic potential of proteins is, of course, extremely variable, according to the amino acid sequence and manner of tertiary fold, the Hcp monomers which comprise the T6SS inner sheath have been shown to be largely neutral overall. This then enables these systems to convey all manner of protein cargoes, without any 'selectivity'. Chapter 3 on page 95 replicates this process in the context of the PVCs sheath proteins, to take a first look, albeit *in silico* at this stage, at the sheath structures and any potential relation to their cargo.

The inner and outer sheaths interact primarily electrostatically, with a small triangular region of negative charge on the outer sheath (on 2 'attachment' helices that protrude upward) corresponding to a triangular region of positive charge on the inner sheath. This causes the inner sheath that corresponds to a given tier/disk in the outer sheath to be offset

upwards by roughly 15 Å, and thus an outer sheath monomer straddles 2 inner tiers. Ge and colleagues showed that this reversible electrostatic interaction is important, as the outer sheath increases in diameter upon contraction, and detaches itself from the inner sheath, enabling it to protrude beyond the end of the outer sheath in order to execute its function and traverse the membrane of a target cell.

All that remains of the structure, the spike complex, baseplate, and tail fibres, were not well resolved in the Ge et al. study unfortunately. They obtain reasonable densities for the proximal baseplate, being able to identify 'spokes' which connect it to the spike, but do not speculate on, nor provide, further detail or its atomistic structure. It is evident from Figure 1.9A on page 25, however, that the baseplate is much stripped down versus the T4, mirroring the streamlining that is also seen in the removal of the capsid, long fibres, and replicative machinery, serving purely as a mounting point for the tail fibres seemingly. As with the baseplate, for the spike complex, detail was lost as a result of their averaging process. Fortunately, its density is also easy to identify from Figure 1.9A on page 25, and Ge et al report that they were able to locate a co-ordinated metal ion in the tip (typically iron or zinc), which is a hallmark of gp5-gp27 and VgrG-like spike proteins (Shneider et al., 2013; Kube and Wendler, 2015a; Browning et al., 2012).

In summary, the structure of the R-type pyocins appears to more accurately reflect the simplicity that is seen in the PVC operons, following a streamlining process associated with non-replicative entities. From the studies to date however, pyocins have only ever demonstrated anti-prokaryotic activity, while on the other hand, PVCs have only ever demonstrated anti-eukaryotic activity. Now, while this may be due to not testing each complex against an exhaustive repertoire of prokaryotes and eukaryotic cell types, these specificities seem to fit with what is known of their basic biology. This means that there is still much to be discovered about what makes PVCs different, and allows them to act on various higher order cell types in the few genes that are remaining without fully understood functions.

1.2.3.3 Of PVCs and the *Serratia entomophila* "Antifeeding prophage"

This brings us to possibly nearest cousins of the PVCs - the so-called "Antifeeding Prophages" of *Serratia*. The Afp was, until the advent of the Ge et al. (2015a) paper, the best characterised, closest relative, of the PVCs, and much of what was hypothesised about them was based on analogous experiments on the Afps, borne on a plasmid of *Serratia entomophila* (Rybakova, 1994). As its name suggests, like *Photorhabdus*, *S. entomophila* is another common insect pathogen, and they have been shown to be quite closely related (Duchaud et al., 2003; Sproer et al., 1999; Brillard et al., 2002). *S. entomophila* causes "Amber Disease" in the New Zealand grass grub *Costelytra zealandica* specifically, and has been used for some time now as a biopesticide (Chattopadhyay et al., 2012; Opender Koul, 2011). Electron microscopy studies of purified particles revealed similar morphologies to the Pyocins and PVCs:

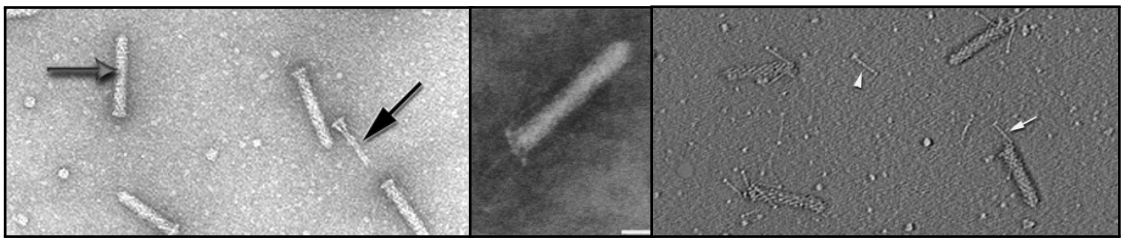


Figure 1.10 | Electron Micrographs of the Antifeeding Prophage of *S. entomophila*.
The left panel shows EMs of the Afps, revealing their "bullet like" shape, and similarity to PVCs. The grey arrow denotes a mature, fully intact Afp particle. The black arrow highlights a "Tube-Baseplate Complex". Adapted and reproduced from Sen et al. (2010). The centre panel shows a close up of an individual mature Afp, it is just possible to make out the skirt-like formation of the baseplate, and a couple of tail fibres, including one pronate against the tube. Adapted and reproduced from Hurst et al. (2007a). The right hand panel shows more intact Afp particles, and particularly reveals the tail fibre like structures, of which multiple can be seen on any one particle, and some can even be seen loose on the grid (white arrows). Adapted and reproduced from Heymann et al. (2013).

The Afps were discovered on the 153,404 bp pADAP ("Amber Disease Associated Plasmid") plasmid (Hurst et al., 2011) due to their pathological effect against *C. zealandica*. The plasmid has been shown to contain other virulence factors such as the *sep* toxins (Hurst et al., 2000), which are the aetiological agents of "Amber disease" (and of which *Photorhabdus* also has analogues - in fact, one such *sepC* analogue is a cognate PVC effector). It was observed in the *sep* studies that another large locus on the plasmid caused a cessation of feeding effect 1 to 3 days after ingestion. In later efforts, this was identified as the "Antifeeding prophage", and hence it earned its name (Hurst et al.,

2004). Over almost 10 years, 4 primary papers were published which steadily elucidated the genetic components and pathology (Hurst et al., 2004), the regulation (Hurst et al., 2007b) and the structural basis of Afp complexes (Sen et al., 2010; Heymann et al., 2013). Additionally, a number of papers were able to identify putative biological roles for some of the more enigmatic proteins in the locus Rybakova et al. (2013, 2015a). The presence of the Afp on the pADAP replicon was fortunate, as it allowed the whole operon to be subcloned in to lab *E. coli* replicons with relative ease (Hurst et al., 2004). This has allowed quite extensive deletion/mutation studies to be conducted, as well as providing the material for structural resolution. To date, no PVC equivalents have been identified on plasmids in *Photorhabdus*, though in *P. luminescens* TT01, 4 PVCs appear tandem to one another, surrounded by conjugation machinery and partitioning proteins such as *mukB*, which may be suggestive of an ancestral recombination event between a large plasmid and the chromosome (Yang et al., 2006).

The Afp operon is comprised of 18 proteins, termed Afp1-18. Analogues to sheath proteins, spike complexes, baseplate proteins and tail fibre proteins were able to be identified bioinformatically upon first sequencing. A number of proteins were matched to *Photorhabdus* orthologues with unknown functions, revealing the close relationship between these 2 loci, though much of the operon remained poorly understood. Efforts by Rybakova and colleagues were able to shed some light on the roles of Afp14 and Afp16 in the control of tail assembly. In 2013, the function of Afp16 was determined to play a role in the tail length termination process, and stabilised the growing tail tube (Rybakova et al., 2013). Full deletion of this protein resulted in aberrant forms of the Afp, with variable lengths, as well as formation of so-called "Tube-Baseplate complexes" (TBCs), which lacked much of the outer sheath, but were able to form a truncated inner sheath and seemingly full baseplate arrangement. Trans-expression of Afp16 did not restore a fully matured morphology to the Afps, suggesting that the expression patterns within the operon itself are also key to the self-assembly process, though exogenously applied purified Afp16 to pseudo-denatured Afps did exhibit some restored assembly - though again, not full length. Truncations of the C-terminus of the protein resulted in an intermediate morphology between the TBCs and a fully matured particle. It is still

unclear at present how these proteins interact in order to produce the 'finished product' however.

In 2015, Rybakova and colleagues were further able to elucidate the role of another enigmatic protein in the formation of the tail tube, this time identifying an analogue of a putative tape measure protein. Truncations of the protein resulted in concomitant shortenings of the elaborated Afp particles, and similarly, elongations of the sequence resulted in particles of increased tail length. Remarkably, there is a near exact linear relationship (R^2 = 0.92) between the length of known tail tubes and their associated tape measure proteins (Rybakova et al., 2015a; Pedulla et al., 2003). Tape measure proteins are difficult to detect via homology alone, as their sequence appears to not be particularly restrictive to function, with no obvious conservation of known phage tape measure domains. The only real hallmarks that have been identified between varied orthologues are a relatively conserved distribution of hydrophobic residues, and atypically high degrees of alpha helical secondary structure.

As with the PVCs, at least one of the genes at the 3'-most end of the operon (Afps 17 and 18) are predicted to encode toxic effectors, though unlike the PVCs, there do no not appear to be many variant forms. This is probably one of the reasons that *S. entomophila* maintains a very specific pathogenicity against the grass grub. The Afps, by virtue of their toxic cargoes have been shown to have an LD_{50} of as little as 500 individual Afp particles (though with the potential for multiple toxins to be present per Afp), against an entire insect (Rybakova, 1994).

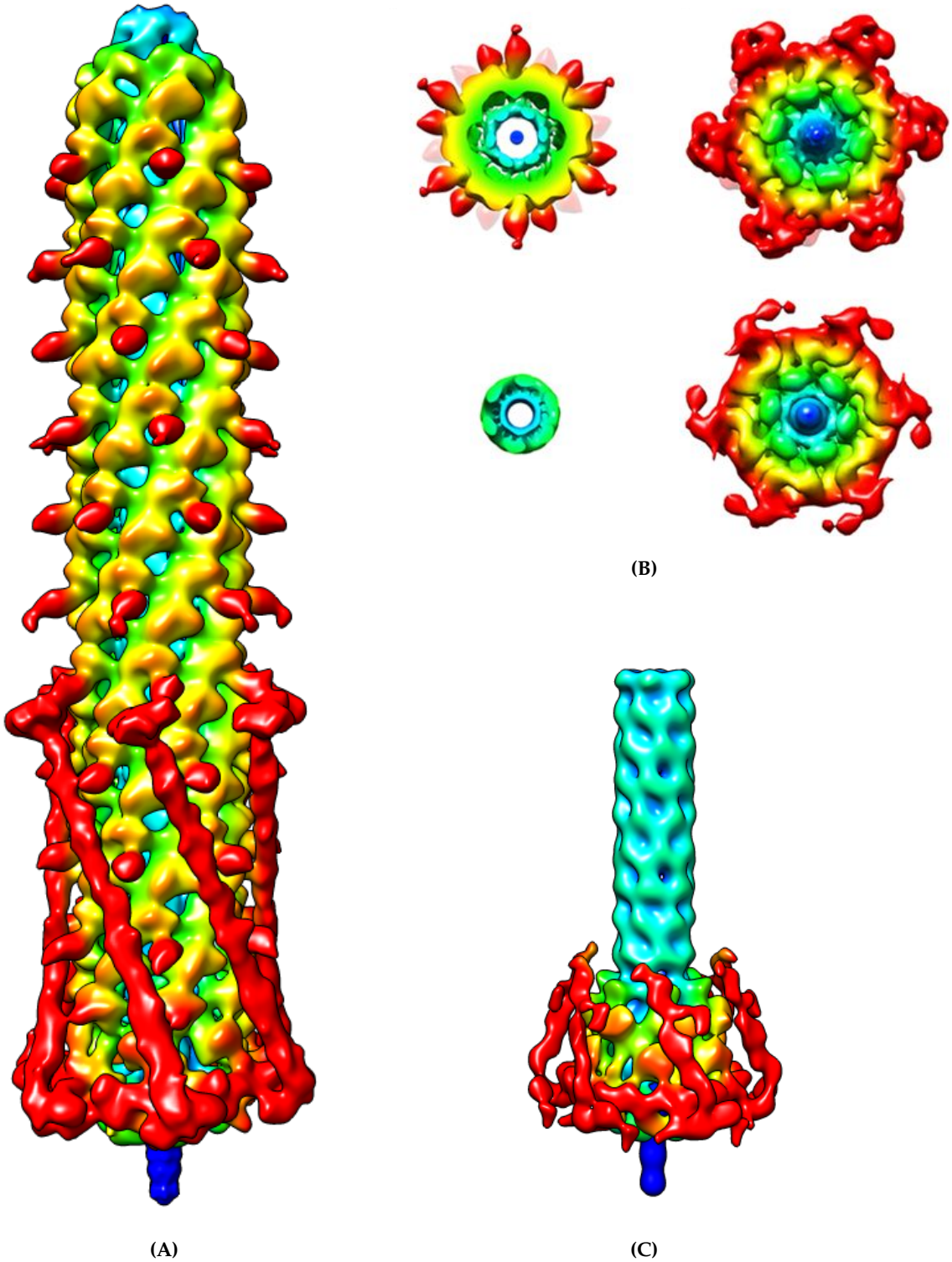


Figure 1.11 | The structures of the Antifeeding prophage from Heymann et al. (2013).
(A) The reconstructed 20 Å electron density map for the *S. entomophila* Antifeeding prophage, based upon Heymann et al. (2013), and reproduced independently from the deposited data under EMDB-2419. All panels in this image are coloured by the distance in Ångstroms from the centre of rotational symmetry (blue ≤ 20 Åto red ≥ 100 Å). Some features of note include the dark read sheath protrusions and the very well defined putative tail fibres folded back against the tube. **(B)** Various orthogonal views of the tube baseplate/spike and inner core. Note, in particular, a density in the lumenal space in the top left panel. Adapted and reproduced from Heymann et al. (2013). **(C)** A "Tube-Baseplate Complex" which was expressed without any exterior sheath proteins in the same study, revealing further detail of the baseplate complex and the inner sheath.

Despite this extensive study, the EM map that was obtained, displayed in Figure 1.11 on page 32, is low resolution (at only 20 Å), and only the gross architecture of the spike and tubes are reliably discernible. This was a substantial improvement over previous iterations however, as the group were able to correct an initial erroneous observation that the tube would have four-fold symmetry, when in actuality, it has six-fold (Sen et al., 2010). Despite this lower resolution, the Afp map does have some unique features, and even advantages over the atomistic R-type pyocin map. As with the other structures that have gone before, the mesh-like structure of the outer sheath is revealed in the obtained density, with the inner sheath visible through 'fenestrations' in the outer sheath structure. A baseplate and spike complex is clearly visible, though with no strongly discernible features at this resolution. Attached to the baseplate however, are incredibly distinctive densities for the putative tail fibres, present in a kind of 'docked' or 'folded' conformation. The fact the tail fibres have been locked in to a prostrate position along the length of the tube, has likely stiffened them, allowing them to be imaged successfully without the averaging effect of tomography blurring them out, as is the case with the structure from Ge et al. (2015a). Indeed, in Figure 1.11C on page 32, the lack of the outer sheath stabilising the distal ends of the fibres has resulted in the commonly seen blurring effect. Even at a 20 Å resolution, it is possible to identify a bulbous region at the distal ends of the tail fibres, which is consistent with the trimeric nature of other, more fully resolved, viral adhesion proteins.

Another striking feature of the Afp structure versus the R-type pyocin and the T4 phage, is that the outer sheath appears to more closely resemble that of T4, due to having long sheath protrusions (visible in dark red in Figure 1.11A on page 32), than it does the R-type pyocin. This structure, combined with the initial suspicion of four-fold symmetry lead Sen et al. (2010) to conclude that the Afps may represent an evolutionarily distinct sub-type of contractile tail structures. Whether or not this is valid in the context of the protrusions being unusual when compared to the R-type pyocin for example, is not clear without a fully resolved atomistic structure. It is clear that the argument from four-fold symmetry is erroneous in light of the more recent and higher resolution studies of Heymann et al. (2013) however. At present, there is no known functional relevance for

these domains.

Finally, there is one particularly interesting feature of the EM densities obtained by Heymann and colleagues. In Figure 1.11B on page 32, in the top left inset panel, a dark blue density can be seen in the lumen of the central tube. This presents a couple of possible explanations. Perhaps the most likely explanation is that it is artifactual from the averaging process, given that this axial region would not move greatly during tomography, and would thus appear as a static, but blurred, part of the structure.

Alternatively, it is possible there is a structural or biological basis for these densities. As was mentioned in the review of T4, caudate structures are proposed to require a tail 'tape measure' protein which extrudes along the length of the growing tail and triggers capping. In T4, this has been proposed to be gp29, and through deletion studies, a similar role was observed for Afp16 (Rybakova et al., 2013; Abuladze et al., 1994; Katsura, 1987). One current theory is that these tape measure proteins exert their effect by lying along the length of the interior of the tube, though there is sparse evidence for this particular mechanism. If this were the case, this density may well correspond to a tape measure protein.

Lastly, the Afps, like the PVCs are thought to package payload effector molecules in to the interior of the tail. This is an intriguing prospect, and would represent the first structural data that attests to this. Given the uniformity of the density along the length of the tube, and its width of only a few Ångstroms however, the former of these 3 theories seems like the most likely given the information at hand.

In summary, the Afps are unsurprisingly very similar to the PVCs given the relatedness of their host genomes. However, as this section as highlighted, they are not without differences, corresponding to potentially drastic differences in selection pressure and deployment in the environment. Chief among the differences are the fact that the Afps are plasmid borne, and the PVCs aren't (though perhaps once were). Moreover, the PVCs are present in various forms, scattered throughout *Photorhabdus* genomes, and this alone has potentially lead to enormously different selection pressures, and potentially morphology - whereas *Serratia* is limited to one example.

1.2.3.4 Of PVCs and Type VI Secretion Systems

Now that the closest cousins of the PVCs have been discussed, in the form of the AFPs and R-type pyocins, moving back up in complexity brings us back to another well studied biological complex - the Type VI Secretion System (T6SS). Bönemann et al. (2010) were the first to draw parallels between the T6SS and the PVCs, realising that the contractile, puncturing mechanism of the system placed it in a 'supergroup' of contractile injection systems.

The T6SS is just one of a family of secretion systems which have come to be recognised in bacteria, as a mechanism for the organism to communicate with, and manipulate, the extracellular environment, including other organisms. At the time of writing, at least 9 "Type *x*" secretion systems have been described (numbered in Roman numerals I to IX), each of which has been studied to a differing degree and there are great number of reviews covering some or all of them to date (Dalbey and Kuhn, 2012; Chang et al., 2014; Bleves et al., 2010; Desvaux et al., 2009; Abby et al., 2016; Costa et al., 2015; Goulet et al., 2004; Remaut et al., 2008; Gerlach and Hensel, 2007; Abdallah et al., 2007; Green and Mecsas, 2015). The "Type *x*" secretions systems are specialised bacterial structures, and are distinct from the Sec and Tat secretion systems which are present in all 3 domains of life, meaning bacteria exhibit a dizzying array of secretion mechanisms (Green and Mecsas, 2015). Discussions of secretion systems are quite 'murky waters' however, as they are not all related, despite being named as if they are 'shades of grey' with respect to each other. As the details of all the secretion systems are not wholly pertinent to discussions of the PVCs, the depth will be left to the aforementioned reviews. This section will highlight the different types and diversity of known secretion systems, and will then proceed to cover the Type VI secretion system in depth.

1.2.3.4.1 The "Type *x*" Secretion System repertoire

Briefly, the T1SS, is a translocator comprised of 3 proteins which is able to secrete a wide variety of bacterial proteins with a wide size range, the one of the largest being the LapA adhesion protein from *Pseudomonas fluorescens*, at an impressive 520 kDa (Boyd et al., 2014). One of the most commonly transported proteins are toxins of the RTX family

(Delepelaire, 2004). Unlike the other secretion systems, Type I is related to the general class of ABC transporters which are ubiquitous efflux pumps for antibiotics and other small molecules, and is entirely independent of the Sec system, not requiring a first translocation of cargo to the periplasm, though cargoes do require a chaperone.

The Type II Secretion system (T2SS) is a common Gram negative secretion system, and has been studied extensively in a number of human pathogens including *Vibrio* and *Pseudomonas*, though it is not ubiquitously present in Gram negatives (Douzi et al., 2012). The system can be divided in to 4 primary components, though the structure as a whole is a large multipartite protein complex. The inner membrane complex and outer membrane complex are connected by a 'pseudopilus' spanning the periplasm, so called as it is made up of a number of proteins resembling pilins (Korotkov et al., 2012). Finally, a crucial hexameric secretion ATPase is associated with the inner membrane complex, and is responsible for the synthesis and dismantling of the pilins, which provides the mechanistic basis for secretion (Lu et al., 2013). Type II is Sec or Tat dependent, and exports a variety of protein cargoes, which often include toxins and degradative enzymes such as proteases and lipases associated with bacterial infection (Korotkov et al., 2012).

An unusual version of a secretion system, and another very well studied apparatus, the Type III Secretion System is quite similar to a PVC in its role as an anti-eukaryotic needle complex delivery system. Homologous to the basal body of the bacterial flagellum (Aizawa, 2001), the T3SS is another molecular syringe, but membrane bound. The Type III is used by bacteria to directly inject effector proteins in to the interior of target eukaryotic cells, making it a potent and widely utilised virulence factor (Abu Hatab et al., 1998) - with examples having been found in *E. coli*, *Shigella*, *Salmonella*, *Vibrio*, *Burkholderia*, *Yersinia*, *Pseudomonas*, as well as a number of plant-associated species such as *Rhizobium*, *Erwinia*, *Ralstonia* and *Xanthomonas*, and more besides. By forming a continuous pore, through the needle bore, from the cytosol of the bacterium to the target cell, Type III is completely Sec/Tat independent. The similarity to the flagella continues in the needle body, as this is homologous to the flagella hook (Lane, 2007). Comparable in complexity to the flagella also, the T3SS is comprised of around 30 distinct proteins, making the T3SS among one of the most intricate secretion systems (Green and Mecsas, 2015).

Like the T3SS, the Type IV Secretion System is a relative of another fundamental bacterial 'appendage' - the conjugation pilus, used by bacteria to exchange genetic material. Unlike the conjugation machinery however, the T4SS is capable of translocating protein (as well as nucleic material). The Type IV system was discovered originally in *Agrobacterium tumefaciens*, the long-used tool for genetic manipulation of plant species, and is the mechanism by which the bacterium actually exerts its modifying effects. Thus, the *A. tumefaciens* system in particular, has become the model for T4SS structure and function studies (Bundock et al., 1995). As with the Type III, the 'injectisome' nature of the T4SS means that it is Sec/Tat independent. However, there are competing theories as to whether the T4SS simply acts as a harpoon, to pull 2 cells in to close register, and translocation occurs in a still as-yet-undetermined manner, or actually forms a continuous channel from cytosol to cytosol, as in the T3SS (Christie et al., 2005; Green and Mecsas, 2015).

Unique among all the secretion systems, The Type V Secretion System is an autotransporter, rather than a channel for other proteins (though it is capable of exporting others as well in some cases), and requires no ATP to function (Thanassi et al., 2005). Proteins that comprise the T5SS class contain a C-terminal region which inserts into the outer membrane (after translocation via Sec), forming a β-barrel, and they then proceed to translocate the N-terminal passenger effector domain, which is proteolytically cleaved. The β-barrel remains in the outer membrane until it is lost or recycled, potentiating the passage of other substrates, once the passenger domain is no longer causing an obstruction. Continuing the theme from the other secretion systems, most known T5SS secreted proteins are virulence factors and host modulators (Green and Mecsas, 2015).

Skipping over the Type 6 for the moment, in favour of a more full review in this section, the Type VII secretion system is unlike the other secretion systems mentioned so far, as it (to date) has only been found in Gram positive bacteria such as the *Corynebacteria* and *Actinobacteria*, and most famously in the *Mycobacteria* (Ates et al., 2016). Gram positive bacteria, due to their (typically) single cell membrane, and thickened cell wall, have different challenges to overcome when secreting molecules in to the extracellular milieu (Green and Mecsas, 2015). It is thought that the T7SS is widespread amongst Gram positives, with Type VII-like operons and orthologues having been detected in

Staphylococcus aureus, and *Bacillus subtilis*, the well known model organism for Gram positives. The structure and function of the complex and its constituent proteins are not yet well understood. There is a large inner membrane complex, formed of at least 5 distinct proteins, which is thought to provide a channel for substrates, though there are a number of additional proteins for which roles have not yet been elucidated. Since the formation of a pore in the membrane would not allow substrate passage beyond the interior face of the cell wall, it seems likely that some or all of these remaining proteins serve to facilitate this last hurdle in some way. Though extremely distant in terms of any genetic relation (if any), there is an interesting parallel between the PVCs, and the T7SS in *M. tuberculosis*; namely, that the *Mycobacteria* harbour up to 5 T7SSs, as *Photorhabdus* habours up to 6 PVCs, and not all of these are present in every genome of the species (Bottai et al., 2017). The T7SS is known to function mostly (though not entirely) in virulence, and this potentially speaks to the same diversification seen in the PVCs, honing multiple copies of highly effective virulence factors to become a more effective pathogen, or cope with varying environmental conditions.

Historically, the Type VIII system has been referred to as the 'extracellular nucleation-precipitation pathway (ENP) and the switch to T8SS was proposed by Desvaux et al. (2009). The structure of the T8SS was resolved by Goyal et al. (2014), and comprised a fairly typical looking membrane 36-strand β-barrel which is embedded in the outer membrane. The T8SS, therefore, is Sec/Tat dependent. Unlike the majority of the other secretion systems, the Type VIII is thought to be limited to a single substrate. It is responsible for secreting proteins known as 'curli' - a primary component of the extracellular matrix of many *Enterobacteriaceae* (Barnhart and Chapman, 2010).

The Type IX Secretion System is one of the most recently discovered secretion systems with only a single structurally resolved component. To date, it has only been detected in certain species of the *Bacterioidetes* phylum, after being originally discovered in the oral pathogen *Porphyromonas gingivalis*. The T9SS has been demonstrated to be implicated in 2 distinct lifestyle roles, both gliding motility and as a pathogenic virulence factor/weapon, though with unknown functional bases. It is dependent on the Sec system, providing only carriage across the outer membrane. Originally termed the PorSS system, 18 proteins

are known to be essential, but roles for all of them remain elusive, while as many as 29 proteins are hypothesised to be involved in some way, making the T9SS comparable in complexity to the Type III and Type VI (Lasica et al., 2017)

So, finally returning to the Type VI secretion system, as with the T4 capsid, this section is not going to dwell extensively on the membrane associated apparatus of the Type 6 Secretion System, since the PVCs appear to be secreted/released by lysis, and thus contain no analogous structures (and the membrane complex is still not well understood). Instead the similarities of the PVCs and the T6SS in terms of their putative translocation role and thus their spikes and tubes (i.e. as contractile nanomachines) will be the focus.

The T6SS has been identified in about 25% of all Gram negative sequences (Basler, 2015b), and despite being the most recently discovered secretion system (first being dubbed the T6SS in 2007) (Nguyen et al., 2018; Pukatzki et al., 2007; Cascales and Cambillau, 2012), it has rapidly become quite well studied, with a significant amount of structural resolution completed to date (Mougous et al., 2006). It has been shown that the T6SS is encoded by a highly conserved 13 gene cassette, which forms the core of the system, with a number of accessory proteins. The presence of these accessory proteins can vary by organism, but are typically well conserved when they are found (Basler, 2015b).

Among all the secretion systems, the Type VI is unique in a couple of primary ways. Firstly, it is the only secretion system with a contractile mechanism, as with T4 and the pyocins etc., as well as being the only system which delivers effectors to both other bacteria, and eukaryotic targets. Thus, the T6SS is an intricate but highly versatile nanosyringe complex - essentially an "upside down myophage in the membrane" - and is employed by a large number of bacteria in their pathogenic, but also community roles (Russell et al., 2014).

Though its role in pathogenesis was determined first by Pukatzki et al. (2006) whereby the T6SS locus of *Vibrio cholerae* was demonstrated as enabling the bacterium to resist predation by the model amoeba species *Dictyostelium discoideum*, the T6SS is deployed predominantly against prokaryotic targets (Green and Mecsas, 2015; Russell et al., 2014; Hood et al., 2010). A wide variety of roles for the T6SS have been postulated, both in antagonism (which is well documented) and in synergism. The diversity of competition

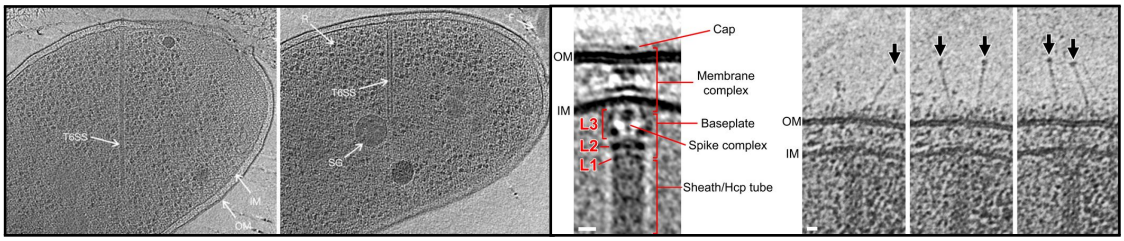


Figure 1.12 | Electron Micrographs of the Type VI Secretion System.
The left panel shows 2 EMs of the T6SS, and displays the enormous magnitude and variability in size of the tube complex, in the left most inset, the T6SS is labelled. In the right inset, the T6SS is labelled again, along with putative ribosomes (R), a flagellum (F), storage granule (SG), and the inner and outer membranes (IM/OM) Adapted and reproduced from Basler et al. (2012). The right hand panel shows an annotated close up of the T6SS membrane complex. Labelled are the Inner and outer membranes (IM/OM), cap, membrane complex, tube, spike complex, baseplate, and the black arrows in the left most tryptic identify 'antennae', purported to be the T6SS equivalent of phage tail fibres. L1-L3 demarcate different layers of EM density. Adapted and reproduced from Chang et al. (2017).

against which the T6SS might be deployed 'in the wild' is thought to underscore the rampant diversity that is seen between homologues. Furthermore, the need for competition between extremely closely related species and even strains, is driving the underlying selective pressure that has resulted in an enormous variety of Type VI effectors (English et al., 2012; Russell et al., 2012). The effector/immunity protein pairs that typify Type VI effectors have been suggested to act in a number of subtle ways, given this diversity. A rather ingenious mechanism has been proposed, wherein target cells which harbour a cognate immunity protein for a given toxin, utilise the toxin-immunity complex as a signalling molecule. Thus, those which have the correct immunity protein receive a signal, whereas those that don't, receive an antagonistic 'message' - as if the bacteria are mailing each other 'booby trapped' messages (Russell et al., 2014). Among other synergistic roles, the T6SS has been implicated in: the determination of self vs. non-self in *Proteus mirabilis* (Gibbs et al., 2008; Wenren et al., 2013); triggering 'assisted suicide' in phage infected cells inter-cellularly, in a manner analogous to that shown for 'classic' Toxin-Antitoxin systems (Hazan and Engelberg-Kulka, 2004); and as a method for overcoming the outer membrane to deliver cell wall remodelling factors which have been shown in other systems to 'resuscitate' neighbouring cells from viable-but-non-cultureable states (Downing et al., 2005; Mukamolova et al., 2006).

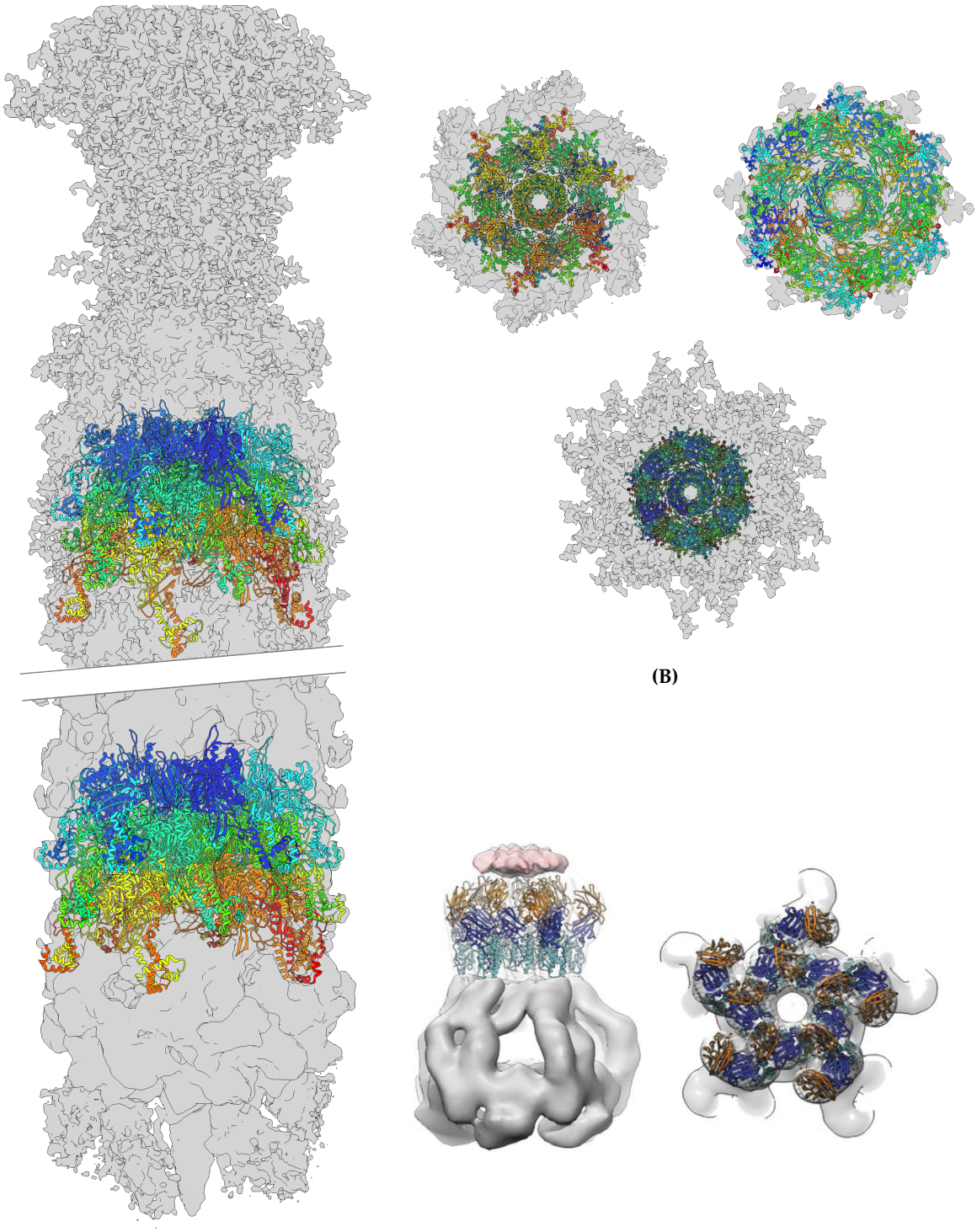


Figure 1.13 | The structures of resolved components of the Type VI Secretion System.
(A) The EM density for the proximal and distal ends of the pre-contraction Type VI secretion tube. Figures were reproduced independently from the deposited data under EMDB-3878 and 3879 from Nazarov et al. (2017). The density shows the spike complex and a distal cap like structure as well as the atomic architecture of the sheath. **(B)** Top left - a bottom view of the proximal end of the tube (spike toward the viewer). Top right - a slice through the centre of the tube, demonstrating well the dodecameric spokes in different planes. Middle bottom - a view from the top of the cap-like protrusion density. All also reproduced independently from Nazarov et al. (2017) EMDBs 3878/3879. **(C)** The 12 Å map of the membrane complex of the system, with a TssM/J complex fitted in to the upper arches. Adapted and reproduced from Durand et al. (2015).

The T6SS is broadly divided in to approximately 4 structural complexes - a contractile tube (one could make the case that the spike complex forms a 5th component), a baseplate complex, a transmembrane domain, and associated soluble proteins. As a contractile tail system it, of course, exhibits a pair of concentric tubes, the inner of which is tipped with a spike complex, as with the other systems discussed so far. The inner tube is comprised of Hcp ("Haemolysin coregulated protein") hexameric toroids; Hcp being the ortholog of gp19/Afp 1 & 5. The Hcp tube is approximately 80 Å in outer diameter, with an inner lumenal diameter of roughly 40 Å (Mougous et al., 2006). Multiple crystal structures of Hcp orthologues and paralogues have been solved, and reveal the same overall gross architecture, though there is often some flexibility in the secondary structure and even more so in sequence. The inner tube appears more similar to that of R-type pyocins and the PVCs as it doesn't exhibit a helical turn, instead just being direct stacks of Hcp (Silverman et al., 2012; Mougous et al., 2006; Osipiuk et al., 2011). It was originally thought that Hcp itself was a secreted molecule, though without a known function, though it is now realised that the Hcp proteins may dissociate when puncturing in to the interior of another cell, and the filament of the tube is exposed to the extracellular environment during contraction. Thus, any dissociated Hcp monomers are essentially secreted 'inadvertently', though this could obviously not be determined in the early experiments. In the T6SS of *Edwardsiella tarda* it was also observed that the VgrG spike protein is secreted, but if either *Hcp* or *VgrG* homologues were knocked out, neither could then be detected in the supernatants of T6SS$^+$ cultures. The generally accepted, and likely, explanation for this is that Hcp tubules only assemble and secrete/puncture through the membrane if first polymerised off the VgrG baseplate hub analogous structure, as is the case in the T4 phage. Similarly, VgrG can only be detected in supernatants if Hcp is present to form the tube to which VgrG binds, and then is carried across the cell envelope by 'riding' the tube out of the cell (Pukatzki et al., 2007; Zheng and Leung, 2007; Hachani et al., 2011). High resolution microscopic studies with the help of fluorescent constructs have been able to demonstrate that multiple T6SS complexes can be carried by a cell at once, and the tube complexes can extend many hundreds of nanometers, deep in to the cytosol of the cell - even reaching the opposite cell envelope (Nguyen et al., 2018; Chang et al., 2017).

Unlike the other systems mentioned up to this point, the exterior sheath of the T6SS differs quite substantially in structure, despite still forming a contractile system. As can be seen in the upper right panel of Figure 1.13B on page 41, the outer sheath is actually a dodecameric toroid, comprised of 2 different proteins: TssB and TssC. Type 6 secretion proteins have been given alphabetic nomenclature preceded by "Tss" (Type six subunits), as such, the outer 2 sheath proteins for example, are termed TssB/C, and though unrelated, would be the structural orthologues of gp18 in T4. TssB is a smaller protein of around 18 kDa, and TssC comprises the bulk of the tube at ≈55 kDa. Together they form an alternating dodecameric ring which is approximately 240 Å across in the extended state and 290 Å in the contracted form, with an inner diameter of 80Å before contraction, and approximately 110 Å post-contraction (in the *V. cholerae* T6SS structure) (Cascales and Cambillau, 2012; Wang et al., 2017a; Kube et al., 2014b). If one were to think of the ring as an analog clock face, a TssB monomer would occupy all the odd numbers, and a TssC monomer would occupy all the even positions (Kube et al., 2014b). In the extended form, the sheath has a 38 Å rise, and a 23.6° twist, and this shifts to a 15.8 Å rise and 29.4° twist after contraction (Wang et al., 2017a). Both subunits appear to contribute to protrusions around the exterior of the tube which form a left-handed helical series of ridges. This is not a trivial observation either, as this is the opposite handedness to the T4 phage tail. Such a dramatic rearrangement in structure further underscores the hypothesis/observation that T6SS outer sheaths are unrelated in origin to T4 phage (Kube et al., 2014b). As is the case with the PVCs and Afps, one of the most conserved proteins appears to be a ClpV AAA+ family ATPase. These are typically protease enzymes which take their name from being "ATPases Associated with various cellular Activities", and are a group of enzymes/chaperones which are able to cause conformational changes in an enormous variety of cellular proteins (Hanson and Whiteheart, 2005). It has been demonstrated that the ATPase is not entirely essential for T6SS function (at least in *V. cholerae*) but its role has recently been determined to be in recycling of the tube after contraction, utilising these sheath protrusions, ameliorating some of the significant cost to the 'cellular economy' of building such an enormous and complex structure. Consequently, the ATPase is required for repeated synthesis and discharge of the same T6SS complex, though cells are perfectly

capable of continuing Type VI mediated secretion, though an entirely new T6SS must be generated, essentially becoming 'single use' Basler (2015a). The ATPase itself forms a hexameric ring, and interacts with an α-helix near the N-terminal of TssC using its central pore, and dissociates it from TssB only in the contracted state, this causes the outer sheath to 'disintegrate', freeing up the subunits for recycling (Costa et al., 2015); in the extended conformation, the helix is obscured preventing premature depolymerisation. The presence and role of the ATPase within Afp and PVC operons remains a mystery, since there is less rationale for a need to recycle the components which are released from the cell completely. Due to its high level of conservation and readily identifiable domain family, despite not being entirely essential, ATPase presence and recycling is now considered a hallmark of Type VI secretion (Nguyen et al., 2018).

Sitting atop the tube complex is a PAAR spike-tip protein and VgrG spike complex which is homologous to the gp27-gp5 complex of T4, though lacks the lysozyme domain (a seemingly common 'deletion' outside of T4). Interestingly, in the Type VI, due to the huge diversity of operons in the vast number of sequences studied to date, there is now evidence that in certain cases the T6SS also employs so-called "evolved VgrGs" (Pukatzki et al., 2007; Suarez et al., 2010; Hood et al., 2010; Cascales and Cambillau, 2012). A slightly clumsy term perhaps, but the premise is that different VgrG spikes have, in effect, acquired domains for alternative enzymatic functions that can exert an effect on the target cell once they're translocated in to the interior. By doing so, the Type VI is able to deliver a 'double whammy' of delivered effectors from within the tube lumen, as well as a functionalised 'warhead'. Clearly, the VgrG is not just a wedge with which to separate the cell envelope, and similarly, the PAAR repeat proteins which sharpen the VgrG apex, are more than simply structural. Discussion of these proteins has been left until now, despite there being orthologues in most of the structures discussed so far (Sarris et al., 2014), as the T6SS appears to have among the most interesting collection of these tiny proteins, and some of the better characterised experimental data. PAAR proteins take their name from "Proline-Alanine-Alanine Repeats", which in concert with a coordinated metal atom, confer on the protein a triangular pyramidal shape. The T4 phage analogue of this protein is gp5.4, and they were initially identified for T6SS by Schneider et al., by

examining all small proteins (<23 kDa) associated with gp5 bearing genomes (Shneider et al., 2013). Amazingly, in many cases these PAAR spikes present a single amino acid side chain at the tip, making it as sharp as just a single atom or two (e.g in the PDB ID 4JIV, a lysine sidechain sits at the apex). This is not just a simple honing process to improve the T6SS puncturing efficiency however, the PAAR spike tip proteins have been shown in at least 2 studies to be essential for T6SS function (Shneider et al., 2013; Cianfanelli et al., 2016). The claim was made earlier that PAAR repeat proteins are more than simply structural, and indeed that is the case. A growing body of evidence has been able to identify a number of toxins which are bound to the PAAR spike tip, in a similar fashion to those associated with VgrG Hachani et al. (2014); Ma et al. (2017). Cianfanelli et al. (2016) additionally showed that VgrGs and PAAR proteins are not completely interchangeable, with certain combinations having a clear preference for one another, and moreover they had a specificity for the types of effectors they carried, to the extent that they define distinct 'versions' of the Type VI. All in all, this means that the Type VI, as well as being a 'loaded needle' can also act like a 'poison arrow', discharging a lethal tip, upon injection.

The baseplate complex of the T6SS has not been well studied to date, but is suggested to continue the homology to that of phage T4. The baseplate is known to comprise the proteins TssE, F, G, and K. TssK forms a trimer, and has had its structure resolved recently. Unusually, the protein loosely resembles a tail fibre like structure, with a 'head' and 'shoulder' region, connected by a neck/shaft-like region, though this has yet to shed any light on an actual role (Desmyter et al., 2015; English et al., 2014). It has been suggested, that the baseplate is formed of subcomplexes, akin to the baseplate wedge complexes seen in Figure 1.6A on page 20, and most likely follows a hexameric symmetry. A homolog of gp25, a protein which interfaces the spike hub complex and the tube proper in T4 has been identified in a Type VI locus from *E. coli* corresponding to TssE, which is essential for tube biogenesis (Nguyen et al., 2018; Brunet et al., 2013; Leiman et al., 2009). TssF and G are further proposed to form a complex reminiscent of gp6-gp53, which do form a significant part of the wedge assembly and would fit with the hypothesis that the baseplate will also exhibit 6-fold symmetry (Nguyen et al., 2018). Attempts to determine this by stoichiometry analyses have been frustrated so far however, and there is conflicting

information (Nguyen et al., 2018; Nazarov et al., 2017). It will simply be a matter of time before structural data of sufficient quality is obtained to answer this once and for all, and given the intense interest and rapid pace of research in the area in the last decade or two, it seems unlikely that it will be much of a wait.

The membrane bound components of the T6SS include TssL, M and J. TssL is present in the inner membrane, and TssJ is situated at the periplasmic face of the outer membrane. TssM is a large protein (1100 residues) and has been shown to interact with both TssL and TssJ, meaning that its most likely configuration is spanning the periplasmic space to bind the inner and outer components together, though any conclusive structural data is still lacking (Zheng and Leung, 2007; Ma et al., 2009; Felisberto-Rodrigues et al., 2011; Nguyen et al., 2018). Some gross architecture was uncovered in 2015, when a 12 Å map of the membrane complex was determined (Durand et al., 2015). The complex consists of 5 'pillars' and a TssM-J complex was able to be fitted in to the density to reasonable accuracy, though much of the structural information for the inner membrane proximal region is still lacking. Interestingly, this means that the T6SS displays 3-fold (VgrG complex), 6-fold (Hcp tube), 12-fold (outer sheath) and 5-fold (membrane complex) symmetries, which is unusual compared to the other systems studied so far, which are all 3 or 6-fold. There is some suggestion that this 5-fold symmetry may be an aberration however, since the recently resolved TssA protein, a putative sheath cap, was shown not to bind to C5 complexes, and conferred a 6-fold symmetry to the membrane complex via displacement (Zoued et al., 2016). As with the baseplate, it is unlikely that the architecture of the membrane components will remain a mystery for long.

In summary, the current operating hypothesis is that the T6SS has an overall architecture where the contractile sheath and spike complex is surrounded in the membrane space by a sort of pear-shaped, buttressing cage of struts. This membrane complex scaffolds the central tube core of the system, and complexes with the baseplate at the interface of the cytosol and inner membrane, though structural data relating to this interaction is still missing. 12 of the 13 identified 'core components' have been localised, if not structurally resolved, on at least a preliminary basis. The 13th, the ATPase, is a known cytosolic protein, which it needs to be to exert its disassembling role.

1.2.3.5 Of PVCs and their extended family

The known role, based on the earliest experiments on the PVCs, is as toxin delivery systems (Yang et al., 2006). However, as seen in the Type VI Secretion System, contractile tail structures are not limited to this function. In recent years there have been a number of unusual related systems discovered which demonstrate extremely diverse ecological roles, though 'secreted' caudate structures have only ever been implicated in lethality to date (Shikuma et al., 2014). This section will explore some further examples of enigmatic 'second cousins' of the PVCs. Many of these apparatuses are not well characterised and this is unlikely to be an exhaustive list, but these are some of the more unique and better studied examples which have evidence beyond simply matching in database queries like BLAST.

1.2.3.5.1 In *Pseudoalteromonas luteoviolacea*

As alluded to in the last paragraph, until very recently, 'tailocins', Afps, and the PVCs were the only known examples of 'secreted' caudate structures (not including phage) - and all of them have been observed to exert a lethal effect in, in one form or another, against the targeted cells. In 2014, this changed, as Shikuma et al. published structural data of a remarkable contractile complex produced by the marine bacterium *Pseudoalteromonas luteoviolacea*. Termed the "Metamorphosis Associated Contractile" Structures or "MAC" complexes, they observed that these incredible assemblies were the cryptic causative agent that drove the differentiation of the larvae of the marine tubeworm, *Hydroides elegans*, in to its juvenile form (Shikuma et al., 2014). This discovery was astounding for 2 particular reasons. Firstly, their discovery represents the first example of a beneficial interaction between a type of contractile structure, and a target organism. Prior to this finding, it had been observed that many marine organisms respond to bacterial associations, but the underlying mechanism(s) had not been uncovered (Hadfield, 2011). Shikuma and colleagues were able to demonstrate a differentiable phenotype with purified MAC complexes, unambiguously confirming its role. When they probed the structure of the MAC complexes, the second astounding discovery was made. Not only are the MAC complexes caudate contractile structures, but they are actually formed from an interlaced

hexagonal array of tail tubes, which are 'secreted' (released from the cell by lysis), and, in effect, form a kind of 'bed of nails'. The tails are all 'up-ended', such that the spikes face away from a substrate they are attached to, and the tails essentially interlock their arms, with 6 tail-fibre like proteins from each tube interacting with the 6 adjacent tubes, and each tube therefore contributes 6, and receives 6, points of contact with its neighbours.

In order to metamorphose, the tubeworm must lie on this bed of contractile nails, at which point contraction is triggered (by an as yet undetermined mechanism), and differentiation factors are delivered in to the larval worm. These differentiation factors still remain to be concretely identified, but in a followup paper, Shikuma and colleagues were able to narrow the possibilities down to a short stretch of sequence (≈8.2 kb), comprising 6 proteins sequences, in close proximity to the MAC operon, which were able to induce Mitogen Associated Protein Kinase (MAPK) based signalling cascades (Shikuma et al., 2016).

1.2.3.5.2 In Amoebophilus asiaticus

More recently, another example of a MAC-like structure was structurally elucidated, but this time in an amoeboid symbiont, *Amoebophilus asiaticus*. Unlike the MAC complex, this complex which the authors identify as an arrayed T6SS ("Type VI Secretion System$^{subtypeIV''}$"), is purely membrane associated. Nevertheless, it still resembles an interlocked 'bed of nails'. No obvious density could be imaged for any tail-fibre filamentous network like that observed in the MACs, though its possible that being embedded in the membrane like the T6SS, means that the collar/membrane complex region could be held in tight register by other means (Böck et al., 2017). The structure forms a tightly packed hexagonal array, putatively joined at the baseplates at the cytoplasmic face of the inner membrane, though the complexes are somewhat smaller. In the MACs, it was not uncommon to see arrays of 100 tail tubes, whereas the average for these T6SS^{IV}s is only 8.

However, Nguyen et al. (2018) observes that this structure appears to be a 'stunted' T6SS, as the tube is much reduced in length. Some other curiosities include the fact that it contains a tape measure protein, which a canonical T6SS does not, it has no known effectors, and is not recycled by an ATPase - most of which are considered hallmarks of

a 'true' Type VI. Thus, Nguyen et al disagree with the authors that this represents a new type of T6SS, and instead suggest it more closely resembles a membrane bound Afp. The criticism from Nguyen et al. is probably well founded, as even the authors note that the system is much more closely related to Afps/MACs and a similar complex in another intracellular mutualist, *Cardinium hertigii*, and lacks any real similarity to the T6SS at the sequence level.

A role for the *A. asiaticus* complex has not been determined fully, though its similarity to the *C. hertegii* structure, and the fact that both organisms share an intracellular lifestyle is thought to suggest they play a similar role. The *C. hertegii* complex is discussed in the next section.

1.2.3.5.3 In *Cardinium hertegii*

Very little is known about the structure of the Afp-like island that Penz et al. (2012) identified in the genome of the endosymbiont (of parasitic wasps) *Cardinium hertegii*. However, a putative role has been suggested. Bacterial symbionts of insects are well known, and perhaps the best understood example is *Wolbachia*. These symbionts are capable of exerting large scale physiological and developmental effects on the host (Hedges et al., 2008; Oliver et al., 2003). One of the best studied effects is "Cytoplasmic incompatibility". This has serious reproductive consequences; when a male harbouring the endosymbiont reproduces with an uninfected female, the embryos become non-viable and die very early in gestation. By doing so, a fitness cost is conferred against uninfected individuals thus promotes the survival of the endosymbiont (Werren et al., 2008).

In *Wolbachia*, a Type VI Secretion System is responsible for mediating cytoplasmic incompatibility, suggesting that factors derived from synthesis inside the cytoplasm of the endosymbiont are important causative agents of this phenomenon, and thus they must be translocated to the cytosol of the host. Penz et al. (2012) observed that there are no known secretion systems in *C. hertegii*, but they do harbour 16 Afp-like genes, though fragmented in to 5 different loci, rather than a single cassette. It is not known whether there are membrane-bound associated proteins which would be able to present these Afp-like genes in a more 'conventional secretion system' form, though this seems probable,

since it is unlikely that there is much need to secrete a whole tailocin like structure, when the bacterium is already within the cell. At present, there are no known toxins or other substrates for the putative secretion system either, but this does suggest another non-pathogenic utility to contractile tail structures. Furthermore, given the diverse pathways intracellular symbionts are known to be able to manipulate, it seems likely that this may represent another general purpose secretion system (Werren et al., 2008).

To summarise, it's evident that caudate structures appear to by widespread in all bacterial species, potentially somewhat 'enriched' in marine species, and are capable of serving a wide variety of ecological functions. Figure 1.14 on the following page shows a schematic overview of the similarities between these structures and their targets, collecting all the information of the past sections.

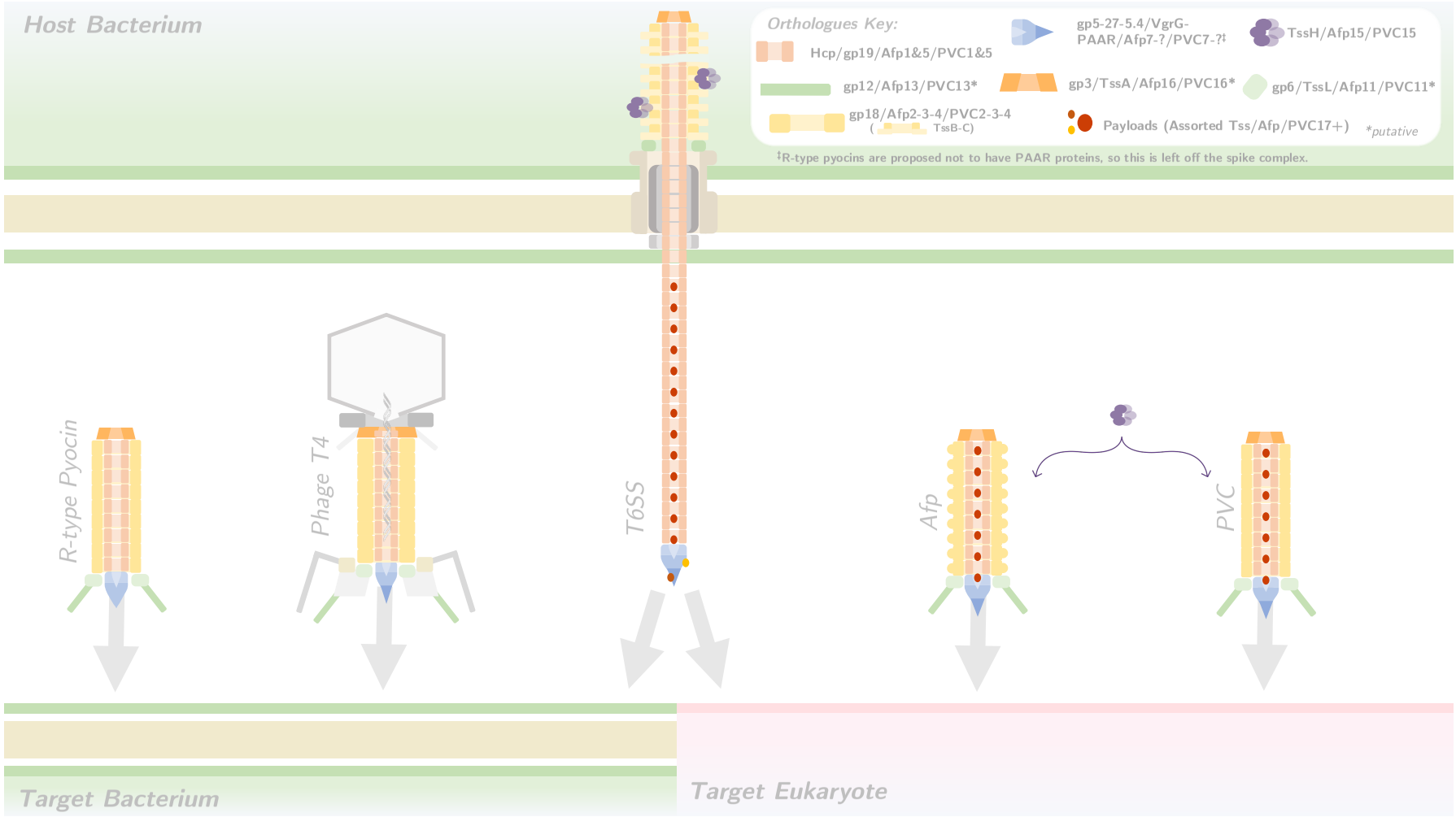


Figure 1.14 | Schematic of conserved architecture in various caudate structures.

A diagrammatic comparison between the conserved components of PVCs and other caudate structures discussed in this introduction. The inset key depicts structural orthologs, though they may not be ancestrally related. Neutral/gray colours identify proteins which are not shared between structures. Items in the key with asterisks are putative homologues. A question mark indicates that a homologous structure has been identified, but the protein is not yet known. Not to scale.

1.2.4 Mechanism of Action

The actual mechanism of contraction has been the subject of significant debate and speculation in recent years. It has been generally accepted that the pre-contraction state of the tube complex represents an energetically tensioned system (meaning that the often seen of references to this conformation as 'relaxed', is in fact, the exact opposite), and the prevailing hypothesis has been that the conformational change in the outer sheath proceeds in a wave-like fashion from the proximal baseplate, toward the capsid, and is driven by solvation free energy gain (Brackmann et al., 2017; Kube and Wendler, 2015a; Kube et al., 2014a; Moody, 1973).

As the pre-contraction state is therefore an energetically unfavourable conformation, it begs the question of how contractile tail structures are able to be assembled seemingly against the laws of thermodynamics. While there is still no definitive answer to this, despite the wealth of data now available, the generally accepted mechanism is that the baseplate is produced first, in a static form (not changing in the contraction process), and serves as a 'nucleation site' and a sort of intrinsic chaperone, allowing the first tier of the tube to polymerise off it. In effect, the baseplate is analogous to the pin in a mousetrap or a grenade, holding the conformation of the first tier in its a 'locked' and tensioned state. The logic then follows that the tensioned tier 1 of the tube acts as an 'auto-chaperone' allowing further tensioned forms of the tube hexamers (or dodecamers in the case of T6SS) to polymerise off it, also in a tensioned form (Kube and Wendler, 2015b).

For the T4 phage, it is now understood that the contraction of the tube is triggered by conformational changes transduced through the tail fibres, leading to large scale rearrangements in the baseplate. A contractile trigger mechanism for the Type 6 remains to be discovered, though the cryptic 'antennae' that can be seen in the right panel of Figure 1.12 on page 40, may suggest that the T6SS is prompted to contract in a similar manner, which would also allow the cell to sense when a target is in close proximity, such that the T6SS is not discharged aberrantly, causing significant cost to the cell to regenerate.

The contraction process is an enormously energetic process as a result of the release of the sheath tension. Not only is there a lateral translocation of the inner tube to provide the eversion, but the helical nature of the outer sheaths also applies a rotational torque

to the sheath, quite literally drilling it in to the target cell envelope (Kube and Wendler, 2015b). This leads some very impressive statistics. Wang et al. (2017b) calculate that for contraction of a "1 μm long sheath, composed of 260 rings, [the sheath] would push the inner tube by 420 nm and rotate it by 4.2 turns [...]. The overall amount of energy released during a single sheath contraction could be close to 44,000 kcal mol^{-1}". Vettiger et al. (2017) report that the entire contraction of the sheath occurs in just a couple of frames, even at 500 frames per second, and they conclude that the contraction speed is therefore >800 nm per millisecond. The Type VI is something of a special case given the lack of restriction on tube length, but Wang et al. (2017b) calculate, on a per-subunit basis, that "the free energy gained during contraction is 28.5 kcal mol^{-1} $subunit^{-1}$, more than 9.1 kcal mol^{-1} $subunit^{-1}$ calculated for [the] R-type pyocin". Exact contraction kinetics have not been observed for the other contractile mechanisms, but even the extrapolation here demonstrates that a significant amount of energy is expended by contractile machines to traverse these membranes (Brackmann et al., 2017).

1.2.5 The *status quo* of PVC genetics

With the superficial resemblances to analogous structures covered, this section will briefly highlight the state of knowledge about the gene modules within the PVCs specifically when this project began. This provides the 'jumping off point' for the upcoming Chapter, where the roles for many of these genes were probed further bioinformatically.

1.2.5.1 The PVC tail tube and sheath

As mentioned in previous sections, the tube and sheath components of the PVCs have long been among the best annotated genes within the operons, though this does not necessarily say much. Typical annotations, for instance from the first annotation of the published *P. asymbiotica* ATCC43949 genome, for the "LopT" PVC operon include "phage tail sheath" and "phage tail region" proteins. Belying the diversity of the PVCs which this thesis will continue to unpick, however, many of the operons still only picked up "conserved hypothetical protein" annotations, even for these proteins. Of some approximately 350 CDSs made up from 16 operons identified in 3 genomes (see Figure 1.4 on page 14),

around 250 of them had no functional information whatsoever, and those that did were almost exclusively the cognate toxins and some tube proteins. Furthermore, in the TT01 genome, every single locus for every single PVC operon had the descriptor "hypothetical protein", with no useful functional annotations whatsoever.

Nevertheless, with further inspection via BLAST and other tools, a good understanding of much of the PVC structure was able to be elucidated by Yang et al. (2006). The phage tail tube proteins were unambiguously identified, though this wasn't true for the whole operon. Some additional observations could be made, including the deletion of a sheath protein in 3 operons. This raises questions about why the PVCs maintain 2 inner sheath and up to 3 outer sheath proteins, when at least one is evidently non-essential (assuming all PVCs are fully functional), and other caudate structures typically do not.

1.2.5.2 The spike complex and baseplate

The original genome annotations tell a similar story for the putative baseplate complex. What has subsequently been identified as the PVC's VgrG spike protein homologue, has only previously been annotated as hypothetical. Other structural components of a putative baseplate were detected however, with several operons displaying annotations for "phage baseplate assembly proteins", "similar to baseplate protein gp25". Though once again, the underlying variability in the PVCs means that a number of operons were left with uninformative annotations still. It is likely that much of the 'dark matter' in the middle of the operon contributes to the baseplate apparatus, but with little to no orthology to anything previously detected in the databases.

1.2.5.3 The operon core

What is being termed here as the 'operon core' describes a number of single copy genes located in the centre of the operon, which appear relatively conserved, and potentially have important non-structural roles. Though there are only a couple of useful annotations for this region, the Yang et al. (2006) paper was able to identify some compelling orthologues. Firstly, a gene which putatively has a role in transcript regulation resides in approximately locus position 10 (though this varies if operons have deleted upstream

genes of course), but little else is known of its role, and the sequence identities are low (Waterfield et al., 2009). Immediately downstream of this is an unknown protein, with no functional information whatsoever.

The next gene, typically in locus position 13, is the putative tail fibre gene for the PVC complex. In the PVClumT operon of strain ATCC43949, this was attributed adenoviral tail fibre orthology. This is unusual, given the hypothesised bacteriophage origin of the structure, and there are no other known examples of this non-phage-like sequence similarity in other caudate structures, making the PVCs potentially unique. It has been demonstrated experimentally that the PVCs only exert toxic effects against eukaryotic targets (insect haemocytes), and have no antimicrobial activity whatsoever. It is possible that the reason for this that the PVCs have evolved anti-eukaryotic target recognition tail fibres, and thus they no longer bind to/work against prokaryotic targets. These unusual tail fibres are explored much more extensively in Chapter 5 on page 138.

Positions 14 and 16 in the operon core also elaborate proteins with no readily apparent functions. The best hypothesis for these proteins, based on the experimental data and synteny with the *Serratia* Afp, from the Hurst et al. experiments, is that they are the PVC equivalents of tape measure proteins and some kind of tube terminator or cap (Rybakova et al., 2013, 2015b). Chapter 3 on page 95 speculates on these proteins a little further.

Lastly, locus 15, despite not being well annotated in the originally available genomes, is readily identifiable as a AAA ATPase, like that of the T6SS (though from a disting phylogenetic family (Frickey and Lupas, 2004)). Despite its ease of identification, there is currently no further experimental information available to suggest a role as covered in section Section 1.2.3.4 on page 35. The main hypothesis to date had been that the ATPase maybe served as a 'loading pump', passing the PVC payloads in to the lumen of the tube, though this is purely speculation.

1.2.5.4 The hypervariable payload region

Finally, a hallmark of the PVCs which has been alluded to a few times in this introduction already, is the hypervariable toxin payload region. As there are multiple operons for each PVC, they each carry at least 1 unique toxic effector. This has not been observed with any

other related structures, with the partial caveat of the fact that Type *x* Secretion Systems are known to have various substrates, but they are not necessarily encoded *cis* to the system itself. The carriage of *cis* encoded toxins is true for the Afps as well, thought it appears it may only have a single toxin. Similarly, while there are no shortage of caudate structures which have been identified in the literature, in studies like Sarris et al. (2014), the PVCs remain unusual for this particular feature, and therefore potentially do not entirely fit in to the classifications others are attempting to apply to them.

The toxins encoded in the PVC loci are typically easy to identify, and have been reasonably well annotated, in part because they generally appear to be effectors which are already well known bacterial toxins from other systems. Pnf, for example, is and unequivocal orthologue of cnf1 from *E. coli*, and was annotated as such in the original ATCC43949 genome. Similarly, the lopT operon in the same genome, harbours 3 different toxins, all of which have been seen in other instances. The lopT toxin itself takes its name from the yopT cysteine protease class of toxins in *Yersinia*, an rtxA toxin is also present which is named for the family of toxins to which it belongs ("repeats-in-toxin"); a family that is well represented in other organisms, such as *Vibrio* (Lin et al., 1999). Lastly, it also harbours a TccC domain protein, which are the toxic components of the Tc toxin complexes, which are actually elaborated by *Photorhabdus* itself, separately, and were recently structurally resolved (Bowen and Ensign, 1998; Meusch et al., 2014) and are yet another staggeringly impressive toxin delivery mechanism of the bacteria. Nevertheless, following the trend in this section, not all operons were given useful or informative annotations, even for the toxins.

1.2.6 PVC myths

Despite the increasing wealth of information appearing, there are several papers which, in their attempts to group the PVCs with other structures, appear to come to spurious conclusions, and this section will briefly draw attention to these.

Firstly, Zhang et al. (2012), suggest that the ATPase which is distinctive within the PVCs has a role in cleavage/delivery of toxin molecules that is in some way separate from that of the PVC structure. There is little to no experimental evidence for this, and the

majority of the paper disregards the actual syringe complex which is arguably the most important component. They do suggest that the ATPase may have a role in recycling the PVCs, as it has been shown to in the T6SS. However, as all the evidence to date points to the release of the PVCs, to act at a distance in a 'torpedo' like fashion, there appears to be no evolutionary need/advantage to recycling the structure. That said, two possible explanations for its persistence, if recycling is its actual role, are that it may be vestigial, though this seems unlikely given the level of conservation. Given the fact that these ATPases are defined by their roles in various cellular pathways however, it could be that sufficient selection pressure to maintain the ATPases is being exerted based on their activity outside the PVC operons (Iyer et al., 2004). Alternatively, the ATPase may recycle the tube subunits continually so that they do not build up unnecessarily inside the cell, before mature PVCs are ready to be deployed 'in anger'. Since so many more of these proteins are required per syringe, its possible that they're made in significantly higher proportions, and to offset the metabolic cost of building such a structure aberrantly, they are being turned over continually to replenish cellular concentrations of amino acids and other substrates.

They speculate that the N-termini of the toxin effector molecules contain a distinctive metallo-peptidase domain/activity, though the paper is not clear on what sequences were used to arrive at this conclusion. From our own studies (as yet unpublished), it has been demonstrated that the N-termini of the toxins for several PVC effectors have a stabilising/chaperone-like role, possibly with a syringe loading signal. However, it is more or less impossible to construct a meaningful multiple sequence alignment for all but the most closely related effectors, much less to define a characteristic domain structure for all PVC toxins, which calls in to question some of the conclusions of the paper, at least for the sequences they are attributing to PVCs.

Despite being an otherwise excellent review, Kube and Wendler make the statement that PVC sheath proteins are most like T4 sheath proteins, but pyocin proteins are most like phage P2. This appears only partly true however (Kube and Wendler, 2015b). Chapter 3 on page 95 examines this further, but it appears that actually only one of the sheaths (the inner) is T4 like, whereas the outer is pyocin- (and therefore P2) like, also resembling the

T6SS. In the same paper, the authors also make the statement that the Afp cluster lacks any lysis systems, which are seen in phage and pyocins. While it is true that evidence for lysis-based release of the PVCs and Afps is scarce as the authors point out, the PVCs can often be found with lysis associated proteins. For example, downstream of the PVCpnf ATCC43949 operon, beyond the payload region, but on the same strand and in close register to the rest of the operon, several bacteriophage lysis proteins and lysozymes can be detected, though it is true that this cannot be said for all of the operons, at least at the present level of sequence identification. It seems the likely do harbour general lysis systems, though some of them may be comparatively enigmatic.

Similarly, though it is also an excellent study, the Sarris et al. (2014) paper makes many claims about the PVCs in attempting to group them with other "Phage-Like-Translocation-Systems", though they appear to be only considering a single PVC example from *P. luminescens* ("Unit2"), and some of these statements may not hold true for all PVCs. One example of an erroneous claim is in their discussion of synteny conservation. While there is undoubtedly a great deal of synteny conservation, they state that the sheath proteins are typically located downstream of baseplate genes. It's not clear whether this is simply a syntactical or typographical error, but it's readily apparent, including from the figures in their own paper, that the reverse is true. Since they are basing a level of significance on these similarities for identifying similar structures elsewhere, this synteny rearrangement would potentially have consequences for how sequences are grouped and ancestry inferred. Another objectionable conclusion is their readiness to include many, potentially distantly related, operons in to this "PLTS" family, without any actual consideration being paid to whether the operons harboured any payloads which are translocatable. This is important, as they identify R-type pyocins, which are non-translocating structures, as a separate 'clade'. Thus, whether or not a candidate caudate structure should be considered more like an R-type pyocin or an Afp/PVC cannot be decided from sequence alone due to the huge diversity, and functional relatedness is therefore a key factor. Additionally, Sarris and colleagues also fall foul of the same observation as Kube and Wendler (2015b), in stating that there are no lysis proteins associated with PVCs. Had they considered more than one PVC example in their analysis,

they may have observed that this doesn't appear to be true.

1.3 Summary and Thesis Aims

Despite this richness of data for related systems - the cumulative product of centuries of study - the picture is, perhaps unsurprisingly, still far from complete for the PVCs, being comparatively understudied. Their unique role as bacterial secretion systems that act at a distance means there is much left to be understood about what makes them different. This thesis attempts to tackle this in a number of ways:

- Firstly, an up-to-date exploration of the structural similarities and differences with the benefit of time and more advanced bioinformatics resources/databases versus the original description of PVCs can be found in Chapter 3 on page 95. This chapter aims to improve understanding of the poorly characterised genes for all the proteins in the operons, and generate hypotheses for testing in the lab and for future work.

- Secondly, a phylogenetic study of the PVCs which attempts to shed light on the microevolution within the operons, examining the variability and loss of genes in this context, can be found in Chapter 4 on page 101. All the comparative genomic studies that have been covered in this introduction are keen to place the PVCs in a wider context, whereas an inward-looking study trying to better understand why so many PVC variants exist, and why they are so diverse has been lacking.

- Key proteins in the mechanistics of (at least) non-membrane-bound caudate structures are the tail fibres. They are responsible for triggering the contractile mechanisms, and also conferring the 'target spectrum' that the structures are able to act against. For the PVCs, sequence similarities in these proteins were weak at the outset, though curiously, some were able to pick up annotations against Adenoviral motifs. Chapter 5 on page 138 explores the tail fibres in more detail, and represents possibly the first experimental studies of naturally occurring chimeric tail fibres.

- Chapter 6 on page 185 explores efforts to understand the natural expression patterns of the PVCs, as well as attempts to heterologously clone and express the PVC

operons. Experimental work to date had been conducted with a cosmid library, but there were several issues with this approach. The PVCs were still under the control of their native promotors, and this made them unstable and difficult to work with in the lab. A key question for this chapter, therefore, is to try and probe any population heterogeneity in how the PVCs are deployed naturally.

Chapter 2

Materials & Methodology

All methods from all chapters are collected here in detail, for clarity. Where appropriate, the methods have been reiterated at a higher level, in the context of the experimental workflow in their respective chapters.

2.1 Bacterial Culture Techniques

The vast majority of this project, as a molecular and synthetic biology research project, involved microbial culture and heterologous expression work. Despite being a thesis on the study of *Photorhabdus*, almost all of the work conducted was in *E. coli*. As a member of the *Enterobacteriaceae*, *Photorhabdus* is fairly closely related to *E. coli*, thus much of the genetic work can be conducted in the considerably more tractable lab strain with few, if any, complications.

2.1.1 Strains

A number of specialist and host strains were used for various purposes and they are detailed in Table 2.1 on the next page, along with their purpose, and if available, their genotypes. With the exception of BL21(DE3) "NiCo21"'s, which were purchased from New England Biolabs, and DY380 which was a gift from Donald Court[1], all strains were present in the lab freezer stocks. Their original sources are provided in the table.

[1] https://redrecombineering.ncifcrf.gov/

Table 2.1 | *E. coli* strains used throughout this work, their available genotypic data, and their originating source.

Strain	Genotype	Purpose	Reference
	Cloning/Plasmid Maintenance Strains		
DH5-α	F- *endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG purB20 ϕ80dlacZΔM15 Δ(lacZYA-argF)U169, hsdR17(rK-mK+), λ-* HB1100 derivatised strain from Bethesda Research Laboratories	High transformation efficiency general purpose cloning strain. Cloning and plasmid maintenance	(Glover, 1995)
DH10-β ("TOP10")	F- *endA1 deoR+ recA1 galE15 galK16 nupG rpsL Δ(lac)X74 ϕ80lacZΔM15 araD139 Δ(ara,leu)7697 mcrA Δ(mrr-hsdRMS-mcrBC) StrR λ-* MC1061 derivatised strain	High transformation efficiency general purpose cloning strain, reported to be more tolerant of large inserts/constructs. Maintenance of cosmids and large constructs	Invitrogen
EC100	F- *mcrA Δ(mrr-hsdRMS-mcrBC) ϕ80dlacZΔM15 ΔlacX74 recA1 endA1 araD139Δ(ara, leu)7697 galU galK λ- rpsL nupG*	High transformation efficiency cloning strain for exceptionally large constructs (cosmids/BACs etc.) Used in this study to harbour cosmid library	Epicenter (Lucigen)
S17λ*pir*	*TpR SmR recA, thi, pro, hsdR-M+RP4: 2-Tc:Mu: Km Tn7 λpir*	*E. coli* DH5-α strain for maintenance of conjugable plasmids	Biomedal
	Expression Strains		
NEB "NiCo21" BL21(DE3)	*can::CBD fhuA2 [lon] ompT gal (λ DE3) [dcm] arnA::CBD slyD::CBD glmS6Ala ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δ nin5* Derivatised BL21(DE3) with reduced proteases/IMAC contaminating proteins	IMAC optimised BL21 expression strain lysogenised with the DE3 prophage for T7-polymerase driven expression via IPTG induction	New England Biolabs
	Recombineering Strains		
DY380	F- *mcrA λ(mrr-hsdRMS-mcrBC) ϕ80dlacZ M15 ΔlacX74 deoR recA1 endA1 araD139 Δ(ara, leu) 7649 galU galK rspL nupG [λcI857 (cro-bioA) <> tet]* Derivatised DH10-β strain with defective λ prophage and temperature sensitive cI875 repressor	Recombineering strain with the β, γ and *exo* proteins chromosomally located. Can be derepressed by temperature shift to 42 °C. Used in this study to modify cosmids and overcome plasmid shortcomings.	(Lee et al., 2001)
BW25113	F- Δ(araD-araB)567, lacZ4787(Δ)::rrnB-3, LAM-, rph-1, Δ(rhaD-rhaB)568, hsdR514 Derivative of K12 strain BD792	Keio Collection WT Parent strain. A Δ*rfaH* strain was used in this study for regulation analysis experiments, and the wild type was retained as a control.	(Baba et al., 2006)

2.1.2 Culture Conditions

2.1.2.1 Media

2.1.2.1.1 LB Routine culture of *E. coli* and *Photorhabdus* was conducted in standard Lysogeny Broth (LB) liquid media and agar plates, at 200 RPM in a shaking incubator (or static incubator for plates). The media is supplemented with 0.1% pyruvate when culturing *Photorhabdus*. For *P. luminescens* strains, cultures were grown at 28 °C due to their temperature intolerance.

2.1.2.1.2 SOC Super Optimal Media with catabolite repression (SOC), is a high glucose medium routinely used in the recovery culture phase of bacterial transformation. It is designed to be a rich media which reduces stress on the transformed cells, allowing them to optimally uptake the target DNA. In particular, the high glucose content in comparison to standard LB media is useful as it represses the pBAD promotor system (used quite extensively in this study - see **??** on page ??), helping to clone otherwise potentially toxic/recalcitrant targets.

2.1.2.2 Antibiotics & Media Supplements

Various antibiotics and media supplements were used during this project. Table 2.3 on the next page shows concentrations of compounds used.

Table 2.2 | Antibiotics and other media supplements, and the final concentrations for use.

Supplement	Working Concentration
Antibiotics	
Ampicillin	100 μg mL^{-1}
Kanamycin	25 μg mL^{-1}
Chloramphenicol	25 μg mL^{-1}
Gentamycin	10 μg mL^{-1}
Tetracycline	10 μg mL^{-1}
Growth Supplements	
Pyruvate	0.1 % (w/v)
Expression	
Arabinose	0.2% (w/v)
IPTG	2 mM
Tetracyline	0.2 μM
Repression	
Glucose	0.2% (w/v)

2.2 Molecular Techniques - Nucleic Acid Methods

2.2.1 Purification of Nucleic Acids

DNA isolation was a frequent task in the course of this work. Replicon DNA in the form of plasmids and cosmids was required for screening, cloning and expression purposes. Genomic DNA was purified for PCR templates and for assessment of recombination. This was performed exclusively via commercial kit. Manufacturers protocols were followed in every case, with some minor modifications, which are detailed in this section. In all cases, DNA once purified was stored at -20 °C.

2.2.1.1 Genomic DNA

Genomic DNA (gDNA) is isolated with the Qiagen "Blood and Tissue" extraction kit, with the following modifications for bacterial culture:

5 - 10 mL of overnight culture is set up in appropriate conditions (i.e. with selection if possible). Cells are pelleted at 7,000 RCF, 4 °C for 10 minutes. Pellets are resuspended

in 180 μL of the manufacturer supplied ATL buffer, with 20 μL of the supplied Proteinase K mix added. RNAse H is optionally added if the DNA is to be used for next generation sequencing. From here the protocol proceeds directly to the manufacturers step 2, and follows the standard procedure until elution. Elution was conducted in 2 x 17.5 μL washes in AE buffer (unless it is to be used for sequencing, then H_2O or EB Buffer is used).

2.2.1.2 Replicon DNA

2.2.1.2.1 Plasmids For plasmid isolation the Qiagen Miniprep Spin Kit was used according to the manufacturers instructions. 5 - 10 mL overnights of culture are prepared in appropriate conditions (e.g. for plasmids with selection, add antibiotics - see Section 2.1.2.2 on page 63). 10 mL of culture is used for lower copy number plasmids, to ensure adequate DNA recovery. Elution was conducted in 2 x 17.5 μL washes, with molecular grade water instead of a single 50 μL buffer wash. For isolation of cosmids, and plasmids in excess of $\approx$ 10 kbp, the same miniprep kit is used, but with the manufacturers suggested optional optimisations, namely: the optional wash with PB buffer is conducted, and elution buffer/water is preheated to 70 °C. 2 x 17.5 μL washes are conducted as in plasmid preparation.

2.2.2 Plasmids and Cosmids

All plasmids were either bought, gifted or created in this study. Recombineering plasmids pKD46/pCP20/pJET-FRT-Cm/pJET-FRT-Kan were a kind gift from Dr. Helge Bode at Goethe University, Frankfurt. pET29a was received from Jenny Goodman, a fellow PhD student at Warwick.

Table 2.4 on the next page details all the existing plasmids used in this study that have been previously constructed and/or published. Table 2.5 on page 68 details all the constructs produced during the course of this study.

Table 2.4 | Existing plasmids used as the bases for derivations listed in Table 2.5 on the next page. All plasmids were either gifted, existed in lab stocks already, or purchased.

Plasmid Designation	Purpose	Reference
	Cloning/Expression Plasmid Bases	
pBAD30	Basic inducible expression vector. Arabinose inducible via *araBAD* system, glucose repressible. Ampicillin resistant, with a p15 *ori* (compatibility group B) and f1 *ori* (compatibility group A).	(Guzman et al., 1995)
pET15b	Inducible expression vector. IPTG inducible via *lac*/T7 polymerase system, glucose repressible. Ampicillin resistant, with ColE1 *ori* (compatibility group A). The plasmid contains an N-terminal hexa-histidine tag with a thrombin cleavage site for in-frame tagging of recombinant protein and cleavage after purification.	Novagen
pET29a	Inducible expression vector. IPTG inducible via *lac*/T7 polymerase system, glucose repressible. Kanamycin resistant, with ColE1 *ori* (compatibility group A). The plasmid contains a C-terminal hexa-histidine tag with a thrombin cleavage site for in-frame tagging of recombinant protein and cleavage after purification. Additionally contains an N-terminal Streptavidin tag.	Novagen
pGAG1	Promoterless GFP reporter 'empty' vector. Conjugative plasmid requiring *λpir E. coli* for propagation.	CITATION NEEDED
pAGAG	Promoterless GFP bearing plasmid, without GFP start codon for promotor fusion reporter construct creation. Conjugative plasmid requiring *λpir E. coli* for propagation.	CITATION NEEDED
	Recombineering Plasmids	
pKD46	*λ*Red plasmid bearing *β*, *γ*, and *Exo* recombineering enzymes, under the arabinose inducible control of the *araBAD* system. Ampicillin resistant, with the temperature sensitive *ori* 101ts (compatibility group C)	(Datsenko and Wanner, 2000)
pCP20	FRT "flippase" bearing plasmid for excision of regions surrounded by FRT recognition sites (typically after successful recombineering with pKD). Ampicillin and Chloramphenicol resistant, with the temperature sensitive *ori* 101ts (compatibility group C).	(Datsenko and Wanner, 2000)
pJET-FRT-Cm	Recombineering knockout cassette template plasmids bearing an FRT-flanked Chloramphenicol cassette. Ampicillin and Chloramphenicol resistant, with a ColE1 *ori* (compatibility group A).	Helge Bode (derivatised Thermo Scientific Vector)
pJET-FRT-Kan	Recombineering knockout cassette template plasmids bearing an FRT-flanked Kanamycin cassette. Ampicillin and Kanamycin resistant, with a ColE1 *ori* (compatibility group A).	Helge Bode (derivatised Thermo Scientific Vector)

Table 2.5 | Cloned and/or derivatised plasmids created during the course of this study.

Plasmid Designation	Insert	Backbone	Function/Purpose
		Expression Constructs	
pET15b_*pnf13*	PVC*pnf* Putative Tail Fibre Gene	pET15b	PVC*pnf13* Tail fibre cloned in-frame with the N-terminal hexa-histidine tag and thrombin cleavage site, for expression and purification via IMAC
pET15b_*lumt13*	PVC*lumt* Putative Tail Fibre Gene	pET15b	PVC*lumt13* Tail fibre cloned in-frame with the N-terminal hexa-histidine tag and thrombin cleavage site, for expression and purification via IMAC
pET29a_*pnf13*	PVC*pnf* Putative Tail Fibre Gene	pET29a	PVC*pnf13* Tail fibre cloned in-frame with the C-terminal hexa-histidine tag and thrombin cleavage site, for expression and purification via IMAC
pET29a_*lumt13*	PVC*lumt* Putative Tail Fibre Gene	pET29a	PVC*lumt13* Tail fibre cloned in-frame with the C-terminal hexa-histidine tag and thrombin cleavage site, for expression and purification via IMAC
		Reporter Constructs	
pAGAG_PB68.1PVCpnf	*P. asymbiotica* Thai PB68.1 PVC*pnf* promoter	pAGAG	*P. asymbiotica* strain Thai PB68.1 PVC*pnf* operon promoter fused to GFP
pAGAG_PB68.1PVClopT	*P. asymbiotica* Thai PB68.1 PVC*lopT* promoter	pAGAG	*P. asymbiotica* strain Thai PB68.1 PVC*lopT* operon promoter fused to GFP
pAGAG_PB68.1PVCcif	*P. asymbiotica* Thai PB68.1 PVC*cif* promoter	pAGAG	*P. asymbiotica* strain Thai PB68.1 PVC*cif* operon promoter fused to GFP
pAGAG_PB68.1PVCU1	*P. asymbiotica* Thai PB68.1 PVC*Unit1* promoter	pAGAG	*P. asymbiotica* strain Thai PB68.1 PVC*Unit1* operon promoter fused to GFP
pAGAG_TT01PVCU4	*P. luminescens* TT01 PVC*Unit4* promoter	pAGAG	*P. luminescens* strain TT01 PVC*pnf* operon promoter fused to GFP
pAGAG_TT01PVClopT	*P. luminescens* TT01 PVC*lopT* promoter	pAGAG	*P. luminescens* strain TT01 PVC*pnf* operon promoter fused to GFP
pAGAG_TT01PVCcif	*P. luminescens* TT01 PVC*cif* promoter	pAGAG	*P. luminescens* strain TT01 PVC*pnf* operon promoter fused to GFP
pAGAG_TT01PVCU1	*P. luminescens* TT01 PVC*Unit1* promoter	pAGAG	*P. luminescens* strain TT01 PVC*pnf* operon promoter fused to GFP

2.2.3 PCR

2.2.3.1 Primers

All primers used in this study were purchased from Integrated DNA Technologies (IDT).

Table 2.6 | Primer sequences used in this study for simple amplification and detection purposes - no sequence modifications. Annealing temperatures are given as per the IDT Oligoanalyzer's reported value, or, in the case of values in square parentheses, those given by NEB Tm Calculator (with 500 nM primer concentration and Q5 product group parameters).

Primer Name	Function	Sequence (5′ → 3′)	Tm (°C)	Length (bp)
no1_F	Detection of pJET	CGCACTTCCAGACCCAGATC	57.9	≈1200
no2_R		GATGGAGTAAATGGTACCTTGGG	55.1	
Gam_Bet_F	Detection of pKD/pCP	TTTCACAGCTATTTCAGGAGTTC	52.9	1112
Gam_Bet_R		CATGCTGCCACCTTCTG	53.8	
T7_Prom_F	T7 Sequencing Primer	TAATACGACTCACTATAGGG	46.5 [58]	Varied
T7_Term_R		GCTAGTTATTGCTCAGCGG	46.5 [58]	
rfaH_5′_SP_F	rfaH Knockout Sequencing Primer	CAACTTCACGCAGCG	51.4 [62]	Varied
rfaH_3′_SP_R		TATGACATTGCTGGAGCC	52.2 [62]	

Finish populating all primer tables

Table 2.8 | Primers for specialist purposes, harbouring modifications, including restriction sites for cloning and overlap homologies for recombineering and Gibson Assembly. Restriction Sites are shown in **bold**. Overlap homology is shown underlined Annealing temperatures shown in [ ] are specific to NEB's Q5 Polymerase. F: Forward Primer, R: Reverse Primer, bp - Base Pair.

Primer Name	Function/Target	Sequence (5′ → 3′)	Tm (°C)	Length (bp)
		Classical Cloning		
PVCpnf13-NdeI_F	*pnf* Tail Fibre	GAGTTA**CATATG**AACGAAACTCGTTATAATGC	[67]	1548
PVCpnf13-BamHI_R		TTTTC**AGGATCC**TTAAAGCTTTATGATGAAAGC	[67]	
PVCpnf13-KpnI_R		TTTTC**AGGTACC**AAAAGCTTTATGATGAAAGC	[67]	
PVClumt13-NdeI_F	*lumt* Tail Fibre	GCCGGA**CATATG**GACAACAAAAATAAC	[67]	675
PVClumt13-BamHI_R		TTACTT**GGATCC**TTACACAACCTTAATCATATAG	[67]	
PVClumt13-KpnI_R		TTACTT**GGTACC**AACACAACCTTAATCATATAG	[67]	
rfaH_EcoRI_F	*rfaH*	ACAGGT**CATATG**GAATTAAATGAGTTAACTAACAAATT	[71]	500
rfaH_SalI_R		CTA**GTCGAC**TTAGAGTTTGCGGAACTCG	[71]	
		Gibson Assembly		
pBAD30frag_F	pBAD30	ATGTAATTAATTCAACCATCACGGAGAGTTTATCAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCC	[72]	4791
pBAD30frag_R		CTTGTAGACATAAAAGCCCCTTTTTAGACAAAAAATAGCCCAAAAAAACGGGTATGGAGAAACAGTAGAG	[72]	
PNFfrag1_F	PVC*pnf* 1-8	CTCTACTGTTTCTCCATACCCGTTTTTTTGGGCTATTTTTTGTCTAAAAAGGGGCTTTTATGTCTACAAG	[72]	7181
PNFfrag1_R		GTGTCAGTATTTGATTTTCCATTCATCGTCACCTTTCATTGGGTAAGATTAATTTTTGCGCCTTTGATTT	[72]	
PNFfrag2_F	PVC*pnf* 9-12	AAATCAAAGGCGCAAAAATTAATCTTACCCAATGAAAGGTGACGATGAATGGAAAATCAAATACTGACAC	[72]	6039
PNFfrag2_R		TCTTGTACAGTTGCATTATAACGAGTTTCGTTCATGATTAACTCCAGAAAACATATTTAATTCAACATCA	[72]	
PNFfrag3_F	PVC*pnf* 13-15	TGATGTTGAATTAAATATGTTTTCTGGAGTTAATCATGAACGAAACTCGTTATAATGCAACTGTACAAGA	[72]	5514
PNFfrag3_R		TTATTGACATCAATAATAGTTTGCGTGTTTAACATAAAAAACCTCTCTTAAATTATATCGTGATAACTTT	[72]	
PNFfrag4_F	PVC*pnf* 16-18	AAAGTTATCACGATATAATTTAAGAGAGGTTTTTTATGTTAAACACGCAAACTATTATTGATGTCAATAA	[72]	4416
PNFfrag4_R		GGGGAGACCCCACACTACCATCGGCGCTACGGCGTTGATAAACTCTCCGTGATGGTTGAATTAATTACAT	[72]	

Recombineering Primers					
ψendA_no1_F	*endA* Deletion site	AAACAGCTTTCGCTACGTTGCTGGCTCGTTTTAACACGGAGTAAGTGATGCGCACTTCCAGACCCAGATC	[72]	1100	
ψendA_no2_R		GTTAACAAAAAGAATCCCGCTAGTGTAGGTTAGCTCTTTCGCGCCTGGCAGATGGAGTAAATGGTACCTTGGG	[72]		
speB_no1_F	*speB*	GTTTTACCCGTGCGCATCGCATCTGGTGCTTACTCGCCCTTTTTCGCCGCCGCACTTCCAGACCCAGATC	[72]	1100	
speB_no2_R		GACGCGGAAGGGTTTTTTTATATCGACTTTGTAATAGGAGTCCATCCATGGATGGAGTAAATGGTACCTTGGG	[72]		
hyfC_no1_F	*hyfC*	GTTTTACCCGTGCGCATCGCATCTGGTGCTTACTCGCCCTTTTTCGCCGCCGCACTTCCAGACCCAGATC	[72]	1100	
hyfC_no2_R		GACGCGGAAGGGTTTTTTTATATCGACTTTGTAATAGGAGTCCATCCATGGATGGAGTAAATGGTACCTTGGG	[72]		

2.2.3.2 *Taq* & Colony PCR

For colony PCR, and applications where sequence fidelity was not absolutely necessary (e.g. band shift assessment), *Taq* polymerase was used, purchased from Invitrogen. Typical reaction composition and cycling parameters are laid out in Table 2.9. The enzyme was used largely according the the manufacturers protocol.

For rapid screening of transformed bacteria and detection of sequences in colonies/-culture, colony PCR was used. Single colonies, or 5 μL of liquid culture, are resuspended in 50 μL of molecular grade water, boiled at 100 °C for 10 minutes and pelleted at 16,000 RCF for 1 minute. 5 μL of supernatant can then be used in place of the standard 1 μL of DNA template (offsetting the volumes with reduced water content in the PCR), to give good amplification.

Table 2.9 | PCR set up for use with *Taq* polymerase. Subtable **(a)** shows typical thermocycling conditions. Subtable **(b)** shows a typical reaction composition. For colony PCR, 5 μL of DNA template is substituted and offset against the final volume of water added.

(a)

Step		Temperature (°C)	Time (m:s)
Initial Denaturation		94	3:00
Denature	29X Cycles	94	0:45
Anneal	29X Cycles	*Tm* - 3	0:30
Extend	29X Cycles	72	1:30 kb^{-1}
Final Extension		72	10:00
Hold		4	Indefinitely

(b)

Reagent	Volume (μL)
Taq Buffer	5
$MgCl_2$	0.75
dNTPs	0.5
Primer 1	1.25
Primer 2	1.25
Template	≈ 1
Polymerase	0.3
H_2O	to 25

2.2.3.3 Q5

For all cloning experiments and use cases where sequence fidelity was crucial, the high-fidelity enzyme Q5 was used, from New England Biolabs. Reactions were performed as per the manufacturers protocol, however reaction size was reduced, proportionally, to 20 μL.

Annealing temperatures for reactions when using Q5 are non-standard. As such, annealing temperatures are recalculated with the online tool provided by NEB[2]. Annealing

[2] `http://tmcalculator.neb.com/#!/`

temperatures are given in the primer table in Section 2.2.3.1 on page 69.

Table 2.10 | PCR set up for use with Q5 polymerase. Subtable (a) shows typical thermocycling conditions. Subtable (b) shows a typical reaction composition.

(a)

Step		Temperature (°C)	Time (m:s)
Initial Denaturation		98	0:30
Denature	39X Cycles	98	0:15
Anneal	39X Cycles	Tm - 3	0:15
Extend	39X Cycles	72	0:30 kb^{-1}
Final Extension		72	10:00
Hold		4	Indefinitely

(b)

Reagent	Volume (μL)
Q5 Buffer	2.5
dNTPs	0.75
Primer 1	1.25
Primer 2	1.25
Template	≈ 1
Polymerase	0.25
H_2O	to 20

2.2.3.4 Post-PCR Clean-up

After PCR, gel extraction, and restriction digests, it is necessary to clean up nucleic acid samples, to remove residual buffers, additives, enzymes and DNA fragments. PCR clean up in this study was performed with the GE Healthcare "illustra GFX" PCR DNA and Gel Band Purification Kit, as per the manufacturers instructions. The same elution modification is made as detailed in **??** 2.2.1.2.1 on page 65

2.2.3.5 Quantification

2.2.3.5.1 Platereader Routine nucleic acid quantification was performed by measuring absorbance at 260 nm on the BMG Labtech SPECTROstar Nano microplate reader with the LVis plate insert. 1-2 μL of sample or blank is pipetted on to the plate in duplicate, absorbance measured and an average of the 2 returned values was taken as the DNA concentration of the sample.

2.2.4 Agarose Gel Electrophoresis

For sequences of between approximately 1-4 kb 1% gels (w/v) were used. Larger DNA fragments were typically run on 0.8% gels. For a "mini-gel", 0.5 g of agarose powder is added to 50 ml of 1X concentration Tris-Acetate-EDTA (TAE) buffer (and scaled appropriately for larger gels). The mixture is microwaved until the agarose is melted and the solution is completely clear. SYBR®-safe gel stain is added to the mixture at a 1:10,000X

dilution. The liquid gel is poured in to casting trays with the appropriate comb for the number of wells required, and left to set for approximately 30 minutes. Gels were run in tanks containing 1X TAE, at 100 Volts for between 30-40 minutes, or until the loading dye cloud reached the bottom of the gel. Visualisation was performed using the GelDoc transillumination cabinet. To size the bands and act as a positive control for imaging, samples were run with Bioline Hyperladder 1kb, 100bp or NEB 2-log DNA ladders.

- 50X Stock TAE Buffer (pH 8.2-8.4):
  - 2 M Tris Base ($C_4H_{11}NO_3$)
  - 57.1 mL Glacial Acetic acid (CH_3COOH)
  - 50 mM EDTA ($C_{10}H_{16}N_2O_8$)

2.2.4.1 Gel Extraction

Gel extraction was used for isolating correct length fragments among mixed populations, or for separating products from their templates to avoid carry through of plasmids etc. A normal agarose gel is prepared, and then visualised by eye on a blue light or ultraviolet transillumination box after running. A scalpel is used to slice out the required band, and added to purification buffer from the PCR clean-up kit as detailed in Section 2.2.3.4 on page 73. Extraction of the DNA from the agarose is performed as per the manufacturers instructions.

2.2.5 Classical Cloning

2.2.5.1 Restriction Enzyme Digestions

Restriction enzymes are selected for cloning by ensuring their restriction sites are not present within the insert used, and whenever possible, 2 different enzymes are chosen for orientation and compatibility with the vector of choice, as well as ideally having complementary incubation temperatures and buffers. All restriction enzymes used in this study were purchased from New England Biolabs, and are detailed in the primer table in Table 2.7 on page 69, highlighted in bold.

Restriction digests were conducted mostly according to the manufacturers instructions. The reaction mixes were prepared as follows, to 40 μl total reaction volume: 4

μL reaction buffer, 0.4 μL of respective pairs of nucleases (unless specifically directed), a volume of DNA preparation to give between 700-1000 ng total DNA and lastly, nuclease-free water to the final reaction volume. NEB's website is referred to in order to work out buffer compatibility for enzyme pairs. When enzymes do not have the same buffer compatibility or incubation conditions, serial incubations were performed and a PCR clean up is performed between the first and second enzyme incubation.

Digestions were typically left for $\approx$ 4 hours at 37 °C (unless specifically directed otherwise be the manufacturer. Overnight digestions were used when enzymes had no star activity, at a room temperature.

2.2.5.2 Vector Dephosphorylation

For difficult or low efficiency cloning, dephosphorylation of the vector to avoid self-ligation and recircularisation was conducted. Antarctic Phosphatase from NEB was used, according to the manufacturers instructions, as it can be inactivated by heating at 80 °C for 2 minutes, meaning that a subsequent clean up was not needed, preserving concentrations of DNA for transformation.

2.2.5.3 Ligation

Ligation reactions were performed using T4 ligase from NEB in 10 μL total volume at room temperature for 1 hour as per manufacturer instructions, or overnight for less efficient reactions. Routinely, 3 different ligation insert:vector molar ratios (1:1, 3:1, 10:1) were used as it is often not possible to know ahead of time which will perform optimally. The mass of insert to use which corresponds to the above ratios is calculated like so:

$$ng_{Insert} = R \times \left(\frac{ng_{Vector} \times bp_{Insert}}{bp_{Vector}} \right) \tag{2.1}$$

Molar Ratio Ligation Calculation

where R is the ratio you intend to use, ng_{Vector} is the nanogram amount of vector DNA used (usually 30 ng); and bp_{Insert}, bp_{Vector} are the sizes in basepairs of the insert and vector respectively.

2.2.6 Gibson Assembly

Gibson assembly was performed using the NEBuilder HiFi Gibson Assembly mastermix, as described by the manufacturer. Exceptions to the manufacturers 'one-pot' protocols were made for sequential assembly - i.e. for multiple fragment assembly, adjoining pairs ("A", "B" & "C", "D") were assembled in 2 reactions for 1 hour, then reactions combined to join fragments "AB" with "CD" for an additional hour to try and achieve a full size "ABCD" fragment. DNA amounts and concentrations were altered from the manufacturers specifications in an experiment-dependant manner, and were optimised each time.

Gibson primers were always designed with 35 basepair overlap homology on each side of a fragment joint. These could be easily PCR'd with Q5 at a Tm of 72 °C.

2.2.7 Transformation

2.2.7.1 Creation of Chemically Competent Cells

Overnight cultures were used to inoculate 100 mL LB media at a dilution of 1:100. Cultures were grown to exponential phase when the optical density at 600 nm (OD_{600}) reached between 0.4-0.5 and were then placed on ice for 10-15 minutes. The bacteria were then prepared for chemical competency via pelleting by centrifugation (4000 RCF and 4 °C for 10 minutes) resuspending in 20 ml of ionic Solution I (see below). Samples were kept on ice for 10 min and re-centrifuged. The pellet was then resuspended in 4 ml of Solution II for storage. Aliquots (50 μL) were placed on dry ice and stored at -80 °C for later use.

- Solution I (pH 5.6-6):
  - 10 mM Sodium Acetate ($C_2H_3NaO_2$)
  - 50 mM $MnCl_2$
  - 5 mM NaCl
- Solution II (pH 5.6-6):

- 10 mM Sodium Acetate ($C_2H_3NaO_2$)
- 5% glycerol
- 70 mM $CaCl_2$
- 5 mM $MnCl_2$

2.2.7.2 Heat-shock Transformation of Chemically Competent Cells

Transformation of commercial chemically competent bacteria was performed exactly as detailed in the manufacturers accompanying instructions. For 'homemade' competents, cells were thawed on ice and 3 μL of plasmid or ligation reaction mix added. The cell/DNA mix was incubated on ice for 20 min and then heat shocked at 42 °C for 1 minute, followed by chilling on ice for 5 further minutes. After which, 1 mL of SOC media is added to the cells and they are recovered at 37 °C (except in the case of temperature sensitive replicons where 30 °C is used) for 1 hour. After recovery, 100 μL is plated on to appropriate selection plates (see **??** 2.2.1.2.1 on page 65 and Section 2.1.2.2 on page 63 for details). The remaining culture is pelleted at 13,000 RCF and resuspended in 100 μL SOC, then secondarily plated. Plates are then incubated at 37 °C unless temperature sensitive. Closed vector was routinely used as a transformation efficiency control. Successful transformation was checked for incorporation of inserts by colony PCR and/or diagnostic restriction digest. Successful clones were sent for sequencing, to assess the fidelity of the insert (see Section 2.2.9.1 on page 80).

2.2.7.3 Electrocompetent Cells

For recombineering protocols (see Section 2.2.8 on page 79), recalcitrant transformations, and transformation of large plasmids/cosmids, cells were typically transformed via electroporation instead of heat shock. Additionally, *Photorhabdus*, does not tolerate the process of induced chemical competence, and so electroporation was a routine transformation protocol in these cases.

2.2.7.3.1 ***E. coli*** For *E. coli*, an appropriate number of 100 mL LB cultures for the intended number of transformations and controls, were inoculated from overnight cultures, at a 1:100 dilution. 100 mL provides approximately 100 μL of final cell volume. Cultures

are incubated until they reach OD_{600nm} 0.4-0.6. At this point, cells were chilled on to ice for 20 minutes. The culture was split in to 2 x 50 mL falcon tubes and centrifuged at 4000 RCF, 4 °C, for 10 minutes. Each pellet was then resuspended in 1 volume equivalent (50 mL) of ice cold sterile water to wash. Cells were re-pelleted as before. The ice cold water wash is repeated a further 2 times, each time in half the volume equivalent of the previous step. Cells were finally resuspended in 100 μL of ice cold sterile water ready for electroporation. Biorad 0.2 cm gap, long electrode cuvettes were used specifically, and were kept chilled right up until electroporation.

Electroporation was conducted at 2.5 kV, 25 μF, 200 Ω. Time constants between 4.9 - 5.4 were typically indicative of a successful electroporation. Immediately after pulsing, 1 mL of SOC media was added to the cells and they were recovered for 1 hour, shaking, at 37 °C (unless transforming temperature sensitive replicons). After recovery, cultures were plated on to appropriate selection media, as described earlier.

2.2.7.3.2 ***Photorhabdus*** For *Photorhabdus*, no protocols currently exist that are sufficiently optimised to reliably produce chemically competent cells. In order to create competent *Photorhabdus* electroporation was used as the routine method of transformation. Cultures were grown to OD_{600} 0.2 and then chilled on ice for 90 minutes, followed by centrifugation at 4000 x g and 4 °C for 10 minutes. In the same manner as for *E. coli* described above, cells were washed 3 times with HEPES buffer (1 mM HEPES, pH 7.0, 5% sucrose) in 1:1 volume equivalent, followed by a 0.5 equivalent, pelleting as above, between each round. The cells were finally resuspended HEPES in 0.001 volume of original culture, and chilled on ice for electroporation. Electroporation cuvettes were also chilled on ice and 50 μL of cells added to each, with 10 μL of DNA for transformation. Electroporation parameters were: 2.3 kV, 200, 25 μF and Ω. Cells were then transferred to 1 mL of LB media containing 0.1% pyruvate and 1 mM $MgCl_2$, and incubated at 30 °C for 2-3 hours followed by re-plating on to selective LB plates, incubated at room temperature in the dark.

- *Photorhabdus* Electroporation Wash Buffer (pH 7.0):
  - 1 mM HEPES ($C_8H_{18}N_2O_4s$)
  - 5% Sucrose $C_{12}H_{22}O_{11}$)

2.2.8 Recombineering

The recombineering protocol devised in this work is a modified version of the electroporation protocol, with an included induction step for the recombineering enzymes.

2.2.8.1 Preparation of Linear Oligonucleotides

Primers to create the recombination cassette can be seen in Table 2.8 on page 70. Primers are designed such that they contain 50 nucleotides of homology to the target insertion site, with 20 nucleotides of cassette priming nucleotides 3' to the homology. Primers with a total length of 70 basepairs are amplified exclusively with Q5 (see Table 2.10 on page 73). In order to ensure there is no carry through when using helper plasmids as template, all linear oligos were gel extracted, treated with DpnI digestion, and finally PCR purified.

2.2.8.2 Electroporation-Recombination using λ-Red Bearing Plasmids

E. coli harbouring plasmid pKD46 or pKD56 (see Table 2.4 on page 67 were cultured overnight at 30 °C, as they contain a temperature sensitive origin of replication. The following day, 100 mL cultures are set up with a 1:100 dilution of the overnight. Cultures were grown (at 30 °C) until OD_{600nm} 0.1, split in half, and one half was induced with 0.2% Arabinose (for pKD46) or 0.2 μM tetracycline (for pKD56). The remaining half was retained as a non-induced control. The cultures continue to be grown until OD_{600nm} 0.4-0.6 is reached, at which point the cells are chilled, washed and prepared for electroporation in the same manner as in Section 2.2.7.3 on page 77.

Electroporation was conducted with the previously detailed parameters for *E. coli*, using $\approx$ 400 ng of linear DNA - though this may need optimisation on a per-experiment basis. For co-electroporation of linear modifying oligo with the target replicon, a low replicon to cell ratio is used and a maximum of 1 ng of replicon DNA.

To calculate this, 100 μL of cells from 50 mL of OD 0.4 culture should be approximately 1.6×10^{10} cells in total (or $1.6\times10^{8}\mu L^{-1}$), and double stranded DNA copy number can be calculated as follows:

$$dsDNA\ (mol) = \left(\frac{mass\ (g)}{Length\ (bp) \times 617.96\ g\ mol^{-1} + 36.04\ g\ mol^{-1}} \right) \times \mathrm{N}_A \qquad (2.2)$$

Conversion from mass and length of DNA to copy number

where N_A is Avogadro's Constant.

2.2.8.3 Electroporation-Recombination with λ-Red Chromosomal Strains

The same workflow as detailed above in Section 2.2.8.2 on page 79 is followed for the recombineering strain DY380 which bears all the recombineering enzymes within the chromosome (see Table 2.1 on page 62), with the exception that induction is conducted at $\mathrm{OD}_{600\mathrm{nm}}$ by heat-shock at 42 °C for 15 minutes, and selection for the strain is done with tetracyline.

2.2.9 Sequencing

2.2.9.1 Di-deoxy-chain-termination (Sanger) Sequencing

Routine short amplicon (≤1400 bp) sequencing of cloning constructs and for validation purposes was performed via the departmental outsourcing service to GATC Biotech, Germany. As Sanger sequencing is error prone, especially near the ends of an amplicon, sequence-sensitive applications were sequenced several times over. Primers used for routine sequencing/confirmation can be seen in Section 2.2.3.1 on page 69.

2.2.9.2 Next Generation

As the PVC operons are larger than is amenable to the vast majority of routine techniques (constructs might be anywhere from 20-50 kb), they were occasionally sequenced in-house on the Illumina MiSeq platform. They could be easily assembled in to single contigs after discarding contaminating host reads - see **??** on page ??. Libraries were prepared according to the manufacturers specifications, using the paired-end 2x250bp Nextera XT kit.

2.3 Molecular Techniques - Protein Methods

2.3.1 Expression

Expression from pET vector constructs was trialled under a couple of standard conditions (Chapter 5 on page 138). The T7 polymerase dependence of these plasmids meant that after construction in *E. coli* DH5-α, the plasmids were miniprepped and transferred in to the NEB strain, NiCo21(DE3) in order to be expressed. This strain is optimised for the expression of proteins which are poly-histidine tagged, as a number of common proteins which contaminate affinity chromatography procedures have been engineered to reduce affinity, or tagged to allow secondary removal (Bolanos-Garcia and Davies, 2006; Robichon et al., 2011).

Strains bearing the relevant construct were grown overnight in a small flask to provide enough volume for subsequent dilution. On the second day, up to 6 $\times$ 2 L flasks with 1 L of fresh media were inocculated at a 1:100 dilution. For any single purification round, only 2 L worth of pellet was used, but the remaining culture could be pelleted and flash frozen for use at a later time. The 1 L cultures were allowed to grow to an OD_{600nm} of 0.4-0.6, at which point they were induced by addition of IPTG to a final concentration of 2 mM. Cultures were left to grow overnight at a reduced temperature of 25 °C.

2.3.2 Harvesting

On the day following large scale culture, cells are harvested by centrifugation at 5,000 RCF for 20 minutes in appropriate large volume centrifuge bottles, using a high-speed centrifuge. 6 x 1 Litre cultures were reduced by pelleting in to 3 pellets derived from 2 litres each. At this point, pellets could be flash frozen for long term storage, which was typically done with 2 of the 3 pellets, proceeding directly to lysis and purification with the remaining one.

2.3.3 Lysis

The retained pellet is stored on ice whenever possible during purification. Each pellet was resuspended in 30 mL of lysis buffer, with EDTA-free total protease inhibitors. Cells

are lysed and protein released by sonication with a needle sonicator via repeated 1 minute sonication cycles 3 to 5 times. Alternatively, cells can be lysed with a homogenised, french press or other technique. After lysis, the solution is centrifuged at high speed (50,000 RCF) for 30 minutes at 4 °C to remove cellular debris. The clarified supernatant is retained, ready for column loading.

At the same time as preparing the lysis buffer, an elution buffer is also prepared:

- Lysis Buffer (pH 7.4):
  - 500 mM $NaCl_2$
  - 20 mM $NaPO_4$
  - 10 mM Imidazole
  - 10% (v/v) Glycerol
- Elution Buffer (pH 7.4):
  - 500 mM $NaCl_2$
  - 20 mM $NaPO_4$
  - 500 mM Imidazole
  - 10% (v/v) glycerol

Additional additives, if compatible with the columns to be used, can be supplemented in to these buffers. In this project, 2 mM di-thio-threitol (DTT), 2 M urea and 2 mM Zinc Chloride were additionally tested to remove further impurities - see Chapter 5 on page 138.

2.3.4 Purification

2.3.4.1 Immobilised Metal-ion Affinity Chromatography

The expressed proteins were poly-histidine tagged to allow for various follow up techniques such as western blots, nano-bead immobilisation, but also for purification via Immobilised Metal ion Affinity Chromatography (IMAC). For this, Hi-Trap 5 mL IMAC columns were purchased from GE Healthcare. Columns were maintained and prepared as per the manufacturers instructions, and in this case, were charged with Nickel-II Chloride as the metal ion.

The lysate from sonication is cycled through the column via a peristaltic pump (or chromatography apparatus). The longer the lysate is cycled through the pump the better retention of proteins, but care is taken to avoid adding air bubbles on to the column which would ruin the chromatogram. As a minimum, it was ensured that all the lysate was cycled through the column at least once.

Once the column was loaded, it was processed on an Akt Pure 2 Fast Protein Liquid Chromatography machine, or an Agilent High Performance Liquid Chromatography machine as soon as possible. The column was first washed by pumping ≥ 4 column volumes of buffer A (lysis buffer) through to remove loosely bound impurities. A gradient elution was then used, whereby buffer B (elution buffer listed above) is mixed steadily with the flow of buffer A, until the flow is 100% buffer B, causing ionically bound proteins to disassociate with the Nickel at varying points depending on the strength of the association. The gradient was collected in a fractionator, and fractions responding to high UV_{280nm} traces were taken for subsequent examination of purity via SDS-PAGE/Western blot.

Alternatively, for quicker, but slightly less pure preparations, an "assisted gravity flow" resin purification can also be used. "cOmplete" His-tag purification resin from Sigma-Aldrich was purchased and used in conjunction with glass chromatography gravity flow columns from Bio-Rad. The resin was washed several times with multiple column volumes of ethanol, followed by deionised water. Depending on the volume of culture and expected protein yield, up to $\approx$ 3-4 mL of resin was added to the column and equilibrated by mixing with lysis buffer (as per the previous section). The buffer is allowed to drain from the column or can be "assisted" by connecting a syringe to add back-pressure. The clarified lysate from high-speed centrifugation was then mixed with the resin by rotating the resin-lysate mixture end-over-end in a sealed falcon tube, in a cold room for a minimum of 1 hour (it can be left overnight for better yields). The resin was then added back to the column, and a low imidazole concentration wash buffer ($\approx$ 20 mM) passed through the column resin-protein matrix. Finally, elution buffer was passed through the column and collected. This was repeated 2 or 3 times to ensure as much protein as possible was collected.

2.3.4.2 Gel Filtration

For subsequent polishing of protein purifications, gel filtration was performed using the same chromatography apparatus. A simplified lysis buffer was used for gel filtration, the high salt content is retained for protein stability, but the imidazole and glycerol were removed. Fractions were once again collected which corresponded to peaks in the UV_{280} trace.

2.3.4.3 Concentration/Dialysis

Concentration of protein samples from gel filtration and IMAC was performed by centrifugation at 7,000 RCF in Amicon filter columns. Appropriate molecular weight cutoffs were chosen for the theoretical size of the protein to ensure maximum retention of just the protein of interest. Concentration of these volumes was typically a slow process, requiring several concentration cycles of 30 minutes to an hour at 4 °C, though this is indicative of high protein concentrations. Dialysis can also be performed using Amicon columns, by cycles of centrifugation, resuspension/dilution, and washing in the new buffer. Alternatively, dialysis was performed with Thermo "Slide-A-Lyzer MINI" dialysis tubes, placed on an orbital shaker at low speed. Depending on the application, the buffer was changed a number of times over the course of approximately 48 hours.

2.3.5 Quantification

Routine quantification of protein samples was performed with a nanospectrophotometer, measuring absorbance at UV_{280nm}. Fluorescence dyes such as those used in the Qubit spectrometer were found to precipitate the proteins studied in this work and could not be relied on.

2.3.6 SDS-PAGE

Sodium-Dodecyl-Sulphate Polyacrylamide Gel Electrophoresis was used routinely to estimate the purity of protein samples and to gauge their size to ensure correct expression and assembly. Commonly, precast gels were used with various well numbers/sizes. In cases where a large number of gels were required, gels were 'homemade'. Precast gels,

namely Biorad TGX Mini-protean 4-15% gels were used for more sensitive applications such as Western Blots, and run as per the manufacturers instructions.

Gels were prepared via standard methods, routinely using 12 and 15 % v/v resolving gels for visualisation. Briefly, for a single 12 % resolving gel, 1.5 M Tris-HCl pH 8.8 is mixed with 29 % (w/v) acrylamide, water and 10 % (w/v) sodium dodecyl-sulphate. 10 % (w/v) ammonium persulphate is added and mixed thoroughly. The casting frame is set up and tested for leaks. Once ready to pour the gel, tetramethlyethylenediamine is added to catalyse the polymerisation of the gel. The gel is overlaid with a thin layer of isopropanol to ensure a straight interface for the stacking gel. For the stacking gel, the process is the same, but 0.5 M Tris-HCl is used at pH 6.8, and the volumes of the reagents change. For full details of the reaction proportions, see Table 2.11 below. The stacking gel is poured in the the frame over the resolving gel once it is set, and an appropriate well comb is added. The gel is left to set for approximately an hour.

Table 2.11 | Reaction composition for creation of SDS-PAGE stacking and resolving gels. Abbreviations: SDS - Sodium Dodecyl Sulphate, APS - Ammonium Persulphate, TEMED - TEtraMethylEthyleneDiamine

	1 Resolving Gel			1 Stacking Gel
Reagent	12 %	15 %	Reagent	
1.5 M Tris-HCl pH 8.8	1.41 mL	1.41 mL	0.5 M Tris-HCl pH 6.8	1.25 mL
29 % (w/v) acrylamide	2.3 mL	2.6 mL	29 % (w/v) acrylamide	0.5 mL
Water	1.9 mL	1.3 mL	Water	3.25 mL
10 % (w/v) SDS	57.5 μL	57.5 μL	10 % (w/v) SDS	50 μL
10 % (w/v) APS	57.5 μL	57.5 μL	10 % (w/v) APS	50 μL
TEMED	5 μL	5 μL	TEMED	7.5 μL

2.3.7 Staining

Staining was performed by shaking overnight in a Coomassie blue solution, or for approximately 1 hour in Instant-Blue (Expedeon). Destaining was performed using an 80% ethanol, 20% acetic acid solution, mixed 1:1 with water, until the desired de-colouration was observed.

2.3.8 Western Blotting

For Western blots, gels were run as just described. Upon removal of the gel from the running tank, it was washed thoroughly in water. The gels band were transferred to

polyvinylidene fluoride (PVDF) membranes via a Biorad "TransBlot Turbo" electroblotter, using the 7 minute turbo protocol. For washing, antibody binding and visualisation, the Pierce Fast Western Blot kit from Thermo was used according to the manufacturers protocol. A rabbit anti-his monoclonal antibody from Cell Signalling was used as the primary. The secondary was included with the Pierce kit and was a horseradish peroxidase conjugate, which could be visualised in the GelDoc transillumination cabinet upon addition of luminol.

2.4 Bio-physical Techniques

2.4.1 Fluorescence microscopy

For fluorescence microscopy time course studies, cultures were sampled across the growth curve. 2 μL of each time point normalised to an optical density of 0.05 was added to GeneFrames from Thermo, according to the manufacturers instructions. Images of the frames were collected on a Leica DMi8 microscope fitted with a Hamamatsu Flash4 Camera under phase contrast, and with a FITC filter cube for GFP fluorescence.

2.4.2 Circular Dichroism

Circular Dichroism was performed using the JASCO 1500 instrument. Ideal protein concentrations to obtain appropriate HT voltages (not exceeding 600 V) were determined empirically at the time of use by taking single spectral traces at 20 °C and diluting 2-fold as necessary from a 1 mg mL^{-1} stock solution, dialysed in a 0.5 M Sodium Fluoride buffer. NaF is used as a salt substitute in place of $NaCl_2$. Chlorides strongly absorb at around 190 nm, which impedes spectra collection. Similarly, the buffer pH is balanced with acetic acid so as to avoid the chloride group in hydrochloric acid. Measurements were taken between 185 - 260 nm, at a data pitch of 0.2 nm, 1 nm bandwidth, and a scanning speed of 100 nm min^{-1}. Each spectrum was accumulated 6 times and averaged. A buffer only baseline was also run for 6 accumulations and subtracted from the sample spectra after the run.

Once ideal conditions for individual traces are identified, a temperature ramping

gradient experiment was set up, increasing by 2 °C min^{-1}, to a final temperature of 90 °C, with spectra accumulated every 5 °C.

Spectral data were analysed with the online tool `Dichroweb` (Whitmore and Wallace, 2004). Details of reference sets and other analysis parameters are discussed in Chapter 5 on page 138 due to it requiring some empirical experimentation, and results are presented there.

2.4.3 Crystallography

Initial crystallographic screens were set up using ≈ 150 μL of purified protein at between 10 and 15 mg mL^{-1}. Crystallisation conditions were screened in picolitre drop volumes using the mosquito crystal screening robot from TTP Biotech, and several commercially available 96-well format buffer plates; namely, the "Wizard" 1, 2, 3, 4, "SG1", and "Morpheus" screens from Molecular Dimensions. In total, around 400 conditions could be screened in a manner of hours. Progress of crystallisation was checked every few days, and each well of the plate was examined via microscope.

If promising preliminary conditions were identified, the corresponding buffer was made up in larger volume, and an increased buffer and protein crystal "sitting drop" was set up to obtain fully sized crystals for diffraction testing.

2.5 Bioinformatics Methods

Bioinformatics workflows often require a great many different tools for different purposes, and it is beyond the scope and remit of this thesis to discuss the intricacies of all of them. Here, an overview of their purpose in this study is given, and where necessary/relevant, the concept underlying the tool. Where specified parameters may have influenced the result of the computation, those parameters are provided here. Any specific scripts for file manipulation, analysis, or visualisation are given in **??** on page ??. As is conventional in computer science and bioinformatics fora, names of scripts and programs will be given in `monospaced font`. Except where explicitly stated otherwise, software was used with its default/recommended parameters. Work was performed mostly on our local server (a ProLiant DL385p Gen8, with 32x AMD Opteron 6380s, 377 GB DDR3 RAM).

For structural simulation, we fortunately had access to a pre-public early beta version of the now-completed MRC CLIMB infrastructure (Cloud Infrastructure for Microbial Bioinformatics)(Connor et al., 2016) (more information in Section 2.5.9 on page 92.

For general file manipulation and miscellaneous tasks, various bespoke `bash` and `python` scripts were used. For inter-conversion of bioinformatic file formats, `BioPython` was a primary tool (Cock et al., 2009).

2.5.1 Quality Control

Short read sequencing obtained from MiSeq runs was assembled in-house. The retrieved sequences are examined for quality before assembly. Raw reads were first analysed with `FastQC v0.10.1` and optionally trimmed with `seqtk v1.0-r31`.

2.5.2 Assembly

Sequence files passing quality control were *de novo* assembled using `SPAdes v2.5.1`, with the `--careful` flag, to reduce errors (Bankevich et al., 2012). Optionally, the resulting contigs (if not a single sequence) were reordered to published reference genomes, mainly for visualisation purposes, using `progressiveMauve v2.3.1` (Darling et al., 2010).

2.5.3 Mapping

Mapping for examining coverage etc. was performed with `bwa v0.7.5a-r405` (Li and Durbin, 2009). Coverage and quality estimates were calculated from these alignments, and visualised with, `QualiMap v.0.7.1` (García-Alcalde et al., 2012)

2.5.4 Annotation

Annotation was performed with the prokaryotic annotation pipeline `prokka v1.11` (Seemann, 2014). A set of preferred/trusted annotations was provided with the `--proteins` option, compiled from the published *P. luminescens* TT01, *P. asymbiotica* ATCC43949 and *P. asymbiotica* Kingscliff genomes as they contained some bespoke annotations from legacy use within the lab.

2.5.5 Alignment

Multiple sequence alignments were generated with `Clustal Omega v1.2.0` (Sievers et al., 2011).

2.5.6 Phylogenetics

Initial trees were computed with `RAxML v7.0.3` (Stamatakis, 2006) with 500 rapid bootstraps in a single run ("`-f a`"). Seeds were arbitrarily set at "12345" for all runs, for reproducibility purposes.

Consensus trees were computed with `ASTRAL-II v4.7.12` (Mirarab and Warnow, 2015). As there were at most 16 taxa in the trees provided to ASTRAL, it was run with the "`--exact`" flag for improved accuracy.

2.5.7 Congruency

Congruency was estimated in 2 ways. The Adjusted Wallace Coefficient was used, but required an element of subjectivity. With this in mind, the data was also tested with a less powerful, but entirely objective method - the Normalised Robinson-Foulds distance.

The Adjusted Wallace Coefficient (AWC) builds on a tool which compares how data is clustered via different methods. Much more information about the various metrics, and the technique's use in sequence typing can be found at the Comparing Partitions website[3] (Pinto et al., 2008; Severiano et al., 2011a,b; Carriço et al., 2006). Since the manner in which it was used in this study is valid, but not the norm, some time will be spent explaining the process:

In the case of experimental sequence typing, it is common to predict STs from one or more experimental techniques (e.g eletrophoretic restriction enzyme tests with 2 different restriction enzymes), and researchers commonly wish to see how well they agree on their predicted ST. The Comparing Partitions web-server takes as an input, a matrix where one scores how each taxon label clusters in each tree. This had to be created manually by visually inspecting the clustering behaviour of each tree, for a given taxon label. An arbitrary cluster label is assigned (1 to n), and the taxa that are within that cluster are

[3] `http://www.comparingpartitions.info/index.php?link=Tut8`

assigned its number. Absent taxon labels were just assigned a unique cluster identifier (equivalent to not clustering at all). As this has a large subjective component, clustering was corroborated by several other individuals in a 'blind' manner (no knowledge of how anyone else clustered the trees). While subjective in the manner in which it was used here, the AWC has greater resolution, in that it is an asymmetric measurement. The metric captures some of the discriminatory power of one tree versus another, that is to say, *how clear* a given cluster is. Under the normal use case, one can think of this as how 'definitive' one typing method is versus another.

In brief, the data of interest is clustered by 2 methods of choice. In this case, the clusters would be 2 different phylogenetic trees (clusterings), of operons where the 'method' would be use of different genes (under normal usage, the clusters would be sequence types, and the 'method' might be pulsed field gel eletrophoresis for example). This is repeated for all pairs of clustering methods. A contingency table is constructed from this information, which effectively describes how often the same cluster is predicted by each technique. The AWC then describes the fraction of all the times the same cluster is found between 2 methods, out of all the clusterings in which the taxa appeared. The "Adjusted" part of the metric comes from the added step, in which the "coefficient under independence" is subtracted from the Wallace Coefficient. This is a kind of normalising step, which removes the effects of clusters occurring randomly, leaving only the contributions from meaningful clusterings. Therefore, the final value of the AWC is given by the equation below. For a more detailed explanation, see the link and the references provided at the start of this section. The values returned from this equation are those depicted in resulting figures.

$$AW_{A\to B} = \frac{\mathrm{W}_{A\to B} - W_{i(A\to B)}}{1 - W_{i(A\to B)}} \tag{2.3}$$

Adjusted Wallace Coefficient Definition

The Normalised Robinson-Foulds metric was calculated simply with the inbuilt

`compare` function from the `ETE3 v3.0.0b36` (Huerta-Cepas et al., 2016) toolkit in an all-vs-all pairwise manner for every tree, using the "`--unrooted`" flag. The metric is defined as below, and this is the value used in the resulting figures. In short, the RF metric simply measures the minimum number of topological transformations required to maximise the congruency between 2 trees. The RF metric is one of the most widely used and probably easiest to intuitively understand, as well as being computationally efficient (linear or $O(n)$ running time) (Pattengale et al., 2007). The normalised RF metric is the same calculation, but normalised against the maximum distance 2 trees could have ($2(n-3)$ as there are always 3 fewer nodes than the number of leaves in a tree, if n is the number of taxa present in both trees). RF ignores unshared branches, which is also advantageous for this study due to some gene deletions.

$$nRF(T_1, T_2) = \frac{1}{2}\left(\frac{|B(T_1) - B(T_2)| + |B(T_2) - B(T_1)|}{2(n-3)}\right) \tag{2.4}$$

Normalised Robinson-Foulds Metric Definition

where T_1 and T_2 are 2 trees, and $B(T_i)$ is the set of bipartitions (splits) of Tree i.

2.5.8 Ortholog Detection

For structural studies, the `HHSuite` of programs has proven to give useful and accurate predictions of protein structural orthologs, especially those which have only low confidence or distant homologies known. `HHSuite v2.0.15` (Remmert et al., 2011; Söding et al., 2005) was used extensively in this work, being run iteratively over all the protein sequences of interest at various points, to identify new homologies detected as the databases are continually expanded and improved. The program was invoked with comparatively relaxed parameters in an effort to gather even low quality hits which may be more informative than "hypothetical protein", using a minimum probability of 60, minimum E value of 1×10^{-3}, and was run against the latest version of the PDB70 database available from the HHSuite database site[4]. Results were parsed into a tabular format

[4] `http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/`

using a custom parser (scripts provided in supplementary information, and at GitHub[5]).
Add MultiGeneBlast, if time allows

2.5.9 Structure Prediction

Due to the private state of the MRC CLIMB infrastructure at the time of carrying out this work, almost unilateral access to the Warwick node compute power was available. For the simulations, 12 virtual machines, each of 32 vCPUs (Intel Xeon E5-4610 v2s) and 96 GB of RAM (a total of 384 vCPUs, and 1,152 GB RAM) were used to spawn multithreaded jobs for ≈330 individual protein sequences. It is worth noting however, that structure simulation jobs are seemingly primarily processing speed/thread limited, as the memory requirements for a 32 core server running at or near full compute capacity, only requires about 45 GB of the 96 available, meaning the same workload could be achieved with probably <500 GB of memory.

A local installation of the `I-TASSER v4.4` structural prediction pipeline was implemented on each of the servers, and the ≈330 sequences to be simulated were distributed equally amongst the servers (Yang et al., 2014; Roy et al., 2010; Zhang, 2008). `I-TASSER` is consistently ranked as one of, if not the best, structural prediction suites available in the CASP competitions (Moult et al., 2015).

2.5.10 Structural Analysis

For analysis and visualisation of structures generated in this work, UCSF `Chimera v1.12` (and to a lesser degree the experimental `ChimeraX v0.5`) were used primarily (Pettersen et al., 2004; Goddard et al., 2018). For automated large scale analysis, the commandline implementation of chimera modules `pychimera v0.2.2+3.g3b96991` was used (Rodríguez-Guerra Pedregal and Maréchal, 2018).

Of particular relevance is the `MatchMaker` function within `Chimera`, which was used for calculation of the Root Mean Square Deviation (RMSD) between structures, to assess accuracy, in the default 'best-chain-pair' mode. Scripts for these analyses are also available on GitHub.

[5] `https://github.com/jrjhealey/bioinfo-tools/blob/master/tabulateHHpred.py`

2.5.11 Repeat detection

Repeats in protein sequences were detected automatically via the EMBL-EBI's `Rapid Automatic Detection and Alignment of Repeats program (RADAR) v1.1.1.1`[6]. The most frequent or longest set of repeats were retained after a default number of iterations.

2.5.12 Data Visualisation

Various programs were used to visualise data for different tasks/purposes during this project.

For examining sequence information, `Artemis v16.0.0` (Rutherford et al., 2000) genome browser was a mainstay. For visualisation of smaller data sets, such as operons and plasmids, SnapGene[7] was used. SnapGene was also the standard tool for *in silico* cloning design and plasmid/operon maps in this thesis are rendered from the software. Phylogenetic trees were prepared with `FigTree v1.4.3`[8]

For visualisation of plotted data, such as heatmaps and line graphs, the `ggplot2 v2.2.1.9000` package (Wickham, 2009) within `R/RStudio v3.3.3` was used (RStudio Team, 2015; R Core Team, 2014). Scripts for these are provided in the supplementary information and on GitHub[9]

General figures such as schematics and diagrams were typically prepared with Microsoft Powerpoint, or BioRender[10].

Finish primer tables

flesh out table/figure captions and titles

[6] `https://www.ebi.ac.uk/Tools/pfa/radar/`
[7] GSL Biotech LLC
[8] `http://tree.bio.ed.ac.uk/software/figtree/`
[9] `https://github.com/jrjhealey/bioinformatics/tree/master/Rscripts`
[10] `https://biorender.io/`

Part II

Computational Results

Chapter 3

Structural Bioinformatics of PVC Proteins

3.1 Introduction

Add a suitable epigraph

Make very clear that the sequence is NOT THE WHOLE STORY. We must try to get structural simulation data as structures diverge slower than sequence

be sure to write a section around the claim in this paper https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4137710/ about sequences diverging faster than sequence

Use quote from Leiman et al. (2010) "The evolutionary relationship cannot be detected in their amino acid sequences" and "The crystal structure of the N-terminal fragment the Escherichia coli CFT073 VgrG protein encoded by ORF c3393 shows a significant structural similarity to the gp5-gp27 complex, despite only 13% sequence identity [84]"

Make a high-level chart of whether genes are more like phage or T6ss e.g. T4Like ¡——- PVC1 ———¿ T6SS

As large multi-partite biological complexes, there are far too many proteins involved in the biology of PVCs to study them all in detail experimentally within a single thesis. However, due to the increasing availability of high performance computing resources, it is possible to study all the genes and proteins, to some extent and to reasonable confidence, using bioinformatics approaches. This chapter attempts to glean as many clues as possible

about structure and function of the proteins involved in construction of PVCs, by 'brute force'. In this particular case, a pre-public beta version of the Medical Research Council Cloud Infrastructure for Microbial Bioinformatics (MRC CLIMB) HPC was used and over 400 compute cores were available. By querying the sequences against structure databases and conducting simple structure simulations, hopefully new insights can be discovered which can inform lab experiments.

Since there are numerous PVC operons that have been identified, and each one contains >16 proteins, the dataset quickly becomes unmanageable for the experimentalist. A rigorous exploration of the hypothetical structures and functions of the proteins can collapse this dataset somewhat, and reveal subtleties of the structures which are interesting or relevant for further study.

As the databases for gene function and protein structure are continuously updated and improved, this chapter is also a 'revisit' to the now outdated genome annotations that were first put together when the strains were sequenced. By re-running these analyses at various points, new putative functions and homologies may be discovered.

Chapter Aims

This chapter aims to confirm, and generate new, hypotheses about the structural elements of the PVCs via structural bioinformatics approaches. In the absence of a true resolved structure, simulations based upon related structures can be remarkably close to the real thing. This chapter also serves as a 'tour' of the PVCs in depth, providing some context for the subsequent chapters. Simulations allow us to get our first real 'look' at the gross structure of the PVCs, and to propose hypotheses for experimental testing.

- Explore the structural orthologies shared by PVC proteins which currently lack informative annotations.
- Use structural simulation approaches to get a 'first look' at the potential structure of the PVCs.
- Examine any high quality simulations for physical characteristics of the PVCs.

3.2 Methods

rename this section

Many proteins in the existing *Photorhabdus* annotations are listed as 'hypothetical', and in the case of some older genome annotations for some of the more diverse PVC elements, this can be the entire operon (Duchaud et al., 2003). Since an annotation of 'hypothetical' leaves nothing to inform experiments, short of blind cloning/deleting and structural resolution attempts, a logical first step seemed to be to assess each CDS within 16 PVC operons for any structural similarity (even at comparatively low scores) to glean as much information as possible and form further hypotheses about their roles. Not only does this provide better functional predictions that the existing ones, but simply querying against a more up to date database often turns up previously unseen similarities between proteins.

3.2.1 Annotation

From a previous project, a number of *Photorhabdus* genomes were sequenced, and a number of existing sequences in NCBI were reassembled and re-annotated along with them for consistency. In all of the work conducted, we utilised the consistent, re-annotated sequences and any given locus tags will correspond to these. Genomes were annotated with a database of existing *Photorhabdus* proteins, utilising Prokka (Seemann, 2014) (see Chapter 2 on page 61, Section 2.5.4 on page 88). All current annotated genomes are provided in supplementary information for this chapter.

Discuss differences arising from reannotation?

3.2.2 Hidden Markov Model Homology Searching

As the Protein DataBank and other databases are frequently updated, Hidden Markov Model searches were run repeatedly throughout the course of the project, usually picking up at least 2 or 3 improved structural annotations, with each new run. This was performed using `HHsearch` from the HHsuite of tools (Remmert et al., 2011). Searches for 312 proteins were run via a commandline implementation of HHsearch v 2.0.15 on a Ubuntu server, with the following parameters: E-value cutoff = 1×10^{-3}, Probability cutoff = 60, and returned the top 10 hits. The searches were queried against the PDB database in each instance, having downloaded the latest version before each run.

Hidden Markov Models ("HMMs") are a sensitive way of searching for sequence similarity, that can outperform tools such as BLAST in certain situations. A Markov Model can be thought of as representing each position in the sequence as being one of many different amino acid possibilities, which are weighted. This arise from the fact that not all amino acids are equally likely to appear adjacent to one another - for instance, a stretch of amino acids, all of which can form β-sheets, are more likely to appear near one another than would an amino acid which contributes to helices, and thus HMMs capture domain information very well.

Test for bimodal distribution of HHpred E-value scores?

The latest full table of results for each gene can be found in the supplementary information.

3.3 Exploration of the structure of PVCs by functional unit

It has been made abundantly clear that it is insufficient to consider sequence similarity alone when comparing structural proteins. Sequences are at liberty to diverge, and if the structure they give rise to is particularly robust, the 'space' that the sequence has to drift in is even larger. This is a generally observed phenomenon, but appears to be particularly true for many of the proteins in contractile tail structres. One postulate for this is that phage represent an extremely ancient domain of life, and spend a significant amount of their life cycle outside of the protective environment of the cell they infect. Thus their proteins have evolved over aeons to become particularly stable and robust. The arms race associated with infection cycles has also no doubt driven the diversification of these proteins in an effort to avoid immune mechanisms of their hosts. It has been observed many times in the related literature that, for example, the vgrG/gp5-gp27 spike complex of these caudate systems look almost identical structurally, with many of the same domains identifiable such as the OB-fold, yet may share as little as 12% protein sequence identity, and due to the slower evolution of proteins sequences attributable codon redundancy, the corresponding DNA sequences may be even less similar.

Consequently, we must make efforts to study the structure as best as possible. In silico methods are improiving all the time, and with more computer power than ever,

simulations are becoming routinely feasible. threading approaches are not ideal as they are still too dependent on first identifyin sequence similarity. ab initio approaches allow the structures to be refined without a dependence on the sequence, which should offer an improvement. Structurally conserved proteins with a high degree of robustness should therefore naturally coalesce toward the same structure.

3.3.1 The PVC tube

Image of PVC1-5 in locus position

Among the better annotated genes at the outset of this work, the first 5 loci of the PVCs, are predicted to match phage tail tube proteins, though the existing annotations were not much more informative than this (the vast majority of which were "hypothetical proteins"). After re-annotation, these genes are consistently annotated as T4-like virus tail tube or baseplate proteins (orthologs of gp6/gp19) and sheath proteins from the recently resolved *Pseudomonas aeruginosa* R-type pyocin. From the resolved structure databases and literature, gp19 is known to be the inner sheath of the T4 bacteriophage (as can be seen in PDB IDs 5IV5 and 5W5F (Taylor et al., 2016b; Zheng et al., 2017)), and the outer sheath of PDB ID 3J9Q which corresponds to the resolved pyocin tube structure (Ge et al., 2015b). Over several iterations of homology searches with the HHpred suite, these 3 recent PDB depositions have come to be the most highly similar structures predicted, though in past results, the best hits have included Type 6 Secretion System components from *Edwardsiella tarda* (for the outer sheath proteins).

These proteins comprise the bulk of the PVC, with electron microscopy estimates suggesting that they are about 200 nm in length, and therefore probably include as many as 50 hexameric 'donuts' of each, taking in to account the measurements in **??** on page ??

The upcoming figures demonstrate the simulated structure similarities to the published known structures. In each case, as there are up to 5 simulated models per locus, and up to 16 alleles, so for simplicity, the 'best' model, from the best fitting is, and an conservation map by overlaying the multiple sequence alignments in Appedix BLAH to encapsulate the variability, as well as a strucuture overlay.

For the outer sheath, there are 3 loci, for there are 1 loci, so each of these are shown

separately

Immunogenicity profiling of exterior sheath

Electrostatic comparisons of interior tube proteins + general comparisons

comparisons of PVCs2,3,4 to try and understand their paralogy?

table of HHpred matches

3.4 Discussion

PVCs are a hybrid between T4 and pyocin like structures, with an inner sheath most resembling the former, and an outer sheath the latter.

Table of HHpred results in appendix

Correlation between sequence similarity and structure similarity

Chapter 4

Comparative Phylogenetics of PVC Operons

"You should use more mathematics, like we do."

Richard P. Feynman

4.1 Introduction

The PVCs are complex operons for which the paradoxical idiom "the same but different" very much applies. Of the 16 operons observed in the 3 strains most commonly studied in the lab, there few real 'hard-and-fast' rules that can applied to all of them - other than that they elaborate the same ultimate structure. Just with some simple 'sequence-gazing', quite drastic differences can be identified easily.

There have been quite extensive studies of analogous systems to the PVCs, such as phage (see (Yap and Rossmann, 2014b) for a good review), R-type pyocins/tailocins (Ge et al., 2015a; Ghequire and De Mot, 2015), and membrane bound secretion systems (Cascales and Cambillau, 2012), that can be found in the literature (Sarris et al., 2014; Kube and Wendler, 2015b) (these are just illustrative examples, a more exhaustive literature search can be found in Chapter 1 on page 2). However, these types of comparison studies tend to focus on the common features between these systems, without paying much, if any, attention to what it is that makes them different (e.g. identifying them all as

contractile mechanisms). Given the diversity seen among PVC elements within even the same genome, it seems clear that the Devil is in the detail, and it's actually what sets each PVC apart from one another that is of most interest, given the 'effort' *Photorhabdus* is going to, to maintain 5-6 highly paralogous sequences.

To date, there has been no real attempt to perform a systematic analysis of all of the operons, and much of what is known of the functions of genes within has been predicted from (now aged) genome annotations and simple BLAST studies 'by hand'. The Sarris et al. paper attempted to do a systematic study of contractile tail structures across many genera, but at the expense of studying any of them in great detail, and again, focussed on the common details, defining a 'consensus operon'. In this chapter, this is addressed within the scope of PVCs specifically, highlighting the micro-evolution that sets these operons apart from one another, and from related structures in other genera.

The micro-evolution within the operons was examined here via a phylogenetic congruency workflow. Genes within the PVC operons are compared for their sequence divergence and ability to accurately represent the known phylogenetic history of the genus. Those which are found to be incongruent are inferred to be evolving differentially. The chapter speculates, based on the clustering of PVCs with their effectors, how interchangeable PVC components may be, versus whether they are honed in some way to each of their cognate effectors. Additionally, this chapter attempts to define the hallmarks of PVCs, such that contractile tail like systems in as yet unstudied genomes can be identified, and demarcated from other contractile tail like structures.

Chapter Aims:

- Create a systematic, comparative analysis of genes within PVC operons.
- Establish the likelihood and extent of recombination within the operons.
- Establish a criteria/framework for identifying PVC-like elements in additional genomes.

4.2 Phylogenetics and Congruency Analysis

This section describes, at a higher level, the workflow and concepts required for the analysis conducted. Specific details of algorithmic parameters, software versions and other technical details are reserved for Chapter 2 on page 61.

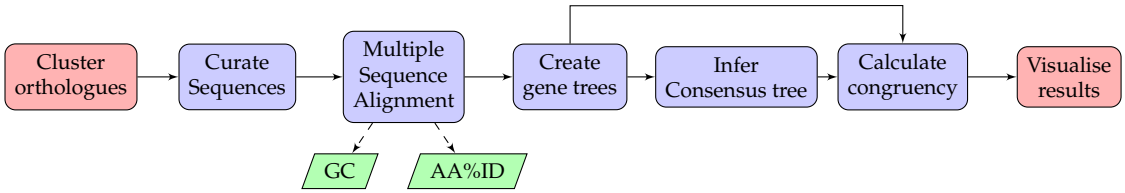


Figure 4.1 | A flowchart to demonstrate, at high level, the steps involved in the process of the congruency analysis presented in this chapter.

4.2.1 Syntenic Clustering of Orthologs

In order to analyse each gene, they had to be separated out in to syntenic clusters. Since we elected to use 16 operons, totalling around 300 genes, the list was curated manually. This was preferable in this instance versus a computational approach (e.g. syntenic clustering with programs), due to the differences between operons where genes may be missing or unique, which complicates the process. It was also not possible to classify the sequences on sequence alone, as several of the genes within a PVC operon are direct paralogs of one another, and would thus be combined in to the same cluster if done by sequence alone, resulting in the comparison of locus 1 and locus 5 for example. Table 4.1 on page 106 shows the clustering and nomenclature arrived at which remained consistent for the duration of the analyses. In the cases where a gene appeared to have been deleted/lost, the functional predictions from Chapter 3 on page 95, sequence similarity, locus length, and synteny with the neighbouring genes was examined. If the loss looked to be a legitimate deletion (i.e. neighbouring genes obviously belonged in other clusters), it was marked down as absent - the analysis was set up in such a way as for this not to matter however.

Additionally, for congruency analysis on a gene-by-gene basis, only the first 16 genes of the operon are used (hereafter, PVC1 to PVC16). There were a number of reasons for

this. Firstly, in each PVC, there are toxin genes in the region downstream of PVC16, but there aren't always the same number, and each toxin can be completely different (they are comparatively well characterised and can be seen to be non-orthologous). Being unrelated, their alignments and resulting trees would most likely be spurious, and the resulting trees would have as few as 2 members, which is obviously not possible.

Moreover, the fact that each of the toxin genes is known to be different between each PVC (to the extent that they are used to differentiate between operons), means that the question of whether this region of the operon is recombinant seems to be answered from the outset. Secondly, between PVC16 and the toxin genes is an extremely variable region. The additional information used for classifying genes earlier in the operon is lacking for genes in the PVC16+ region, thus it was not possible to sufficiently well disentangle this region of the operon, as the genes have no known functions and no ontological information in existing databases due to their lack of similarity to anything currently known. Many of these genes are unique, with no analogous genes. The workflow described here is tolerant to the deletion of members of a group, as long as there are other members within the group to compare with (a deletion is penalised as an incongruency). It was decided it would make for a simpler and more robust analysis to disregard these genes.

4.2.1.1 Curation of the anomalous lumt operon

Curation of orthologs for the lumt operon proved to be more complicated, as the operon architecture is more distinctive; it has lost a couple of genes (one more so in ATCC43949 than in Kingscliff) and gained several others. Lumt was curated last, once all the other operons were clustered effectively. Because of this a few additional CDSs were discarded from the operon for this analysis. Firstly, there is an additional 5' preceding gene, referred to as PVC0 (refer to Chapter 3 on page 95, which belongs only to those operons. It is unclear as yet what, if any, role this protein has. Recent structural similarity searching explored in Chapter 3 on page 95 has found high confidence hits, but without any clear indication of its involvement in the PVC structure/function. With no equivalent orthologs, it cannot be included in this workflow.

Both lumt operons harbour an additional paralogue of PVC11, which appears to be

similar to the gp6 phage baseplate - one of these paralogues for each lumt operon was retained as the representative for locus 11. Additionally the lumt operon has several genes toward the 3' end which do not match well to clusters in any of the other operons, have no well defined functions/orthologies, and throw out the numbering scheme commonly used for all the other operons. Specifically, in orthologous pairs: PAU02194 & PAK02000, PAU02193 & PAK01999, and PAU02192 & PAK01998. Each of these genes are present with their counterpart in the lumt operons from the USA and Kingscliff strains (respectively), but with no equivalent representative in any of the other PVCs.

Table 4.1 | The final clustering of CDS features for phylogenetic analysis. Where a cell is blank, a gene deletion was observed.

	"PNF"		"CIF"			"LOPT"			"UNIT4"		"UNIT2"		"UNIT1"	"UNIT3"	"LUMT"	
ORF	43949	Kingscliff	43949	Kingscliff	TT01	43949	Kingscliff	TT01	43949	TT01	Kingscliff	TT01	TT01	TT01	43949	Kingscliff
PVC1	PAU03392	PAK03203	PAU01961	PAK01787	PLT02568	PAU02074	PAK01896	PLT02424	PAU02775	PLT01696	PAK02606	PLT01736	PLT01758	PLT01716	PAU02206	PAK02014
PVC2	PAU03391	PAK03202	PAU01962	PAK01788	PLT02567	PAU02073	PAK01895	PLT02425	PAU02776	PLT01695	PAK02607	PLT01735	PLT01757	PLT01715	PAU02205	PAK02013
PVC3	PAU03390	PAK03201	PAU01963	PAK01789	PLT02566			PLT02426	PAU02777	PLT01694	PAK02608	PLT01734	PLT01756	PLT01714		PAK02012
PVC4	PAU03389	PAK03200	PAU01964	PAK01790	PLT02565	PAU02072	PAK01894	PLT02427	PAU02778	PLT01693	PAK02609	PLT01733	PLT01755	PLT01713	PAU02204	PAK02011
PVC5	PAU03388	PAK03199	PAU01965	PAK01791	PLT02564	PAU02071	PAK01893	PLT02428	PAU02779	PLT01692	PAK02610	PLT01732	PLT01754	PLT01712	PAU02203	PAK02010
PVC6	PAU03387	PAK03198	PAU01966	PAK01792	PLT02563	PAU02070	PAK01892	PLT02429	PAU02780	PLT01691	PAK02611	PLT01731	PLT01753	PLT01711	PAU02202	PAK02009
PVC7	PAU03386	PAK03197	PAU01967	PAK01793	PLT02562	PAU02069	PAK01891	PLT02430	PAU02781	PLT01690	PAK02612	PLT01730	PLT01752	PLT01710	PAU02201	PAK02008
PVC8	PAU03385	PAK03196	PAU01968	PAK01794	PLT02561	PAU02068	PAK01890	PLT02431	PAU02782	PLT01689	PAK02613	PLT01729	PLT01751	PLT01709	PAU02200	PAK02007
PVC9	PAU03384	PAK03195	PAU01969	PAK01795	PLT02560	PAU02067	PAK01889	PLT02432	PAU02783	PLT01688	PAK02614	PLT01728	PLT01750	PLT01708	PAU02199	PAK02006
PVC10	PAU03383	PAK03194	PAU01970	PAK01796	PLT02559	PAU02066	PAK01888	PLT02433	PAU02784	PLT01687	PAK02615	PLT01727	PLT01749	PLT01707	PAU02198	PAK02005
PVC11	PAU03382	PAK03193	PAU01971	PAK01797	PLT02558	PAU02065	PAK01887	PLT02434	PAU02785	PLT01686	PAK02616	PLT01726	PLT01748	PLT01706	PAU02197	PAK02004
PVC12	PAU03381	PAK03192	PAU01972	PAK01798	PLT02557	PAU02064	PAK01886	PLT02435	PAU02786	PLT01685	PAK02617	PLT01725	PLT01747	PLT01705	PAU02196	PAK02002
PVC13	PAU03380	PAK03191	PAU01973	PAK01799	PLT02556				PAU02787	PLT01684	PAK02618	PLT01724	PLT01746	PLT01704	PAU02195	PAK02001
PVC14	PAU03379	PAK03190	PAU01974	PAK01800	PLT02555	PAU02063	PAK01885	PLT02436	PAU02788	PLT01683	PAK02619	PLT01722	PLT01745	PLT01703		
PVC15	PAU03378	PAK03189	PAU01975	PAK01801	PLT02554	PAU02062	PAK01884	PLT02437	PAU02789	PLT01682	PAK02620	PLT01721	PLT01744	PLT01702	PAU02191	PAK01997
PVC16	PAU03377	PAK03188	PAU01976	PAK01802	PLT02553	PAU02061	PAK01883	PLT02438	PAU02790	PLT01681	PAK02621	PLT01720	PLT01743	PLT01701	PAU02190	PAK01996

4.2.2 Curation of Sequences

There is legacy sequencing data published in NCBI for the 3 strains used for this analysis, *Photorhabdus luminescens* TT01, *P. asymbiotica* ATCC43949 (referred to here also as "USA"), and *P. asymbiotica* Kingscliff. These strains were used as they harbour the originally discovered PVC sequences as published by Yang *et al.*(Yang et al., 2006), they are used routinely in the lab for experimental work, and most is known about them. Re-annotated sequences were used as mentioned in Section 3.2.1 on page 97, and any locus tags referred to in this thesis are from the new annotations. There was some slight variation in the re-annotated operons, particularly in the prediction of fewer CDS features within a couple of the PVCs. The features predicted only in the older annotations were likely to be spurious as they were short, lacked similarity to known sequences when BLAST-ed, and were not always identified in all operons. Each CDS feature was extracted as a nucleotide fasta and organised in clusters according to Table 4.1 on page 106.

As a further note on the existing confusing nomenclature; 2 operons were renamed in this study for clarity. Specifically, the PVC operons with the naming system "Unit #" were named as such when discovered, due to their syntenic arrangement within the genome of *P. luminescens*, where 4 PVC cassettes are positioned in tandem, one after another directly (this arrangement is not present in *P. asymbiotica* genomes). When the equivalent PVC was discovered in the *P. asymbiotica* genomes, they were not given consistent names (instead being given the "Unit 1" designation, indicating the first of its type found in that genome). In this study they are renamed based on their homology to the *P. luminescens* counterparts. To state it plainly:

- "PVC Unit 1" in *Photorhabdus* ATCC43949, is most similar to "Unit 4" in *P. luminescens*, and was thus renumbered to be consistent with *P. luminescens* - "PAU_U4"
- "PVC Unit 1" in *Photorhabdus* Kingscliff, is most similar to "Unit 2" in *P. luminescens*, and was thus renumbered to be consistent with *P. luminescens* - "PAK_U2"

4.2.3 Sequence Alignment and Phylogenies

Nucleotide sequences for each CDS cluster were multiply aligned with Clustal Omega (ClustalO) (Sievers et al., 2011) and bootstrapped trees calculated with RAxML (Stamatakis, 2014). Figures 4.5 to 4.19 on pages 110–117 show the resultant phylogenies obtained. All the trees are shown midpoint rooted, with nodes displayed in descending order for consistency and clarity, the trees themselves are unrooted. The equivalent amino acid alignments are given in **??** on page ?? for visualisation.

4.2.3.1 GC Content and CDS Identity Within Operons

With the sequences curated for each PVC locus and alignments produced, basic sequence statistics such as GC content and identity for each position were also gathered, for reference.

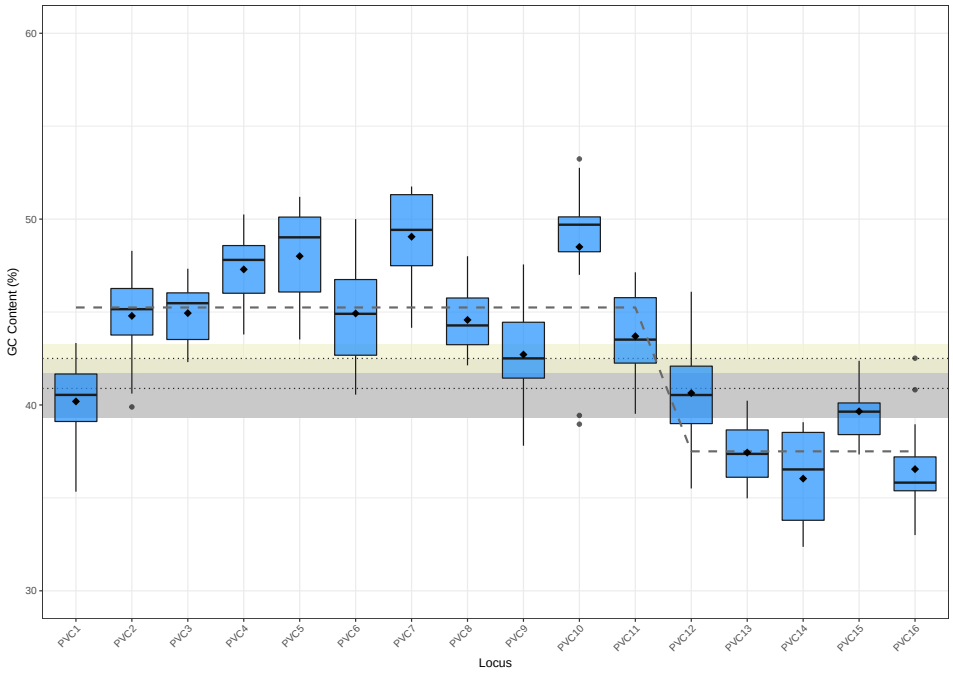


Figure 4.2 | The GC content (%) of PVC genes. There is a trend toward significantly lower GC content at the 3′ end of the operon. ♦ denotes the mean, the • symbols denote extreme outliers outside 1.5× the interquartile range for the sample. The beige box surrounding the upper dotted line shows the mean and standard deviation of the genome GC content. The grey box and lower dotted line depict the same information, but for just the operons.

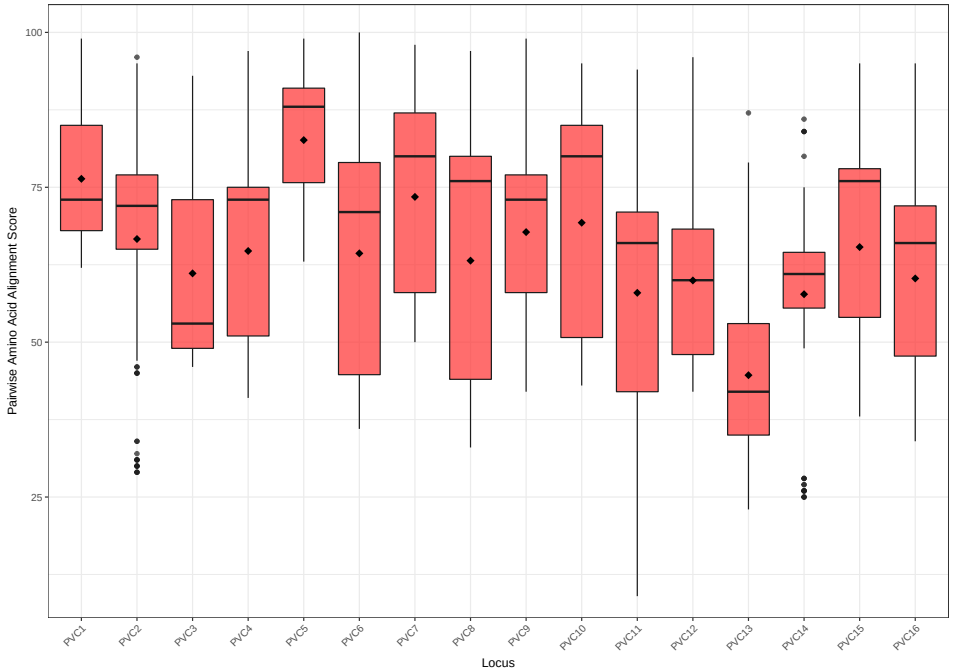


Figure 4.3 | Boxplots depicting the averaged pairwise amino acid identity scores of a multiple sequence alignment of all the sequences within a given syntenic position. ♦ denotes the mean, the • symbols denote extreme outliers outside 1.5× the interquartile range for the sample.

4.2.4 Gene trees

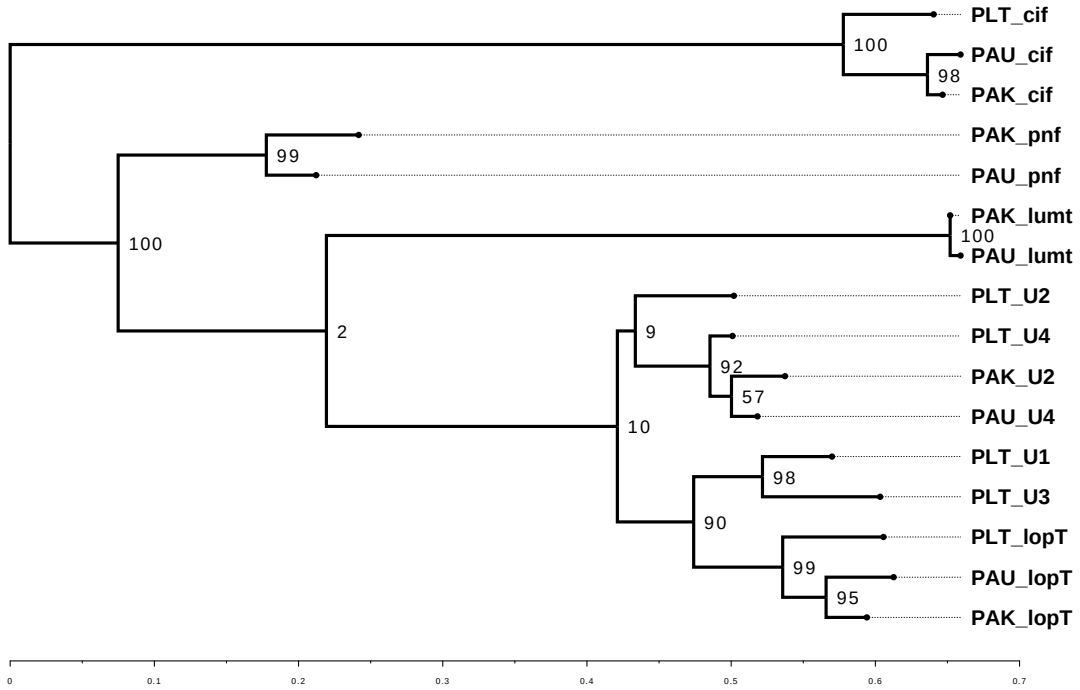


Figure 4.4 | Phylogeny of the locus position (PVC1) from each operon.

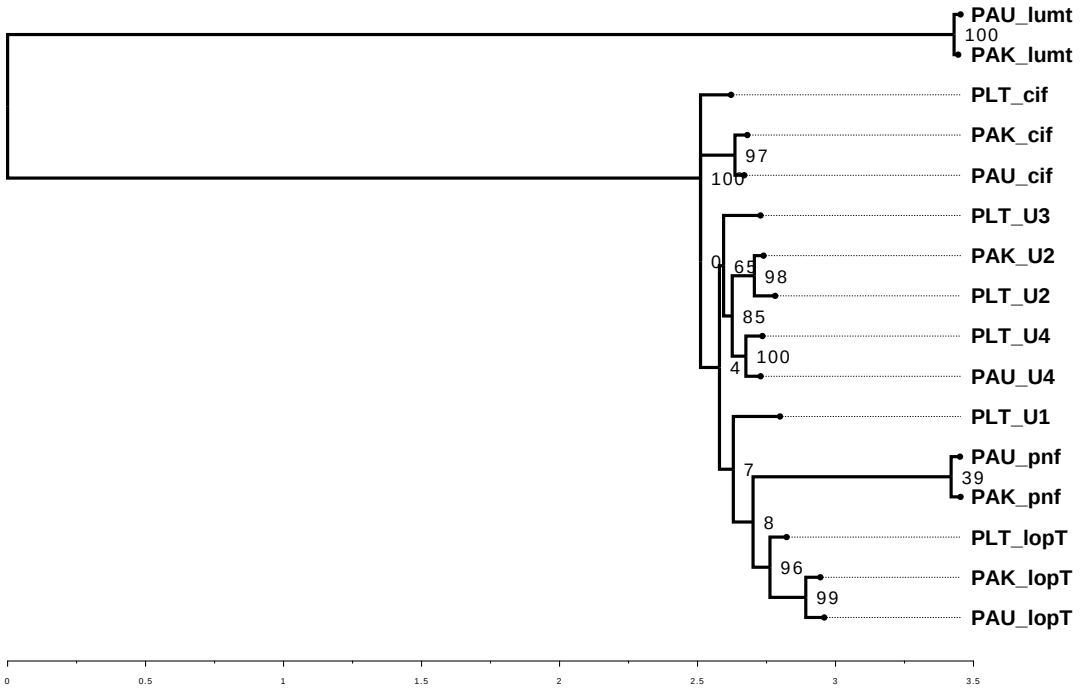


Figure 4.5 | Phylogeny of the locus position (PVC2) from each operon.

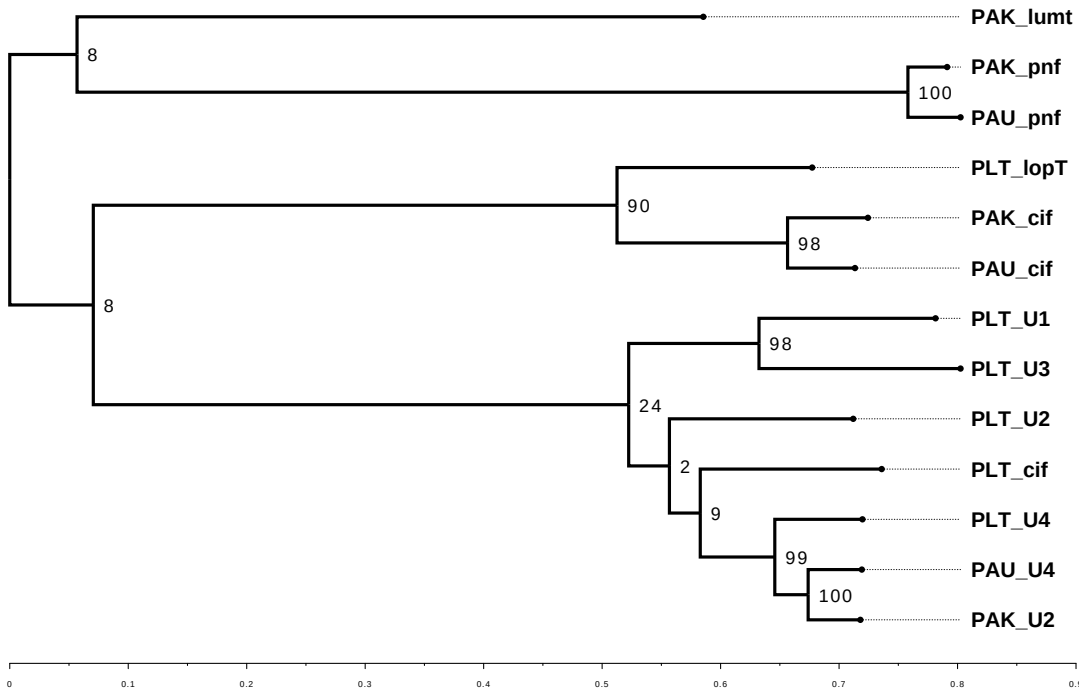


Figure 4.6 | Phylogeny of the locus position (PVC3) from each operon.

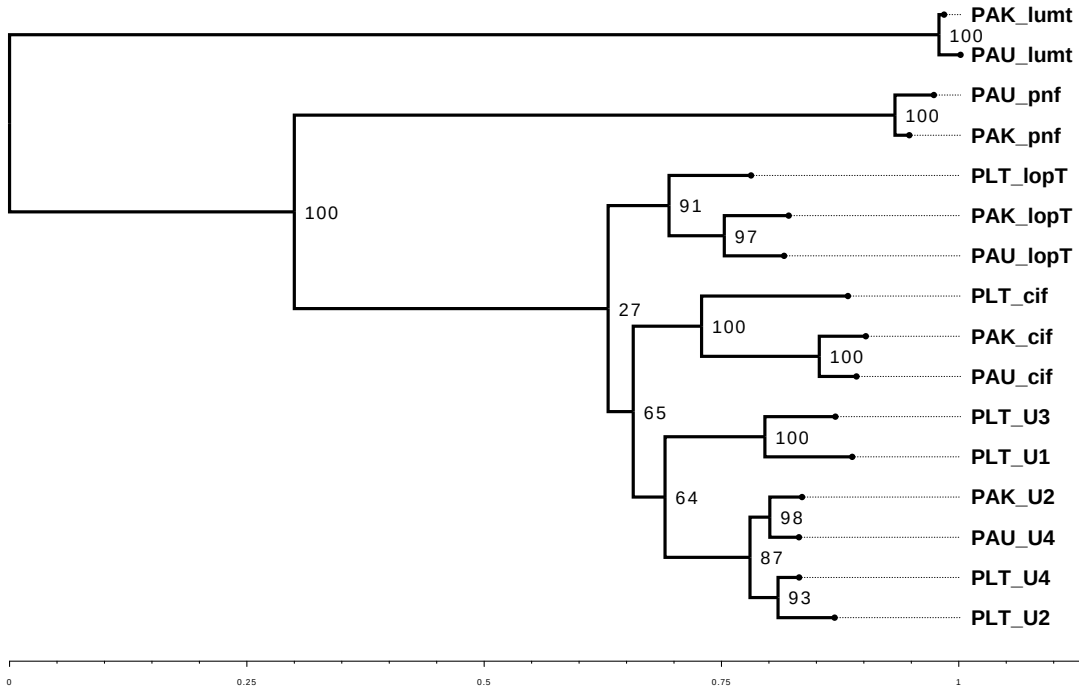


Figure 4.7 | Phylogeny of the locus position (PVC4) from each operon.

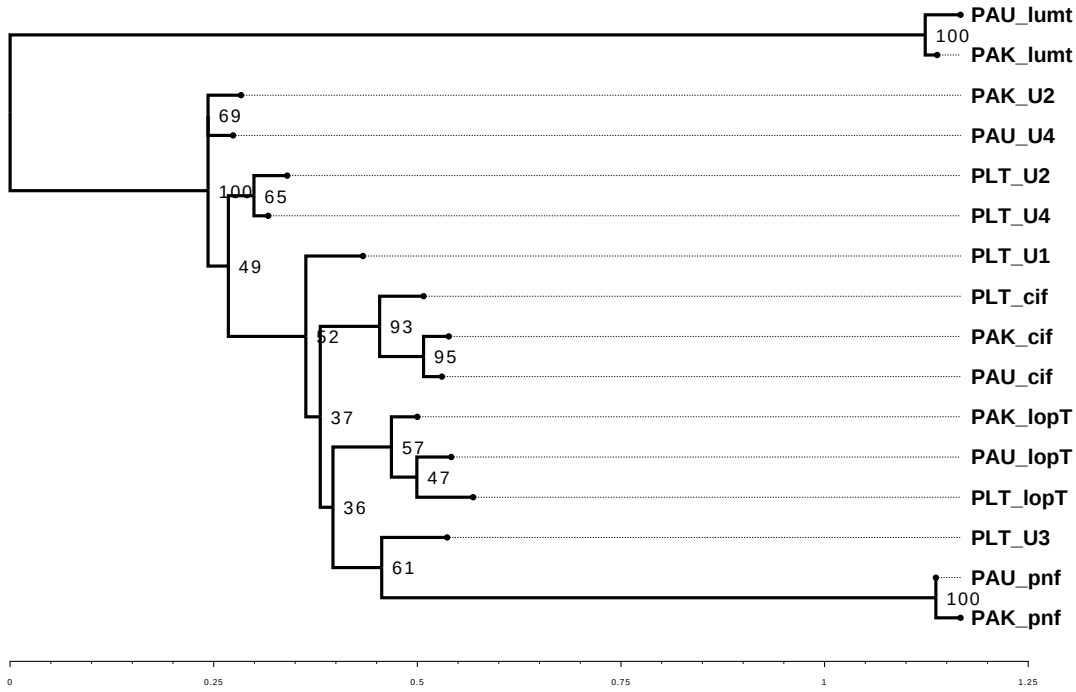


Figure 4.8 | Phylogeny of the locus position (PVC5) from each operon.

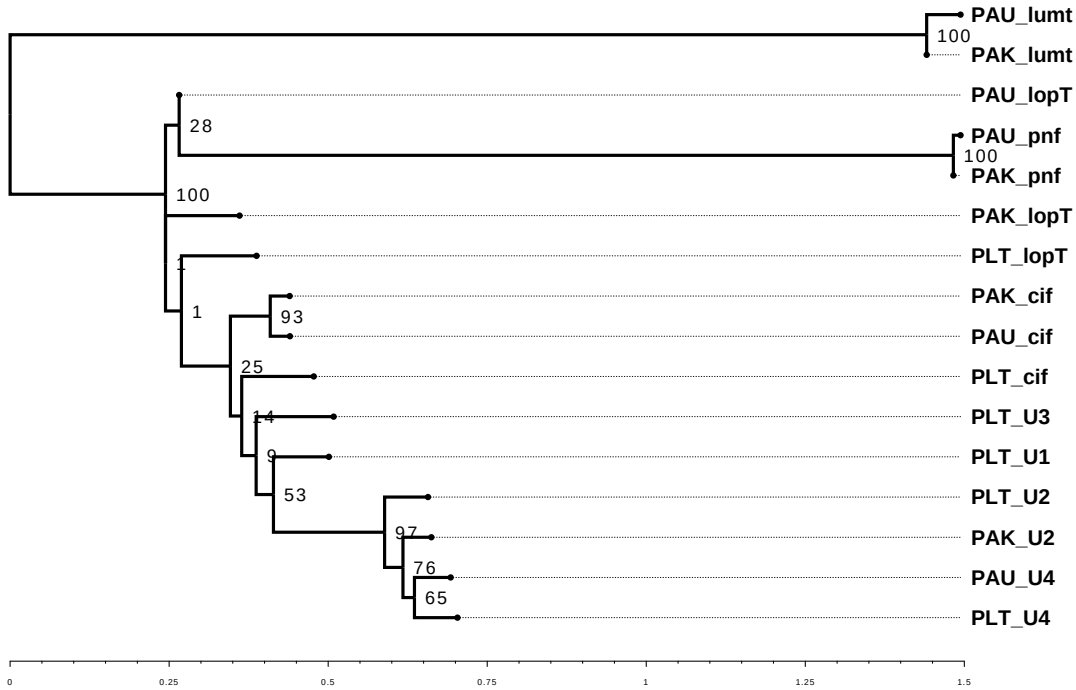


Figure 4.9 | Phylogeny of the locus position (PVC6) from each operon.

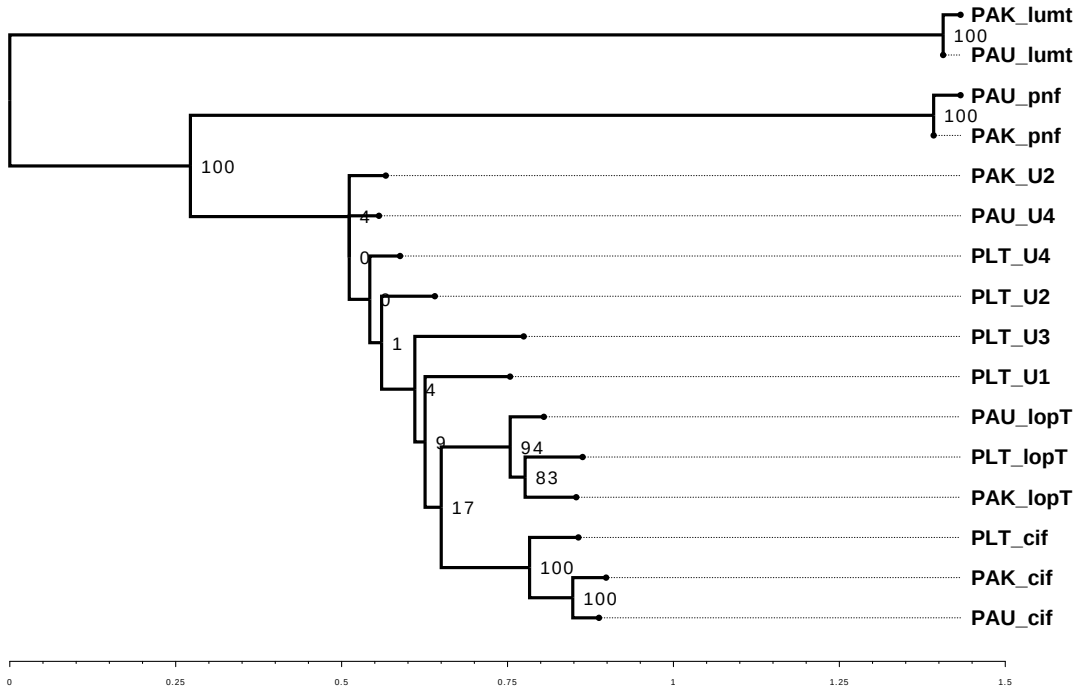


Figure 4.10 | Phylogeny of the locus position (PVC7) from each operon.

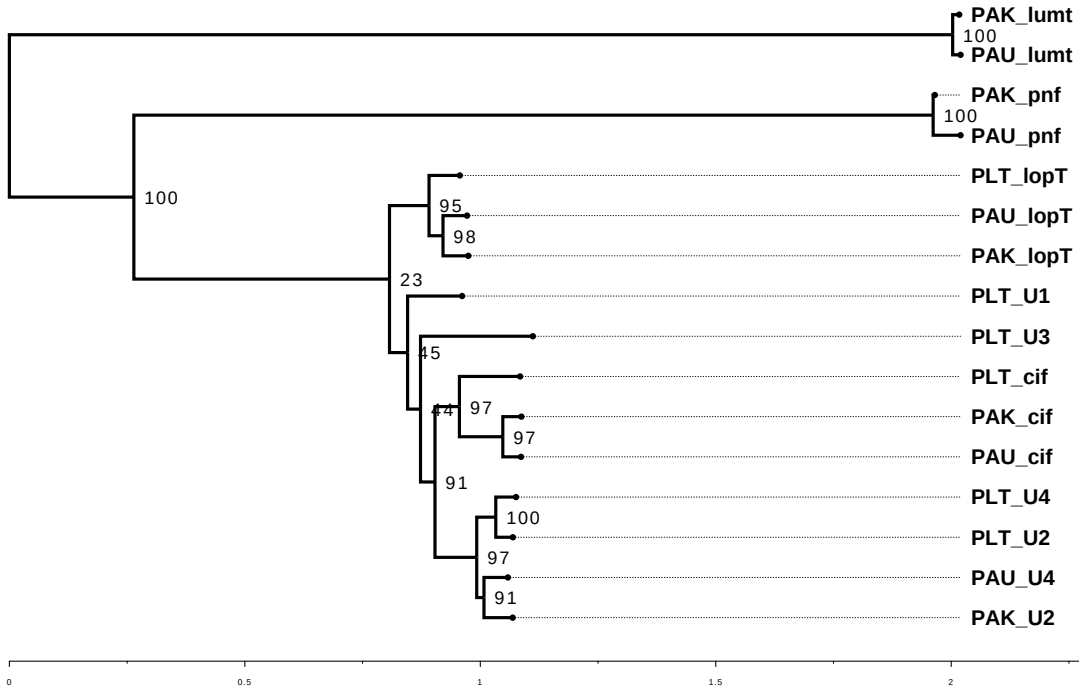


Figure 4.11 | Phylogeny of the locus position (PVC8) from each operon.

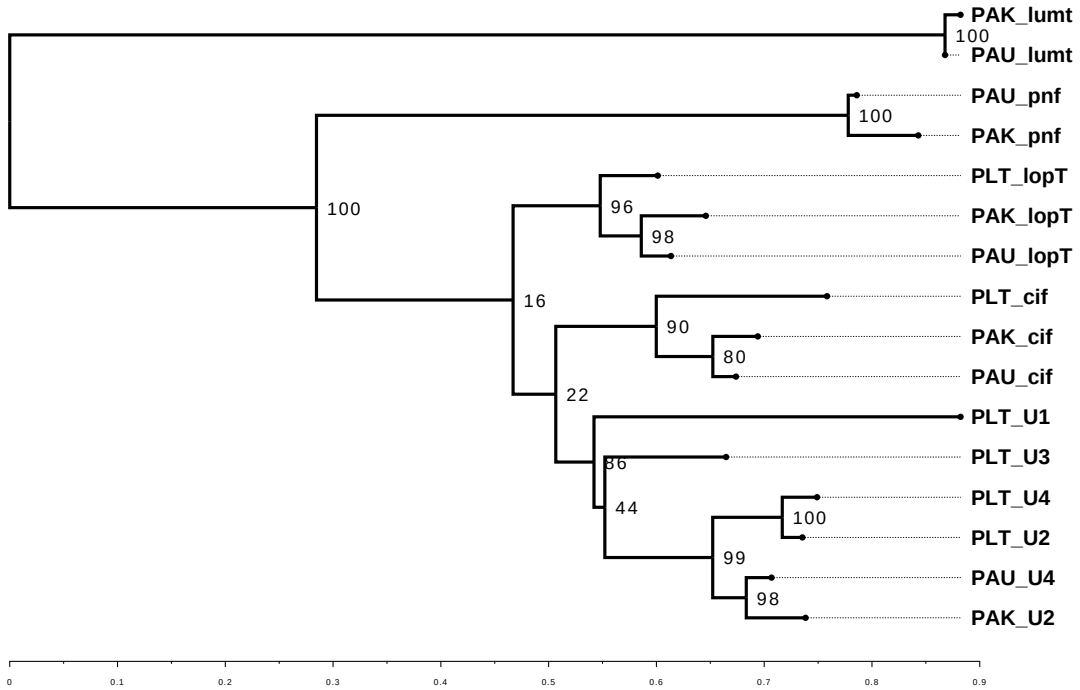


Figure 4.12 | Phylogeny of the locus position (PVC9) from each operon.

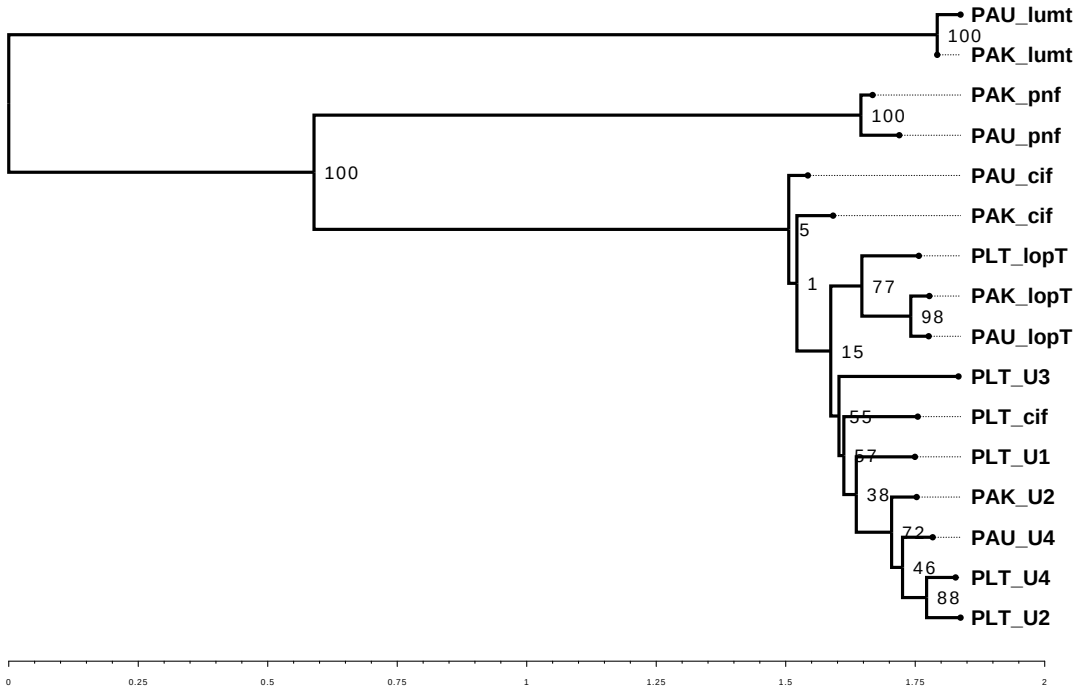


Figure 4.13 | Phylogeny of the locus position (PVC10) from each operon.

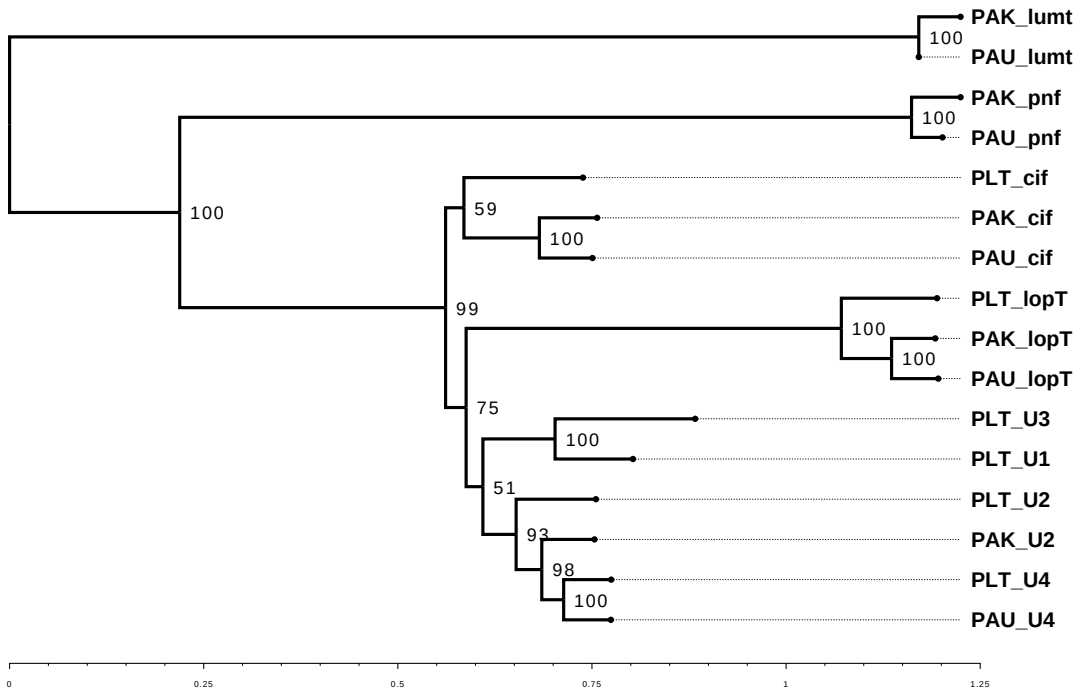


Figure 4.14 | Phylogeny of the locus position (PVC11) from each operon.

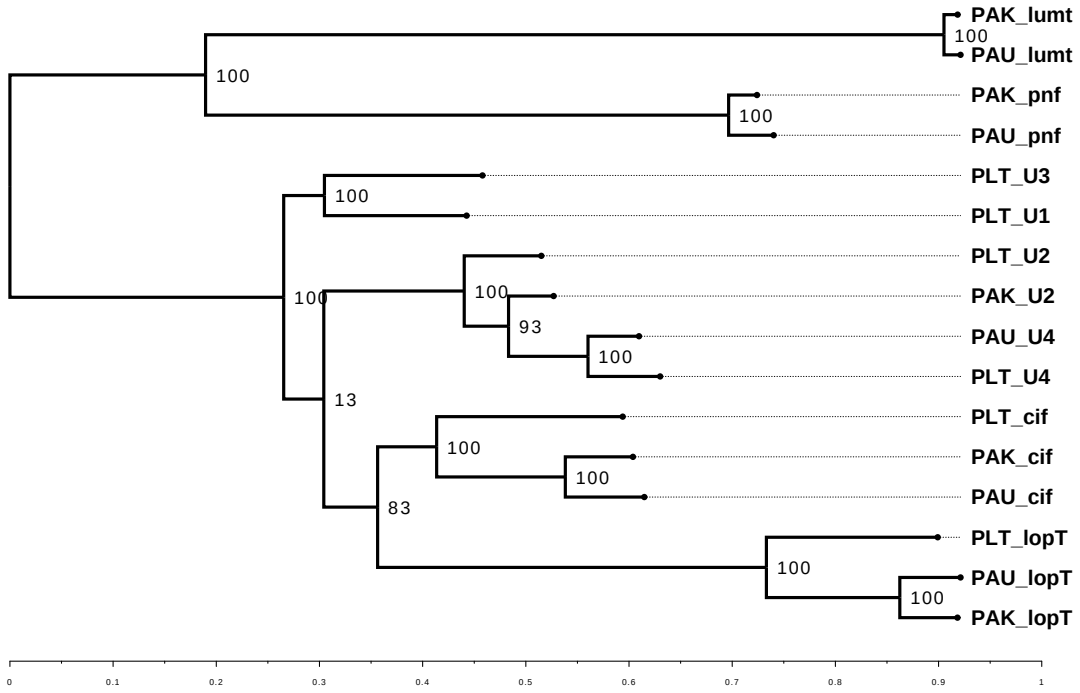


Figure 4.15 | Phylogeny of the locus position (PVC12) from each operon.

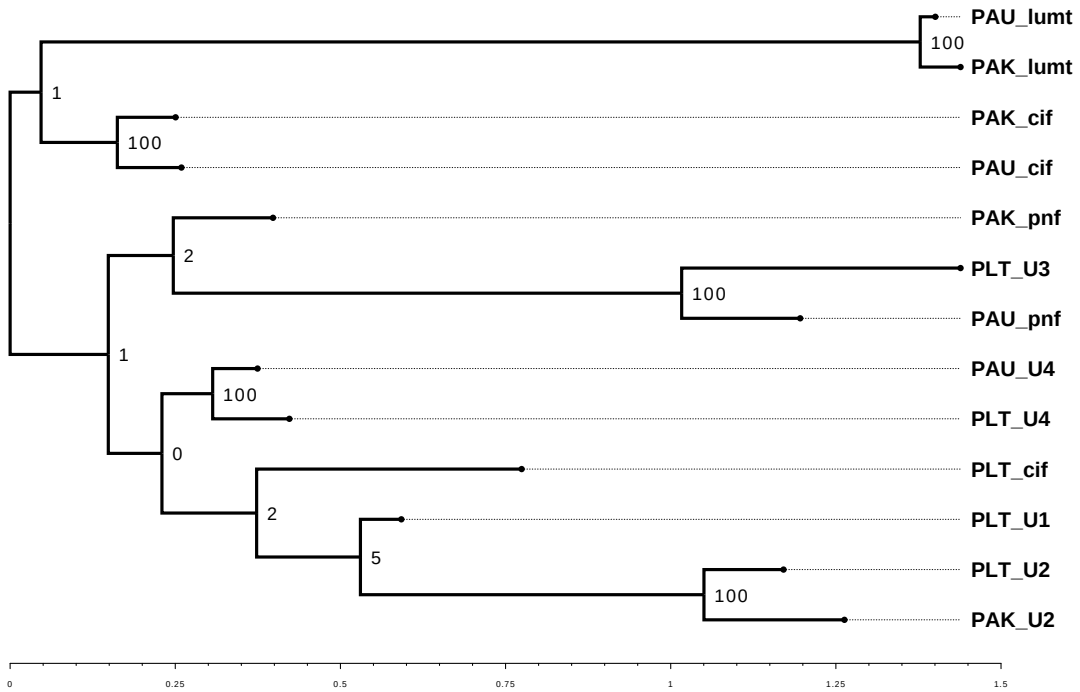


Figure 4.16 | Phylogeny of the locus position (PVC13) from each operon.

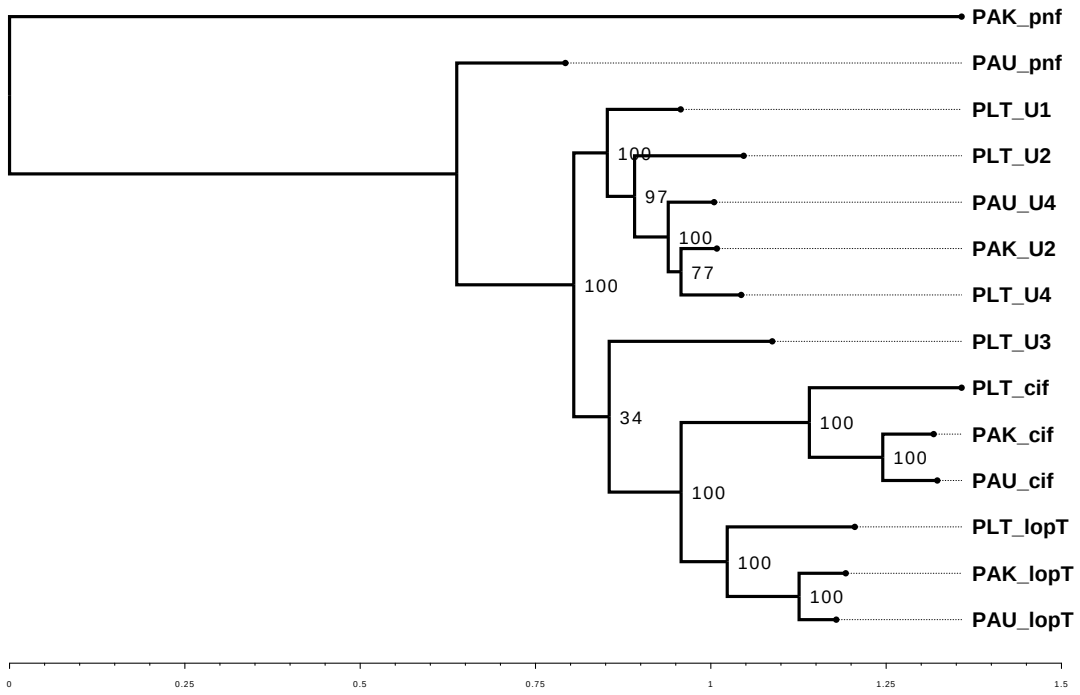


Figure 4.17 | Phylogeny of the locus position (PVC14) from each operon.

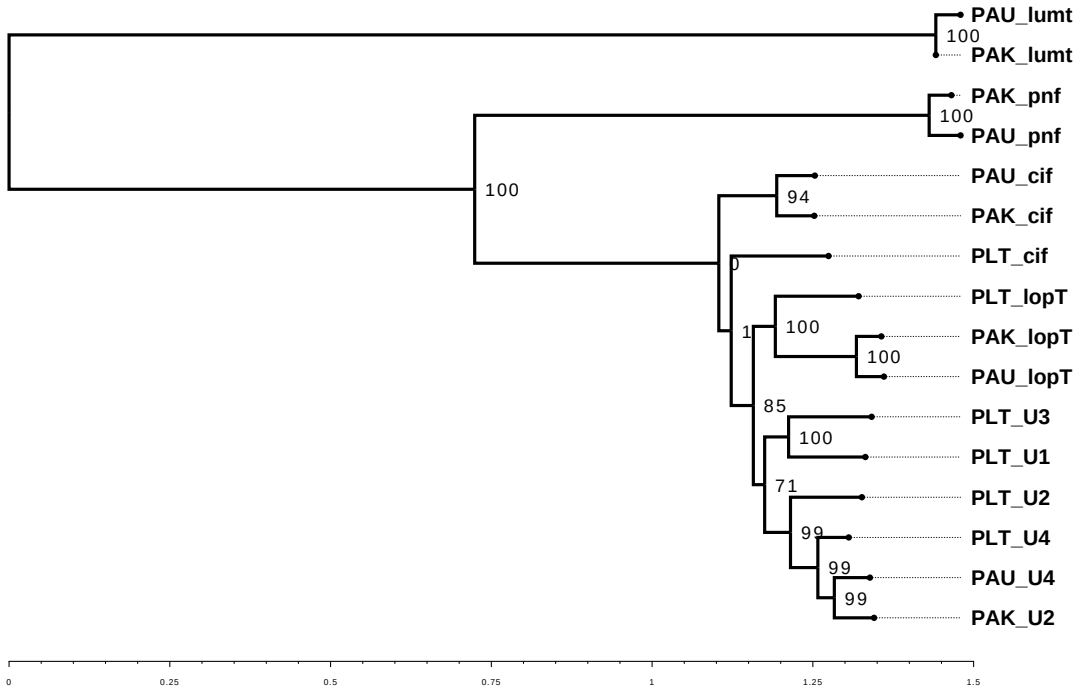


Figure 4.18 | Phylogeny of the locus position (PVC15) from each operon.

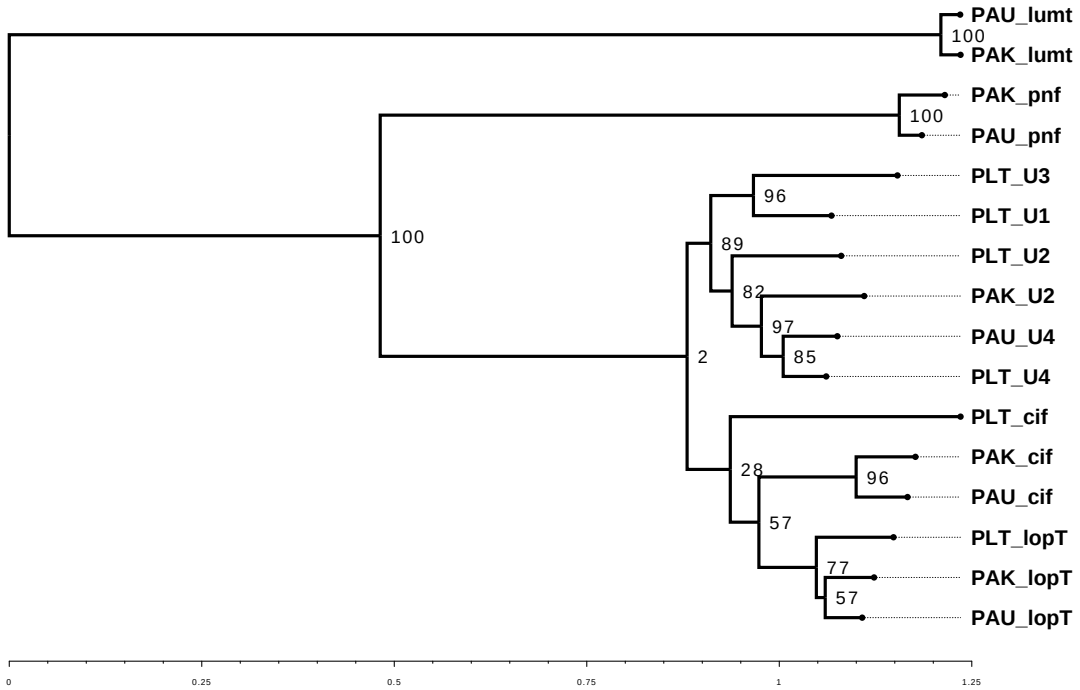


Figure 4.19 | Phylogeny of the locus position (PVC16) from each operon.

4.2.5 Consensus Tree Inference via ASTRAL-II

The penultimate step of the congruency work flow was to infer the consensus tree from just the sequences within the PVC operons. By doing so, we can determine which patterns of evolution from genes within the operon most and least closely follow the known species phylogeny during the congruency analysis. Figure 4.20 shows the inferred phylogeny output by ASTRAL-II (Mirarab and Warnow, 2015). ASTRAL was run with the bootstrap gene trees from RAxML. The software arbitrarily selects a taxa to root from, in this case PLT_U2. The tree is otherwise depicted in decreasing node order for clarity and consistency with the gene trees.

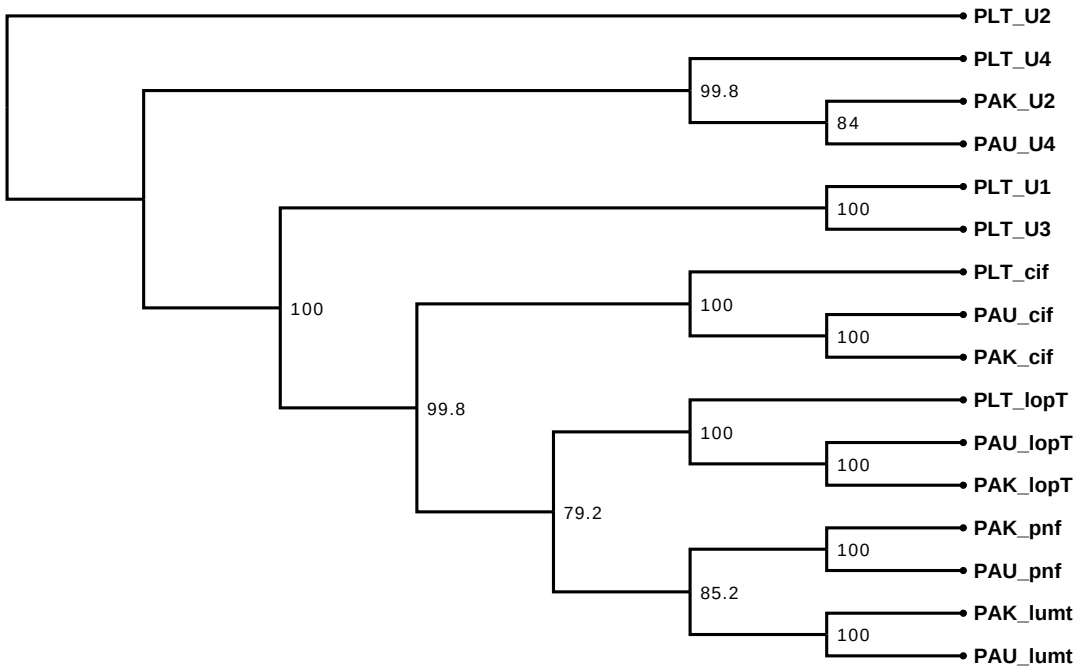


Figure 4.20 | The inferred consensus tree from the genetrees PVC1-16. Branch lengths are arbitrary for this tree and therefore it has been depicted as an cladogram with no scale.

The tree shows well supported branches for all splits, and consistently clusters orthologous operons together (e.g. all Pnfs, Cifs etc.).

4.2.6 Congruency Analysis

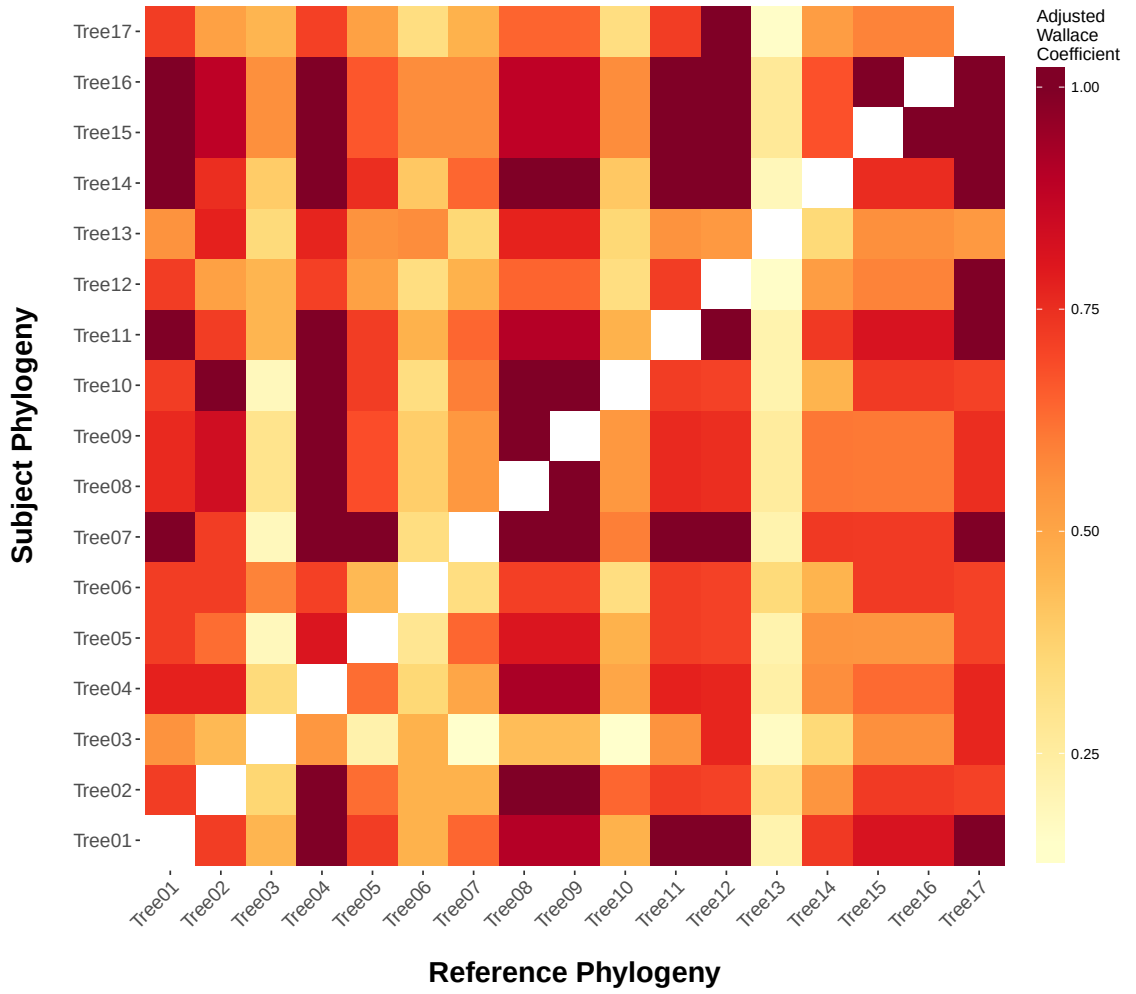
With each gene tree output from RAxML and the 17th tree as the inferred tree from ASTRAL-II, the pairwise congruency between all trees was calculated. Evaluation of congruency metrics ultimately means one is able to put a single number on to a subject tree with respect to a reference, to say how similar the 2 trees are - and ultimately whether they follow the same evolutionary pattern.

4.2.6.1 Adjusted Wallace Coefficient

Congruency was initially tested utilising a metric called the Adjusted Wallace Coefficient (AWC). The Wallace coefficient is one of many used in the study of clustering concordance, but has advantages over others such as the well known Rand metric (Rand, 1971), in that it has a 'directional component' (Wallace, 1983). The Wallace coefficient can be thought of as saying "what is the probability that some data is classified together in test B, knowing that it also was in test A". Details of how the AWC is calculated can be found in Chapter 2 on page 61, and at the associated references.

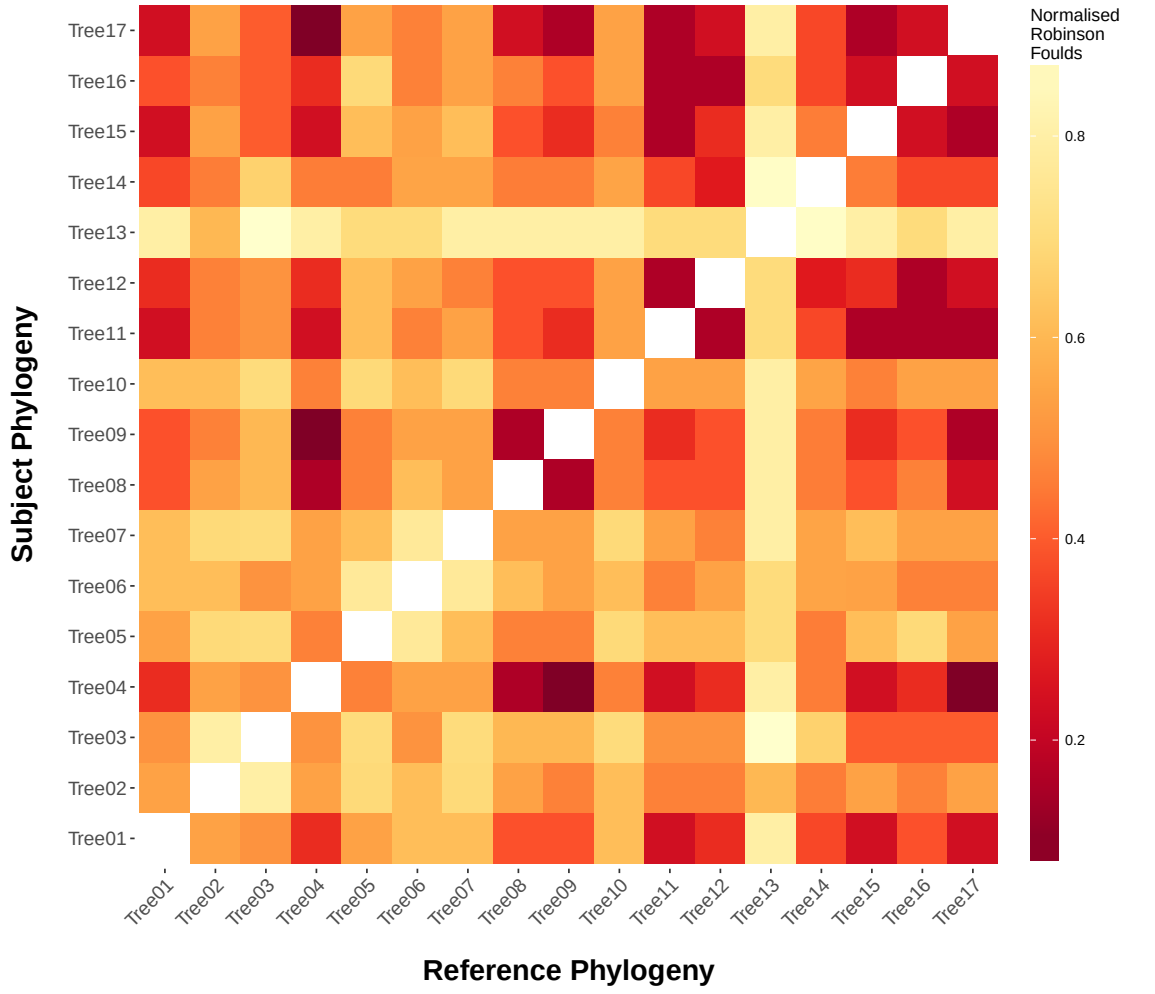
4.2.6.2 Normalised Robinson-Foulds

To address the issue of subjectivity in the previous clustering method, an unbiased although lower resolution technique was used to corroborate the trends - the topological transformation metric developed by Robinson and Foulds ("RF")(Robinson and Foulds, 1981). The RF distance is useful specifically for unrooted trees such as these, since it makes no assumptions about any particular nodes/leaves, it simply calculates the minimum number of topological transformations required to make 2 trees maximally congruent. Because of this, the RF metric is a symmetric one (transforming A = B, is an equivalent number of transforms to make B = A, but reversed with respect to one another).



Tree01	Tree02	Tree03	Tree04	Tree05	Tree06	Tree07	Tree08	Tree09	Tree10	Tree11	Tree12	Tree13	Tree14	Tree15	Tree16	Tree17
Congruency Sums (Reference vs Subject) ($ADW_{A \to B}$)																
12.74	11.75	6.13	14.26	10.12	6.71	8.11	13.53	13.53	7.25	12.74	13.43	3.67	9.11	11.01	11.01	13.43
Congruency Sums (Subject vs Reference) ($ADW_{B \to A}$)																
11.86	10.74	7.12	10.12	9.36	9.69	12.23	10.41	10.41	10.83	11.86	8.91	8.95	12.05	12.55	12.55	8.91

Figure 4.21 | All pairwise comparisons of congruency as measured by the Adjusted Wallace Coefficient. The darker the colour, the better the congruency is. Adjusted Wallace Coefficients of 1 indicate good agreement. The cumulative summed congruency for each locus is displayed below, to quickly show numerically the most and least congruent. There are 2 sets of values for the row and column sums due to the asymmetry of the Adjusted Wallace Coefficient.



Tree01	Tree02	Tree03	Tree04	Tree05	Tree06	Tree07	Tree08	Tree09	Tree10	Tree11	Tree12	Tree13	Tree14	Tree15	Tree16	Tree17
Congruency Sums																
7.05	9.01	9.36	5.99	9.62	9.28	9.74	7.01	6.47	9.28	5.95	6.32	12.26	7.63	6.42	6.46	5.64

Figure 4.22 | All pairwise comparisons of congruency as measured by the Normalised Robinson-Foulds metric (nRF). The closer the nRF is to 0 (as depicted by the darker colours in this heatmap), the better the congruency is. Note that the metric scale is reversed relative to the previous heatmap, but the colourscale has been inverted to maintain colour consistency (i.e. darker colours are more congruent for both heatmaps). The cumulative summed congruency for each locus is displayed below, to quickly show numerically the most and least congruent.

4.2.7 Detecting PVC Homologs

Run MGB for a selection of closely related genomes, and for photorhabdus. Identify all PVCs from all photo genomes to date. Identify orthologs in a selection of closely related species

Zhang et al (Zhang et al., 2012) posit that PVCs are present in MANY species outside the enterobacteria - examine this further?

4.3 Discussion

The data for the congruency methods reveals several trends and unambiguously confirms hypotheses about PVC13, the putative phage tail-fibre like gene. Multiple sequence alignments (MSAs) for all the data generated here are given in **??** on page ??, as they take up considerable space.

Proceeding consecutively, PVCs 1 and 2, which comprise part of the inner and outer tube of the needle complex respectively, both score well in general for congruency. This is as expected; these proteins are present in every PVC, and are always the first 2 genes in the operon (with the exception of "PVC0" that was discussed in Section 4.2.1 on page 103). As can be seen from Figure 4.3 on page 109, PVC1 is generally quite well conserved, though some interesting gross architecture begins to emerge. It seems that the various "Unit #" operons, and "LopT", cluster together quite neatly, but the remaining PVCs begin to segregate somewhat. This is potentially suggestive of the inner sheath adaptations the PVCs are undergoing to accommodate their cognate payloads. Since the exterior sheath serves essentially the same purpose in all the PVCs, regardless of their payload, it doesn't seem surprising that they almost all cluster considerably closer to one another, with short branch lengths and low bootstraps in places. The very obvious exception to this which is apparent in both the multiple sequence alignment (MSA) **??** on page ??, and the tree, is that the "lumt" operons from both the *P. asymbiotica* genomes vary enormously. These sequences align quite well locally to the other sequences, but contain many more residues versus the other sequences (though Kingscliff's "lumt" gene has a truncated C-terminus). One hypothesis to explain this is that the additional residues form extra surface loops, as shown in Chapter 3 on page 95, , potentially altering the target organisms' immune response to the complex, while maintaining the contractile functionality.

add reference to figure once made

PVC3 scores generally lower (more apparent in Figure 4.21 on page 120, than Figure 4.22 on page 121), and this too is unsurprising as this gene is missing from 3 of the PVC operons (which in this workflow is penalised as an incongruency). PVC3 itself looks to be an external sheath protein, a paralogue of PVC2 and possibly PVC4. PVC2, 3, and 4 frequently match homologs of the external sheath structure of tail-tube structures when querying databases (see **??** on page ??). If it is assumed that all the PVCs are fully functional, it's not clear why some have lost this gene but others retain it - though this suggests that a single copy of the sheath proteins is sufficient to produce the PVC complex. It's possible that this is a 'snapshot' in the active evolution of these structures, and they may all be capable, or in the process of, losing PVC3 without any deleterious effects (since it is paralogous). If we consider that some of the PVCs could be defunct, they may have become so because of the loss of PVC3. Given the number of copies of the inner and outer sheath proteins required to produce a single needle complex (6 of each per stratum of the PVC), compared to the other proteins involved, suggests that multiple copies of the protein may simply be present for stoichiometric reasons. The sequence differences among the PVC3s that are observed are quite pronounced, with large and varied INDELs appearing in almost every sequence, but with largely conserved C-terminal ends. As with PVC2, it may be the case that these modifications manifest on the surface of the PVC tube, resulting in modified immune responses by the target immune system. A question that can't yet be answered however, is in what ratio these paralogs are incorporated in to the final structure (i.e. why could a PVC2 monomer get incorporated instead of a PVC3, or *vice versa*).

PVC4's gene tree resembles PVC2, though with longer branch lengths and some internal node reordering, and does seem to score better for congruency overall. As a paralogue of PVC2, it seems likely that the 2 genes would follow similar evolutionary patterns, though the average pairwise amino acid similarity scores shown in Figure 4.3 on page 109, shows PVC4 to be better conserved (fewer extremes) overall than PVC2, despite having similar mean and median values. Hirst *et al.* has suggested that Afp4 in the *Serratia* Antifeeding prophage (and thus orthologue of PVC4) may be a slightly variant form, which actually comprises part of the collar, rather than the tube proper.

This may account for its maintenance along side PVC2, whilst PVC3 is free to be deleted.

PVC5 is a direct paralog of PVC1, the inner sheath proteins. Interestingly, it (and as mentioned, its paralog) is always present in all the operons studied, despite there being as many as 12 copies of these genes within a given genome. Its amino acid sequence is also very well conserved (as is PVC1, though to a lesser degree - see Figure 4.3 on page 109). Both genes, despite performing the same function, and remaining present in all operons, also have disparate GC content. One likely explanation is that both proteins are needed to maintain the stoichiometry of the tube, as mentioned earlier in this section, but their direct paralogy has allowed them to drift in sequence, with the protein tertiary structure evidently being reasonably robust to sequence change. It is interesting that the 2 proteins do not have perfect congruence, thus it is possible the proteins are divergent for a reason that we do not yet understand. The same question remains about the selective incorporation of one paralog over the other though - it's unclear whether both proteins are there for purely stoichiometric reasons, and the resulting tube structure is a 'patchwork' of different proteins or not, though this seems like the simplest explanation.

PVC6 has no good known homologs at present. Given its position within the structural region of the PVC operon, the best guesses at this point are that it is some sort of additional baseplate component that would form the collar around the spike, or potentially complexes with the spike itself, which is another nearby gene. The known T4 baseplate and collar is an extensive structure made up of many different proteins (Kostyuchenko et al., 2003). PVC6 and it's downstream neighbour PVC7 are both enigmatic proteins with unknown functions, and by this analysis look to be moderately diverse, certainly more so than most of the other structural proteins. Given that, at present, it is unknown where the tail-fibre like binding arms of the PVCs 'dock' with the needle tube complex, and the tail fibres (PVC13) are demonstrably highly variable, a potential hypothesis is that PVC6 and 7 may be responsible for anchoring the tail-fibres to the tube collar. It would make sense, therefore, that as the tail-fibre sequences have drifted and evolved differentially, that they proteins responsible for making them 'compatible' with the rest of the structure may also have to have changed over time to accommodate. As with many of the genes, both Pnf" and Lumt are notably different in sequence compared to the other PVCs.

PVC8 is a well conserved protein, though with the latter ≈ 250 residues generally aligning better than the beginning of the sequence with a much higher number of 100% identical residues at a given position (**??** on page ??). As demonstrated in Chapter 3 on page 95, it is a homolog of the valine-glycine-repeat protein vgrG, which forms the spike structure at the tip of the inner sheath, also analogous to the gp7-gp25 complex of the T4 bacteriophage. Though appearing 'stripped down' somewhat in comparison, apparently lacking the lysozyme domain that the T4 bears, thus more closely resembling the *E. coli* c3393 gene product (PDB ID 2P5Z). The general structure of the vgrG spike proteins in all tail-tube like structures studied to date is almost exactly equivalent even with as little as 12% homology at the sequence level: a homotrimer base with each monomer having a protruding beta strand intertwined with it's partners to form a triangular prism-like shape, as shown in **??** on page ??. This has been demonstrated in the literature frequently (Leiman et al., 2009). As a required and single copy gene, it is unsurprising that PVC8 is well conserved and one of the more congruent of the proteins studied. The bases for adaptation/variation of this spike for its varied jobs is not well understood, but the PVCs inactivity against prokaryotic targets speaks to the lack of the lysozyme domain within the protein itself.

However, PVC9 shows a similar pattern of congruence as PVC8, with similarities to tail lysozyme domains as shown in **??** on page ?? (Arisaka et al., 2003) and the "gp5" gene product (PDB ID 2IA7). It's not clear at present whether or not this is a functional enzymatic lysozyme domain because, as mentioned in the previous paragraph, PVCs theoretically have no need of one. It may be that this is simply a structural vestige, or that this was once the lysozyme domain of the vgrG gene which is in close register to this gene and could have undergone a gene split. Their proximity and potentially intertwined roles mean that a similar pattern of congruence is probably to be expected.

PVC10 is another enigmatic gene and the functional predictions for the gene vary wildly in match score and putative function. Two of the most likely candidate functions which are hit at varying levels of confidence are a so-called "PAAR-repeat domain" spike protein, and potentially another structural component, gp6 (see Chapter 3 on page 95). Given the comparatively conserved nature of gp6 homologs, it's less likely for this to

be the case, as PVC10 is the second most diverse (least congruent) gene in this analysis, and PVC11 looks to be the PVC equivalent of gp6. Since no other candidate genes exist to cover the role of a PAAR protein, and they are known to be diverse in the literature (Shneider et al., 2013), it seems likely that this analysis has detected the variable gene, and the spikes are just as diverse among PVC elements.

PVC11 is reasonably congruent within this analysis, and appears to have a well conserved functional role, though the sequences comprise extremely diverse, and extremely well conserved localised regions. At different points in the MSA, "lumt" has significant deletions relative to the other sequences, but only the sequence from the Kingscliff genome has a dramatically truncated N-terminus. As mentioned in Section 4.2.1.1 on page 104, the operons actually harbour a second PVC11 paralogue, though only 1 was used (the most similar) for this analysis. Both of these paralogs draw gp6 phage baseplate like homology via HHSuite, as do the PVC11s from the other operons, but they appear to be sequentially very distinct, having dropped a large span of sequence in the middle starting from $\approx$ 180 residues in, before becoming similar again at the C-terminus. Pnf similarly lacks a significant proportion of sequence in the middle of the protein. "lopT" on the other hand, has an approximately 450 amino acid extension to its C-terminus. Structural prediction suggests this protein is likely a T4 gp6 orthologue, and potentially a baseplate or collar structural protein but with a defined role within the PVC as yet unknown, and the relevance of these large deletions and extensions remain a mystery (Cardarelli et al., 2010; Aksyuk et al., 2009b).

PVC12 seems to show a similar pattern of congruence as PVC11, perhaps due to their proximity and both being among the larger of the genes within the operon. PVC12 appears to be very well conserved in particular regions with all sequences sharing runs of many identical residues, even among the more diverse, such as "pnf" and "lumt". This presumably speaks to the maintenance of active sites rather than purely structural domains, since it's evident that PVC structures are maintained even if the sequence drifts in other genes. There are few, if any, reliable structural homologies predicted for this protein, so its role can only be speculated about at present. It appears that the protein may be responsible for binding nucleotides, as previous searches have suggested it may

contain a GGDEF domain (which binds cyclic di-GMP) and matches weakly to certain ATP-binding transcriptional regulators in HHpred results (Paul et al., 2004). It seems that this protein is required for the PVC's structure or function, being so well maintained, especially for such a large protein, but this role is as yet unknown.

The next gene in the operon, PVC13, is of special interest in the context of general PVC mechanistics. Until quite recently, the quality of hits retrieved when querying services such as BLAST, HHSuite and so on were poor. Typically the hits would either give poor results, or good results but to small regions of the proteins. Proteins from different operons often retrieved different best hits, so it was difficult to come to a consensus about their definite function. The types of hits that would typically be found, were matches to adenoviral motifs, and so for a while it was unclear as to whether these hits were spurious. Figure 4.3 on page 109 shows the marked decrease in average identity for PVC13 clearly. The hypothesis, therefore, is that these were tail-fibre like domains, akin to those of bacteriophages (gp34-38) (Bartual et al., 2010; Leiman et al., 2010), used to bind the PVCs to their targets. They have demonstrated homology to both T4 like domains, and non-bacteriophage viral domains (such as those of Adenoviruses as mentioned), which is shown in **??** on page ??. In this workflow PVC13s are clearly the least congruent genes within the operon with Figure 4.16 on page 116 not clustering the PVCs well, with low confidence nodes and long branch lengths. These proteins have low overall identity (Figure 4.3 on page 109), and are therefore probably responding to very specific, and potentially very strong selection pressures. The current hypothesis is that the co-opting of eukaryotic viral molecular patterns has allowed *Photorhabdus* to repurpose these structures as a toxin system for use during infection of higher organisms. By recombining or evolving new receptor binding motifs, there may also be incredibly tight specificity for certain eukaryotic cell types, much as phage exhibit for specific bacterial strains/species. To the best of our knowledge, this is the only known example of a natural chimerism between a fibre-like protein of bacterial/phage origin which has recombined with a eukaryotic motif. Studies in the literature have demonstrated that these chimeras can be made experimentally, affirming the uniqueness of this class of proteins within *Photorhabdus* (Papanikolopoulou et al., 2004b). Because of its unusual putative structure, PVC13 was

studied further experimentally to try to confirm or refute this role (see Chapter 5 on page 138).

PVC14 is one of the other remaining mysteries within these operons. Structural predictions and homology searching are ambiguous at best. By alignment, the genes all seem to have a reasonably well conserved C-terminus, and to a lesser extent N-terminus, but with a substantially variable 100 or so amino acids in the middle of the gene. Curiously, the gene representative from the Pnf operon of the Kingscliff genome is substantially different at the N-terminal end, with just a handful of 100% conserved residues; being different even from that of the "pnf" operon in the ATCC43949 (USA) genome, but maintaining the conserved C-terminus (though still to a lesser degree). Based on syntenic position, and gross operon similarity to the Antifeeding prophage of *Serratia entomophila* (Heymann et al., 2013), it's suggested that PVC14, may fulfil the role that Afp14 is demonstrated to have experimentally, controlling the length of the sheath. The need to maintain similar termini, perhaps for binding to the 2 ends of a PVC tube, potentially speaks to this role, whereas the middle may drift in sequence and length (sequences range from 465 - 654 AAs) as it purely acts as a 'chain' between poles of the tube (Rybakova et al., 2015b). The gene is moderately congruent in this analysis, clustering the PVCs by their effector types quite well, this is perhaps suggestive of the notion that PVCs from different genomes bearing the same payload may be approximately the same size, and therefore maintain roughly similar tape measure proteins.

PVC15 has been one of the easiest to identify genes for some time, and its one of the few that is actually identified with a proper gene name locus tag in genomic annotations, typically coming up as *ftsH*, a so-called AAA+ ("ATPases Associated with diverse cellular Activities") ATPase and metalloprotease. Clustering based on this gene, as with PVC14, demarcates the PVCs by payload quite well, and shows "lumt" and "pnf" to be relative outliers once again. Nevertheless, this particular sequence demonstrates the longest uninterrupted runs of 100% sequence identity of any studied within the operon, and a high proportion of all column positions within the MSA are identical. This is typical of this class of ATPases, as they're known to comprise a conserved ≈ 250 amino acid domain which is the case here too (Hanson and Whiteheart, 2005). With all that said, however,

the role these ATPases play in PVC mechanistics is still unknown. They are known to hydrolyse ATP in order to exert effects on macromolecular complexes (Erzberger and Berger, 2006), and in the case of the T6SS, it has been shown that it is responsible for proteolysis and recycling of the triggered T6SS tube, so that it can be rebuilt (Bönemann et al., 2009; Forster et al., 2014). Since the PVCs are not membrane bound, but instead act as 'torpedoes' at a distance, there is theoretically no apparent need to recycle them. Similarities between PVC15s and Afp15 have been demonstrated previously but there is currently no known role for the analogous Afp15 from *Serratia* either (**?**). This opens up other potential theories, such as the ATPase maybe having some role in either the loading of payloads (if they need to be partially unfolded first for example as with the bacterial flagella (Muskotál et al., 2006), or in triggering the contractile machinery itself. There generally good congruency and conservation suggests that this is quite a constrained structure, since it requires >200 amino acids to form the active domain, consequently the protein is likely slow to evolve, especially in comparison to other PVC proteins. Structural homologs are known to form hexameric rings, as with the actual PVC tube structure, which would suggest it potentially sits atop the complex which could speak to its role in loading the syringe itself.

Lastly, the final structural gene which is consistently present between all operons is PVC16, but is without good homologs or a known role. It maintains a reasonably well conserved N-terminus, with many positions identical across all sequences, but becomes variable in the latter half of the CDS, particularly in the case of the protein from "Unit 2" of Kingscliff, which has a significant truncation, as does "Unit 3" from TT01, though to a slightly lesser degree. The gene is similarly congruent to PVC14 and 15, likely down to proximity once more. One hypothesis, based on the same logic as PVC14 (synteny to Afp), is that PVC16 may be a tail tube terminator protein (Rybakova et al., 2013).

4.3.1 Correlation between PVC structural proteins and their payloads

Though the effectors of the PVCs are not specifically handled within this congruency workflow, as mentioned in Section 4.2.1 on page 103, they are known for each of the sequence sets used here, and are the discriminators between operons within a genome.

Superficially, the PVCs look to be elaborating the same structures, and the original hypothesis within the group was that effectors may be promiscuous and capable of being utilised with any PVC. While this is still not (dis)proven experimentally, and work is ongoing in the lab, on closer inspection, it seems that the different PVC operons are genetically less similar than it initially appeared. This may be suggestive of a 'honing' process, whereby some, but not total, interchangeability could be possible, but that different PVCs now have some (mechanistically unknown) specificity/preference for particular effector types (this is reminiscent of the specificity seen in T6SS for VgrG-PAAR-payload complexes). To explore this, the frequency of clustering of PVC tube protein sequences with the same effectors can be examined from the gene trees.

For the inner tube proteins, in the case of PVC1 (Figure 4.5 on page 110, all of the sequences cluster according to their effectors (all the cifs group together, as do the lopTs and so on). There are substantive out-groupings of the cif, pnf and lumt sequences compared to the others, with much longer branch lengths. This possibly points to these 3 PVC operon types having undergone some particular adaptations for their payloads, however unpublished data from our own lab has shown the pnf toxin to be promiscuous in it's ability to be secreted from Type 3 systems, and its N-terminus strongly promotes 'cross-packaging' in to other PVC types, so any degree of 'bespoke-ness' required for the pnf PVC may not be wholly explained by its cargo. The pnf operon does also house an additional toxin however, with homology to the cyaA adenyl cyclase from *Bordatella pertussis* (see the HHSuite results in Chapter 3 on page 95), which is perhaps a little less versatile than pnf. The various "Unit#" operons cluster reasonably well together, and also include the lopT sequences, though with a couple of lower confidence ancestral nodes. This may indicate a degree of greater interchangeability between these proteins, or much more subtle sequence modifications giving rise to any effector preference. The tree for PVC5, the paralogue of PVC1, (Figure 4.8 on page 112) is markedly different in the branch lengths (note also the different scale), but reasonably congruent (ADW scores of 0.77 and nRF of 0.53 vs Tree 1). Tree 5 suggests that lumt and pnf have substantially different inner core proteins once again, but now demonstrates much less difference between the remaining PVCs, this is also reflected in Figure 4.3 on page 109 where PVC5 is the highest

identity locus. The structure of PVC5 therefore, may be more discriminatory in terms of why pnf and lumt have developed different tube sequences, since the locus is clearly being preserved, but to a lesser degree in those loci, suggesting a pressure, rather than drift, which is driving the sequence change.

In most PVC operons there are 3 putative outer sheath proteins (PVC2, 3 and 4). In the case of lopT, both *P. asymbiotica* genomes have lost one of these, while the *P. luminescens* equivalent persists. In the lumt operons, only the USA *P. asymbiotica* strain is missing the gene. Initially it was assumed that these sequences were all the same, and each operon had triplicate paralogues. On closer inspection of sequence similarity, it appears that PVC4 is less like the other 2, and this suggests that the operons which have a gene deletion, are actually lacking PVC3, retaining the 2 variant forms. This current thinking is potentially backed up by unpublished preliminary findings from Hurst *et al.*'s lab, where PVC4/Afp4 is now thought to be a slightly modified tube protein which is serving as a collar protein or part of the baseplate complex. It would make sense, therefore, that PVC3 is able to be deleted without abolition of PVC production due to the paralogy, but this is not the case for PVC4. This is also borne out by the congruency analysis which shows PVC4 to have better congruency overall. The gene tree for PVC4 clusters PVCs by their associated effectors perfectly.

This does raise further questions as to why 2 copies of the inner sheath proteins are always present (and possibly required), since the stoichiometric ratio of inner to outer sheath proteins should be close to 1:1, yet a single exterior sheath appears sufficient; if it assumed all the PVCs are functional.

In **??** on page ??, operons are once again clustered largely according to their effector designations, with pnf and lumt yet again appearing to be among the most diverse with comparatively long branch lengths. Unusually, pnf is placed as a shallow internal node, where more commonly it is seen as an outgroup or deep split. The various "Unit#" operons also cluster together closely, but with a low confidence ancestral node. Figure 4.6 on page 111 reveals a different topology, and disregarding the deletions, uncommon splits occur: such as PLT_cif being placed well away from its *P. asymbiotica* counterparts. While lumt in Kingscliff does contain a PVC3, it is radically reduced in protein length (at only

86 amino acids, versus approximately 480 for all the other PVC3s). PVC3, therefore, is not strongly characteristic of PVC 'identity'. Its comparative lack of similarity within the cluster, as well as versus the paralogue of PVC2 (see the alignments in **??** on page ??), and the fact that it has been deleted from several operons, may suggest that the protein is in the process of disappearing from all the operons.

The functional basis for why the "Unit#" operons have remained comparatively similar to one another is unknown. It's possible that at least some of these PVCs may be primarily involved in symbiotic interactions with the nematode host rather than direct toxic effects against prey.

The Unit 4 operon from TT01 carries halovibrin-like effectors which are known in the literature to be a mediating factor in the ability of *Aliivibrio harveyi* and *fischeri* to colonise the light organ of the bobtailed squid (*Euprymna scolopes*) - the original model for quorum sensing and symbiosis (Ruby and McFall-Ngai, 1999; Verma and Miyashiro, 2013). The strict relationship between *Photorhabdus* and *Heterorhabditid* nematodes would mean a relatively stable ecosystem and potentially conservation of the associated PVCs further adding to this hypothesis.

Taken together, this may be indicative of the PVCs evolving alongside their payloads, rather than retaining an absolute 'one-size-fits-all' syringe complex. Experimental work has shown that PVCs are capable of trans-packaging alternative payloads, but it may not be possible to incorporate all toxins in to all variants of the syringe. Further experimental combinations will need to be tested to answer this once and for all. The consensus tree would also speak to this hypothesis - all the PVC operons are grouped well by effector molecule, which is suggestive of some co-evolution of payloads with structural components. Furthermore, it demonstrates that, despite speciation, all "pnfs" are more like one another across the genera, than any 2 PVCs within a genome are like one another.

4.3.2 Identifying the PVC blueprint in other locations

At present, there are only a handful of *Photorhabdus* genomes available, so there is almost certainly as yet unsampled diversity, but the patterns demonstrated here may be useful for identifying other PVC elements in additional genomes. The hallmarks that can be picked

out of this data can help find these extra elements. Sarris *et al.*'s analysis was similar, however they were interested in finding all contractile tube like elements, which meant that much of what specifically groups the PVCs is disregarded when it is too prescriptive of PVCs only. This section attempts to lay out a framework or criteria for identifying and curating additional PVC elements in future study.

Based on the gene trees and congruency analysis, plus what's known of contractile mechanisms at present, the following criteria could be used for reference:

- Tube proteins
  - Presence and comparatively high conservation of the inner sheath proteins (PVCs 1 and 5) appears required for PVC architecture.
  - 1 or more copies of an outer sheath protein, with an additional variant paralogue (PVC4) which will likely match to the same structural homologs, but with lower scores due to its putative role as a collar/baseplate subunit. A deletion (PVC3) may be observed here in some cases.
- The spike complex
  - There may be 1 or more unknown loci at PVC5 and 6, immediately followed by 2 well conserved and easily identifiable loci for the tube spike, a vgrG homolog and a phage tail lysozyme-like domain. The role for the lysozyme domain in phage is well characterised (Arisaka et al., 2003), though its function in a PVC is unknown, it remains a consistent feature.
  - Immediately following the spike and lysozyme, is a third part of the spike complex, the putative PAAR-repeat spike tip protein (Shneider et al., 2013). However, these proteins are notoriously variant (the second least congruent gene after PVC13), and homologies are weak. Detectable homologies to PAAR may be useful in confirmation if they arise, but probably should not be relied upon as they are not consistently hit when databases are queried.
- Operon core

  - Beyond PVC10, the genes increase in size and are almost always single copy. PVC11 strongly resembles another gp6 protein, and given its size, is likely to be a major structural component of the PVC collar assembly. It shares similar congruency to PVC4, the other hypothesised collar protein, which suggests that this is the case. Both PVC11 and 12 cluster PVC sequences concordantly by effector, and so are also likely to be good 'hallmark' PVC proteins.
  - PVC13 is an unusual case, as previously discussed. As the gene is extremely incongruent and diverse, and also entirely missing from lopT operons it is not a good marker for PVC structure. It is not yet understood how the PVCs function without a tail fibre-like protein, though a region of low identity within the middle of the operon may also be a smoking gun in many cases (though should not be relied on). Given the PVCs activity against eukaryotic targets, the PVC13 tail fibres are very much responsible for the uniqueness of the needle complex's activity, and should probably not be discounted all together when on the hunt for new examples.
  - PVC14 is not a well characterised operon, though it has conserved C-termini, and to a lesser extent N-termini. If the suggestion that this protein is a tape measure protein (since it's only discerning characteristic seems to be variation in length by ≈ 100 amino acids), it is likely that this gene, as with the PVC13s and PAAR proteins will be present, but may not be easy to identify. It's absence from the lumt operons could be artifactual if the sequence is simply so low in identity that it did not appear to belong to the cluster. PVC14 is therefore unlikely to be a reliable marker for PVC identification.
  - The AAA+ ATPase is a hallmark of most if not all contractile tail mechanisms, and is identified easily in genome annotations. It is clear that the PVCs are required to have one, given its presence and degree of conservation, though its mechanistic role in the PVCs is not. Any putative sequence should therefore contain an orthologue, though it is not yet known if the PVC ATPase is markedly different in any characteristic way at present.

- Identification of PVCs will also be contingent on being coupled to a payload region

at the 3′ end. Carrying one or more toxins in this region is a defining feature of the operons, however they are incredibly variant, making automated identification of the full width of the operon more difficult.

- Lastly, a recurring pattern with the PVC operons is a notably reduced GC content at the 3′ end, as demonstrated in Figure 4.2 on page 109. It was this GC signature that lead to the PVCs being found in the first place, when the repeating GC pattern in the 4 tandem *P. luminescens* genomes was spotted as unusual. Quickly calculating the GC trace across a putative PVC operon may also provide some confirmation, though this may not be unique to PVCs.

Given the diversity of the operons however, any putative operons that may be identified automatically, will almost certainly still need visual inspection before they could be unambiguously labelled as such - with particular attention being paid to the 3′ payload region effector types.

In summary, this analysis suggests the PVC sequences to be more ancestral/less mobile than first anticipated, though they almost certainly originate from co-opted phage mobile elements; this has also been proposed as the origins for the related contractile mechanisms (T6SS, R-type pyocins, etc.) In combination with **??** on page ??, diving in to the structural and phylogenetic bioinformatics has revealed potential new roles for previously unknown proteins, and identified the key regions of proteins which currently have no known roles. This will hopefully be invaluable for elucidating their function as further structures and domains are discovered and databases updated. PVC13 has been unambiguously identified as the single most variant gene within the operons, and in the next chapter, the structure and function is explored experimentally.

Nabil: Add a schematic of PVC presence and absence between strains to intro

Make a Sarris/Easyfig type thing to show mosaicism?

Include more details of the parameters of clustering the genes (ID/function etc)

Explain the clustering table, or change to presence/absence and move locus tag info to supp.

add more detail to tree, GC and AAID figure captions (mention median line in GC/AAID

- Consider collapsing low support nodes in trees/colour coordination/annotation/show significance of splits
- consider merging trees in to 4x4 panels
- rename SPECIES Tree to 'consensus' tree
- Move explanation of AWC to methods - make it clear its not my own work
- annotate heatmaps/figures with the functions of the PVC# numberings (e.g. "tube"/"spike")
- Create more summary figures for the discussion e.g. table of PVC identification criteria
- Elaborate on the limitations of the method (e.g. this is parsimony, but could be wrong)
- General proof read/strip down of language etc

Part III

Experimental Results

Chapter 5

Structure and Function of PVC Tail Fibre-like Genes

"It is very easy to answer many of these fundamental biological questions; you just look at the thing!"

Richard P. Feynman

5.1 Introduction

Bacteriophages are ubiquitous viruses of bacteria, and are the most abundant organisms in the biosphere by a considerable margin (Clokie et al., 2011; Bartual et al., 2010). Having been studied for over a century, and gaining increased interest in recent years as we combat the "antibiotic apocalypse", we now understand much of the biology of a great many phages. Less well known however, are the numerous phage-like elements which are scattered through the genomes of most if not all bacteria studied to date (Sarris et al., 2014). While they appear, at first glance, to be (often defunct) prophages, in actual fact many of these elements have been co-opted by their hosts and are 'weaponised', resulting in lethality against prokaryotic or eukaryotic targets. The pyocins discussed in 1 are an excellent example of this.

A key protein or protein complex for many bacteriophages, such as the *Myoviridae* family, which includes the model T4 phage, are the so called 'tail fibres'. In the T4 phage

there are both 'long' and 'short' fibres which have subtly different roles (Leiman et al., 2010). The long tail fibres are typically laid back along the length of the virion in 'free-living' phages, though some may be loose. The long fibres are responsible for the initial stages of target recognition, which. Once a sufficient number of the tail fibre proteins have bound to the surface of a target cell, the conformational change induced in the baseplate complex extends the short tail fibres. The long fibre binding stage reduces the distance between the virion and the target cell, enabling the short fibres to come in to play. Short tail fibre binding is irreversible, unlike the long fibres, and therefore provides anchoring to prevent the extrusion of the inner sheath from pushing the virion back off the cell surface. in the case of the T4 phage, is the OmpA protein or lipopolysaccharides of *E. coli* (Granell et al., 2014a; Taylor et al., 2016b; Riede, 1987). It's the fine structure of the short fibres that provides the exquisite selectivity that phages demonstrate for their hosts.

The evolution of co-opted phage loci, particularly against eukaryotic targets, raises the interesting question of how their structures, or at least their binding fibres, have evolved to be able to facilitate this. With the initial discovery of the PVCs, a putative tail-fibre like gene within the PVC was suggested (Yang et al., 2006). It was hypothesised that the PVCs should possess an equivalent structure since target recognition is a key feature of caudate complex mechanistics. Literature suggests that a conformational change in the peripheral baseplate of the T4 phage is transduced through the baseplate to the sheath in order to trigger contraction (Taylor et al., 2016a), and that the fibres are responsible for the initially conveyance of a conformational change to the peripheral baseplate proteins. Taylor et al. (2016a) have posited that a "similar sequence of events is likely to occur in any such system, regardless of the complexity of its peripheral baseplate or tail fibre network", and thus, as tail complexes, the PVCs likely have an analogous mechanism of conductance of the signal through the tail fibres and baseplate complexes.

To date, several tail fibre structures have been resolved, though typically only as sub-domains. Figure 5.1 on page 141 shows a collection of the structures available in the Protein DataBank, at the time of writing. Primarily, the structures correspond to different domains of the long and short fibres of the T4 phage, though Figures 5.1E and 5.1D depict Adenoviral domains. A striking feature of all of these structures regardless

of origin, is that they are made of interwoven trimers which form a manner of triple helix in many places. In the figures, separate chains are shown in red, green and blue, with any metal ions depicted in orange, and their coordinating side chains shown in purple. The coordinated metal atoms in the core serve to "cross-brace" the strands of the tail fibres. The interwoven nature and contributions of any coordinated metal ions leads to very high stability, with the T4 tail fibers (and other structural proteins) known to be heat and protease resistant (Bartual et al., 2010; Granell et al., 2014a). The structures appear to be primarily comprised of β-sheet and disordered turns, with only a few, if any, stretches of α-helix. Interestingly, the 'disordered' turns, are actually organised in a very regular fashion, being stabilised by nearby secondary structures, and by the counterpart stretches of disordered turn in the other chains, but do not contain much 'within' strand interaction/structure.

It was decided to study the putative tail fibres of the PVCs further experimentally for a few reasons. Firstly, existing genome annotations and homology searches had attributed fibre-like orthology to the sequences, but typically with low confidence and low overall sequence coverage. As mentioned in previous sections, the putative PVC tail fibres had also shown a curious similarity to Adenoviral motifs, which, it appears, has not been seen in phage-like tail fibres before. Moreover, the putative fibre genes showed a great deal of variation, even between PVC operons. As explained in Chapter 4 on page 101, they are the least congruent of all the genes studied, across all the operons, and this also results in inconsistent database hits between different genes. Combined, this meant that it was unclear whether these were meaningful orthologies or merely artifactual/spurious hits; and therefore some experimental investigation would be valuable.

Though there are resolved tail fibre structures, they have been difficult to study structurally in the past. This is primarily due to them only being anchored to the main tube apparatus at one end, such that the target recognition site of the fibre at the distal end is freely mobile in order to bind the cell surface. Consequently, when averaging multiple images in electron microscopy reconstructions, the tail fibres may not consistently overlap, and are averaged out, as in the case in the (Ge et al., 2015b) paper for instance. In Heymann et al. (2013) (see Figure 1.11 on page 32), there appear to be tail fibre like

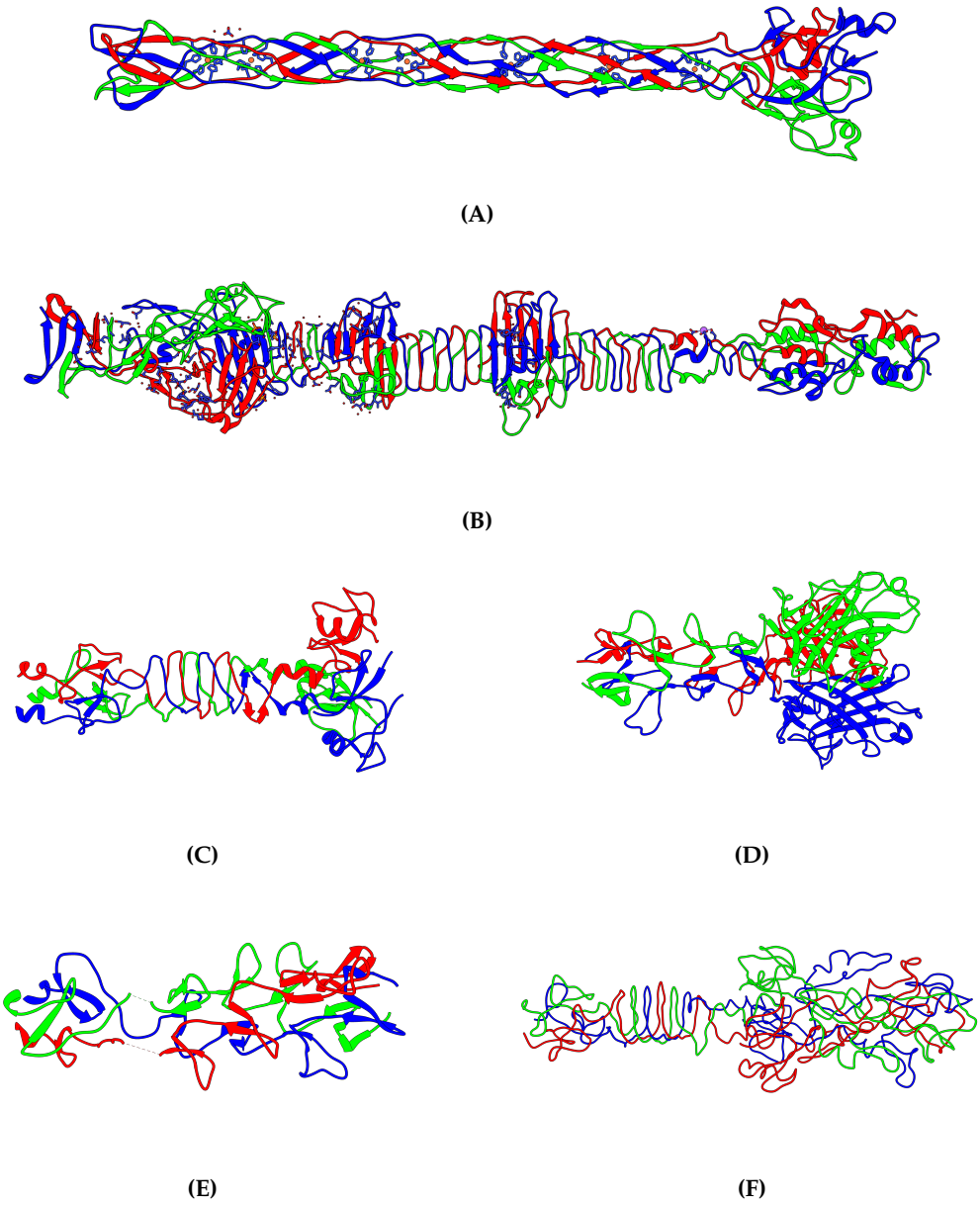


Figure 5.1 | Crystal structures for resolved tail fibre-like proteins.

Reference structures for other tail fibre like structures, some of which occur in the results of HHPred homology searches for PVC13 proteins. Note in-particular, the conserved twisted trimeric nature of the structures, and the co-ordination of metal atoms in a number of the structures (denoted by the side chains being present). Structures are not scaled relative to one another. **(A)** The crystal structure of the T4 phage long fibre receptor binding tip (gp37 residues 785 to 1026) (PDB ID 2XGF), as determined in Bartual et al. (2010). **(B)** The crystal structure of the T4 phage long fibre shaft (gp34 residues 95 to 1289) (PDB ID 5NXF), as determined in Granell et al. (2014b). **(C)** The crystal structure of the "heat and protease resistant fragment" of the T4 phage short fibre (PDB ID 1H6W), as determined in van Raaij et al. (2001). **(D)** The crystal structure of the human Adenovirus C serotype 2 shaft region (PDB ID 1QIU), as determined in van Raaij et al. (1999). **(E)** The crystal structure of an artificial chimeric tail fibre protein, comprised of human Adenovirus C serotype 2 shaft and T4 phage fibre domains (PDB ID 1V1H), as determined in Papanikolopoulou et al. (2004b). **(F)** The crystal structure of the short tail fibre C-terminal region as fitted in to the T4 baseplate in Kostyuchenko et al. (2003) (PDB ID 1PDI).

structures laid back long the length of the sheath in some form of docked, latent, position and this appears to have sufficiently immobilised them such that their densities can be visualised. However, this conformation appears to be an exception, rather than the rule in the structures studied to date; and though this is the most probable explanation for those densities, the map is only ≈20 Å, and is thus not entirely conclusive.

Existing computational methods such as the simulations described in **??** on page ?? can fail to produce plausible structures in many cases. *Ab initio* methods are still difficult to implement for large multi-chain structures, as, without approximation via coarse graining. Calculations simply scale too poorly due the worse-than-exponential increase in atomic interactions as a system grows in size. Not to mention, molecular dynamics is an entire field of its own, and not readily amenable to most researchers. Threading webservers such as I-Tasser (Yang et al., 2014; Zhang, 2008; Roy et al., 2010) (used in Chapter 3 on page 95) and Phyre2 (Kelly et al., 2015) have addressed this by making easy to use job submission interfaces for homology modelling. Threading approaches sometimes produce workable structures, however accuracy can become compromised for proteins with split domain structure (chimeras), where the best template protein is different for each domain, without artificially breaking the gene up and then having to 'stitch together' the resultant models. Inclusion of a molecular dynamics-based refinement step, may actually worsen a simulated monomer if it's ordinarily part of a non-globular multimeric structure, by allowing the model to relax (e.g. in energy minimisation/solvation), as it will cease to be constrained to the threaded chain.

This leaves experimental structural determination as the best option, though it remains non trivial. Structural resolution via Nuclear Magnetic Resonance is unlikely to be feasible for most tail fibres. There is an upper bound on the size that NMR studies can resolve (≈35 kDa), which even some of the smallest tail fibres exceed, due to their trimeric nature. It is possible that sub-cloning the proteins on a domain-by-domain basis may be within the range NMR could be applicable to, but this would become laborious and raise issues about correct folding and so on. Cryo-EM is also a possibility; as seen in the Heymann paper, it has the resolution to image the tail fibres in complex with the tail tube, but with the caveat that they are immobilised by the tube itself to prevent class averaging losing the

densities. Additionally, the fibres may only be a few atoms thick at their thinnest point, which without high-end equipment could be extremely difficult to image effectively due to low contrast. This leaves X-ray crystallography as the main viable option, but this is also not without caveats. In a number of papers (such as those producing the structures in Figure 5.1 on page 141), crystallisation was only achievable when expressing sub-cloned domains of the the protein, and often required the presence of multiple chaperones too.

The variability, above and beyond that of the rest of the operons, in the tail fibre proteins of the PVCs, combined with the PVCs known role as anti-eukaryotic toxin systems, suggests that their putative tail fibre proteins may be of particular interest. Understanding the existing target binding spectrum may shed some light on the role of the PVCs in virulence and pathogenesis during *Photorhabdus*' infection cycle in the various hosts it can infect. In future we would also like to be able to explore rational modification and engineering of the tail fibres, potentially controlling their tropisms. This is effectively a case of recapitulating similar studies which have been done for re-targeted R-type pyocins (Scholl et al., 2009), whereby tail fibres were swapped or fused, conferring new target cell spectrums to pyocins that would ordinarily have no effect on the cell type of interest. In the case of the PVCs, a natural repertoire of diverse tail fibres exists, providing an as-yet-unstudied library of motifs to potentially target different cell types - though a crucial difference being that these cell types will be eukaryotic, and will convey a cargo, unlike the Pyocins.

In summary, this chapter examines tail fibres in isolation from the difficult-to-manipulate PVCs; in order to shed the first 'experimental light' on the true nature of their structure, orthology, and function.

Chapter Aims:

- Clone and express tagged versions of putative tail-fibres from different PVC operons.
- Devise and optimise a purification strategy.
- Probe tail fibre structure via biochemical/biophysical/crystallographic methods.
- Exploit functionalised tail-fibre complexes for binding studies.

5.2 Experimental Procedures

5.2.1 *in silico* examination of tail fibre sequences

The putative tail fibre hereafter called 'pnf13' was cloned from the *P. asymbiotica* ATCC43949 (a.k.a. "USA") PVC-pnf operon; similarly, the putative tail fibre 'lumt13' was cloned from the PVC-lumt operon of the same genome. They both carry the designation '13' from their general syntenic position within the operon, however the lumt operon does have an upstream deletion. Operon organisation is covered in previous chapters, but Figure 5.2 and Figure 5.3 show the fibres in their genomic/gene cluster context.

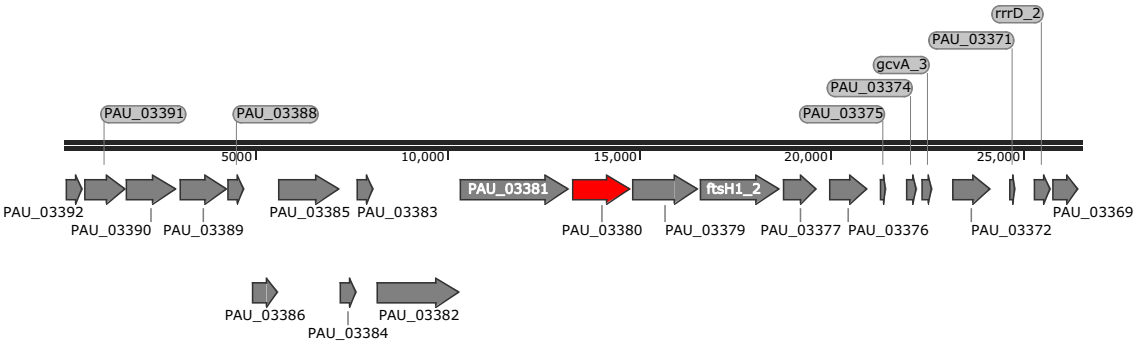


Figure 5.2 | The "pnf" operon from *P. asymbiotica* ATCC43949.
The PVC*pnf* operon, with the putative tail fibre gene (PVCpnf13) that was cloned in red. Locus tags correspond to the most recent genome annotation used throughout this study.

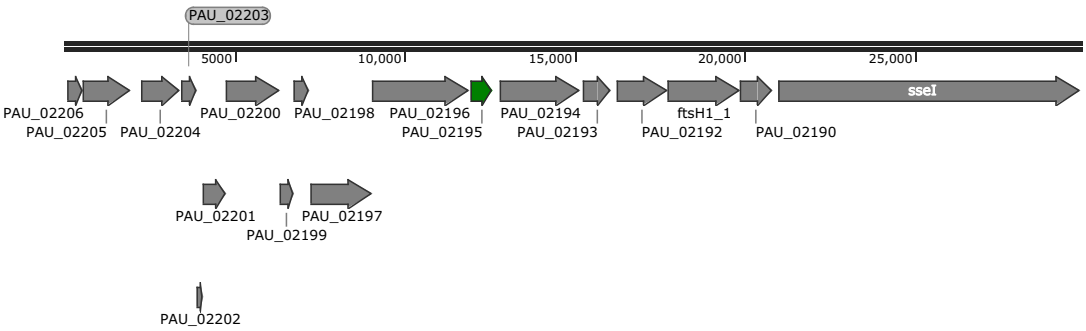


Figure 5.3 | The "lumt" operon from *P. asymbiotica* ATCC43949.
The PVC*lumt* operon, with the putative tail fibre gene (PVClumt13) that was cloned in green. Locus tags correspond to the most recent genome annotation used throughout this study.

These 2 particular tail-fibres were chosen from the many choices as they are both unique to the human pathogenic strains but from disparate operons, and would allow us to explore differential tissue culture activities etc. Though the operons group together in Figure 4.20 on page 118, it can be seen from Figure 5.2 and Figure 5.3 that the 2

genes, despite having putatively the same function, are significantly different sizes and therefore could have quite different higher order structure. In the case of *lumt13*, the CDS encodes a 23.6 kDa protein (as a monomer), and *pnf13* is twice as large at 51.7 kDa. This is suggestive, therefore, of 2 tail fibres which are both evolved to act against eukaryotic targets, but potentially with each honed in a different manner, to exploit different molecular mechanisms.

5.2.1.1 Domain structure

Probably the most interesting aspect of these 2 proteins (and tail fibres from PVC operons in general) is their apparent chimerism in domain structure as demonstrated in the HHPred results from Section 3.2.2 on page 97. Table 5.1 on the next page summarises the HHPred hits for these 2 tail fibres. They match distinct PDB entries in different regions, typically from either T4 bacteriophages or human Adenovirus fibre proteins. Chapter 3 on page 95 delves in to the matches detected for other PVC tail fibres in more detail.

This is suggestive of *Photorhabdus* co-opting Adenoviral motifs (or something at least resembling Adenovirus motifs) in the environment or perhaps during infection of mammalian systems. PVCpnf and PVClumt are unique to the mammalian pathogenic strains providing both 'motive' and opportunity for recombination with a mammalian virus to have occurred. However, this is also true of some *P. luminescens* operons too, and thus may point to an ancestor that was capable of mammalian infection, and that *P. luminescens* has subsequently lost this capability. An alternative suggestion is simply convergent evolution, and both Adenoviruses and PVC fibres have evolved to a specific receptor or receptor family. Some evidence for this can be taken from the fact that both Coxsackie and Adenoviruses bind to the so-called "CAR" or "Coxsackie and Adenovirus Receptor", despite both being evolutionarily disparate viruses (the former is a (+)ssRNA virus (Group IV) and the latter a dsDNA (Group I) virus). Since these 2 distinct viruses have both evolved to exploit the same receptor, it is not a big leap to suppose that PVCs may also have done so.

Since the PVCs themselves are similar in structure to T4 phage tails, it is likely that the split domain structure maintaining a T4 region is required for 'mounting' the fibre

on to the sheath complex. This implicitly suggests 2 points; firstly that the tail fibres do need to maintain a phage like domain within the protein, albeit somewhat diversified, to enable 'mounting' to the tube of the PVC, much as T4 fibres would need to mount to the tail tube of the phage. However, the homologies are lower in comparison which would suggest that the PVC tail fibres and the tube itself may be drifting together to become more 'bespoke'. There is an additional subtlety to consider; the protein structure databases are rich in T4 and Adenovirus structures, to the point of bias, so it is small wonder that they match these structures when searched. But, if the suggestions of Sarris et al. (2014) are to be believed, PVCs and other tail-tube like elements may have diverged in deep time, in which case their tail fibres may be fairly unique, perhaps representing a convergently evolved structure, rather than a chance recombinant one.

Table 5.1 | The top 5 HHPred structural homologies detected for pnf13 and lumt13.

#	PDB Hit	Prob	E-Value	P-Value	Score	Hit Descriptor
				pnf13 (PAU_03380)		
1	3IZO	98.1	2.8×10^{-9}	7.3×10^{-14}	107.6	Fiber; pentameric penton base, trimeric viral protein; 3.60 Å{Human Adenovirus 5}
2	3IZO	97.6	1.3×10^{-7}	3.4×10^{-12}	95.6	Fiber; pentameric penton base, trimeric viral protein; 3.60 Å{Human Adenovirus 5}
3	1OCY	97.6	1.8×10^{-7}	4.7×10^{-12}	80.4	Bacteriophage T4 short tail fibre; 1.5 Å{Bacteriophage T4}
4	1V1H	96.1	0.00012	3×10^{-9}	60.4	Fibritin, fiber protein; chimera; 1.9 Å{Human Adenovirus type 2}
5	1V1H	95.9	0.0002	5.1×10^{-9}	59.1	Fibritin, fiber protein; chimera; 1.9 Å{Human Adenovirus type 2}
				lumt13 (PAU_02195)		
1	1V1H	97.7	9.4×10^{-8}	2.4×10^{-12}	71.3	Fibritin, fiber protein; chimera; 1.9 Å{Human Adenovirus type 2}
2	1V1H	96.3	5.9×10^{-5}	1.5×10^{-9}	56.3	Fibritin, fiber protein; chimera; 1.9 Å{Human Adenovirus type 2}
3	1QIU	95.8	0.00028	7.2×10^{-9}	60.4	Adenovirus fibre; fibre protein, triple beta-spiral; 2.4 Å{Human Adenovirus 2}
4	3IZO	94.7	0.0025	6.4×10^{-8}	59.1	Fiber; pentameric penton base, trimeric viral protein; 3.60 Å{Human Adenovirus 5}
5	1OCY	66.7	1.3	5.3×10^{-7}	33.1	Bacteriophage T4 short tail fibre; structural protein, 1.5 Å{Bacteriophage T4}

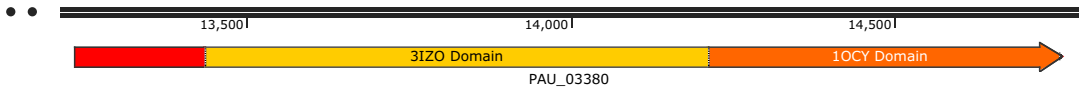


Figure 5.4 | The PVCpnf13 tail fibre with dominant domain homologies depicted
A map of the dominant protein domain splits within the PVCpnf13 tail fibre protein according to HHPred. Toward the 3′ end there is a region which matches well to the PDB ID 3IZO, a fibre protein-penton base from the human Adenovirus 5, as determined in Liu et al. (2011). At the 5′ end, a domain match to the Bacteriophage T4 short tail fibre (PDB ID 1OCY , as determined in Thomassen et al. (2003)) can be seen. The protein therefore resembles a natural chimera/fusion protein of an Adenoviral motif and a phage motif. Interestingly, PVCpnf13 also shares similarities to PDB ID 1V1H, which is an artificial Adenovirus-phage fibre chimera, created in the study by Papanikolopoulou et al. (2004b).

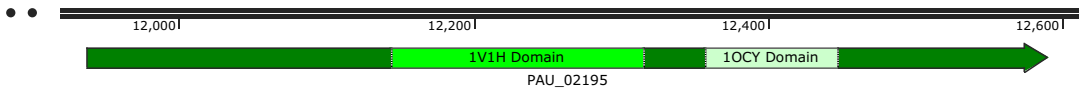


Figure 5.5 | The PVClumt13 tail fibre with dominant domain homologies depicted
A map of the dominant protein domain splits within the PVClumt13 tail fibre protein according to HHPred. Toward the 3′ end there is a region which matches to the PDB ID 1V1H, an artificial fibre fusion protein from the human Adenovirus 5 and the T4 phage Papanikolopoulou et al. (2004b). At the 5′ end, a domain match to the Bacteriophage T4 short tail fibre (PDB ID 1OCY, as determined in **?**) can be seen. The protein therefore resembles a natural chimera/fusion protein of an Adenoviral motif and a phage motif.

5.2.1.2 Sequence characteristics

As well as identifying the orthologues of the sequences using HMMs like those in Table 5.1 on page 146, a hallmark of putative tail fibre sequences is the coordination of metal atoms, like those seen in several of the structures in Figure 5.1 on page 141. For example, in the tail fibre tip structure PDB ID 2XGF solved by Bartual et al. (2010), it was observed that iron was present in the crystal structures, most likely in the Fe^{2+} oxidation state. They were able to identify 7 iron sites within the crystal, and this matched the frequent occurrence of His-*x*-His motifs within the protein sequence. The authors were also able to obtain improved stability of expressed proteins when supplementing growth media with Manganese (II) chloride, and though they did not identify any Manganese in the final crystals/structures, they concluded that this is indicative of the need for metal ions in a 2+ oxidation state. Continuing the theme, in van Raaij et al. (2001), they were able to identify a set of conserved repeats, in the receptor binding tip of the T4 short tail fibre. To examine if a similar sequence pattern might be evident in the PVC tail fibres, sequences were analysed with the Rapid Automatic Detection and Alignment of Repeats program from

the EMBL (Heger and Holm, 2000). For the 2 proteins being studied here, the results are displayed in Tables 5.2 to 5.3 on the next page. For the results for the other PVC tail fibres, see Chapter D on page 219. Repeats are identifiable in all but one of the tail fibre proteins, and typically runs of 4-6 repeats can be detected to varying degrees of sequence identity. PVCpnf is unusually repetitive, even among the tail fibres, with 10 very conserved repeat patterns, each separated by 2 amino acids (see Table 5.2 on the next page). The next most repetitive proteins are PLT_01746, the tail fibre from the *P. luminescens* TT01 "Unit 1" operon, and PAK_02618, from the *P. asymbiotica* Kingscliff "Unit 1" operon, both of which have 7 tandem repeats (note: these 2 operons are not orthologues, despite the naming scheme). Given the likely role of the repeats in forming the shaft region of the phage fibres, this suggests that the pnf tail fibre should be, potentially substantially, longer than the lumt one. The significance that this might have is not yet understood however.

Belying their diversity, there does not appear to be one particular structural motif across all the tail fibre proteins - there are no motifs such as the T4 fibre's H-x-H which seems to predominate any particular fibre. In the multiple sequence alignment which has been reproduced from the Appendices in Figure 5.6 on page 150. It is possible to identify 4 seemingly conserved domains, but there appears to be no obvious conservation of repeats, despite (nearly) all of the tail fibre proteins having repetitive stretches.

Table 5.2 | The largest stretches of sequence repeats within the PVCpnf13 tail fibre.
This table shows the sequences and statistics for the repeat detection from RADAR, for PVCpnf13. A well conserved set of 10, 14 amino acid stretches can be found, which are rich in valines, lysines, serines and glycines, and each 15 amino acid stretch is separated by 2 amino acids.

# Repeats	Total Score	Length	Diagonal	BW-From	BW-To	Level
10	298.40	15	15	200	214	1

Repeat Indices	Alignment Score/Z-score	Repeat Sequence
149 - 163	(25.32/10.54)	VKVSANKGLSVDSSG
166 - 180	(29.64/13.67)	VKVNTDKGISVDGNG
183 - 197	(29.67/13.68)	VKVNTSKGISVDNTG
200 - 214	(31.92/15.31)	VIANASKGISVDGSG
217 - 231	(33.00/16.10)	VIANTSKGISVDGSG
234 - 248	(30.39/14.21)	VIANTSKGISVDNTG
251 - 265	(31.92/15.31)	VIANASKGISVDGSG
268 - 282	(33.00/16.10)	VIANTSKGISVDGSG
285 - 299	(31.82/15.24)	VIANTSKGISVDSSG
302 - 316	(21.72/ 7.93)	VKVKANGGIKVDANG

Table 5.3 | The largest stretches of sequence repeats within the PVClumt13 tail fibre.
This table shows the sequences and statistics for the repeat detection from RADAR, for PVClumt13. lumt is a shorter protein, thus less propensity for long tandem repeats is possible, but 3 stretches relatively abundant in valine, isoleucine, glycine and aspartic acid are found.

# Repeats	Total Score	Length	Diagonal	BW-From	BW-To	Level
3	72.06	14	28	81	94	1

Repeat Indices	Alignment Score/Z-score	Repeat Sequence
81 - 94	(22.51/10.89)	LQVKAGAGVDIDNN
97 - 110	(24.55/12.40)	ITIKSGHGIKVDGN
112 - 125	(24.99/12.73)	ISVKPGSGIKVDSN

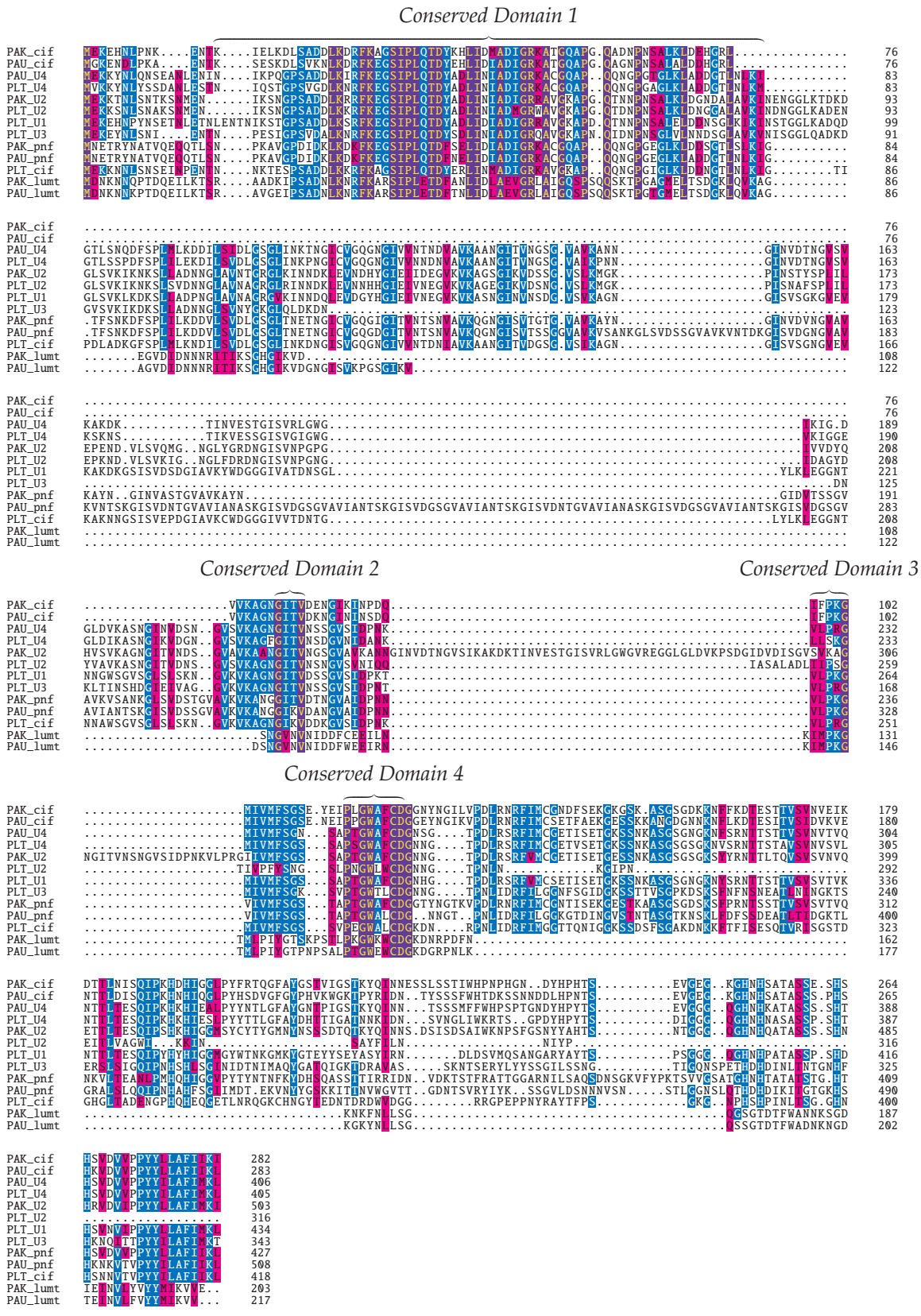
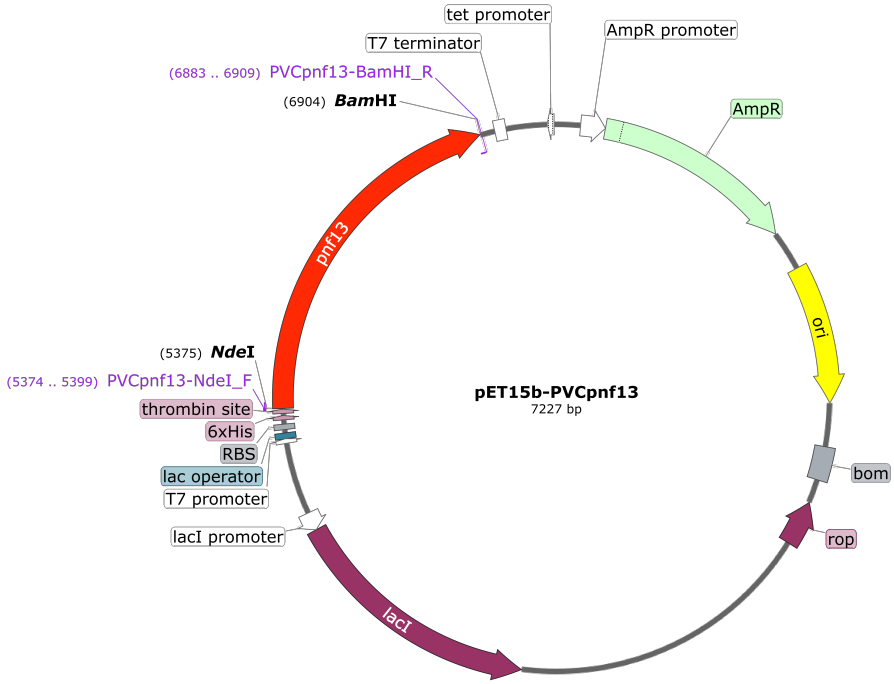


Figure 5.6 | Annotated multiple sequence alignment for putative PVC tail fibres

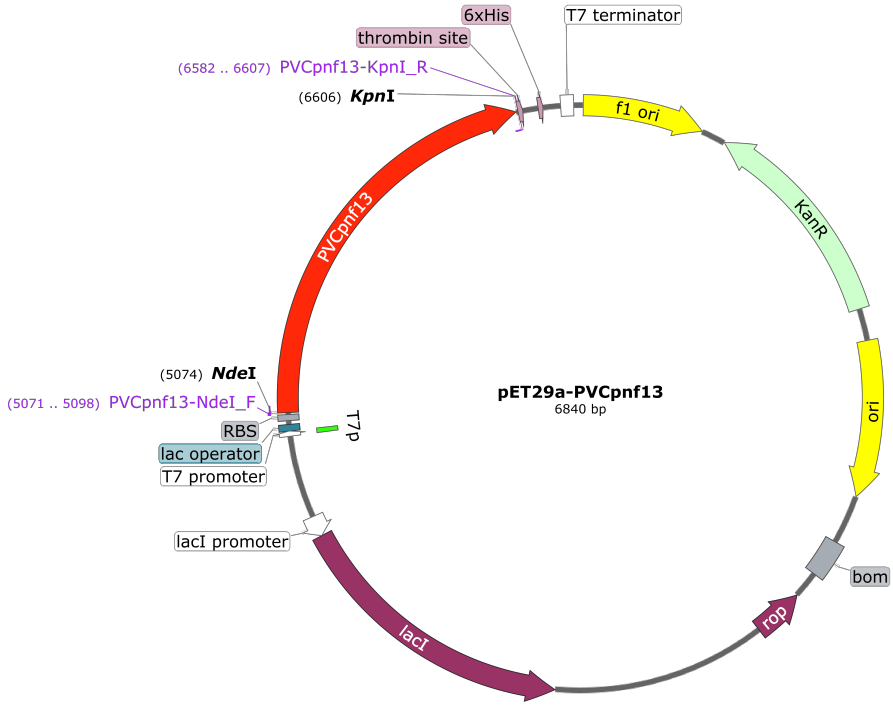
Alignment reproduced from Appendix **??** on page ??. The MSA was generated with Clustal Omega, and visualised here via the `TexShade` LaTeXpackage. Residues are colour coded by residue similarity, and identifiable conserved domains are annotated.

5.2.1.3 *in silico* cloning

From the sequence analysis alone, it was not possible to tell categorically which region of the tail fibres would be responsible for binding to the host, and binding to the rest of the tail structure. It was decided to clone both of the fibres studied C- and N- terminally hexahistidine tagged in case one tag was found to interfere with the protein downstream. Primers were designed to insert the 2 genes in-frame with the histidine tags in pET15b (N-terminal) and pET29a (C-terminal), yielding pET15b-pnf13, pET15b-lumt13, pET29a-pnf13 and pET29a-lumt13 (see Table 2.5 on page 68 and Chapter 2 on page 61 for technical detail such as primer sequences etc.). For each gene, the same forward primer bearing an NdeI restriction site was used for both pET15b and pET29a, with the ATG of the restriction site serving as the start codon for the gene. Reverse primers had to be redesigned for each vector, utilising a BamHI site for pET15b, and a KpnI site for pET29a. Also for pET29a, 2 additional bases were added to the histidine tag linker region to bring the tag in-frame. Construct maps can be found in Figure 5.8 on page 153.



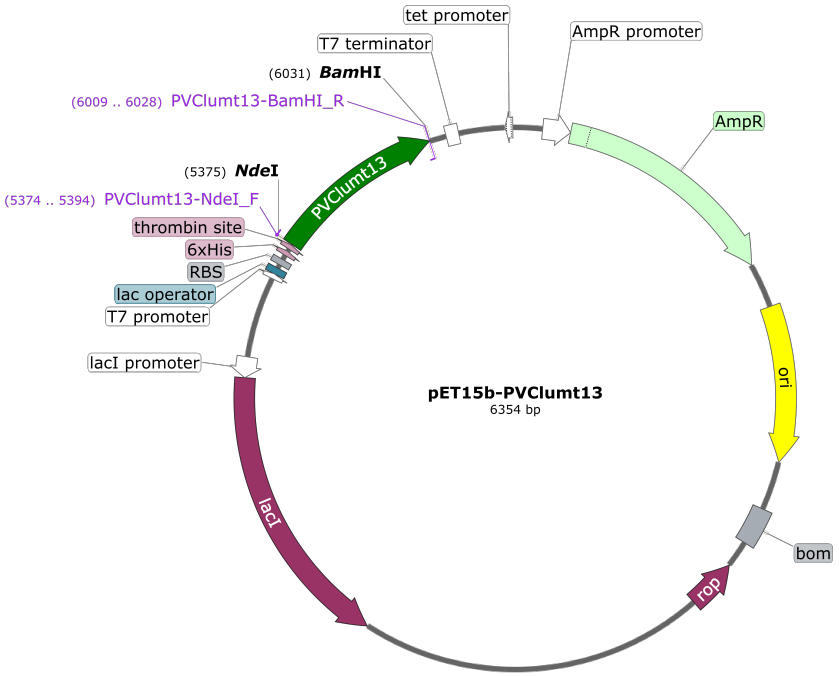
(A)



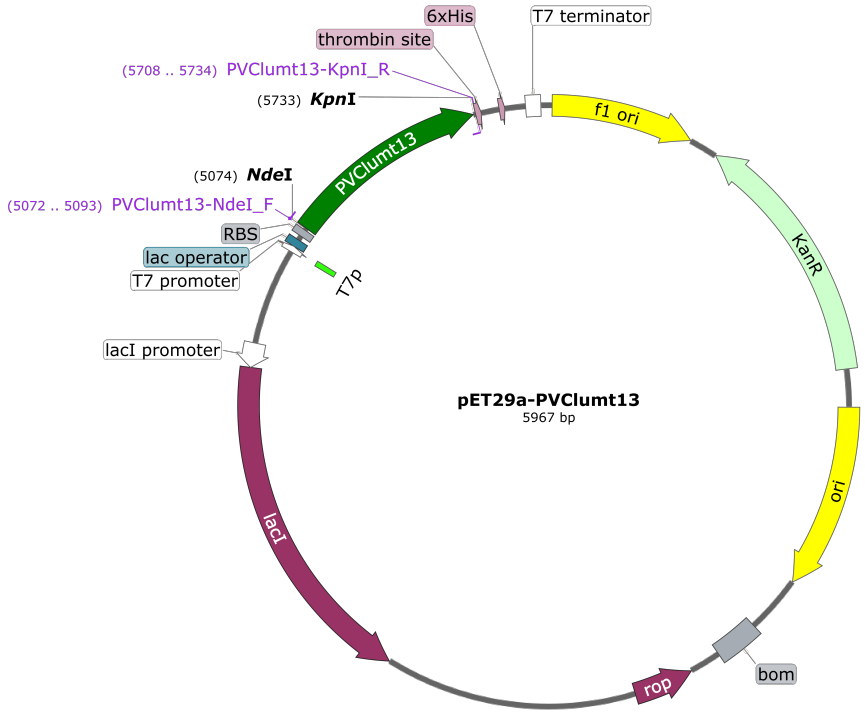
(B)

Figure 5.7 | Fusion construct maps for tagged PVCpnf13 tail fibres.

Plasmid maps for the tail fibre-hexahistidine tag fusion proteins of the PVCpnf operon from *P. asymbiotica* ATCC43949, used in this study for purification and functionalisation. The insert sequences are annotated as red CDSs, the primers are labelled in pink, restriction sites in black, fusion tags in light pink oval boxes. **(A)** The tail fibre from the PVCpnf operon fused N-terminally to a hexahistidine tag in the vector pET15b.**(B)** The tail fibre from the PVCpnf operon fused C-terminally to a hexahistidine tag in the vector pET29a.



(A)



(B)

Figure 5.8 | Fusion construct maps for tagged PVClumt13 tail fibres.
Plasmid maps for the tail fibre-hexahistidine tag fusion proteins of the PVClumt operon from *P. asymbiotica* ATCC43949, used in this study for purification and functionalisation. The insert sequences are annotated as green CDSs, the primers are labelled in pink, restriction sites in black, fusion tags in light pink oval boxes. **(A)** The tail fibre from the PVClumt operon fused N-terminally to a hexahistidine tag in the vector pET15b.**(B)** The tail fibre from the PVClumt operon fused C-terminally to a hexahistidine tag in the vector pET29a.

5.2.2 Experimental cloning, expression and purification

To generate the constructs for protein expression, PCRs were conducted following standard manufacturers procedures, using the primers as per Table 2.8 on page 70, and the PCR conditions outlined in Table 2.10 on page 73, with the proofreading Q5 enzyme. Genomic DNA was prepared using the Qiagen Blood and Tissue kit (see Section 2.2.1.1 on page 64). High-fidelity New England Biolabs restriction enzymes used for both constructs had compatible incubation conditions and thus cloning was achieved by direct double digest of inserts and vectors, heat inactivation and proceeding directly to ligation and transformation all according to manufacturers specifications. All constructs were confirmed by Sanger sequencing.

Both tagged proteins were able to be expressed well in an overnight culture, using a derivatised *E. coli* BL21(DE3) strain from NEB ("NiCo21") when induced at an OD_{600nm} of 0.4-0.6. Figures 5.9 to 5.10 on the following page show Western blots using an anti-HIS primary antibody and an anti-mouse/rabbit Horseradish Peroxidase (HRP) conjugate secondary antibody from a time course expression trial. It was not possible to see a Western signal from the C-terminally tagged (pET29) construct for pnf13, and in both cases, greater expression was seen from the pET15b, N-terminal constructs.

It remains unclear why no signal could be seen with the N-terminally tagged pnf13. The most likely explanations are that the His tag lead to malformed protein which may have formed inclusion bodies, been rapidly degraded, or potentially that the C-terminus in this particular tail fibre is buried. The N-terminal constructs were scaled up and used for all further purifications and analyses.

5.2.2.1 IMAC Purification and Polishing

Purification was performed via Immobilised Metal ion Affinity Chromatography ("IMAC"), with $Nickel^{2+}$ as the metal ion, and sample polishing was done with a Superdex200 gel filtration column. Each sample was able to be purified well, particularly in the case of lumt13, where it was not uncommon to recover in excess of 100 mg of protein from 2 litres of bacterial culture. Purification was performed using an Äkta FPLC system and a

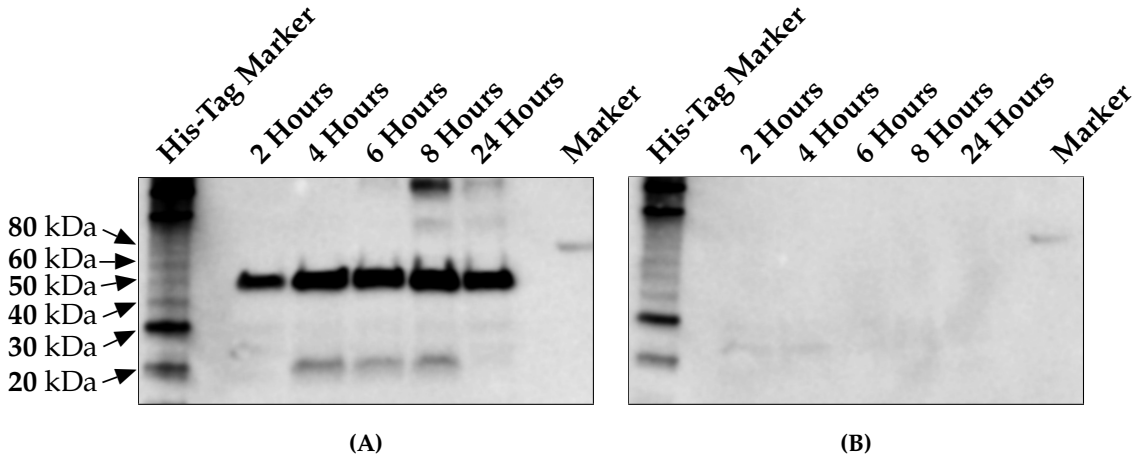


Figure 5.9 | A Western blot of pnf13 expression in BL23(DE3) cells after induction. **(A)** Western blot of inductions of pnf13 from the pET15b vector, with an N-terminal Hexahistidine tag. **(B)** Western blot of inductions of pnf13 from the pET29a vector, with a C-terminal Hexahistidine tag. Good yields can be seen as early as 2 hours for the N-terminal tag – No expression of C-terminally tagged pnf13 could be observed. Subsequent time points have been normalised to the same optical density, showing a roughly equivalent amount of protein on the gel, but higher yield in culture due to increased cell numbers.

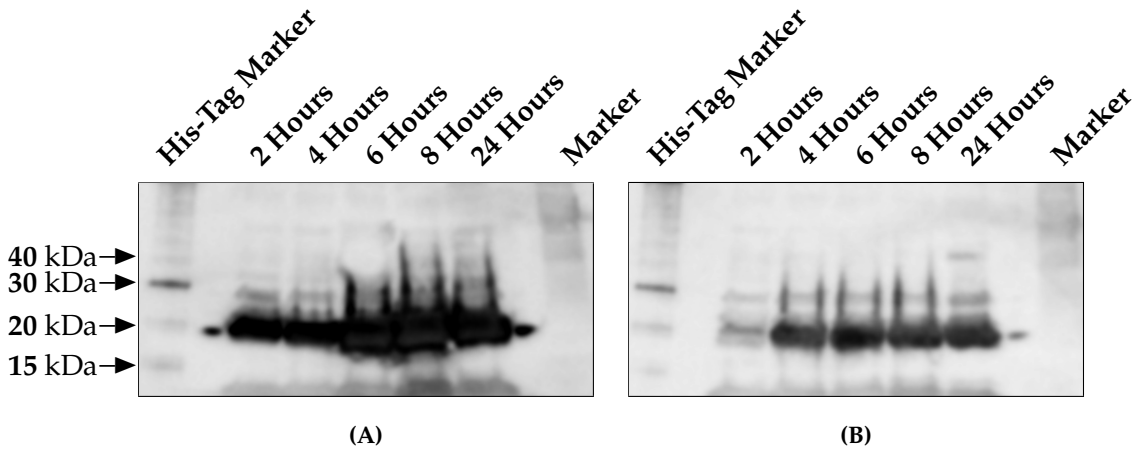
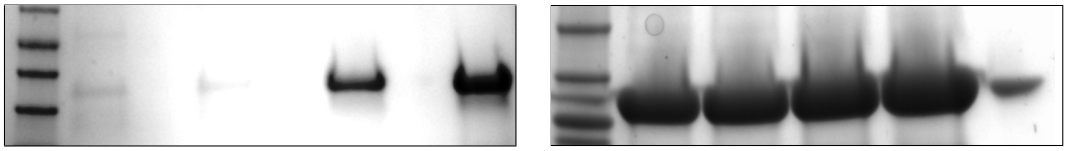


Figure 5.10 | A Western blot of lumt13 expression in BL23(DE3) cells after induction. **(A)** Western blot of inductions of lumt13 from the pET15b vector, with an N-terminal Hexahistidine tag. **(B)** Western blot of inductions of lumt13 from the pET29a vector, with a C-terminal Hexahistidine tag. Good yields can be seen as early as 2 hours for the N-terminal and C-terminal tags, with greater expression from the N-terminal (pET15b) clones. Subsequent time points have been normalised to the same optical density, showing a roughly equivalent amount of protein on the gel, but higher yield in culture due to increased cell numbers.

gradient elution (or gravity flow resin chromatography). Over the course of this project, multiple rounds of purification were performed, but the chromatogram trace was not highly reproducible, and had broad peaks; as a result, SDS-PAGEs were run on candidate fractions to identify the purest ones. Given the potential trimeric nature of the tail fibres, 3 hexahistidine tags will be present per final protein structure. A potential explanation for the unusual chromatograms could, therefore be, that stochastically, some proteins manage to bind 1, 2, or 3 histidine tags, resulting in differences in binding strength. It might be expected, in this case, that 3 peaks would result as the affinity of each multiple binding is reached and subsequently eluted, though this wasn't commonly observed, so something more complicated may be occurring. It is also possible that the tail fibres putative 'extruded' shape may impede their flow into and out of the column, even when the dissociation point is reached, causing an extended elution peak. One final potential explanation could be that, given the putative nature of the tail fibres as binding structures, they may be quite 'sticky' and are therefore interacting with the column or other binding partners as yet unknown.



(A) **(B)**

Figure 5.11 | SDS-PAGE gels of the PVC tail fibres purified from pET15b.
(A) Stained SDS-PAGE gel of pnf13 expressed from pET15b, from several fractions across the elution profile from IMAC, after concentration with Amicon centrifugal columns. Sample is approaching purity. **(B)** Staining of SDS-PAGE gel of lumt13 expressed from pET15b, from several fractions from across the elution profile, after concentration with Amicon centrifugal columns. Samples are approaching purity. The difference in expression levels between pnf13 and lumt13 is apparent. Final polishing was conducted via gel filtration with a Superdex 200 Increase column.

5.2.3 Structural Analyses

5.2.3.1 Trimerism of PVC tail fibre proteins

During routine SDS-PAGE while running expression and purification experiments, it was observed that the tail fibres often did not readily migrate in to the acrylamide (this can be seen in Figure 5.12 on the next page). A standard SDS-PAGE set up included boiling the sample in the presence of gel loading dye, containing DTT, β-mercaptoethanol and

SDS, which ordinarily would be more than enough chemical and physical disruption to denature most proteins. Better, though in the case of pnf13, not complete, denaturation could be coerced with the presence of urea at roughly 8 M, and the inclusion of EDTA in the loading dye. This thermal/chemical stability is a known hallmark of β-stranded fibre proteins and is a valuable indication of the true structure and correct fold of these proteins (Papanikolopoulou et al., 2008a,b)

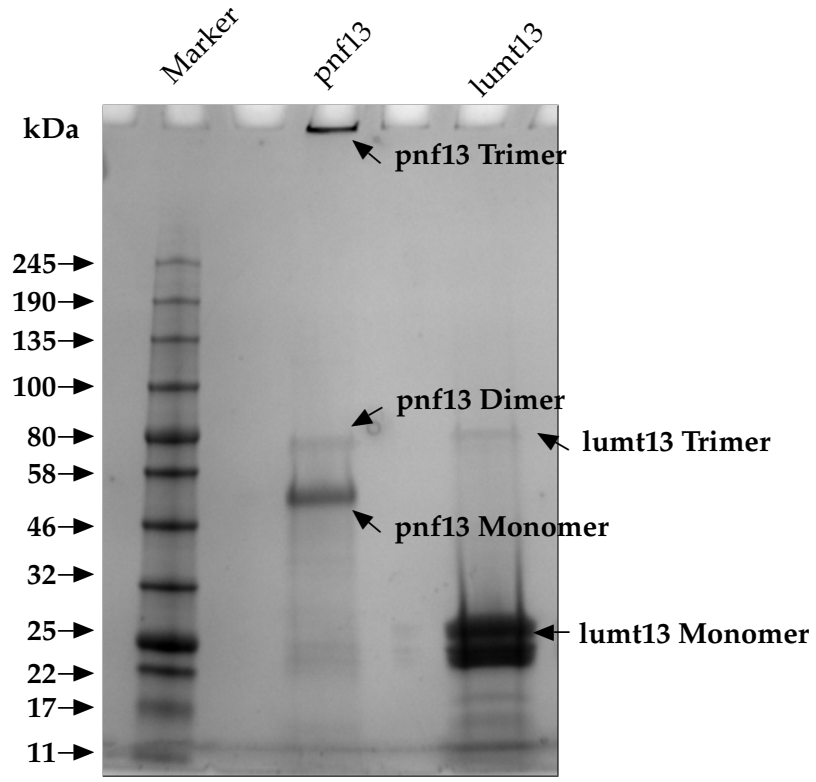


Figure 5.12 | Trimerism of PVC tail fibres revealed via SDS-PAGE

An example of a routine SDS-PAGE gel for semi-purified (post IMAC) tail fibre proteins (pnf13 and lumt13). Despite being a denaturing gel, where the the input samples were boiled in SDS loading dye with urea, multimeric forms of the proteins can be seen, demonstrating the stability of the tail fibres. Trimeric, dimeric and monomeric forms of pnf13 are identifiable (the trimeric form remains in the well at the top of the gel). For lumt13, monomeric and trimeric forms are apparent. Dimeric forms are seemingly sufficiently unstable that they entirely denature completely to monomers, if the trimers denature at all.

5.2.3.2 Thermal stability and secondary structure studies via Circular Dichroism

Upon observing the stability of the tail fibres in denaturing conditions, it was decided to examine this thermal stability further via temperature ramping CD experiments, from which it is also possible to get secondary structure and to get an indication of whether the folding of the proteins is occurring correctly. Figures 5.13 and 5.14 show a composite of 15 spectra each, acquired at 5 °C increments and coloured by temperature. Each spectrum was acquired 6 times (technical replicates) in each run, and for each protein, 3 runs were performed at separate times, each with a different protein preparation to ensure consistency of purifications (biological replicates). For pnf13, spectra were run at 0.1 mg $mL^{-1}$and for lumt13, at 0.25 mg mL^{-1}. Concentrations for each run were determined empirically prior to setting up the temperature ramp, by sequentially 2 fold diluting a 1 mg mL^{-1}stock of each protein until the CD spectrometer HT voltage did not exceed ≈600 V at 190 nm. While spectra were collected from 260 to 185 nm, without extremely pure buffers and high quality light sources (typically synchrotrons), CD data becomes very noisy at lower wavelengths as many molecules begin to absorb around the 190 nm region. Consequently, only data down to 190 nm was included for analysis. All spectra are baseline subtracted against the buffer control (Sodium fluoride).

A transition can be seen as the 2 extremes of temperature are separated on the graph. Characteristically, this occurred at approximately 65 °C, for pnf13 - putatively signalling the start of unfolding. At higher temperatures, the β-sheet signal actually intensified. For lumt13, the major collapse of secondary structure appears between 50 and 60 °C, but no other structure seems to appear at higher temperatures. In both cases, even up to 95 °C, secondary structure seemingly persists as the signal is not abolished completely, though the structure is almost certainly no longer in its native form.

5.2.3.3 Secondary structure prediction via Dichroweb

As mentioned, one of the primary reasons to conduct circular dichroism studies is to gather information about the secondary structure of a protein. Through use of tools like Dichroweb, input spectra can be deconvoluted and compared to the CD spectra for other proteins of known structure. By doing so, the secondary structure for the unknown

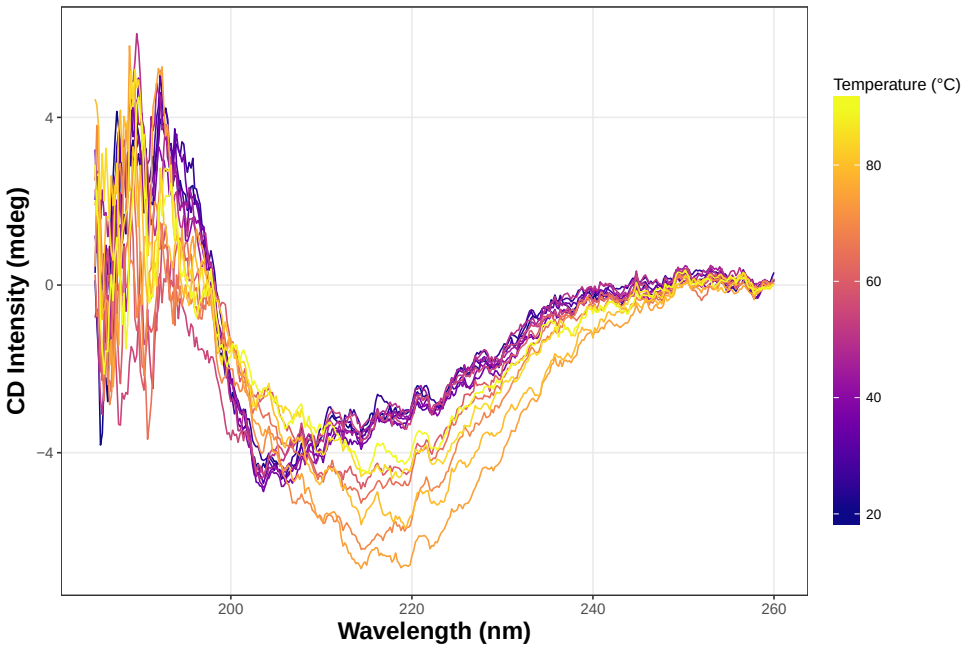


Figure 5.13 | CD TEMPERATURE RAMPING SPECTRA FOR PNF13

A composite of 15 CD melt spectra from the temperature ramping experiment for pnf13 (from 20 °C to 95 °C in 5 °C increments). These are average spectra from 3 biological replicates (each of which in turn is an average of 6 technical replicate spectra accumulations). Cooler colours (purple) correspond to lower temperature spectra, and warmer colours (yellow-orange) correspond to higher temperature spectra.

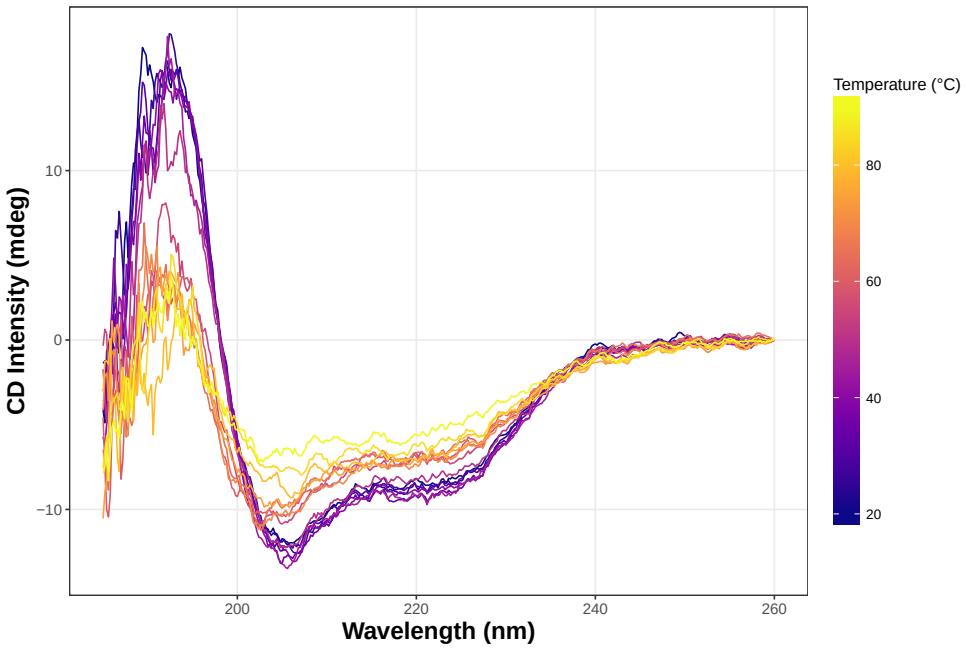


Figure 5.14 | CD TEMPERATURE RAMPING SPECTRA FOR LUMT13

A composite of 15 CD melt spectra from the temperature ramping experiment for lumt13 (from 20 °C to 95 °C in 5 °C increments). These are average spectra from 3 biological replicates (each of which in turn is an average of 6 technical replicate spectra accumulations). Cooler colours (purple) correspond to lower temperature spectra, and warmer colours (yellow-orange) correspond to higher temperature spectra.

candidate protein can be approximated (Whitmore and Wallace, 2004; Lobley et al., 2002). Each of the 45 spectra for each protein (15 spectra per biological replicate) were analysed, and the resulting secondary structure proportions average for each temperature between the 3 runs, thus reporting the average secondary structure across the replicates and temperature curve.

5.2.3.3.1 Algorithm and reference set selection

For calculation, the CDSSTR algorithm (Compton and Johnson, 1986; Sreerama and Woody, 2000; Manavalan and Johnson, 1987) was chosen for a number of reasons. Firstly, it is cited as being one of, if not the most accurate algorithms for circular dichroism (having superseded a number of the others), but with the tradeoff of increased run-time, though that was not a concern for this analysis. Secondly, it is compatible with the spectra reference set chosen (see the following section), and the wavelengths captured (some algorithms require sub-190 nm data, which was available, but considerably more noisy as seen in Figures 5.13 to 5.14 on page 159). The other options for the dataset range available: SELCON, CONTIN and K2D all fail to match the accuracy of CDSSTR in testing here. K2D doesn't require a reference set but provided the worst NRMSD values by a significant margin (see Figure 5.15B on page 162), and only analyses spectra to 200 nm.

Fitting quality was trialled with a number of reference sets compatible with the scan parameters and algorithms available (this immediately limited choices to only a couple of reference sets). It was decided to proceed with reference set 7, as it contains the largest number of non-specialist proteins (i.e. non-membraneous etc.), and also because it contained spectral information for denatured proteins, which for the denaturing gradients seemed likely to give the best representation of the spectra (Sreerama and Woody, 2000; Sreerama et al., 2000). Full details of all the spectra can be found at the Dichroweb site[1]. Reference set 4 also gave good results in testing, but as Set 7 contains all of Set 4's proteins in addition to extras, including denatured forms as mentioned, it was adopted instead. Set 6 was also able to give decent spectra fits, but uses the full 185 nm data range; without extremely high quality experimental materials and access to a synchrotron, it is typically considered unwise to analyse beyond 190 nm.

[1] http://dichroweb.cryst.bbk.ac.uk/html/userguide_datasets.shtml

As an example of the improved results obtained from use of the CDSSTR and the chosen reference set, spectra are shown in Figure 5.15 on the following page for one of the input spectra tested under several models. Dichroweb further provides and NRMSD (Normalised Root-Mean-Square-Deviation) statistic (Mao et al., 1982), to quantitatively assess the least squares goodness of fit.

5.2.3.3.2 Secondary structure predictions

With optimal parameters for secondary structure calculation through Dichroweb identified, all spectra were analysed for their relative secondary structure proportions. Figure 5.16 on page 163 shows the relative proportions according to Dichroweb, plotted as stacked bars. The increasing temperatures are plotted along the y-axis, and the percentage of each secondary structure type long the x-axis. Dichroweb recognises 6 classes of secondary structure, including 2 types of both α-helix and β-sheet. Respectively these are: "Helix1" - regular α-helix, "Helix2" - 'distorted' α-helix, likewise for "Strand" 1 and 2, and finally unstructured turn and unordered regions.

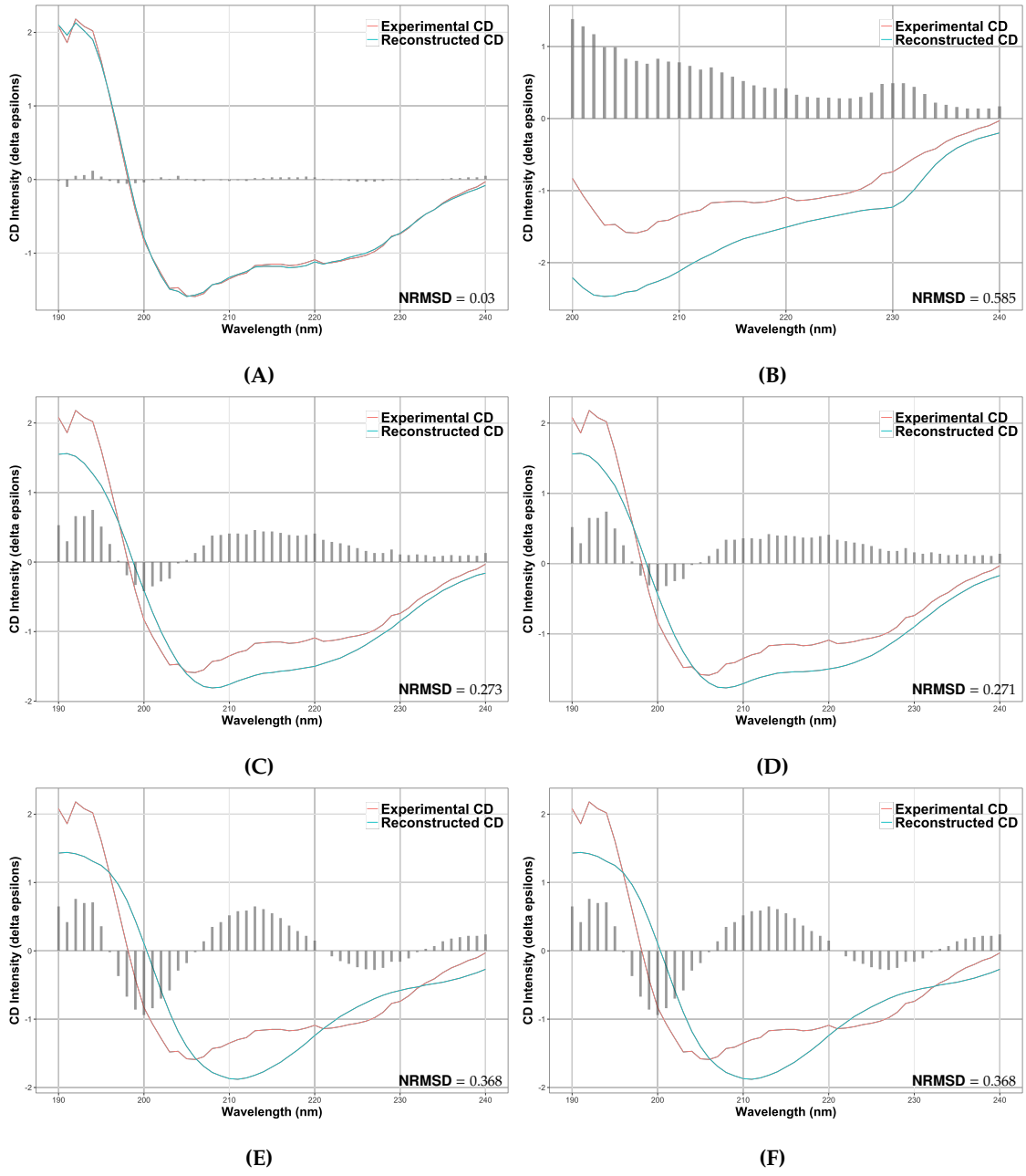


Figure 5.15 | Comparisons of optimal Dichroweb algorithms and reference sets.
Each of these charts shows a comparison of a different algorithm and reference set for identifying the optimal parameters for estimation of secondary structure proportions from the acquired spectra. As an example, each spectra shows the result for the 20 °C spectra for lumt13 when analysed with a selection of compatible reference sets and algorithms. Pink lines are the experimental spectral data, and cyan lines are the reconstructed reference data. light grey bars depict the residual difference between the 2 line spectra at that point to highlight the disparity. **(A)** The optimal solution from this testing, of the lumt13 spectra analysed using CDSSTR and reference set 7. **(B)** Analysis result from the K2D algorithm, which does not require a reference set. **(C)** Result of spectral analysis using the SELCON algorithm and reference set 4. **(D)** Result of spectral analysis using the SELCON algorithm and reference set 7. **(E)** Result of spectral analysis using the CONTIN algorithm and reference set 4. **(F)** Result of spectral analysis using the CONTIN algorithm and reference set 7. Note there is little to no difference in the use of set 4 or set 7.

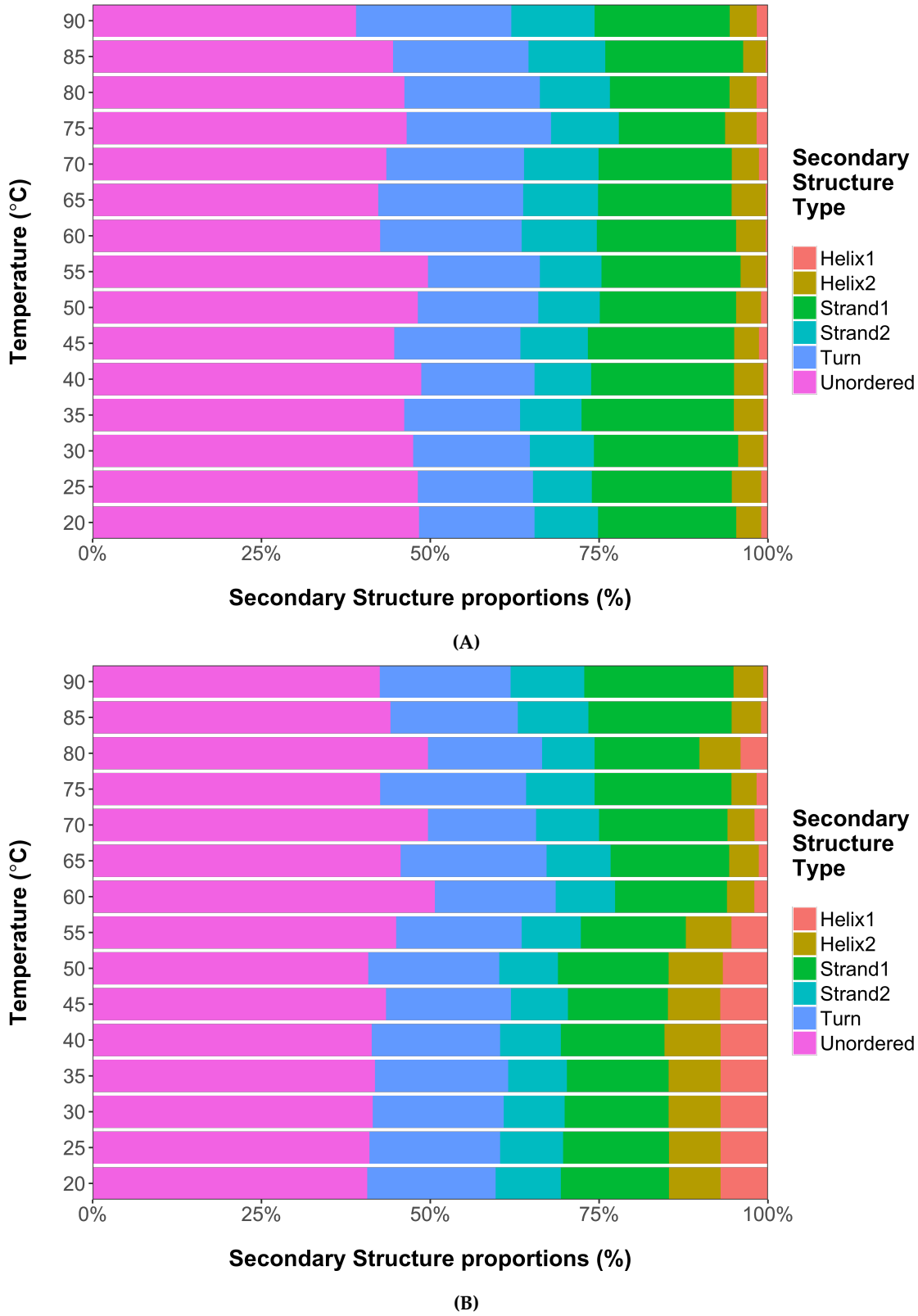


Figure 5.16 | CD MELT SECONDARY STRUCTURE PROPORTIONS FOR TAIL FIBRE PROTEINS.
Stacked bar charts showing the proportions of secondary structure as estimated by Dichroweb, for averages of the 3 replicate spectra, for each protein at 15 different temperatures in the melting gradient experiment. **(A)** Secondary structure proportions for pnf13. **(B)** Secondary structure proportions for lumt13. "Helix1" - regular α-helix, "Helix2" - 'distorted' α-helix, "Strand 1" - regular β-sheet, "Strand 2" - distorted β-sheet, "Turn" - turns/loops, "Unordered" - No canonical secondary structure.

5.2.3.4 Comparisons with known structures

The cloned tail fibres from the PVCs appear to be dominated by unordered, turn, and β-sheet motifs. The assumption is made at this point that the tail fibres are likely to be folding correctly in to their native structures for the following reasons. Firstly, the 'knitted' and interwoven trimeric nature of known tail fibres is unlike that of trimers of typical globular proteins, where they simply form 3 identical monomers which each have a functioning 'lone' structure, and simply complex together. For tail fibres, the functional structure *is* the trimeric form, and all three monomers have to contribute to form the structure. Each monomer on its own would not be capable of maintaining the extruded structure which will have energetically unfavourable regions exposed. The trimerism of known tail fibre structures is apparent from Figure 5.1 on page 141, and for these tail fibres is reinforced by Figure 5.12 on page 157. Secondly, the stability that was seen in the CD melt studies is characteristic of phage proteins, and tail fibre like proteins in particular. Thirdly, the ability to probe and purify via the histidine tag suggests that the proteins are not simply malformed and creating inclusion bodies etc., if that were the case, it would be expected that the histidine tags would be buried within the inclusions and purification would likely have failed.

However, to compare this directly to the published structures for validity, the secondary structure proportions for several existing tail fibre proteins was examined in 2 ways. Firstly, the 'raw' secondary structure proportions were calculated directly from the PDB crystal structures via a bespoke script, using PyChimera (Rodríguez-Guerra Pedregal and Maréchal, 2018) and UCSF Chimera (Pettersen et al., 2004), which in turn assigns secondary structure using the well-known DSSP algorithm (Kabsch and Sander, 1983). Secondly, another web service from the groups behind Dichroweb, "PDB2CD"[2], simulates circular dichroism spectra from resolved structures, and thus the secondary structure of analogous proteins is compared here.

By way of example, the secondary structure proportions for the structures shown in Figure 5.1 on page 141, and domain homologies detected by HHPred are reproduced in Table 5.4 on the next page. Note, these results are extracted from DSSP assignments as

[2] http://pdb2cd.cryst.bbk.ac.uk/

mentioned in the last paragraph, however DSSP only recognises 3 classes of secondary structure (thus the percent helix according to DSSP represents the approximately the combined Helix 1 and Helix 2 that Dichroweb reports for the same structure, and so on).

Table 5.5 shows the same secondary structure calculations performed on the same set of structures, but instead, uses the data output by Dichroweb. While this abstracts the data from the crystal structure slightly, it makes the spectra more directly comparable to the data for the PVC tail fibre proteins. In order to obtain this data, circular dichroism spectra are simulated from the PDB depositions, using the webserver PDB2CD (Mavridis and Janes, 2017), and in turn passed back through Dichroweb. In this case, the CDSSTR algorithm was used, however PDB2CD uses the SP175 reference set (Lees et al., 2006), so this was also used with Dichroweb.

Table 5.4 | DSSP secondary structure proportions for resolved tail fibre proteins.
The secondary structure proportions for various tail fibre like proteins with resolved atomic structures in the PDB database, as determined by calculation directly from the atomic structure. The corresponding structures can be found in Figure 5.1 on page 141 and in Table 5.1 on page 146, with the exception of PDB ID 1PDI, which, for an unknown reason, fails to have secondary structure assigned by DSSP/UCSF Chimera.

PDB ID	% Helix	% Sheet	% Other
2XGF	2	16	82
5NXF	5	9	86
1QIU	7	26	67
1H6W	7	6	87
1V1H	4	18	78
3IZO	13	18	70
1OCY	5	2	93

Table 5.5 | Dichroweb secondary structure proportions for resolved tail fibre proteins.
The secondary structures proportions for various tail fibres with resolved atomic structures in the PDB database, calculated via the PDB2CD and Dichroweb webservices.

PDB ID	% Helix 1	% Helix 2	% Sheet1	% Sheet 2	% Turn	% Other	NRMSD
2XGF	2	8	23	13	12	42	0.027
5NXF	0	6	28	14	11	40	0.065
1QIU	0	6	26	14	11	42	0.048
1H6W	8	11	16	11	14	39	0.032
1V1H	0	6	26	14	11	42	0.035
1OCY	11	12	13	10	14	40	0.03
3IZO	5	10	17	11	14	42	0.033

Exploring the secondary structure of the resolved structures reveals that they are also dominated by β-sheet and 'Other' secondary structure forms, though the agreement between direct calculation and simulated circular dichroism proportions is quite variable. Overall, α-helical structural spans appear very limited in known tail fibres, and this trend is also seen in the tail fibres cloned from the PVCs, contributing to only around 10-15% of the overall structure. Moreover, despite differing substantially in length, sequence and also having somewhat different melting profiles, the secondary structures of the PVCpnf13 and PVClumt13 fibres are roughly equivalent. This is therefore indicative of a robust 'tail fibre' blueprint, in which the macrostructure is important, but the sequence specifics appear free to drift - potentially significantly. For instance, it appears the coordination of one or more metal ions is common (though maybe not obligatory), and yet, the sequences don't appear to preserve a distinct binding pattern, possibly suggesting that many different ions held by many different amino acids are all 'valid solutions' to the problem of creating a tail fibre type protein.

5.2.3.5 Crystallography

Since the tail fibres were able to be expressed to reasonable quantities, some crystallographic screens were attempted, as it's the approach with greatest previous success, as mentioned in Section 5.1 on page 138.

With lumt13, crystals were obtained in 12 conditions, in under a week. Since it is not uncommon for crystallisation screens to result in no crystals at all, even after months or years of incubation, it seemed that crystallisation was a promising approach for these proteins. Table 5.6 on page 168 shows the buffer conditions for which crystals could be seen. Figure 5.17 on page 169 shows a selection of the morphologies obtained. Unfortunately, the reduced yield and purity of the pnf13 tail fibre meant that it was not possible to obtain a sufficient amount of high quality protein for screening.

5.2.3.5.1 *In-situ* partial proteolysis

Despite obtaining a good number of crystals in several conditions in the standard screens, when the largest crystals were extracted to test diffraction it was observed that the samples

were only in a semi-crystalline state, with a gelatinous quality. Consequently, no diffraction was observed with these crystals. Additionally, it was noted that, while the tail fibres appeared to readily crystallise, they often formed numerous small crystals rather than fewer large ones (Figure 5.17 on page 169). It was suspected that this was due to the crystals beginning to successfully form, but not packing closely enough. Protein surface loops which hold the protein molecules apart or contaminating proteins are a likely cause. To this end, a repeat screening was conducted, but this time using *in-situ* partial proteolysis. Proteases are added in at low concentration in to the crystal screening drop, which digest contaminating proteins that do not pack in to the crystal, and also remove some surface loops allowing tighter crystal packing. Partial proteolysis has been shown by the Structural Genomics Consortium to increase the success rate for crystallisation studies of recalcitrant proteins by 10-15% (Dong et al., 2007; Wernimont and Edwards, 2009). Since the "Wizard 1-4" buffer screens yielded most initial crystals, only these 2 were repeated for *in situ* proteolysis. Through this approach, crystals for lumt13 were obtained in another 10 conditions in just 24 hours, some of which overlapped with conditions identified in the first screen. Crystal conditions identified in both cases are shown in Table 5.6 on the next page.

Table 5.6 | Buffer conditions yielding crystals for the lumt13 PVC tail fibre protein in both 'standard' and *in situ* proteolysis screens

Buffer Screen	Well	Condition (precipitant, buffer system)
		Standard Screening
"Wizard 1 & 2"	C5	10% w/v PEG-8000, 200 mM Sodium Chloride, 100 mM CHES/Sodium Hydroxide pH 9.5
	E8	10% w/v PEG-8000, 200 mM Sodium chloride, 100 mM Potassium phosphate monobasic/Sodium phosphate dibasic pH 6.2
	G10	10% w/v PEG-8000, 100 mM Imidazole/Hydrochloric acid pH 8.0
	H7	10% w/v PEG-8000, 200 mM Magnesium chloride, 100 mM Tris base/Hydrochloric acid pH 7.0
"Wizard 3 & 4"	B8	10% w/v PEG-6000, 100 mM HEPES/Sodium hydroxide pH 7.0
	C10	10% w/v PEG-6000, 100 mM bicine/Sodium hydroxide pH 9.0
	F4	15% w/v PEG-550 MME, 100 mM MES/Sodium hydroxide pH 6.5
"Morpheus"	A11	10% w/v PEG-4000, 20% w/v glyercol, 0.03 M divalent cations, 0.1 M Bicine/Trizma base pH 8.5
"SG-1"	B2	12% w/v PEG-20000, 0.2 M Magnesium acetate tetrahydrate, 0.1 M MES pH 6.5
	H9	10% w/v PEG-8000, 0.2 M Sodium acetate trihydrate, 0.1 M Imidazole pH 8.0
		in situ proteolysis screen
"Wizard 1 & 2"	C5	10% w/v PEG-8000, 200 mM Sodium Chloride, 100 mM CHES/Sodium Hydroxide pH 9.5
	D3	20% w/v PEG-1000, 100 mM Sodium Phosphate dibasic/citric acid pH 4.2
	G10	10% w/v PEG-8000, 100 mM Imidazole/Hydrochloric acid pH 8.0
"Wizard 3 & 4"	A11	20% v/v 1,4-butanediol, 100 mM MES/Sodium hydroxide pH 6.0
	B8	10% w/v PEG-6000, 100 mM HEPES/Sodium hydroxide pH 7.0
	B11	15% v/v Reagent alcohol, 100 mM Imidazole/Hydrochloric acid pH 8.0
	C10	10% w/v PEG-6000, 100 mM bicine/Sodium hydroxide pH 9.0
	F4	15% w/v PEG-550 MME, 100 mM MES/Sodium hydroxide pH 6.5
	H1	1 M Potassium-Sodium tartrate, 100 mM Tris/Hydrochloric acid pH 7.0

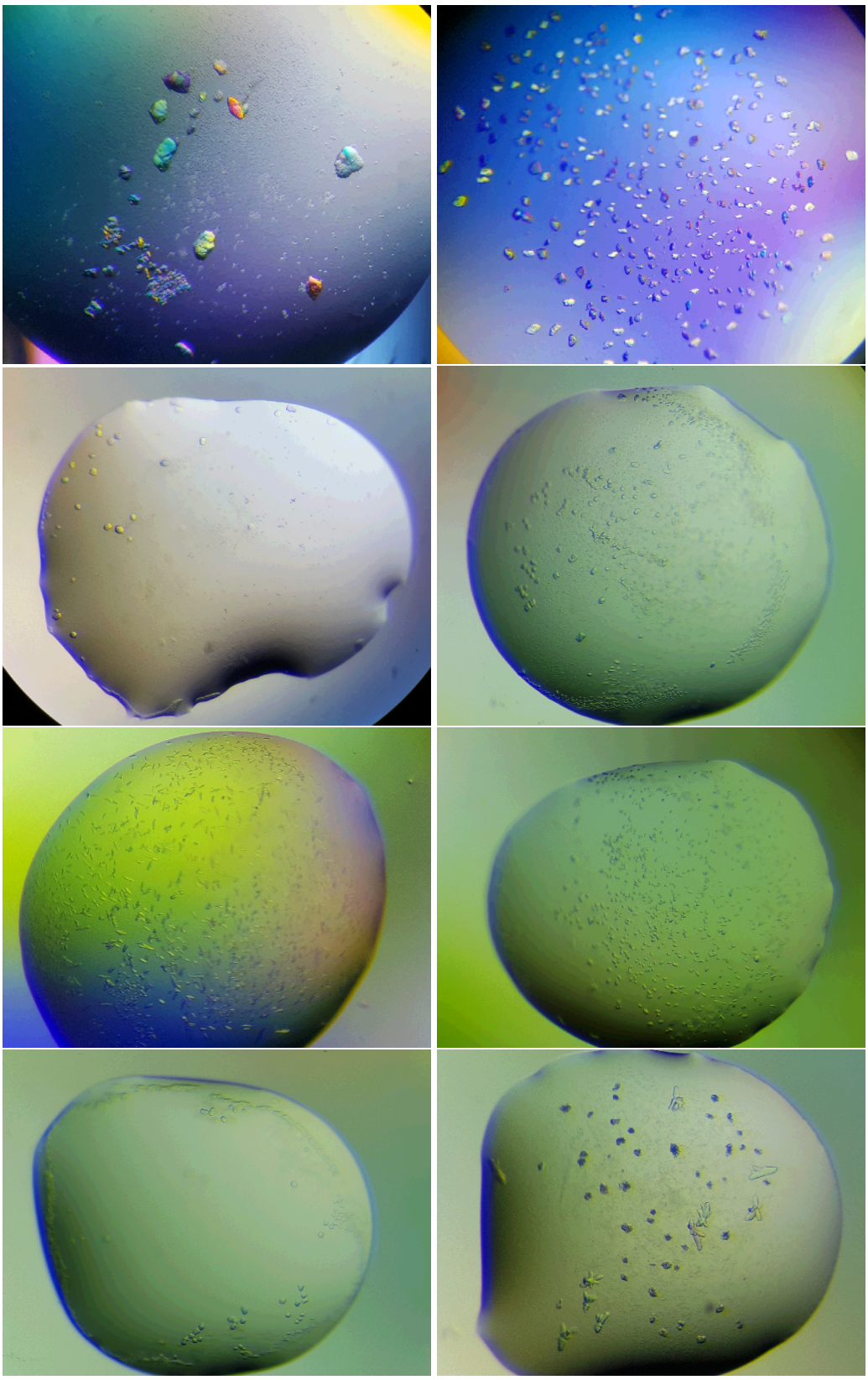


Figure 5.17 | A selection of lumt13 crystal morphologies.
A selection of the crystal morphologies obtained from crystal screening with lumt13 and the Mosquito robot, corresponding to some of the conditions in Table 5.6 on page 168.

5.2.4 Finding binding partners for tail fibre proteins

A key long term aim is to be able to specifically identify the tissues and the molecular partners that the tail fibres are preferentially binding to. A rationale for the diversity seen within the tail fibre proteins is that they may have diverged to target specific cell or tissue types as part of the PVCs role in virulence. Cloning and purifying the tail fibres in isolation from the rest of the PVCs was done in order to enable a suite of downstream bioassays without the complications of the large PVC component itself. The polyhistidine tags that the PVC tail fibres exhibit from cloning are useful 'functional handles', and several assays were designed around their use. This work was conducted in collaboration with another PhD student, who was responsible for devising the methods, so only the preliminary data obtained from these assays and their conceptual basis will be discussed. Nanoparticle conjugation of proteins is a well studied process however, and the reader is directed to Sperling and Parak (2010) and Hainfeld et al. (1999) for a good review and discussion of the mechanism.

5.2.4.1 Iron nanoparticle protein pulldowns

Magnetic (iron) nanoparticle (commonly and commercially known as "Dynabeads") pull down protocols were developed, such that the tail fibres could be incubated with whole protein extracts from tissues of interest to broadly identify candidate binding partners. Briefly, these assays simply required conjugation of the polyhistidine tagged tail fibres to iron nanoparticles which were coated in the chelator nitrilotriacetic acid (NTA), which in turn coordinates Nickel (see Figure 5.18 on the next page). This is the same chemistry as that used in the IMAC purification process.

After incubation of the nanoparticle-tailfibre complex with cellular lysates from mammalian cell lines, the particles are pulled from solution magnetically. The particle complexes can then be washed and have the nanoparticles eluted with imidazole (in the same way as IMAC again). Once the tail fibres and any proteins they have bound are free of the iron nanoparticles, they are processed via Orbitrap mass spectrometry to identify peptides. Proteins which have bound to the tail fibres should be enriched in the pulldown samples. Candidates from preliminary studies with the PVClumt13 tail fibres incubated

with lysates from A549 lung epithelial carcinoma cell lines are shown in Table 5.7 on the following page. These candidates were shown as statistically significantly enriched in peptide numbers matching to these proteins between sample and controls. They have also been filtered to remove likely spurious or uninformative hits (i.e., common contaminants have been discarded, and only proteins with plausible cell surface localisation have been retained). These pulldown studies are still preliminary; as more tail fibres can be tested with more cell lines, patterns in the binding activity of the tail fibres may begin to emerge, though some promising results are detected.

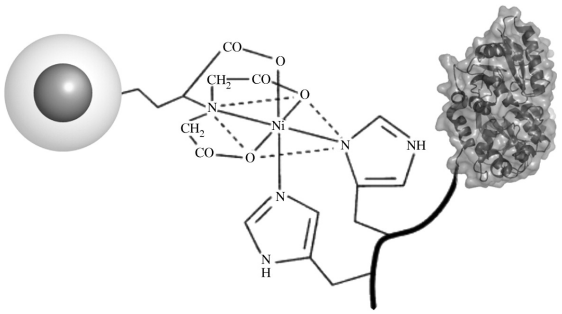


Figure 5.18 | The binding mechanism for nanoparticles interacting with polyhistidine tagged proteins.

A structural and illustrative diagram of the interaction between metal nanoparticles, Nickel nitrolotriacetic acid chelating coordination groups, and the polyhistidine tracts of a target protein. Image adapted and reproduced from Sperling and Parak (2010), which in turn is adapted from Hainfeld et al. (1999).

Table 5.7 | Preliminary candidate proteins enriched in tail fibre Dynabead pulldowns. This table shows the enriched candidate binding partners from PVClumt13 binding studies. The dataset has been filtered to remove likely contaminant proteins, and to retain only statistically significant, and likely cell surface markers or other proteins with plausible localisations so as to be physiologically relevant to the role of the tail fibres.

Protein	Gene Name	Localisation	Putative Role
Protein FAM184A	FAM184A	Cell surface/extracellular	Unknown function.
Lipocalin-1	LCN1	Cell surface/extracellular	Binds a wide range of ligands.
Lactotransferrin	LTF	Nuclear/cell surface/extracellular	Among many functions, LTF promotes binding of species C Adenoviruses to epithelial cells.
Serpin B12	SERPINB12	Cytoplasmic	Protease inhibitor.
Desmoplakin	DSP	Plasma membrane	Desmosome component (cell adhesion).
Desmocollin-1	DSC1	Plasma membrane	Desmosome component (cell adhesion).
Desmoglein-1	DSG1	Plasma membrane	Desmosome component (cell adhesion).
Dermcidin	DCD	Extracellular	Antimicrobial peptide with proteolytic activity.
Cystatin-A	CSTA	Cytoplasmic	Role in desmosome adhesion in lower levels of the epidermis.

5.2.4.2 Sugar binding studies via glycan arrays

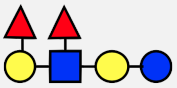
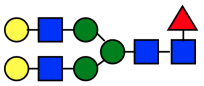
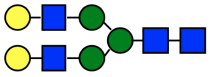
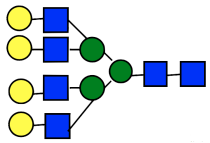
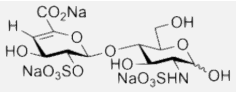
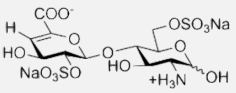
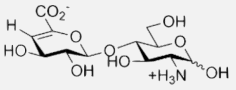
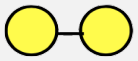
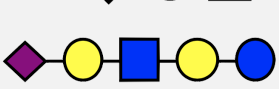
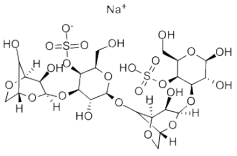
Since the previous 2 approaches primarily aimed at identifying cell/tissue types and proteins which were preferentially binding the tail fibres, an additional assay utilising glycan arrays was designed to screen for sugar binding. The main motivation for this is that there is significant precedence in the literature for other tail fibre proteins binding surface sugars. The T7 phage short fibre has been shown to bind kojibiose for example, and many phage fibres are known to bind LPS and other surface glycans, (Simpson et al., 2015; Le et al., 2013) as well as outer membrane proteins (e.g. LamB and OmpA) (Chatterjee and Rothenberg, 2012; Morona et al., 1984). For PVC fibres, the hypothesis however, is that Adenoviral motifs have replaced the distal region of the proteins, which theoretically means that the phage tropisms should not be so relevant. Adenovirus binding targets are reasonably well understood, with examples of binding to the CD46 (human Adenovirus

B) (Gaggar et al., 2003) and the Coxsackie-Adenovirus Receptor (CAR). Guardado-Calvo et al. (2010) have demonstrated however, that certain types of Adenoviral motifs do not bind the canonical receptors, and in fact, contain galectin domains resulting in tropisms for cell surface sugars.

Glycan arrays were purchased from Dextra UK, which have 104 unique glycans printed on to the slides. Binding was studied by use of a fluorescent FITC Anti-HIS antibody. Due to quantity of protein available, the glycan studies have only been conducted on PVClumt13 to date. While the results are still preliminary, 3 array tests were run, and spots with a fluorescence intensity fold change of at least 1 between control and sample were counted. Table 5.8 on the next page shows the glycans which were identified as bound hits.

Table 5.8 | Array glycan hits for PVClumt13 tail fibre binding.

The glycans from the array which were identified as putatively binding PVClumt13, along with their origins/roles and structure following the "Symbol Nomenclature For Glycans" (where available) (Varki et al., 2015). Gal = Galactose, Glc = Glucose, Man = Mannose, GalNAc = N-Acetyl-Galactosamine, GlcNAc = N-Acetyl-Glucosamine, Neu5Ac = N-Acetyl-Neuraminic Acid, Fuc = Fucose, UA - Uronic Acid.

Glycan	Glycan Provenance	Glycan Structure
Lacto-*N*-difucohexaose I	Lactose based "O"-glycans	
Asialo galactosylated, fucosylated biantennary	Complex type N-glycans	
Asialo, galactosylated, biantennary		
Asialo, galactosylated, tetranatennary, N-linked		
ΔUA→2S-GlucNS	Heparin/Chondrotin derived oligosaccharide	
Heparin unsaturated disaccharide I-H		
Heparin unsaturated disaccharide IV-H		
α1-6/α1-4 mannobiose	oligomannose core structures	
Gal-β1-6-Gal	Tumour antigens and oligosaccharide core structures	
Galβ1-3-GalNAc-β1-4-Gal-β1-4-Glc	*N*-acetyllactosamine analogues	
3'-sialyllactosamine	sialylated oligosaccharides	
LS-tetrasaccharide C (LSTc)		
Neocarratetraose-4^{1,3}-di-O-sulphate (Na+)	Neutral and sulfated Galacto-oligosaccharides	

5.3 Discussion

Tail fibres are an integral part of the mechanism underlying phage life cycles, and more broadly, 'free-living' caudate structures. If the literature that suggests 'trapped' caudate structures like the T6SS have 'antennae' is correct (see Figure 1.12 on page 40 and Chang et al. (2017)), it may be the case that tail fibres are an essential part of most if not all caudate structures. At the beginning of this work, various seemingly unusual orthologies for the PVC tail fibres were detected, including assorted phage and Adenoviral fibre motifs. It was unclear whether these were meaningful or spurious since the matches often did not cover the whole protein, the same domains were not always detected between different putative tail fibres, and often had poor similarity statistics. An appealing hypothesis was formed however, namely, that the proteins may represent natural chimeras between 'anti-eukaryotic' viral binding moieties (Adenoviral motifs), and more T4 phage-like domains to maintain a mounting interface with the rest of the phage-like tube. If this hypothesis proves correct, these proteins represent, to our knowledge, the first natural example of chimerism between viral sequences of prokaryotic/phage origin, and those of viruses from higher organisms. This chapter set out to shed some of the first experimental light on these proteins, to examine if this split domain architecture and putative similarity to Adenoviridae was valid.

5.3.1 Cloning, purification, and characterisation of PVC tail fibres

Fortunately, the tail fibres appeared amenable to tagging and purification overall, though PVClumt13 was significantly easier to work with. It was observed that no signal resulted from a C-terminally tagged PVCpnf13 tail fibre Western blot after expression (Figure 5.9 on page 155). The most likely explanation for this is that the C-terminus is buried within that particular structure, so it may be the case that the protein is still expressable but simply not detectable, as it was possible to express PVClumt13 in this manner. In the latter case, reduced yield of protein was also observed, though this could be down to subtle differences in the vector behaviour since the proteins were not in identical backbones, though they were both pET vectors.

Similarly, even between constructs with the same vector backbone, there was a rea-

sonable degree of difference in the amount of protein expressed. PVCpnf13 consistently yielded less protein than did PVClumt13. It's possible that this is simply due to PVCpnf13 being approximately twice the mass/size of PVClumt13, which simply means less protein is synthesisable for the same starting raw materials. Nevertheless, both proteins were able to be purified efficiently with a relatively simple metal ion affinity and gel filtration process, directly from crude cell lysates. Now that it has been devised for these tail fibres, this protocol is already being tested on additional fibres from other PVC operons, and hopefully in future it will aid in unpicking the precise molecular interactions of all of these proteins, which will in turn shed light on the manner in which PVCs are deployed 'in the wild'.

It is important to have information about the folded state of any expressed protein. This is particularly so if, as in this study, downstream functional information is desired. Circular dichroism has long been a go-to technique for the cheap and non-destructive structural characterisation of biomolecules, and in particular, proteins. The fact that seemingly intact protein (indicated by its formation of tell-tale trimers (Figure 5.12 on page 157)), could be purified was a positive early indication that the tail fibres may be expressing, assembling, and folding correctly, though this by no means guarantees it - probing the secondary structure with CD was therefore a logical step. Reproducible spectra were obtainable from entirely separate protein preparations which suggests the fold of the proteins is intact and correct, and did not vary from isolation to isolation. Moreover, analysis of the obtained spectra reveals that the putative PVC tail fibres have a secondary structure profile that is consistent with that seen in a myriad of other tail fibre-like structures (Section 5.2.3.4 on page 164). Not only this, but temperature ramping experiments which were consistent with the observation of limited unfolding in SDS-PAGE assays at elevated temperatures, also agrees with the known thermal stability of tail fibre proteins (Papanikolopoulou et al., 2004a, 2008a,b). One plausible explanation for why the PVCpnf tail fibre is much hardier than the PVClumt fibre could be by its increased length, and therefore repetitiveness. If the repeat motifs are indeed responsible for coordinating metal atoms, it stands to reason that pnf13 will coordinate more than lumt13, potentially conferring much increased stability.

In the dichroism temperature ramping studies, a shift of secondary structure was identifiable for both proteins at approximately 50-55 °C. For PVClumt13, this seemed to largely abolish the secondary structure of the protein, with the signal intensity lessening across the spectrum. For PVCpnf13, an additional secondary structure shift occurs at around 60 °C, whereby the spectra actually gains intensity in some areas. The significance of this secondary structure change is unknown, though it likely accounts for the difficulties encountered when attempting to have the pnf13 protein migrate in to SDS-PAGE gels. Two possible speculative explanations for the transitions may include, firstly, that the abrupt shifts in structure correspond to a rapid collapse of the protein. As they are putatively extended, fibrous proteins, this may indicate a collapse of the shaft like regions, and the protein essentially becoming more 'globular'. The second hypothesis may be that this is in some way analogous to the proposed conformational changes that occur in phage tail fibres to transduce the binding signal that then trigger contraction. The latter is probably less likely, and what is being seen is simply the denaturing of the proteins, but as this is the limit of the structural resolution of circular dichroism, a concrete answer will have to wait until their structures are fully resolved in future. The (albeit limited) attempts to identify crystallisation conditions (Figure 5.17 on page 169 and Table 5.6 on page 168), are promising however, and resolving the structures of these enigmatic proteins in future may reveal some novel structural patterns.

Structural homologies to phage proteins varied between different fibre proteins, and included hits to the fibritin 'whisker' proteins, and both the long and short fibres. Previous elucidations of structural domains for the long tail fibres have shown that they are comprised of multiple proteins - gp34/gp35/gp36/gp37, with gp34 as the proximal phage 'mounting hardware' and gp37 at the distal end for receptor recognition. The long tail fibres also require the presence of 2 additional chaperones, gp38 and gp57 in order to ensure correct structural formation (Granell et al., 2014b; Bartual et al., 2010). This suggests that the PVC tail fibres are more reminiscent of the short fibres than the long. This may also be consistent with any specificity the PVCs have, since the short fibres of phage are responsible for the 'fine' and irreversible binding of the phage to its target cell. The phage gp12 short fibre proteins are also known to require the presence of gp38 to

fold (Hashemolhosseini et al., 1996). If it is assumed that the PVC tail fibres are indeed forming correctly, then no such dependence on specific chaperones seems apparent. In the case of the short fibres specifically, Hashemolhosseini et al. (1996) showed that chaperons from phage λ (phage Tail fiber assembly proteins (pTfa) could 'step in' and ensure correct folding of T-even phage tail fibres. Therefore this is perhaps indicative of PVC tail fibres either being chaperone independent, or simply relying on other endogenous Enterobacterial chaperones such as GroEL, since they could be expressed outside of *Photorhabdus* with no additional proteins, and there are no obvious candidate chaperones within the PVC operons themselves.

All in all, this provides the first compelling experimental evidence that the putative tail fibres of PVCs do indeed elaborate proteins with many of the hallmarks of known fibre proteins.

5.3.2 The chimeric/split domain structure of PVC tail fibres

As explored in Section 5.2.1.1 on page 145, domain structure within the tail fibres appears split, making the PVC tail fibres reminiscent of chimeric phage-Adenovirus 'adapters'. Natural phage tail fibres are thought to display some mosaicism with the tail fibres of other, unrelated, phage, meaning that the fibres are essentially recombination hotspots. To date, this recombination appears limited to phage-to-phage recombination, and the exact mechanism is subject to debate (Sandmeler, 1994).

There is literature precedent for a number of artificial phage to eukaryote virus fibre fusions to date (Papanikolopoulou et al., 2004a,b; Krasnykh et al., 2001). Since they all share a similar intertwined "β-spiral" trimeric structure (despite not sharing much, if any, sequence similarity), they appear very amenable to these kind of modifications. Fibrous proteins in nature are well studied at this point, with familiar examples such as collagens and amyloid fibrils among the most intensely studied to date. These fibrous proteins are typically made of α-helical triple spirals and coiled coils (Beck and Brodsky, 1998). Fibres comprised of β-spirals are comparatively less well studied, but nevertheless widespread, with it being the dominant structure in the fibrous proteins of Adenoviruses, Reoviruses, and phage (Papanikolopoulou et al., 2004a; **?**). Thus, any additional examples of fibre

proteins which can be better understood structurally will offer insight in to this class of fibrous domain, particularly for putative chimeric proteins, as in artificial studies the exact fusion regions were not well resolved due to their flexibility.

It appears *Photorhabdus* has added to its 'biological box of tricks' by seemingly creating such a fusion naturally - or something resembling a fusion; it may be the case that this is an example of convergent evolution, aimed at exploiting the same eukaryotic cell surface markers. The engineering of these fibres, including their artificial fusion, has been an active area of research in order to derive new viral vectors with new tropisms for cell and gene therapies (Krasnykh et al., 2001; Li et al., 2006). A particularly interesting example of such a fusion actually appears as a favourable result in the HHPred data shown in Table 5.1 on page 146. The top hits for PVClumt13, and the 4th and 5th best hits for PVCpnf13 are all to the PDB ID 1V1H, which is an artificial fusion of the T4 fibritin 'foldon' domain, and the globular head of the human Adenovirus 2, created by Papanikolopoulou et al. (2004b). Yet more evidence for the tail fibres correct conformation can be taken from the paper, where the authors note that the chimeras they produced had the characteristic heat, SDS, protease resistance of 'normal' fibres.

5.3.3 Candidate binding targets for PVC tail fibres

The glycan array studies conducted with the PVClumt13 fibre, although still very preliminary, show promise for identifying candidate binding targets. Firstly, it was possible to detect signals, proving that the tail fibres have at least some lectin-like activity. This is consistent with existing studies of bacteriophage and Adenoviral tail fibre-like proteins, whereby glycan arrays have also been used to demonstrate binding (and this is by no means an exhaustive list) (Guardado-Calvo et al., 2010; Singh et al., 2015; Lenman et al., 2018; Nilsson et al., 2011).

The hits obtained from the glycan array appear to be relatively rich in galactose moieties. Phage are known to bind a wide selection of different sugars, of which galactose is one, and has been shown to be used by members of the *Siphoviridae* (Bertozzi Silva et al., 2016). Another promising indication is the presence of sialylated sugars in the list of results, as silayl sugars are known receptor binding moieties for a number of

eukaryotic viruses including Influenza, Rotaviruses, and of particular relevance, Adenoviruses. Numerous previous studies have demonstrated the binding relationships between a number of different Adenovirus species in many different organisms. For instance Singh et al. (2015) have shown that the fibres from the turkey Adenovirus 3 utilise sialylated cell surface markers as their recognition sites. Human Adenovirus 52 requires polysialic acid as its cell surface marker. Both of the sialylated sugars identified for PVClumt13 are adjoined to a galactose residue, and in the case of Adenovirus species D, they have been demonstrated to bind preferentially to this conformation (Burmeister et al., 2004). As a final example, the canine Adenovirus 2 fibres mimic SIGLEC (Sialic acid-binding immunoglobulin-type lectins) proteins in structure, despite sharing little to no sequence homology, underscoring the potential evolutionary drive to exploit these motifs (Rademacher et al., 2012).

It is thought that these viruses target these types of sugar due to their near ubiquitous appearance and high abundance on eukaryotic cell surfaces (Varki and Gagneux, 2012). This does raise questions around tissue specificity for the PVCs however. The variability seen in the tail fibres was initially hypothesised to potentially confer differential targeting against specific cell types, though the use of sialyl sugars would run counter to this. That said, this data only considers a single tail fibre, and it may be the case that other tail fibres are honed in different ways. There may be an evolutionary advantage for *Photorhabdus* to posses a variant of the PVCs which are capable of wide toxicity. The fact that sialic acids are extremely well conserved in the innate immune system across eukaryotic domains of life also goes a long way to explaining the capabilities of *Photorhabdus* virulence factors such as the PVCs in both insect and human infection, since they could plausible retain a mechanism of action with no 'additional evolution' required for functionality in a new host.

As mentioned, these results are still preliminary, and will need further replicates to be certain. Additionally, it may be informative to screen other glycan arrays with larger numbers of glycans present (the arrays used here were just over 100 glycans, but arrays are available with in excess of 400 such as that used in the paper by Guardado-Calvo et al. (2010)). This work also needs replicating for the PVCpnf13 tail fibre, and

ultimately in future it would be ideal to be able to screen the full library of tail fibres in this manner, to understand the 'spectrum' of binding for the natural 'library' of fibre proteins. Should the chimeric nature of the tail fibres be borne out in further structural study, there will also be additonal work needed to attempt to unpick whether the glycan specificies detected here reflect the behaviour of Adenovirus-like domains or phage-like domains, since there are some overlaps in the moieties both are able to bind. It seems likely however, particularly with the specificity for sialic acid bearing glycans, that this is the first experimental indication that the tail fibres incorporate a domain that mimics Adenoviral binding mechanisms, as this is one particular glycan type that does not appear to overlap with the binding of phage tails.

This early data also potentially correlates with some of the findings from the proteomic pulldown assays. There are a number of hits which are difficult to reconcile functionally, such as the appearance of the SERPIN protease inhibitor, Dermicidin antimicrobial peptide, and the FAM184A protein which has no known function at present. Lipocalin-1 is known to have an extremely wide binding range with a large number of ligands, likely due to its proposed role in olfaction, and the need to be able to detect many different compounds in the air (Flower, 1996). This wide range of binding activity may account for any association with the tail fibres. It may be the case that its enrichment in the presence of the fibres is as a result of 'promiscuous' binding, rather than any real meaningful or informative interaction.

However, among the putatively enriched proteins are a number of cell surface associated molecules. Most prominent among these are 4 protein components of the desmosome ('binding bodies'), a protein complex responsible for cell-cell adhesion, particularly in epithelia (Delva et al., 2009). There is always the possibility of identifying contaminating proteins in proteomic experiments with epithelial proteins being a particular risk (by far the most common of which are keratins). The enrichment of the desmosome proteins relative to controls (as a significant hit), and the fact that desmosome proteins are not listed as primary contaminating proteins in online databases such as the Repository of Adventitous Proteins[3] suggests that these may be meaningful results. Being a cell surface associated protein which is enriched in a cell lysate may be indicative of the kinds of

[3] https://www.thegpm.org/crap/index.html

targets which tail fibres will bind to, though the precise mechanism will require much more extensive elucidation. It is possible however that the use of cellular lysates has lead to a spurious result, as the desmosomes, *in vivo*, would likely not be readily accessible when tissues are joined together tightly. One possible explanation consistent with the glycan study however, is that both Desmocollin-1 and Desmoglein both contain N-linked N-Acetyl-Glucosamine and N-linked N-Acetyl-Galactosamine residues respectively (Ramachandran et al., 2006). One final caveat to mention is that the preliminary studies have only been conducted with an epithelial cell type (A549 lung epithelial carcinoma cells), and so it may be due to the cell type assayed that proteins abundant in epithelial cells are identified as significant. One might expect however, that there are numerous proteins enriched in epithelial cells which have not been detected by this assay, and so some specificity of interaction can potentially be inferred.

As discussed in the previous paragraphs, the glycan array results are relatively abundant in both of these moieties, particularly galactose. Similarly, lactotransferrin which was also detected in the pulldown assays contains 3 N-linked N-Acetyl-Galatosamine conjugated residues. Interestingly, lactotransferrin is known to promote the binding of species C Adenoviruses to epithelial cell surfaces, when the Coxsackie-Adenovirus Receptor is not readily accessible (Johansson et al., 2007). It may therefore be plausible that the putatively Adenoviral motifs of the PVC tail fibres are also capable of binding these proteins based on their glycan structure, and may even be exploiting the same mechanism as the Adenoviruses.

5.3.4 Summary and future work

As the first experimental studies of these proteins, there is a substantial amount that could be done in future. At the very least, it would be ideal to be able to clone, express, and resolve structures for tail fibres from all of the different PVC operons to better understand the diversity that is so apparent from the results in Chapter 4 on page 101. If these proteins are indeed natural chimeras between phage like tail fibres and Adenoviral fibres, the fact that there are many hypervariable tail fibres in the different operons suggests that the mechanism of 'fusing' or evolving new tail fibres is ongoing in various branches of

the *Photorhabdus* phylogenetic tree. Thus, not only would this represent the first known example of a natural chimerism for any particular tail fibre, but that there are many natural chimeras, another novel finding in itself.

There are a considerable number of proteins to test, and, as the most physiologically relevant cell/tissue types for PVC activity have yet to be determined, there are a great many more cell types to assay in order to bolster this early data, though some promising leads have been identified already. If the PVCs function on very specific cell types to control the virulence process in intricate ways, it may be the case that very particular cell types will return hits to important markers which will go undetected when testing on just a handful of cell types in the lab. Similarly, it would be ideal to test the tail fibres against much wider arrays of glycans, and of course, simply repeating the studies to be more confident in the binding profiles will be necessary. Based on the work here though, future studies should be made more feasible, as it has been demonstrated that tail fibres from PVCs are amenable to expression without the need for chaperones, refolding, or any intricate purification techniques. This should allow the purification of many more of the fibres with techniques ensuring adequate yields for further functional studies.

To begin to answer the tissue specificity and *in vivo*/physiological localisation question an alternative nanoparticle strategy that replaces the iron nanoparticles with gold is currently being developed. Gold nanoparticles exhibit an interesting property in solution, whereby their colour changes due to the phenomenon of having surface plasmons (oscillating electrons in a coherent but delocalised 'shell' at the interface between 2 differently charged materials) in a manner dependent on their size or concentration. Concentrated gold nanoparticle solutions are a pinkish-red colour, and the more diluted they are the 'blue-er' the solution becomes. This allows accurate concentration estimates based on standard curves from the gold particles as they concentrate/aggregate. This opens up the possibility of quantifying the amount of binding to cell surfaces for instance (in a manner essentially like a 'gold-based ELISA'). As gold is not found in natural tissues, the amount of gold present can be quantified very accurately through techniques such as plasma atomic emission spectroscopy. One particularly appealing application of this spectroscopy would be to study localisation in whole animals (insects, for instance) by

dissecting out different organs/regions and then quantifying the retained gold

In summary, the results presented in this chapter represent the very first foray in to the experimental investigation of the PVC tail fibres. The state of knowledge prior to this study was entirely based on bioinformatic inferences and the nature of the proteins was far from conclusively understood. To summarise, these results have shown experimental protein similarities to known tail fibre proteins in their trimerism and thermal stability, secondary structure profiles, lack of requirement for dedicated phage-like chaperons, and have begun to shed some light on the functional basis for the putative sequence similarity observed to Adenoviruses, with some compelling, albeit preliminary hits for binding targets.

Chapter 6

Synthetic & Natural PVC Operon Regulation

Cosmid recombineering attempts Gibson constructs reporter construction Microscopy

6.0.1 Population heterogeneity in PVC activity

Brighten any microscopy images that need it (almost all 24 hour for example)

Complete and run the imageprocessing script to deal with image size/brightness

6.0.2 Discussion

Remember to discuss the enlongation/filamenting of some bacteria - almost always green too) Example panels: TT01 LopT, 5 hours, bottom panel + 72 hour bottom panel TT01 Unit 1, 72 hours, top panel. Thai, 72 hours, bottom panel. Long green, but also long and non green

Part IV

Discussion

Chapter 7

Discussion

Major talking points for discussion:

Size of PVC lumen - effector folding Mosaicism of paralogs Population density/disparity of expression Tail fibre binding candidates/role of diversity etc

Applicability to biotech

Future directions: - tail fibre crystallisation - further binding studies - further in depth study of population heterogeneity/response to different stimuli

Bibliography

Abby, S. S., Cury, J., Guglielmini, J., Néron, B., Touchon, M., and Rocha, E. P. (2016). Identification of protein secretion systems in bacterial genomes. *Scientific Reports*, 6(October 2015):1–14.

Abdallah, A. M., Gey van Pittius, N. C., Champion, P. a. D., Cox, J., Luirink, J., Vandenbroucke-Grauls, C. M. J. E., Appelmelk, B. J., and Bitter, W. (2007). Type VII secretion–mycobacteria show the way. *Nature reviews. Microbiology*, 5(11):883–891.

Abu Hatab, M. a., Stuart, R. J., and Gaugler, R. (1998). Antibiotic resistance and protease production by Photorhabdus luminescens and Xenorhabdus poinarii bacteria symbiotic with entomopathogenic nematodes: Variation among species and strains. *Soil Biology and Biochemistry*, 30(14):1955–1961.

Abuladze, N. K., Gingery, M., Tsai, J., and Eiserling, F. a. (1994). Tail length determination in bacteriophage T4.

Ackermann, H.-W. (1998). Tailed Bacteriophages: The Order Caudovirales. *Advances in Virus Research*, 51:135–201.

Aizawa, S. (2001). Bacterial flagella and type III secretion systems. *FEMS Microbiology Letters*, 202(2):157–164.

Aksyuk, A. a., Leiman, P. G., Kurochkina, L. P., Shneider, M. M., Kostyuchenko, V. a., Mesyanzhinov, V. V., and Rossmann, M. G. (2009a). The tail sheath structure of bacteriophage T4: a molecular machine for infecting bacteria. *The EMBO journal*, 28(7):821–829.

Aksyuk, A. A., Leiman, P. G., Shneider, M. M., Mesyanzhinov, V. V., and Rossmann, M. G. (2009b). The Structure of Gene Product 6 of Bacteriophage T4, the Hinge-Pin of the Baseplate. *Structure*, 17(6):800–808.

Ali, S. A., Iwabuchi, N., Matsui, T., Hirota, K., Kidokoro, S. I., Arai, M., Kuwajima, K., Schuck, P., and Arisaka, F. (2003). Reversible and fast association equilibria of a molecular chaperone, gp57A, of bacteriophage T4. *Biophysical Journal*, 85(4):2606–2618.

Amos, L. A. and Klug, A. (1975). Three-dimensional Image Reconstructions of the Contractile Tail of T4 Bacteriophage. *Journal of Molecular Biology*, 99:51–73.

Arisaka, F., Kanamaru, S., Leiman, P., and Rossmann, M. G. (2003). The tail lysozyme complex of bacteriophage T4. *International Journal of Biochemistry and Cell Biology*, 35(1):16–21.

Ates, L. S., Houben, E. N. G., and Bitter, W. (2016). Type VII Secretion: A Highly Versatile Secretion System. *Virulence Mechanisms of Bacterial Pathogens, Fifth Edition*, pages 357–384.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. a., Tomita, M., Wanner, B. L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology*, 2:2006.0008.

Bager, R., Roghanian, M., Gerdes, K., and Clarke, D. J. (2016). Alarmone (p)ppGpp regulates the transition from pathogenicity to mutualism in Photorhabdus luminescens. *Molecular Microbiology*, 100(4):735–747.

Ballister, E. R., Lai, A. H., Zuckermann, R. N., Cheng, Y., and Mougous, J. D. (2008). In vitro self-assembly of tailorable nanotubes from a simple protein building block. *Proceedings of the National Academy of Sciences of the United States of America*, 105(10):3733–3738.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S. O. N., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. X. A., and Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477.

Barnhart, M. M. and Chapman, M. R. (2010). Curli biogenesis and function. *Annual review Microbiogy*, 60:131–147.

Bartual, S. G., Otero, J. M., Garcia-Doval, C., Llamas-Saiz, A. L., Kahn, R., Fox, G. C., and van Raaij, M. J. (2010). Structure of the bacteriophage T4 long tail fiber receptor-binding tip. *Proceedings of the National Academy of Sciences of the United States of America*, 107(47):20287–20292.

Basler, M. (2015a). Type VI secretion system: secretion by a contractile nanomachine. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1679):20150021.

Basler, M. (2015b). Type VI secretion system: secretion by a contractile nanomachine. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1679):20150021.

Basler, M., Pilhofer, M., Henderson, G. P., Jensen, G. J., and Mekalanos, J. J. (2012). Type VI secretion requires a dynamic contractile phage tail-like structure. *Nature*, 483(7388):182–6.

Baur, M. E., Kaya, H. K., and Strong, D. R. (1998). Foraging ants as scavengers on entomopathogenic nematode-killed insects. *Biological Control*, 12(3):231–236.

Beck, K. and Brodsky, B. (1998). Supercoiled protein motifs: The collagen triple-helix and the α- helical coiled coil. *Journal of Structural Biology*, 122(1-2):17–29.

Bertozzi Silva, J., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. *FEMS Microbiology Letters*, 363(4):1–11.

Birmingham, V. A. and Pattee, P. A. (1981). Genetic transformation in Staphylococcus aureus: isolation and characterization of a competence-conferring factor from bacteriophage 80 alpha lysates. *Journal of bacteriology*, 148(1):301–307.

Bleves, S., Viarre, V., Salacha, R., Michel, G. P., Filloux, A., and Voulhoux, R. (2010). Protein secretion systems in Pseudomonas aeruginosa: A wealth of pathogenic weapons. *International Journal of Medical Microbiology*, 300(8):534–543.

Böck, D., Medeiros, J. M., Tsao, H.-f., Penz, T., Weiss, G. L., Aistleitner, K., Horn, M., and Pilhofer, M. (2017). In situ architecture, function, and evolution of a contractile injection system. *Science*, 717(August):713–717.

Boemare, N. E., Akhurst, R. J., and R., M. G. (1993). Symbiotic Bacteria of Entomopathogenic Nematodes , and a Proposal To Transfer Xenorhabdus luminescens to a. *International Journal of Systematic Bacteriology*, 43(18):249–255.

Bolanos-Garcia, V. M. and Davies, O. R. (2006). Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilised metal affinity chromatography. *Biochimica et Biophysica Acta - General Subjects*, 1760(9):1304–1313.

Bönemann, G., Pietrosiuk, A., Diemand, A., Zentgraf, H., and Mogk, A. (2009). Remodelling of VipA/VipB tubules by ClpV-mediated threading is crucial for type VI protein secretion. *The EMBO journal*, 28(4):315–325.

Bönemann, G., Pietrosiuk, A., and Mogk, A. (2010). Tubules and donuts: A type VI secretion story: MicroReview. *Molecular Microbiology*, 76(April):815–821.

Bottai, D., Gröschel, M. I., and Brosch, R. (2017). *Type VII Secretion Systems in Gram-Positive Bacteria*, pages 235–265. Springer International Publishing, Cham.

Bourdin, G., Schmitt, B., Guy, L. M., Germond, J.-e., Zuber, S., Michot, L., and Reuteler, G. (2014). Amplification and Purification of T4-Like Escherichia coli Phages for Phage Therapy : from Laboratory to Pilot

Scale. *Applied and environmental microbiology*, 80(4):1469–1476.

Bowen, D. J. and Ensign, J. C. (1998). Purification and characterization of a high-molecular-weight insecticidal protein complex produced by the entomopathogenic bacterium Photorhabdus luminescens. *Applied and Environmental Microbiology*, 64(8):3029–3035.

Boyd, C. D., Jarrod Smith, T., El-Kirat-Chatel, S., Newell, P. D., Dufrêne, Y. F., and O'Toolea, G. A. (2014). Structural features of the Pseudomonas fluorescens biofilm adhesin LapA required for LapG-dependent cleavage, biofilm formation, and cell surface localization. *Journal of Bacteriology*, 196(15):2775–2788.

Brackmann, M., Nazarov, S., Wang, J., and Basler, M. (2017). Using Force to Punch Holes: Mechanics of Contractile Nanomachines. *Trends in Cell Biology*, 27(9):623–632.

Brillard, J., Duchaud, E., Boemare, N., Kunst, F., and Givaudan, A. (2002). The PhlA Hemolysin from the Entomopathogenic Bacterium Photorhabdus luminescens Belongs to the Two-Partner Secretion Family of Hemolysins. *JOURNAL OF BACTERIOLOGY*, 184(14):3871–3878.

Browning, C., Shneider, M. M., Bowman, V. D., Schwarzer, D., and Leiman, P. G. (2012). Phage pierces the host cell membrane with the iron-loaded spike. *Structure*, 20(2):326–339.

Brunet, Y. R., Espinosa, L., Harchouni, S., Mignot, T., and Cascales, E. (2013). Imaging Type VI Secretion-Mediated Bacterial Killing. *Cell Reports*, 3(1):36–41.

Buetow, L., Flatau, G., Chiu, K., Boquet, P., and Ghosh, P. (2001). Structure of the Rho- activating domain of Escherichia coli cytotoxic necrotizing factor 1. *Nature structural biology*, 8(7):584–588.

Bundock, P., den Dulk-Ras, A., Beijersbergen, A., and Hooykaas, P. J. (1995). Trans-kingdom T-DNA transfer from Agrobacterium tumefaciens to Saccharomyces cerevisiae. *Embo J*, 14(13):3206–3214.

Burmeister, W. P., Guilligay, D., and Cusack, S. (2004). Crystal Structure of Species D Adenovirus Fiber Knobs and Their Sialic Acid Binding Sites Crystal Structure of Species D Adenovirus Fiber Knobs and Their Sialic Acid Binding Sites. 78(14):7727–7736.

Cardarelli, L., Lam, R., Tuite, A., Baker, L. A., Sadowski, P. D., Radford, D. R., Rubinstein, J. L., Battaile, K. P., Chirgadze, N., Maxwell, K. L., and Davidson, A. R. (2010). The Crystal Structure of Bacteriophage HK97 gp6: Defining a Large Family of Head-Tail Connector Proteins. *Journal of Molecular Biology*, 395(4):754–768.

Carriço, J. A., Silva-Costa, C., Melo-Cristino, J., Pinto, F. R., De Lencastre, H., Almeida, J. S., and Ramirez, M. (2006). Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes. *Journal of Clinical Microbiology*, 44(7):2524–2532.

Cascales, E. and Cambillau, C. (2012). Structural biology of type VI secretion systems. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1592):1102–11.

Chang, J. H., Desveaux, D., and Creason, A. L. (2014). The ABCs and 123s of Bacterial Secretion Systems in Plant Pathogenesis. *Annual Review of Phytopathology*, 52(1):317–345.

Chang, Y., Rettberg, L. A., Ortega, D. R., and Jensen, G. J. (2017). In vivo structures of an intact type VI secretion system revealed by electron cryotomography. *EMBO reports*, 18(7):1090–1099.

Chaston, J. M., Suen, G., Tucker, S. L., Andersen, A. W., Bhasin, A., Bode, E., Bode, H. B., Brachmann, A. O., Cowles, C. E., Cowles, K. N., Darby, C., de Léon, L., Drace, K., Du, Z., Givaudan, A., Herbert Tran, E. E., Jewell, K. A., Knack, J. J., Krasomil-Osterfeld, K. C., Kukor, R., Lanois, A., Latreille, P., Leimgruber, N. K., Lipke, C. M., Liu, R., Lu, X., Martens, E. C., Marri, P. R., Médigue, C., Menard, M. L., Miller, N. M., Morales-Soto, N., Norton, S., Ogier, J. C., Orchard, S. S., Park, D., Park, Y., Qurollo, B. A., Sugar, D. R., Richards, G. R., Rouy, Z., Slominski, B., Slominski, K., Snyder, H., Tjaden, B. C., van der Hoeven, R., Welch, R. D., Wheeler, C., Xiang, B., Barbazuk, B., Gaudriault, S., Goodner, B., Slater, S. C., Forst, S., Goldman, B. S., and Goodrich-Blair, H. (2011). The Entomopathogenic Bacterial Endosymbionts Xenorhabdus and Photorhabdus: Convergent Lifestyles from Divergent Genomes. *PLoS ONE*, 6(11).

Chatterjee, S. and Rothenberg, E. (2012). Interaction of bacteriophage λ with Its E. coli receptor, LamB. *Viruses*, 4(11):3162–3178.

Chattopadhyay, P., Chatterjee, S., Gorthi, S., and Sen, S. K. (2012). Exploring Agricultural Potentiality of Serratia entomophila AB 2 : Dual Property of Biopesticide and Biofertilizer. 2(1):1–12.

Christie, P. J., Atmakuri, K., Krishnamoorthy, V., Jakubowski, S., and Cascales, E. (2005). Biogenesis, Architecture, and Function of Bacterial Type IV Secretion Systems. *Annual review Microbiogy*, 59(29).

Cianfanelli, F. R., Alcoforado Diniz, J., Guo, M., De Cesare, V., Trost, M., and Coulthurst, S. J. (2016). VgrG and PAAR Proteins Define Distinct Versions of a Functional Type VI Secretion System. *PLoS Pathogens*, 12(6):1–27.

Ciche, T. A. and Ensign, J. C. (2003). For the insect pathogen Photorhabdus luminescens, which end of a nematode is out? *Applied and Environmental Microbiology*, 69(4):1890–1897.

Clarke, D. J. and Joyce, S. A. (2008). Photorhabdus: Shedding Light on Symbioses. *Microbiology Today*, pages 1–4.

Clemens, D. L., Ge, P., Horwitz, M. A., Zhou, Z. H., Clemens, D. L., Ge, P., Lee, B.-y., Horwitz, M. A.,

and Zhou, Z. H. (2015). Atomic Structure of T6SS Reveals Interlaced Array Essential to Function Article Atomic Structure of T6SS Reveals Interlaced Array Essential to Function. *Cell*, 160(5):940–951.

Clokie, M. R., Millard, A. D., Letarov, A. V., and Heaphy, S. (2011). Phages in nature. *Bacteriophage*, 1(1):31–45.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and De Hoon, M. J. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Coetzee, H. L., De Klerk, H. C., Coetzee, J. N., and Smit, J. A. (1968). Bacteriophage-tail-like particles associated with intra-species killing of Proteus vulgaris. *The Journal of general virology*, 2(1):29–36.

Compton, L. a. and Johnson, W. C. (1986). Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Analytical biochemistry*, 155(1):155–167.

Connor, T. R., Loman, N. J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., Bull, M. J., Richardson, E., Ismail, M., Elwood-Thompson, S., Kitchen, C., Guest, M., Bakke, M., Sheppard, S. K., and Pallen, M. J. (2016). CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *bioRxiv*, (July 2016):064451.

Costa, T. R. D., Felisberto-Rodrigues, C., Meir, A., Prevost, M. S., Redzej, A., Trokter, M., and Waksman, G. (2015). Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nature Reviews Microbiology*, 13(6):343–359.

Daborn, P. J., Waterfield, N., Blight, M. A., and Ffrench-constant, R. H. (2001). Measuring Virulence Factor Expression by the Pathogenic Bacterium Photorhabdus luminescens in Culture and during Insect Infection Measuring Virulence Factor Expression by the Pathogenic Bacterium Photorhabdus luminescens in Culture and during Insect Infec. *Journal of Bacteriology*, 183(20):5834–5839.

Dalbey, R. E. and Kuhn, A. (2012). Protein Traffic in Gram-negative bacteria - how exported and secreted proteins find their way. *FEMS Microbiology Reviews*, 36(6):1023–1045.

Darling, A. E., Mau, B., and Perna, N. T. (2010). Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6).

Datsenko, K. a. and Wanner, B. L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12):6640–5.

Delepelaire, P. (2004). Type I secretion in gram-negative bacteria. *Biochimica et Biophysica Acta - Molecular Cell*

Research, 1694(1-3 SPEC.ISS.):149–161.

Delva, E., Tucker, D. K., and Kowalczyk, A. P. (2009). The desmosome. *Cold Spring Harb.Perspect.Biol*, 1(1943-0264 (Electronic)):a002543.

Desmyter, A., Spinelli, S., Roussel, A., and Cambillau, C. (2015). Camelid nanobodies: Killing two birds with one stone. *Current Opinion in Structural Biology*, 32:1–8.

Desvaux, M., Hébraud, M., Talon, R., and Henderson, I. R. (2009). Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends in Microbiology*, 17(4):139–145.

D'Hérelle, F. (1917). An invisible microbe that is antagonistic to the dysentery bacillus. *Comptesrendus Acad. Sciences,*, 165:373–375.

Dong, A., Xu, X., Edwards, A. M., Chang, C., Chruszcz, M., Cuff, M., Cymborowski, M., Di Leo, R., Egorova, O., Evdokimova, E., Filippova, E., Gu, J., Guthrie, J., Ignatchenko, A., Joachimiak, A., Klostermann, N., Kim, Y., Korniyenko, Y., Minor, W., Que, Q., Savchenko, A., Skarina, T., Tan, K., Yakunin, A., Yee, A., Yim, V., Zhang, R., Zheng, H., Akutsu, M., Arrowsmith, C., Avvakumov, G. V., Bochkarev, A., Dahlgren, L.-G., Dhe-Paganon, S., Dimov, S., Dombrovski, L., Finerty, P., Flodin, S., Flores, A., Gräslund, S., Hammerström, M., Herman, M. D., Hong, B.-S., Hui, R., Johansson, I., Liu, Y., Nilsson, M., Nedyalkova, L., Nordlund, P., Nyman, T., Min, J., Ouyang, H., Park, H.-w., Qi, C., Rabeh, W., Shen, L., Shen, Y., Sukumard, D., Tempel, W., Tong, Y., Tresagues, L., Vedadi, M., Walker, J. R., Weigelt, J., Welin, M., Wu, H., Xiao, T., Zeng, H., and Zhu, H. (2007). In situ proteolysis for protein crystallization and structure determination. *Nature Methods*, 4(12):1019–1021.

Douzi, B., Filloux, A., and Voulhoux, R. (2012). On the path to uncover the bacterial type II secretion system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1592):1059–1072.

Downing, K. J., Mischenko, V. V., Shleeva, M. O., Young, D. I., Young, M., Kaprelyants, A. S., Apt, A. S., and Mizrahi, V. (2005). Mutants of Mycobacterium tuberculosis Lacking Three of the Five rpf-Like Genes Are Defective for Growth In Vivo and for Resuscitation In Vitro. *Society*, 73(5):3038–3043.

Duchaud, E., Rusniok, C., Frangeul, L., Buchrieser, C., Givaudan, A., Taourit, S., Bocs, S., Boursaux-Eude, C., Chandler, M., Charles, J., Dassa, E., Derose, R., Derzelle, S., Freyssinet, G., Gaudriault, S., Médigue, C., Lanois, A., Powell, K., Siguier, P., Vincent, R., Wingate, V., Zouine, M., Glaser, P., Boemare, N., Danchin, A., and Kunst, F. (2003). The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. *Nature Biotechnology*, 21(11):1307–13.

Dunwell, J. M., Purvis, A., and Khuri, S. (2004). Cupins: The most functionally diverse protein superfamily?

Phytochemistry, 65(1):7–17.

Durand, E., Nguyen, V. S., Zoued, A., Logger, L., Péhau-Arnaudet, G., Aschtgen, M.-S., Spinelli, S., Desmyter, A., Bardiaux, B., Dujeancourt, A., Roussel, A., Cambillau, C., Cascales, E., and Fronzes, R. (2015). Biogenesis and structure of a type VI secretion membrane core complex. *Nature*, pages 25–28.

Durham, S. (2001). No Title.

English, G., Byron, O., Cianfanelli, F., Prescott, A., and Coulthurst, S. (2014). Biochemical analysis of TssK, a core component of the bacterial TypeVI secretion system, reveals distinct oligomeric states of TssK and identifies a TssKTssFG subcomplex. *Biochemical Journal*, 461(2):291–304.

English, G., Trunk, K., Rao, V. A., Srikannathasan, V., Hunter, W. N., and Coulthurst, S. J. (2012). New secreted toxins and immunity proteins encoded within the type VI secretion system gene cluster of Serratia marcescens. *Molecular Microbiology*, 86(4):921–936.

Erzberger, J. P. and Berger, J. M. (2006). Evolutionary Relationships and Structural Mechanisms of Aaa+ Proteins. *Annual Review of Biophysics and Biomolecular Structure*, 35(1):93–114.

Farmer, J. J., Jorgensen, J. H., Grimont, P. A., Akhurst, R. J., Poinar, G. O., Ageron, E., Pierce, G. V., Smith, J. A., Carter, G. P., and Wilson, K. L. (1989). Xenorhabdus luminescens (DNA hybridization group 5) from human clinical specimens. *Journal of clinical microbiology*, 27(7):1594–600.

Felisberto-Rodrigues, C., Durand, E., Aschtgen, M. S., Blangy, S., Ortiz-Lombardia, M., Douzi, B., Cambillau, C., and Cascales, E. (2011). Towards a structural comprehension of bacterial type vi secretion systems: Characterization of the TssJ-TssM complex of an escherichia coli pathovar. *PLoS Pathogens*, 7(11):1–11.

Ffrench-Constant, R. H. and Dowling, A. J. (2014). Photorhabdus Toxins. In *Advances in Insect Physiology*, volume 47, pages 343–388. Elsevier Ltd., 1 edition.

Ffrench-Constant, R. H., Waterfield, N., Daborn, P., Joyce, S., Bennett, H., Au, C., Dowling, A., Boundy, S., Reynolds, S., and Clarke, D. (2003). Photorhabdus: Towards a functional genomic analysis of a symbiont and pathogen. *FEMS Microbiology Reviews*, 26:433–456.

Flower, D. R. (1996). The lipocalin protein family: structure and function. *The Biochemical journal*, 318 (Pt 1:1–14.

Fokine, A., Chipman, P. R., Leiman, P. G., Mesyanzhinov, V. V., Rao, V. B., and Rossmann, M. G. (2004). Molecular architecture of the prolate head of bacteriophage T4. *Proceedings of the National Academy of Sciences*, 101(16):6003–6008.

Fokine, A., Zhang, Z., Kanamaru, S., Bowman, V. D., Aksyuk, A. A., and Arisaka, F. (2013). The Molecular Architecture of the Bacteriophage T4 Neck. *Journal of Molecular Biology*, 425(10):1731–1744.

Forster, a., Planamente, S., Manoli, E., Lossi, N. S., Freemont, P. S., and Filloux, a. (2014). Coevolution of the ATPase ClpV, the Sheath Proteins TssB and TssC, and the Accessory Protein TagJ/HsiE1 Distinguishes Type VI Secretion Classes. *Journal of Biological Chemistry*, 289(47):33032–33043.

Frickey, T. and Lupas, A. N. (2004). Phylogenetic analysis of AAA proteins. *Journal of Structural Biology*, 146(1-2):2–10.

Gaggar, A., Shayakhmetov, D. M., and Lieber, A. (2003). CD46 is a cellular receptor for group B adenoviruses. *Nature Medicine*, 9(11):1408–1412.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., and Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*, 28(20):2678–2679.

Ge, P., Scholl, D., Leiman, P. G., Yu, X., Miller, J. F., and Zhou, Z. H. (2015a). Atomic structures of a bactericidal contractile nanotube in its pre- and postcontraction states. *Nat Struct Mol Biol*, 22(5):377–382.

Ge, P., Scholl, D., Leiman, P. G., Yu, X., Miller, J. F., and Zhou, Z. H. (2015b). Atomic structures of a bactericidal contractile nanotube in its pre- and postcontraction states. *Nat Struct Mol Biol*, 22(5):377–382.

Gerlach, R. G. and Hensel, M. (2007). Protein secretion systems and adhesins: The molecular armory of Gram-negative pathogens. *International Journal of Medical Microbiology*, 297(6):401–415.

Gerrard, J. G., Mcnevin, S., Alfredson, D., Forgan-smith, R., and Fraser, N. (2003). Photorhabdus Species: Bioluminescent Bacteria as Emerging Human Pathogens? *Emerging infectious diseases*, 9(2):251–254.

Ghequire, M. G. K. and De Mot, R. (2015). The Tailocin Tale: Peeling off Phage Tails. *Trends in Microbiology*, 23(10):587–590.

Gibbs, K. A., Urbanowski, M. L., and Greenberg, E. P. (2008). Genetic Determinants of Self Identity and Social Recognition in Bacteria. *Science*, 321(5886):256–259.

Glover, D. M. (1995). *DNA Cloning: A practical approach*. IRL Press, Oxford.

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., and Ferrin, T. E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein science : a publication of the Protein Society*, 27(1):14–25.

Goldberg, E., Tsugita, A., Matsui, T., Griniuviene, B., Tanaka, N., and Arisaka, F. (1997). Isolation and Characterization of a Molecular Chaperone , gp57A , of Bacteriophage T4. *Journal of Bacteriology*, 179(6):1846–1851.

Goulet, V., Britigan, B., Nakayama, K., and Grenier, D. (2004). Cleavage of human transferrin by Porphyromonas gingivalis gingipains promotes growth and formation of hydroxyl radicals. *Infection and Immunity*, 72(8):4351–4356.

Goyal, P., Krasteva, P. V., Van Gerven, N., Gubellini, F., Van Den Broeck, I., Troupiotis-Tsaïlaki, A., Jonckheere, W., Péhau-Arnaudet, G., Pinkner, J. S., Chapman, M. R., Hultgren, S. J., Howorka, S., Fronzes, R., and Remaut, H. (2014). Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature*, 516(7530):250–253.

Granell, M., Namura, M., Alvira, S., Garcia-Doval, C., Singh, A. K., Gutsche, I., Van Raaij, M. J., and Kanamaru, S. (2014a). Crystallization of the carboxy-terminal region of the bacteriophage T4 proximal long tail fibre protein gp34. *Acta Crystallographica Section F:Structural Biology Communications*, 70(7):970–975.

Granell, M., Namura, M., Alvira, S., Garcia-Doval, C., Singh, A. K., Gutsche, I., Van Raaij, M. J., and Kanamaru, S. (2014b). Crystallization of the carboxy-terminal region of the bacteriophage T4 proximal long tail fibre protein gp34. *Acta Crystallographica Section F:Structural Biology Communications*, 70(7):970–975.

Green, E. R. and Mecsas, J. (2015). Bacterial Secretion Systems An overview. *American society for Microbiology*, 4(1):1–32.

Guardado-Calvo, P., Muñoz, E. M., Llamas-Saiz, A. L., Fox, G. C., Kahn, R., Curiel, D. T., Glasgow, J. N., and van Raaij, M. J. (2010). Crystallographic Structure of Porcine Adenovirus Type 4 Fiber Head and Galectin Domains. *Journal of Virology*, 84(20):10558–10568.

Guzman, L. L.-m. M., Belin, D., Carson, M. J., Beckwith, J., LUZ-MARIA GUZMAN MICHAEL J. CARSON, AND JON BECKWITH, D. B., and LUZ-MARIA GUZMAN MICHAEL J. CARSON, AND JON BECKWITH, D. B. (1995). Tight Regulation, Modulation, and High-Level Expression by Vectors Containing the Arabinose PBAD Promoter. *Journal of Bacteriology*, 177(14):4121–4130.

Hachani, A., Allsopp, L. P., Oduko, Y., and Filloux, A. (2014). The VgrG proteins are "à la carte" delivery systems for bacterial type VI effectors. *Journal of Biological Chemistry*, 289(25):17872–17884.

Hachani, A., Lossi, N. S., Hamilton, A., Jones, C., Bleves, S., Albesa-Jové, D., and Filloux, A. (2011). Type VI secretion system in Pseudomonas aeruginosa: Secretion and multimerization of VgrG proteins. *Journal of Biological Chemistry*, 286(14):12317–12327.

Hadfield, M. G. (2011). Biofilms and Marine Invertebrate Larvae: What Bacteria Produce That Larvae Use to Choose Settlement Sites. *Annual Review of Marine Science*, 3(1):453–470.

Hainfeld, J. F., Liu, W., Halsey, C. M., Freimuth, P., and Powell, R. D. (1999). Ni-NTA-gold clusters target His-tagged proteins. *Journal of Structural Biology*, 127(2):185–198.

Hanson, P. I. and Whiteheart, S. W. (2005). AAA+ proteins: have engine, will work. *Nature Reviews Molecular Cell Biology*, 6(7):519–529.

Hashemolhosseini, S., Stierhof, Y. D., Hindennach, I., and Henning, U. (1996). Characterization of the helper proteins for the assembly of tail fibers of coliphages T4 and λ. *Journal of Bacteriology*, 178(21):6258–6265.

Hazan, R. and Engelberg-Kulka, H. (2004). Escherichia coli mazEF-mediated cell death as a defense mechanism that inhibits the spread of phage P1. *Molecular Genetics and Genomics*, 272(2):227–234.

Hedges, L. M., Brownlie, J. C., O'Neill, S. L., and Johnson, K. N. (2008). Wolbachia and virus protection in insects. *Science*, 322(5902):702.

Heger, A. and Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Structure, Function and Genetics*, 41(2):224–237.

Heo, Y. J., Chung, I. Y., Choi, K. B., and Cho, Y. H. (2007). R-type pyocin is required for competitive growth advantage between Pseudomonas aeruginosa strains. *Journal of Microbiology and Biotechnology*, 17(1):180–185.

Heymann, J. B., Bartho, J. D., Rybakova, D., Venugopal, H. P., Winkler, D. C., Sen, A., Hurst, M. R. H., and Mitra, A. K. (2013). Three-dimensional structure of the toxin-delivery particle antifeeding prophage of serratia entomophila. *Journal of Biological Chemistry*, 288(35):25276–25284.

Hood, R. D., Singh, P., Hsu, F., Güvener, T., Carl, M. A., Trinidad, R. S., Silverman, J. M., Ohlson, B. B., Hicks, K. G., Rachael, L., Li, M., Schwarz, S., Wang, W. Y., Merz, A. J., David, R., and Mougous, J. D. (2010). A Type VI Secretion System of Pseudomonas aeruginosa Targets a Toxin to Bacteria. *Cell*, 7(1):25–37.

Hu, K. and Webster, J. M. (2000). Antibiotic production in relation to bacterial growth and nematode development in Photorhabdusâ€“Heterorhabditis infected Galleria mellonella larvae. *FEMS Microbiology Letters*, 189(2):219–223.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6):1635–1638.

Hurst, M. R., Becher, S. A., and O'Callaghan, M. (2011). Nucleotide sequence of the Serratia entomophila plasmid pADAP and the Serratia proteamaculans pU143 plasmid virulence associated region. *Plasmid*, 65(1):32–41.

Hurst, M. R. H., Beard, S. S., Jackson, T. a., and Jones, S. M. (2007a). Isolation and characterization of the Serratia entomophila antifeeding prophage. *FEMS Microbiology Letters*, 270:42–48.

Hurst, M. R. H., Beard, S. S., Jackson, T. a., and Jones, S. M. (2007b). Isolation and characterization of the Serratia entomophila antifeeding prophage. *FEMS Microbiology Letters*, 270:42–48.

Hurst, M. R. H., Glare, T. R., and Jackson, T. A. (2004). Cloning Serratia entomophila antifeeding genesa putative defective prophage active against the grass grub Costelytra zealandica. *Journal of Bacteriology*, 186(15):5116–5128.

Hurst, M. R. H., Glare, T. R., Jackson, T. a., and Ronson, C. W. (2000). Plasmid-located pathogenicity determinants of Serratia entomophila, the causal agent of amber disease of grass grub, show similarity to the insecticidal toxins of Photorhabdus luminescens. *Journal of Bacteriology*, 182(18):5127–5138.

ichi Ishii, S., Nishi, Y., and Egami, F. (1965). The fine structure of a pyocin. *Journal of Molecular Biology*, 13(2):IN5–IN12.

Isao, M. O. F. L. D. L. T. (1990). Department of Biology, College of Arts and Sciences, The Universi˜ of Tokyo, Meguro-ku, Tokyo 153, Japan. *Sciences-New York*, 26:1–18.

Iyer, L. M., Leipe, D. D., Koonin, E. V., and Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *Journal of Structural Biology*, 146(1-2):11–31.

Jacob, F. (1954). BIOSYNTHESE INDUITE ET MODE DACTION DUNE PYOCINE, ANTIBIOTIQUE DE PSEUDOMONAS-PYOCYANEA. In *ANNALES DE L INSTITUT PASTEUR*, volume 86, pages 149–160. MASSON EDITEUR 120 BLVD SAINT-GERMAIN, 75280 PARIS 06, FRANCE.

Jobichen, C., Chakraborty, S., Li, M., Zheng, J., Joseph, L., Mok, Y. K., Leung, Y. K., and Sivaraman, J. (2010). Structural basis for the secretion of evpc: A key type vi secretion system protein from edwardsiella tarda. *PLoS ONE*, 5(9):1–10.

Johansson, C., Jonsson, M., Marttila, M., Persson, D., Fan, X.-L., Skog, J., Frangsmyr, L., Wadell, G., and Arnberg, N. (2007). Adenoviruses Use Lactoferrin as a Bridge for CAR-Independent Binding to and Infection of Epithelial Cells. *Journal of Virology*, 81(2):954–963.

Joyce, S. A., Brachmann, A. O., Glazer, I., Lango, L., Schwär, G., Clarke, D. J., and Bode, H. B. (2008). Bacterial

biosynthesis of a multipotent stilbene. *Angewandte Chemie (International ed. in English)*, 47(10):1942–5.

Kabsch, W. and Sander, C. (1983). Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. 22:2577–2637.

Kageyama, M. (1975). Bacteriocins and bacteriophages in Pseudomonas aeruginosa. *Microbial drug resistance*, pages 291–305.

Kanamaru, S., Leiman, P. G., and Kostyuchenko, V. A. (2002). Structure of the cell-puncturing device of bacteriophage T4. *Nature Letters*, 2428(2001):553–557.

Katsura, I. (1987). Determination of bacteriophage lambda tail length by a protein ruler. *Nature*, 327:73–75.

Katsura, I. and Hendrix, R. W. (1984). Length determination in bacteriophage lambda tails. *Cell*, 39(3 PART 2):691–698.

Kelly, L., Mezulis, S., Yates, C., Wass, M., and Sternberg, M. (2015). The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols*, 10(6):845–858.

Knott, G. and Genoud, C. (2013). Is EM dead? *Journal of Cell Science*, 126:4545–4552.

Korotkov, K. V., Sandkvist, M., and Hol, W. G. (2012). The type II secretion system: Biogenesis, molecular architecture and mechanism. *Nature Reviews Microbiology*, 10(5):336–351.

Kostyuchenko, V. A., Chipman, P. R., Leiman, P. G., Arisaka, F., Mesyanzhinov, V. V., and Rossmann, M. G. (2005). The tail structure of bacteriophage T4 and its mechanism of contraction. *Nature Structural & Molecular Biology*, 12(9):810–813.

Kostyuchenko, V. a., Leiman, P. G., Chipman, P. R., Kanamaru, S., van Raaij, M. J., Arisaka, F., Mesyanzhinov, V. V., and Rossmann, M. G. (2003). Three-dimensional structure of bacteriophage T4 baseplate. *Nature structural biology*, 10(9):688–693.

Krasnykh, V., Belousova, N., Korokhov, N., Mikheeva, G., and Curiel, D. T. (2001). Genetic Targeting of an Adenovirus Vector via Replacement of the Fiber Protein with the Phage T4 Fibritin. *Journal of Virology*, 75(9):4176–4183.

Kube, S., Kapitein, N., Zimniak, T., Herzog, F., Mogk, A., and Wendler, P. (2014a). Structure of the VipA / B Type VI Secretion Complex Suggests a Contraction-State-Specific Recycling Mechanism. *Cell Reports*, 8(1):20–30.

Kube, S., Kapitein, N., Zimniak, T., Herzog, F., Mogk, A., and Wendler, P. (2014b). Structure of the VipA/B type VI secretion complex suggests a contraction-state-specific recycling mechanism. *Cell Reports*, 8(1):20–30.

Kube, S. and Wendler, P. (2015a). Structural comparison of contractile nanomachines. *AIMS Biophysics*, 2(2):88–115.

Kube, S. and Wendler, P. (2015b). Structural comparison of contractile nanomachines. *AIMS Biophysics*, 2(2):88–115.

Kudryashev, M., Wang, R.-R., Brackmann, M., Scherer, S., Maier, T., Baker, D., DiMaio, F., Stahlberg, H., Egelman, E., and Basler, M. (2015). Structure of the Type VI Secretion System Contractile Sheath. *Cell*, 160(5):952–962.

Kühlbrandt, W. (2014). The Resolution Revolution. *Science*, 343:1443–1444.

Lan, M., Klose, T., Plevka, P., Aksyuk, A., Zhang, X., Arisaka, F., and Rossmann, M. G. (2014). Structure of the 3.3 MDa, in vitro assembled, hubless bacteriophage T4 baseplate. *Journal of Structural Biology*, 187(2):95–102.

Landraud, L., Pulcini, C., Gounon, P., Flatau, G., Boquet, P., and Lemichez, E. (2004). E. coli CNF1 toxin: A two-in-one system for host-cell invasion. *International Journal of Medical Microbiology*, 293(7-8):513–518.

Lane, C. E. (2007). Bacterial Endosymbionts: Genome Reduction in a Hot Spot. *Current Biology*, 17(13):510–512.

Lango, L. and Clarke, D. J. (2010). A metabolic switch is involved in lifestyle decisions in Photorhabdus luminescens. *Molecular Microbiology*, 77(6):1394–1405.

Lasica, A. M., Ksiazek, M., Madej, M., and Potempa, J. (2017). The Type IX Secretion System (T9SS): Highlights and Recent Insights into Its Structure and Function. *Frontiers in Cellular and Infection Microbiology*, 7(May).

Le, S., He, X., Tan, Y., Huang, G., Zhang, L., Lux, R., Shi, W., and Hu, F. (2013). Mapping the Tail Fiber as the Receptor Binding Protein Responsible for Differential Host Specificity of Pseudomonas aeruginosa Bacteriophages PaP1 and JG004. *PLoS ONE*, 8(7):1–8.

Lee, E.-C., Yu, D., Martinez de Velasco, J., Tessarollo, L., Swing, D. A., Court, D. L., Jenkins, N. A., and Copeland, N. G. (2001). A Highly Efficient Escherichia coli-Based Chromosome Engineering System Adapted for Recombinogenic Targeting and Subcloning of BAC DNA. *Genomics*, 73(1):56–65.

Lee, F. K. N., Dudas, K. C., Hanson, J. a., Bud, M., Loverde, P. T., Apicella, M. a., and Verde, P. T. L. O. (1999).

The R-Type Pyocin of Pseudomonas aeruginosa C Is a Bacteriophage Tail-Like Particle That Contains Single-Stranded DNA The R-Type Pyocin of Pseudomonas aeruginosa C Is a Bacteriophage Tail-Like Particle That Contains Single-Stranded DNA. 67(2):717–725.

Lees, J. G., Miles, A. J., Wien, F., and Wallace, B. A. (2006). A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics*, 22(16):1955–1962.

Leiman, P. G., Arisaka, F., van Raaij, M. J., Kostyuchenko, V. a., Aksyuk, A. a., Kanamaru, S., and Rossmann, M. G. (2010). Morphogenesis of the T4 tail and tail fibers. *Virology journal*, 7(1):355.

Leiman, P. G., Basler, M., Ramagopal, U. a., Bonanno, J. B., Sauder, J. M., Pukatzki, S., Burley, S. K., Almo, S. C., and Mekalanos, J. J. (2009). Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11):4154–9.

Leiman, P. G., Chipman, P. R., Kostyuchenko, V. A., Mesyanzhinov, V. V., and Rossmann, M. G. (2004). Three-dimensional rearrangement of proteins in the tail of bacteriophage T4 on infection of its host. *Cell*, 118(4):419–429.

Lemaitre, B. and Hoffmann, J. (2007). The Host Defense of Drosophila melanogaster. *Annual Review of Immunology*, 25(1):697–743.

Lenman, A., Liaci, A. M., Liu, Y., Frängsmyr, L., Frank, M., Blaum, B. S., and Chai, W. (2018). Polysialic acid is a cellular receptor for human adenovirus 52. *Pnas*, 115(18).

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, J., Lad, S., Yang, G., Luo, Y., Iacobelli-Martinez, M., Primus, F. J., Reisfeld, R. a., and Li, E. (2006). Adenovirus fiber shaft contains a trimerization element that supports peptide fusion for targeted gene delivery. *Journal of virology*, 80(24):12324–31.

Lin, W., Fullner, K. J., Clayton, R., Sexton, J. a., Rogers, M. B., Calia, K. E., Calderwood, S. B., Fraser, C., and Mekalanos, J. J. (1999). Identification of a vibrio cholerae RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. *Proceedings of the National Academy of Sciences of the United States of America*, 96(3):1071–1076.

Liu, H., Wu, L., and Zhou, Z. H. (2011). Model of the trimeric fiber and its interactions with the pentameric penton base of human adenovirus by cryo-electron microscopy. *Journal of Molecular Biology*, 406(5):764–774.

Lobley, A., Whitmore, L., and Wallace, B. A. (2002). DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. *Bioinformatics*, 18(1):211–212.

Lu, C., Turley, S., Marionni, S. T., Park, Y. J., Lee, K. K., Patrick, M., Shah, R., Sandkvist, M., Bush, M. F., and Hol, W. G. (2013). Hexamers of the type II secretion ATPase GspE from Vibrio cholerae with Increased ATPase activity. *Structure*, 21(9):1707–1717.

Ma, J., Sun, M., Dong, W., Pan, Z., Lu, C., and Yao, H. (2017). PAAR-Rhs proteins harbor various C-terminal toxins to diversify the antibacterial pathways of type VI secretion systems. *Environmental Microbiology*, 19(1):345–360.

Ma, L. S., Lin, J. S., and Lai, E. M. (2009). An IcmF family protein, ImpLM, is an integral inner membrane protein interacting with ImpKL, and its Walker a motif is required for type VI secretion system-mediated Hcp secretion in Agrobacterium tumefaciens. *Journal of Bacteriology*, 191(13):4316–4329.

Manavalan, P. and Johnson, W. C. (1987). Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Analytical biochemistry*, 167(1):76–85.

Mao, D., Wachter, E., and Wallace, B. A. (1982). Folding of the mitochondrial proton adenosinetriphosphatase proteolipid channel in phospholipid vesicles. *Biochemistry*, 21(20):4960–4968.

Mavridis, L. and Janes, R. W. (2017). PDB2CD: A web-based application for the generation of circular dichroism spectra from protein atomic coordinates. *Bioinformatics*, 33(1):56–63.

Meusch, D., Gatsogiannis, C., Efremov, R. G., Lang, A. E., Hofnagel, O., Vetter, I. R., Aktories, K., and Raunser, S. (2014). Mechanism of Tc toxin action revealed in molecular detail. *Nature*, 508(1):61–65.

Michel-Briand, Y. and Baysse, C. (2002). The pyocins of Pseudomonas aeruginosa. *Biochimie*, 84:499–510.

Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Ru, W. (2003). Bacteriophage T4 Genome. *Microbiology and Molecular Biology Reviews*, 67(1):86–156.

Mirarab, S. and Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52.

Moody, M. F. (1973). Sheath of bacteriophage T4. III. Contraction mechanism deduced from partially contracted sheaths. *Journal of Molecular Biology*, 80(4):613–635.

Morona, R., Klose, M., and Henning, U. (1984). Escherichia K12 outer membrane protein (ompA): Analysis of mutant genes expressing altered proteins. *Journal of Bacteriology*, 159(2):570–578.

Morse, S. A., Vaughan, P., Johnson, D., and Iglewski, B. H. (1976). Inhibition of Neisseria gonorrhoeae by a bacteriocin from Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy*, 10(2):354–362.

Mougous, J. D., Cuff, M. E., Raunser, S., Shen, A., Zhou, M., Gifford, C. A., Goodman, A. L., Joachimiak, G., Ordoñez, C. L., Lory, S., Walz, T., Joachimiak, A., and Mekalanos, J. J. (2006). A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. *Science*, 312(5779):1526–1530.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Genetics, M., Pasteur, I., and Bolognetti, F. C. (2015). Critical assessment of methods of protein structure prediction. *Proteins*, 82(0 2):1–6.

Mukamolova, G. V., Murzin, A. G., Salina, E. G., Demina, G. R., Kell, D. B., Kaprelyants, A. S., and Young, M. (2006). Muralytic activity of Micrococcus luteus Rpf and its relationship to physiological activity in promoting bacterial growth and resuscitation. *Molecular Microbiology*, 59(1):84–98.

Mulley, G., Beeton, M. L., Wilkinson, P., Vlisidou, I., Ockendon-Powell, N., Hapeshi, A., Tobias, N. J., Nollmann, F. I., Bode, H. B., Van Den Elsen, J., Ffrench-Constant, R. H., and Waterfield, N. R. (2015). From insect to man: Photorhabdus sheds light on the emergence of human pathogenicity. *PLoS ONE*, 10(12):1–32.

Murzin, a. G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *The EMBO journal*, 12(3):861–867.

Muskotál, A., Király, R., Sebestyén, A., Gugolya, Z., Végh, B. M., and Vonderviszt, F. (2006). Interaction of FliS flagellar chaperone with flagellin. *FEBS Letters*, 580(16):3916–3920.

Naidoo, S., Mothupi, B., Featherston, J., Mpangase, P. T., and Gray, V. M. (2015). Draft Genome Sequence and Assembly of Photorhabdus heterorhabditis Strain VMG, a Bacterial Symbiont Associated with the Entomopathogenic Nematode Heterorhabditis zealandica. *Genome announcements*, 3(5):5–6.

Nakayama, K., Takashima, K., Ishihara, H., Shinomiya, T., Kageyama, M., Kanaya, S., Ohnishi, M., Murata, T., Mori, H., and Hayashi, T. (2000). The R-type pyocin of Pseudomonas aeruginosa is related to P2 phage, and the F-type is related to lambda phage. *Molecular Microbiology*, 38:213–231.

Nazarov, S., Schneider, J. P., Brackmann, M., Goldie, K. N., Stahlberg, H., and Basler, M. (2017). CryoEM reconstruction of Type VI secretion system baseplate and sheath distal end. *The EMBO Journal*, page e201797103.

Nguyen, V. S., Douzi, B., Durand, E., Roussel, A., Cascales, E., and Cambillau, C. (2018). Towards a complete structural deciphering of Type VI secretion system. *Current Opinion in Structural Biology*, 49:77–84.

Nilsson, E. C., Storm, R. J., Bauer, J., Johansson, S. M., Lookene, A., Ångström, J., Hedenström, M., Eriksson, T. L., FräCurrency Signngsmyr, L., Rinaldi, S., Willison, H. J., Domellöf, F. P., Stehle, T., and Arnberg, N. (2011). The GD1a glycan is a cellular receptor for adenoviruses causing epidemic keratoconjunctivitis. *Nature Medicine*, 17(1):105–109.

Ohkawa, I., Kageyama, M., and Egami, F. (1973). Purification and Properties of Pyocin S2. *The Journal of Biochemistry*, 73(2):281–289.

Oliver, K. M., Russell, J. A., Moran, N. A., and Hunter, M. S. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proceedings of the National Academy of Sciences*, 100(4):1803–1807.

Opender Koul, O. (2011). Microbial biopesticides: opportunities and challenges. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, 6(056).

Orozco, R. A., Molnár, I., Bode, H., and Patricia Stock, S. (2016). Bioprospecting for secondary metabolites in the entomopathogenic bacterium Photorhabdus luminescens subsp. sonorensis. *Journal of Invertebrate Pathology*.

Osipiuk, J., Xu, X., Cui, H., Savchenko, A., Edwards, A., and Joachimiak, A. (2011). Crystal structure of secretory protein Hcp3 from Pseudomonas aeruginosa. *Journal of Structural and Functional Genomics*, 12(1):21–26.

Papanikolopoulou, K., Forge, V., Goeltz, P., and Mitraki, A. (2004a). Formation of Highly Stable Chimeric Trimers by Fusion of an Adenovirus Fiber Shaft Fragment with the Foldon Domain of Bacteriophage T4 Fibritin. *Journal of Biological Chemistry*, 279(10):8991–8998.

Papanikolopoulou, K., Teixeira, S., Belrhali, H., Forsyth, V. T., Mitraki, A., and Van Raaij, M. J. (2004b). Adenovirus fibre shaft sequences fold into the native triple beta-spiral fold when N-terminally fused to the bacteriophage T4 fibritin foldon trimerisation motif. *Journal of Molecular Biology*, 342(1):219–227.

Papanikolopoulou, K., van Raaij, M. J., and Mitraki, A. (2008a). *Creation of Hybrid Nanorods From Sequences of Natural Trimeric Fibrous Proteins Using the Fibritin Trimerization Motif*, pages 15–33. Humana Press, Totowa, NJ.

Papanikolopoulou, K., van Raaij, M. J., and Mitraki, A. (2008b). *Nanostructure Design: Methods and Protocols*. Humana Press, Tel Aviv.

Pattengale, N. D., Gottlieb, E. J., and Moret, B. M. E. (2007). Efficiently Computing the Robinson-Foulds Metric. *Journal of Computational Biology*, 14(6):724–735.

Paul, R., Weiser, S., Amiot, N. C., Chan, C., Schirmer, T., Giese, B., and Jenal, U. (2004). Cell cycle-dependent dynamic localization of a bacterial response regulator with a novel di-guanylate cyclase output domain. *Genes and Development*, 18(6):715–727.

Peat, S. M., Ffrench-Constant, R. H., Waterfield, N. R., Marokházi, J., Fodor, A., and Adams, B. J. (2010). A robust phylogenetic framework for the bacterial genus Photorhabdus and its use in studying the evolution and maintenance of bioluminescence: a case for 16S, gyrB, and glnA. *Molecular Phylogenetics and Evolution*, 57(2):728–40.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J. G., Jacobs, W. R., Hendrix, R. W., and Hatfull, G. F. (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*, 113(2):171–182.

Pell, L. G., Kanelis, V., Donaldson, L. W., Lynne Howell, P., and Davidson, A. R. (2009). The lambda phage major tail protein structure reveals a common evolution for long-tailed phages and the type VI bacterial secretion system. *Proceedings of the National Academy of Sciences*, 106(11):4160–4165.

Penz, T., Schmitz-Esser, S., Kelly, S. E., Cass, B. N., Müller, A., Woyke, T., Malfatti, S. a., Hunter, M. S., and Horn, M. (2012). Comparative Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in Cardinium hertigii. *PLoS Genetics*, 8(10).

Persson, O. P., Pinhassi, J., Riemann, L., Marklund, B. I., Rhen, M., Normark, S., González, J. M., and Hagström, Å. (2009). High abundance of virulence gene homologues in marine bacteria. *Environmental Microbiology*, 11(6):1348–1357.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

Pinto, F. R., Melo-Cristino, J., and Ramirez, M. (2008). A confidence interval for the wallace coefficient of concordance and its application to microbial typing methods. *PLoS ONE*, 3(11).

Pukatzki, S., Ma, A. T., Revel, A. T., Sturtevant, D., and Mekalanos, J. J. (2007). Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. *Proceedings of the National Academy of Sciences of the United States of America*, 104(39):15508–15513.

Pukatzki, S., Ma, A. T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W. C., Heidelberg, J. F., and Mekalanos, J. J. (2006). Identification of a conserved bacterial protein secretion system in Vibrio cholerae

using the Dictyostelium host model system. *Proceedings of the National Academy of Sciences*, 103(5):1528–1533.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rademacher, C., Bru, T., McBride, R., Robison, E., Nycholat, C. M., Kremer, E. J., and Paulson, J. C. (2012). A Siglec-like sialic-acid-binding motif revealed in an adenovirus capsid protein. *Glycobiology*, 22(8):1086–1091.

Ramachandran, P., Boontheung, P., Xie, Y., Sondej, M., Wong, D. T., and Loo, J. A. (2006). Identification of N-linked glycoproteins in human saliva by glycoprotein capture and mass spectrometry. *Journal of Proteome Research*, 5(6):1493–1503.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846.

Remaut, H., Tang, C., Henderson, N. S., Pinkner, J. S., Wang, T., Hultgren, S. J., Thanassi, D. G., Waksman, G., and Li, H. (2008). Fiber Formation across the Bacterial Outer Membrane by the Chaperone/Usher Pathway. *Cell*, 133(4):640–652.

Remaut, H. and Waksman, G. (2006). Protein-protein interaction through β-strand addition. *Trends in Biochemical Sciences*, 31(8):436–444.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175.

Richardson, J. (1981). The anatomy and toxonomy of protein structure. *Advan. Protein Chem.*, 34:167–330.

Riede, I. (1987). Receptor specificity of the short tail fibres (gp12) of T-even type Escherichia coli phages. *MGG Molecular & General Genetics*, 206(1):110–115.

Robichon, C., Luo, J., Causey, T. B., Benner, J. S., and Samuelson, J. C. (2011). Engineering Escherichia coli BL21(DE3) derivative strains to minimize E. coli Protein contamination after purification by immobilized metal affinity chromatography. *Applied and Environmental Microbiology*, 77(13):4634–4646.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

Rodríguez-Guerra Pedregal, J. and Maréchal, J.-D. (2018). PyChimera: use UCSF Chimera modules in any

Python 2.7 project. *Bioinformatics*, (January):1–2.

Rossmann, M. G., Mesyanzhinov, V. V., Arisaka, F., and Leiman, P. G. (2004). The bacteriophage T4 DNA injection machine. *Current Opinion in Structural Biology*, 14:171–180.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Ruby, E. G. and McFall-Ngai, M. J. (1999). Oxygen-utilizing reactions and symbiotic colonization of the squid light organ by Vibrio fischeri. *Trends in Microbiology*, 7(10):414–420.

Russell, A. B., Peterson, S. B., and Mougous, J. D. (2014). Type VI secretion system effectors: poisons with a purpose. *Nature reviews. Microbiology*, 12(2):137–48.

Russell, A. B., Singh, P., Brittnacher, M., Bui, N. K., Hood, R. D., Carl, M. A., Agnello, D. M., Schwarz, S., Goodlett, D. R., Vollmer, W., and Mougous, J. D. (2012). A widespread bacterial type VI secretion effector superfamily identified using a heuristic approach. *Cell Host and Microbe*, 11(5):538–549.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000). Artemis: Sequence visualization and annotation. *Bioinformatics*, 16(10):944–945.

Rybakova, D. (1994). *Insights into the assembly and biology of the Serratia entomophila Anti-feeding Prophage*. PhD thesis.

Rybakova, D., Radjainia, M., Turner, A., Sen, A., Mitra, A. K., and Hurst, M. R. H. (2013). Role of antifeeding prophage (Afp) protein Afp16 in terminating the length of the Afp tailocin and stabilizing its sheath. *Molecular microbiology*, 89(4):702–14.

Rybakova, D., Schramm, P., Mitra, A. K., and Hurst, M. R. H. (2015a). Afp14 is involved in regulating the length of Anti-feeding prophage (Afp). *Molecular Microbiology*, pages n/a–n/a.

Rybakova, D., Schramm, P., Mitra, A. K., and Hurst, M. R. H. (2015b). Afp14 is involved in regulating the length of Anti-feeding prophage (Afp). *Molecular Microbiology*, 96(4):815–826.

Sandmeler, H. (1994). Acquisition and rearrangement of sequence motifs in the evolution of bacteriophage tail fibres. *Molecular Microbiology*, 12(3):343–350.

Sarris, P. F., Ladoukakis, E. D., Panopoulos, N. J., and Scoulica, E. V. (2014). A phage tail-derived element

with wide distribution among both prokaryotic domains: a comparative genomic and phylogenetic study. *Genome biology and evolution*, 6(7):1739–47.

Saux, M. F.-l., Viallardt, V., Brunelt, B., Normand, P., and Boemarel, N. E. (1999). Polyphasic classification of the genus Photorhabdus and proposal of new taxa : m luminescens subsp . akhurstii subsp . nov ., m temperata subsp . temperata subsp . nov . and P. *International journal of systematic bacteriology*, 49(1 999):1645–1 656.

Scholl, D., Cooley, M., Williams, S. R., Gebhart, D., Martin, D., Bates, A., and Mandrell, R. (2009). An engineered R-type pyocin is a highly specific and sensitive bactericidal agent for the food-borne pathogen Escherichia coli O157:H7. *Antimicrobial Agents and Chemotherapy*, 53(7):3074–3080.

Scholl, D. and Martin, D. W. (2008). Antibacterial efficacy of R-type pyocins towards Pseudomonas aeruginosa in a murine peritonitis model. *Antimicrobial Agents and Chemotherapy*, 52(5):1647–1652.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069.

Sen, A., Rybakova, D., Hurst, M. R. H., and Mitra, A. K. (2010). Structural study of the Serratia entomophila antifeeding prophage: Three-dimensional structure of the helical sheath. *Journal of Bacteriology*, 192(17):4522–4525.

Severiano, A., Carriço, J. A., Robinson, D. A., Ramirez, M., and Pinto, F. R. (2011a). Evaluation of Jackknife and Bootstrap for defining confidence intervals for pairwise agreement measures. *PLoS ONE*, 6(5):6–8.

Severiano, A., Pinto, F. R., Ramirez, M., and Carriço, J. A. (2011b). Adjusted Wallace coefficient as a measure of congruence between typing methods. *Journal of Clinical Microbiology*, 49(11):3997–4000.

Shi, Y.-M. and Bode, H. B. (2018). Chemical language and warfare of bacterial natural products in bacterianematodeinsect interactions. *Natural Product Reports*, 00:1–27.

Shikuma, N. J., Antoshechkin, I., Medeiros, J. M., Pilhofer, M., and Newman, D. K. (2016). Stepwise metamorphosis of the tubeworm Hydroides elegans is mediated by a bacterial inducer and MAPK signaling. *Proceedings of the National Academy of Sciences*, 113(36):10097–10102.

Shikuma, N. J., Pilhofer, M., Weiss, G. L., Hadfield, M. G., Jensen, G. J., and Newman, D. K. (2014). Marine tubeworm metamorphosis induced by arrays of bacterial phage tail-like structures. *Science (New York, N.Y.)*, 343(6170):529–33.

Shinomiya, T. (1972). Studies on biosynthesis and morphogenesis of R-type pyocins of Pseudomonas aeruginosa. II. Biosynthesis of antigenic proteins and their assembly into pyocin particles in mitomycin C-induced

cells. *Journal of Biochemistry*, 72(March):39–48.

Shneider, M. M., Buth, S. a., Ho, B. T., Basler, M., Mekalanos, J. J., and Leiman, P. G. (2013). PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. *Nature*, 500(7462):350–3.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(539).

Silverman, J. M., Brunet, Y. R., Cascales, E., and Mougous, J. D. (2012). Structure and Regulation of the Type VI Secretion System. *Annual Review of Microbiology*, 66(1):453–472.

Simpson, D. J., Sacher, J. C., and Szymanski, C. M. (2015). Exploring the interactions between bacteriophage-encoded glycan binding proteins and carbohydrates. *Current Opinion in Structural Biology*, 34:69–79.

Singh, A. K., Berbís, M. Á., Ballmann, M. Z., Kilcoyne, M., Menéndez, M., Nguyen, T. H., Joshi, L., Cañada, F. J., Jiménez-Barbero, J., Benko, M., Harrach, B., and Van Raaij, M. J. (2015). Structure and sialyllactose binding of the carboxy-terminal head domain of the fibre from a siadenovirus, Turkey adenovirus 3. *PLoS ONE*, 10(9):1–22.

Söding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(SUPPL. 2):244–248.

Soniak, M. (2012). No Title.

Sperling, R. A. and Parak, W. J. (2010). Surface modification, functionalization and bioconjugation of colloidal inorganic nanoparticles. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1915):1333–1383.

Sproer, C., Mendrock, U., Swiderski, J., and Lang, E. (1999). The phylogenetic position of. (1 999):1433–1438.

Sreerama, N., Venyaminov, S. Y., and Woody, R. W. (2000). Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Inclusion of Denatured Proteins with Native Proteins in the Analysis. *Analytical Biochemistry*, 287(2):243–251.

Sreerama, N. and Woody, R. W. (2000). Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set. *Analytical Biochemistry*, 287(2):252–260.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands

of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stephens, C. (1998). Bacterial sporulation: a question of commitment? *Current biology : CB*, 8:R45–R48.

Suarez, G., Sierra, J. C., Erova, T. E., Sha, J., Horneman, A. J., and Chopra, A. K. (2010). A type VI secretion system effector protein, VgrG1, from Aeromonas hydrophila that induces host cell toxicity by ADP ribosylation of actin. *Journal of Bacteriology*, 192(1):155–168.

Taylor, N. M., Prokhorov, N. S., Guerrero-Ferreira, R. C., Shneider, M. M., Browning, C., Goldie, K. N., Stahlberg, H., and Leiman, P. G. (2016a). Structure of the T4 baseplate and its function in triggering sheath contraction. *Nature*, 533(7603):346–352.

Taylor, N. M., Prokhorov, N. S., Guerrero-Ferreira, R. C., Shneider, M. M., Browning, C., Goldie, K. N., Stahlberg, H., and Leiman, P. G. (2016b). Structure of the T4 baseplate and its function in triggering sheath contraction. *Nature*, 533(7603):346–352.

Taylor, N. M., van Raaij, M. J., and Leiman, P. G. (2018). Contractile injection systems of bacteriophages and related systems. *Molecular Microbiology*, 108(February):6–15.

Thanassi, D. G., Stathopoulos, C., Karkal, A., and Li, H. (2005). Protein secretion in the absence of ATP: The autotransporter, two-partner secretion and chaperone/usher pathways of Gram-negative bacteria. *Molecular Membrane Biology*, 22(1-2):63–72.

Thomassen, E., Gielen, G., Schütz, M., Schoehn, G., Abrahams, J. P., Miller, S., and Van Raaij, M. J. (2003). The structure of the receptor-binding domain of the bacteriophage T4 short tail fibre reveals a knitted trimeric metal-binding fold. *Journal of Molecular Biology*, 331(2):361–373.

Tobias, N. J., Heinrich, A. K., Eresmann, H., Wright, P. R., Neubacher, N., Backofen, R., and Bode, H. B. (2016). Photorhabdus-nematode symbiosis is dependent on hfq-mediated regulation of secondary metabolites. *Environmental microbiology*, pages 1–20.

Uratani, Y. and Hoshino, T. (1984). Pyocin Ri Inhibits Active Transport in Pseudomonas aeruginosa and Depolarizes Membrane Potential. 157(2):1–6.

van Raaij, M. J., Mitraki, A., Lavigne, G., and Cusack, S. (1999). A triple beta-spiral in the adenovirus bre shaft reveals a new structural motif for a brous protein. *Nature*, 461:935–938.

van Raaij, M. J., Schoehn, G., Burda, M. R., and Miller, S. (2001). Crystal structure of a heat and protease-stable part of the bacteriophage T4 short tail fibre. *Journal of Molecular Biology*, 314(5):1137–1146.

Varki, A., Cummings, R. D., Aebi, M., Packer, N. H., Seeberger, P. H., Esko, J. D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., Prestegard, J. J., Schnaar, R. L., Freeze, H. H., Marth, J. D., Bertozzi, C. R., Etzler, M. E., Frank, M., Vliegenthart, J. F., Lütteke, T., Perez, S., Bolton, E., Rudd, P., Paulson, J., Kanehisa, M., Toukach, P., Aoki-Kinoshita, K. F., Dell, A., Narimatsu, H., York, W., Taniguchi, N., and Kornfeld, S. (2015). Symbol nomenclature for graphical representations of glycans. *Glycobiology*, 25(12):1323–1324.

Varki, A. and Gagneux, P. (2012). Multifarious roles of sialic acids in immunity. *Annals of the New York Academy of Sciences*, 1253(1):16–36.

Veesler, D. and Cambillau, C. (2011). A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. *Microbiology and molecular biology reviews : MMBR*, 75(3):423–433.

Verma, S. C. and Miyashiro, T. (2013). Quorum sensing in the squid-Vibrio symbiosis. *International journal of molecular sciences*, 14(8):16386–16401.

Vettiger, A., Winter, J., Lin, L., and Basler, M. (2017). The type VI secretion system sheath assembles at the end distal from the membrane anchor. *Nature Communications*, 8(May):1–9.

Wallace, D. L. (1983). A Method for Comparing Two Hierarchical Clusterings: comment. *Journal of the American Statistical Association*, 78(383):553–569.

Wang, J., Brackmann, M., Castaño-Díez, D., Kudryashev, M., Goldie, K. N., Maier, T., Stahlberg, H., and Basler, M. (2017a). Cryo-EM structure of the extended type VI secretion system sheath-tube complex. *Nature Microbiology*, 2(11):1507–1512.

Wang, J., Brackmann, M., Castaño-Díez, D., Kudryashev, M., Goldie, K. N., Maier, T., Stahlberg, H., and Basler, M. (2017b). Cryo-EM structure of the extended type VI secretion system sheath-tube complex. *Nature Microbiology*, 2(11):1507–1512.

Waterfield, N. R., Ciche, T. A., and Clarke, D. J. (2009). Photorhabdus and a host of hosts. *Annual Review of Microbiology*, 63:557–74.

Waterfield, N. R., Daborn, P. J., and Ffrench-Constant, R. H. (2004). Insect pathogenicity islands in the insect pathogenic bacterium Photorhabdus. *Physiological Entomology*, 29(3 SPEC. ISS.):240–250.

Waterfield, N. R., Sanchez-Contreras, M., Eleftherianos, I., Dowling, A., Yang, G., Wilkinson, P., Parkhill, J., Thomson, N., Reynolds, S. E., Bode, H. B., Dorus, S., and Ffrench-Constant, R. H. (2008). Rapid

Virulence Annotation (RVA): identification of virulence factors using a bacterial genome library and multiple invertebrate hosts. *Proceedings of the National Academy of Sciences of the United States of America*, 105(41):15967–72.

Wenren, L. M., Sullivan, N. L., and Cardarelli, L. (2013). Two Independent Pathways for Self-Recognition in Proteus mirabilis Are Linked by Type VI-Dependent Export. *mBio*, 4(4):1–10.

Wernimont, A. and Edwards, A. (2009). In Situ proteolysis to generate crystals for structure determination: An update. *PLoS ONE*, 4(4).

Werren, J. H., Baldo, L., and Clark, M. E. (2008). Wolbachia: Master manipulators of invertebrate biology. *Nature Reviews Microbiology*, 6(10):741–751.

Whitmore, L. and Wallace, B. A. (2004). DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Research*, 32(WEB SERVER ISS.):668–673.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkinson, P., Paszkiewicz, K., Moorhouse, A., Szubert, J. M., Beatson, S., Gerrard, J., Waterfield, N. R., and Ffrench-Constant, R. H. (2010). New plasmids and putative virulence factors from the draft genome of an Australian clinical isolate of Photorhabdus asymbiotica. *FEMS Microbiology Letters*, 309(2):136–143.

Wilkinson, P., Waterfield, N. R., Crossman, L., Corton, C., Sanchez-contreras, M., Vlisidou, I., Barron, A., Bignell, A., Clark, L., Ormond, D., Mayho, M., Bason, N., Smith, F., Simmonds, M., Churcher, C., Harris, D., Thompson, N. R., Quail, M., Parkhill, J., Ffrench-Constant, R. H., Richard, H., and Ffrench-Constant, R. H. (2009). Comparative genomics of the emerging human pathogen Photorhabdus asymbiotica with the insect pathogen Photorhabdus luminescens. *BMC Genomics*, 10(302):302.

Williams, S. R., Gebhart, D., Martin, D. W., and Scholl, D. (2008). Retargeting R-type pyocins to generate novel bactericidal protein complexes. *Applied and Environmental Microbiology*, 74(12):3868–3876.

Yang, G., Dowling, A. J., Gerike, U., and Waterfield, N. R. (2006). Photorhabdus virulence cassettes confer injectable insecticidal activity against the wax moth. *Journal of . . .* , 188(6):2254–2261.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2014). The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, 12(1):7–8.

Yap, M. L. and Rossmann, M. G. (2014a). Structure and function of bacteriophage T4. *Future Microbiology*, 9(12):1319–1327.

Yap, M. L. and Rossmann, M. G. (2014b). Structure and function of bacteriophage T4. *Future Microbiology*, 9(12):1319–1327.

Zhang, D., de Souza, R. F., Anantharaman, V., Iyer, L. M., and Aravind, L. (2012). Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biology Direct*, 7.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:1–8.

Zheng, J. and Leung, K. Y. (2007). Dissection of a type VI secretion system in Edwardsiella tarda. *Molecular Microbiology*, 66(5):1192–1206.

Zheng, W., Wang, F., Taylor, N. M., Guerrero-Ferreira, R. C., Leiman, P. G., and Egelman, E. H. (2017). Refined Cryo-EM Structure of the T4 Tail Tube: Exploring the Lowest Dose Limit. *Structure*, 25(9):1436–1441.e2.

Zink, R., Loessner, M. J., and Scherer, S. (1995). Characterization of cryptic prophages (monocins) in Listeria and sequence analysis of a holin/endolysin gene. *Microbiology*, 141(10):2577–2584.

Zoued, A., Durand, E., Brunet, Y. R., Spinelli, S., Douzi, B., Guzzo, M., Flaugnatti, N., Legrand, P., Journet, L., Fronzes, R., Mignot, T., Cambillau, C., and Cascales, E. (2016). Priming and polymerization of a bacterial contractile tail structure. *Nature*, 531(7592):59–63.

Appendices

Appendix A

Chapter 3 Appendices

Appendix B

Chapter 4 Appendices

Appendix C

Chapter 5 Appendices

Appendix D

Chapter 5 Appendices

```
>PAK_01799
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      60.81|      18|      93|      48|      65|       1
--------------------------------------------------------------------------------
   48-   65 (33.43/21.54) DIGRKATG.QAPGQADNPN
  144-  162 (27.38/16.39) DFSEKGKGsKASGSGDKKN
--------------------------------------------------------------------------------
>PAK_02001
No repeats found
No repeats found
>PAK_02618
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             7|     144.10|      12|      14|     245|     256|       1
--------------------------------------------------------------------------------
  146-  157 (21.30/ 9.21) GSGIKVDSSGVS
  185-  196 (18.32/ 6.87) GNGLYGRDNGIS
  200-  211 (18.19/ 6.77) GPGIVVDYQHVS
  215-  226 (22.19/ 9.91) GNGITVNDSGVA
  245-  256 (22.59/10.22) NNGINVDTNGVS
  291-  302 (18.24/ 6.81) SDGIDVDISGVS
  306-  317 (23.27/10.75) GNGITVNSNGVS
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             5|     110.09|      15|      29|      83|      97|       2
--------------------------------------------------------------------------------
   65-   80 (19.72/ 8.73) NPNSALKLDgNDALAV
   83-   97 (25.51/13.41) NENGGLKTD.KDGLSV
  100-  113 (20.33/ 9.22) K.NKSLLAD.NNGLAV
  114-  128 (23.82/12.04) NTGRGLKIN.NDKLEV
  129-  143 (20.71/ 9.53) NDHYGIEII.DEGVKV
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             3|     158.95|      34|      74|     380|     413|       3
--------------------------------------------------------------------------------
  380-  413 (56.72/31.94) GKSY.......YRNTTLTQVSVSVNVQETTLTESQIPSHKH
  416-  455 (48.34/26.18) GMSYcytygmnYNSSSDTQTKYQIN.NSDSISDSAIWKNPS
```

```
  457-  486 (53.89/29.99) GSNY.......YAHTSNTGGGQGHNHQATASS....SSHNH
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             3|      66.49|      17|      28|     227|     243|       4
--------------------------------------------------------------------------------
  227-  243 (27.06/12.46) VKAAN.GITVNGSGVAVK
  257-  274 (20.97/ 8.20) IKAKDkTINVESTGISVR
  276-  289 (18.46/ 6.45) ...GW.GVREGGLGLDVK
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      43.58|      15|      21|      26|      41|       5
--------------------------------------------------------------------------------
   25-   40 (22.86/21.64) D........DLKRRfKEGSIPLQT
   41-   63 (20.73/12.94) DyadliniaDIGRR.AVGKAPGQT
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      54.13|      15|      27|     326|     340|       6
--------------------------------------------------------------------------------
  326-  340 (26.70/17.74) RGIIVMFSGSSAPTG
  354-  368 (27.44/18.43) RSRFVMCGETISETG
--------------------------------------------------------------------------------
>PAK_03191
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             4|     179.44|      29|      29|     137|     165|       1
--------------------------------------------------------------------------------
  114-  134 (35.62/13.49) ..NGICV.GQGI........GITVNTSNVAVK
  135-  164 (51.96/22.42) QgNGISVTGTGVA..VKAYNGINVDVNGVAVK
  165-  194 (50.00/21.36) AyNGINVASTGVA..VKAYNGIDVTSSGVAVK
  197-  228 (41.87/16.91) AnKGLSVDSTGVAvkVKANGGITVDTNGVAID
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      65.48|      19|     212|      58|      77|       2
--------------------------------------------------------------------------------
   36-   54 (30.28/16.33) GSIPLQTDFSELIDIADIG
   59-   77 (35.19/27.84) GQAPQQNGPGEGLKLDDSG
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|     117.70|      35|      82|     303|     338|       3
--------------------------------------------------------------------------------
  303-  338 (58.88/43.01) TTVSVSVTVQNKVLTeANLPMHQH.IGGVPYTYNTNF
  388-  423 (58.82/38.30) TSVVGSATGHNHTAT.ATSTGHTHsVDVVPPYYLLAF
--------------------------------------------------------------------------------
>PAU_01973
No repeats found
>PAU_02195
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             3|      72.06|      14|      28|      81|      94|       1
--------------------------------------------------------------------------------
   81-   94 (22.51/11.66) LQVKAGAGVDIDNN
   97-  110 (24.55/13.28) ITIKSGHGIKVDGN
  112-  125 (24.99/13.63) ISVKPGSGIKVDSN
--------------------------------------------------------------------------------
```

```
>PAU_02787
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             6|     121.28|      11|      35|     114|     124|       1
--------------------------------------------------------------------------------
  114-  124 (21.65/11.89) NGICVGQGNGI
  129-  139 (19.16/ 9.67) NDVAVKAANGI
  144-  154 (19.88/10.32) SGVAVKANNGI
  175-  185 (19.01/ 9.55) TGISVRLGWGI
  189-  199 (18.47/ 9.07) DGLDVKASNGI
  204-  214 (23.10/13.18) NGVSVKAGNGI
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|     201.43|      66|      89|     229|     301|       2
--------------------------------------------------------------------------------
  229-  301 (107.35/53.54) LPRGMIVMFSGNSA.P..TGWAFCDGNsgTPDLRSRFimcgeTISETGKSS..............NKASGSGNGKNF..SRNTTSTTVSVNV
  308-  393 (94.07/36.31) LTESQIPKHKHIEAlPyyNTLGFAYGN..TPIGSTKY.....QINNTSSSMffwhpsptgndyhpyTSEVGGGQGHNHkaTASSSSHTHSVDV
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      36.38|      10|      28|      17|      26|       3
--------------------------------------------------------------------------------
   17-   26 (19.02/12.84) INIKPQGPSA
   48-   57 (17.36/11.06) INIADIGRKA
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      62.29|      19|      20|      65|      84|       7
--------------------------------------------------------------------------------
   65-   84 (28.72/24.41) NGPGTGLKLADDgTLNLKIG
   88-  106 (33.57/22.82) NQDFSPLMLKDD.ILSIDLG
--------------------------------------------------------------------------------
>PAU_03380
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
            10|     298.40|      15|      15|     200|     214|       1
--------------------------------------------------------------------------------
  149-  163 (25.32/10.37) VKVSANKGLSVDSSG
  166-  180 (29.64/13.40) VKVNTDKGISVDGNG
  183-  197 (29.67/13.42) VKVNTSKGISVDNTG
  200-  214 (31.92/14.99) VIANASKGISVDGSG
  217-  231 (33.00/15.76) VIANTSKGISVDGSG
  234-  248 (30.39/13.93) VIANTSKGISVDNTG
  251-  265 (31.92/14.99) VIANASKGISVDGSG
  268-  282 (33.00/15.76) VIANTSKGISVDGSG
  285-  299 (31.82/14.93) VIANTSKGISVDSSG
  302-  316 (21.72/ 7.85) VKVKANGGIKVDANG
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      93.80|      35|      67|      15|      77|       2
--------------------------------------------------------------------------------
   15-   77 (37.47/66.35) TLSNpKAVGPDIdkLKDkfkegsiplqtdfneliDI..ADIGrkaCGQAPQQNG..PGEGlkladDG
   85-  123 (56.33/27.90) TFSN.KDFSPLI..LKD.................DVlsVDLG...SGLTNETNGicVGQG.....DG
--------------------------------------------------------------------------------
>PLT_01684
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
```

```
             5|     181.88|      27|      28|     122|     148|       1
--------------------------------------------------------------------------------
   94-  118 (28.66/13.56) ..LILEKDILSVDL....GSG.LINKPNGICV
  122-  148 (42.82/24.12) NGIVVNNDNVAVKA....ANG.ITVNGSGVAI
  152-  179 (37.26/19.98) NGINVDTNGVSVKS....KNStIKVESSGISV
  183-  209 (44.20/25.16) WGVKIGGEGLDIKA....SNG.IKVDGNGVSV
  213-  240 (28.95/13.78) FGITVNSDGVNIDAnkllSKG.MIVMFSG...
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      67.43|      19|      21|      36|      54|       3
--------------------------------------------------------------------------------
   36-   54 (32.92/17.96) GSIPLQTDYADLINIADIG
   59-   77 (34.50/19.14) GQAPQQNGPGAGLKLADDG
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             3|      95.66|      28|      28|     320|     347|       4
--------------------------------------------------------------------------------
  320-  337 (24.81/10.43) ......................IESLPYYTTLGFAYDHTT
  338-  366 (45.88/24.95) IGATNNKIDN..........sVNGLIWKRTSGPDYHPYT
  369-  401 (24.97/10.54) IGGGQGHNHNasasspshthsVDVVPPYYILAF......
--------------------------------------------------------------------------------
>PLT_01704
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             4|     249.30|      60|      63|      79|     141|       1
--------------------------------------------------------------------------------
   30-   77 (46.14/21.67) ...........KEG.SIPL....QtD.YSDLiniADIGRQ.......AV..G.....KAPnQIDN.PNSG.LVLNND.SGL.....A
   79-  141 (94.99/59.09) KVNISGGLQADKDGVSVKI....K.D.KSLL...ADNNGL.......SVNYG.....KGL.QLDK.DNDNKLTINSH.DGIEIvagG
  143-  197 (63.32/31.76) KVKAGNGITVNSSGVS..I....D.P.NTVL...P..RGMivmfsgkSVPTG.....WTL..CDG.NNGTPNLIDR...........
  199-  260 (44.86/20.11) ...ILGGNFSGIDGKSSTTvsgpK.DsKSFN...FNSNEA.......TLNINgktseRSL.SIGQiPN......HSHlSGINI....
--------------------------------------------------------------------------------
>PLT_01724
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             4|     148.39|      26|      27|     108|     133|       1
--------------------------------------------------------------------------------
  108-  133 (44.89/22.80) NNGLAVNAGRG.........LRINNDKLEVNNHHG
  154-  187 (27.97/11.41) .NGVSLKMGKPisnafspliLEPKNDVLSVKIGNG
  192-  217 (43.68/21.98) DNGISVNPGNG.........IDAGYDYVAVKASNG
  222-  243 (31.86/14.02) NSGVSVKAGNG.........ITVNSNGVSVN....
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      37.36|      15|      15|      67|      81|       2
--------------------------------------------------------------------------------
   52-   75 (16.15/ 8.79) GRWAVgKapgqtdnpNSALKLDNG
   76-   93 (21.22/14.31) GALAV.K.....indNGGLKADEN
--------------------------------------------------------------------------------
>PLT_01746
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             7|     143.23|      15|      15|      83|      97|       1
--------------------------------------------------------------------------------
   83-   97 (21.47/ 9.86) GLKIKINSTGG.LKAD
  100-  113 (19.14/ 7.96) GLSVKLKDK.S.LLAD
  116-  128 (17.21/ 6.39) GL..AVNAGRG.VKIN
  161-  173 (20.90/ 9.39) GVSVK..AGNG.ISVS
```

```
  176-  190 (22.62/10.79) GVEVKAKDKGS.ISVD
  193-  208 (19.52/ 8.27) GIAVKYWDGGGiVATD
  237-  249 (22.35/10.57) GVKVK..AGNG.ITVD
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|     101.35|      32|      85|     295|     335|       2
--------------------------------------------------------------------------------
  295-  335 (47.37/37.26) VMcsetisetgKSSNKA.......SGSGNGKNYSRNTTSTTVSVSVTV
  383-  421 (53.98/25.99) VM.........QSANGAryaytspSGGGQGHNHPATASSPSHDHSVNV
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             2|      75.82|      20|      50|     211|     230|       5
--------------------------------------------------------------------------------
  211-  230 (36.65/20.30) GLYLKLEGGNTNNGWSGVSG
  264-  283 (39.17/22.19) GMIVMFSGSSAPTGWAFCDG
--------------------------------------------------------------------------------
>PLT_02556
--------------------------------------------------------------------------------
No. of Repeats|Total Score|Length  |Diagonal| BW-From|   BW-To|   Level
             5|     303.36|      59|      60|     134|     192|       1
--------------------------------------------------------------------------------
    8-   39 (29.29/ 7.50) ..........................SN.....................SEINPENTNNKTESPS....ADDLK.....KRFKAG...SIP
   53-   95 (47.20/15.81) IGRKAV..............GKAPQQNG.....................PGIGLKLDDNGTLNLKigtiPD.LAD....KGFS........
  134-  192 (104.31/42.29) IAVKAANGITVDGSGVSIKAGNGISVSG.....................NGVEVKAKNNGSISVE....PDGIAV....KCWDGG...GIV
  228-  277 (66.87/24.93) ...KAGNGIKVDDKGVSIDP.NKVLPRG.....................MIVMF....SGS.SV.....PEGWAL.....C.DGKdnrPNL
  278-  352 (55.70/19.75) IDRFIMGGTTQNIGGKSSD.....SFSGakdnkkftfisesqtvrisgstDGHGLTADENGPHQHE.....QGETLnrqgKCHNG.......
--------------------------------------------------------------------------------
```

Appendix E

Publications arising from this candidature

The following is a paper published in Angewandte Chemie in 2017. The structural modelling within the paper was performed by me and is the basis of my inclusion as an author. This paper is included as an appendix as it was a publication arising from a side project during my PhD candidature, but is not relevant to my personal thesis research, though the same methods as described in Chapter 3 were used.

Biomaterials

Deutsche Ausgabe: DOI: 10.1002/ange.201706703
Internationale Ausgabe: DOI: 10.1002/anie.201706703

Polyproline as a Minimal Antifreeze Protein Mimic That Enhances the Cryopreservation of Cell Monolayers

*Ben Graham, Trisha L. Bailey, Joseph R. J. Healey, Moreno Marcellini, Sylvain Deville, and Matthew I. Gibson**

Abstract: *Tissue engineering, gene therapy, drug screening, and emerging regenerative medicine therapies are fundamentally reliant on high-quality adherent cell culture, but current methods to cryopreserve cells in this format can give low cell yields and require large volumes of solvent "antifreezes". Herein, we report polyproline as a minimum (bio)synthetic mimic of antifreeze proteins that is accessible by solution, solid-phase, and recombinant methods. We demonstrate that polyproline has ice recrystallisation inhibition activity linked to its amphipathic helix and that it enhances the DMSO cryopreservation of adherent cell lines. Polyproline may be a versatile additive in the emerging field of macromolecular cryoprotectants.*

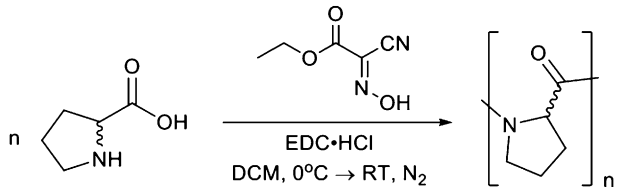
Tissue engineering, gene therapy, therapeutic protein production, and transplantation rely on the successful storage and transport of donor cells.[1] For example, in the production of therapeutic proteins, a specific cell line must be developed for each protein.[2] Given that any in vitro culture will undergo phenotypic and genotypic changes when propagated for long periods of time, it is neither possible nor practical to maintain a continuous culture of cells.[3] The only solution to this is the cryopreservation of cells using significant volumes of cryoprotectants, such as DMSO (dimethyl sulfoxide), which are intrinsically toxic.[4] The repeated use of DMSO has an impact on the epigenetic profile of cells, specifically the alteration of DNA methylation profiles, which results in phenotypic changes.[5,6] There is a real need for robust methods to cryopreserve cells in monolayer (adhered to tissue culture scaffolds) format to provide phenotypically identical cells for assays, obviating the need for replating between freeze–thaw cycles. Formulations containing 5–10% DMSO reduce cryoinjury by moderating the increase in solute concentration during freezing[7–9] but for adhered embryonic stem cells, their use results in just 5% cell recovery.[10,11] A key contributor to cell death during cryopreservation is ice recrystallisation (growth) and additives that can inhibit recrystallisation have the potential to redefine cell storage and hence biomedicine.

Antifreeze (glyco)proteins (AF(G)Ps) are potent ice recrystallisation inhibitors (IRIs), but are unsuitable for cryopreservation applications owing to their potential toxicity/immunogenicity and their secondary effect of dynamic ice shaping (DIS), which leads to needle-like ice crystals that pierce cell membranes.[12] Synthetic polymers that are potent IRIs have emerged as new tools for controlling ice growth.[13] The most studied one is poly(vinyl alcohol) (PVA), which can inhibit ice growth at concentrations below 0.1 mg mL^{-1} and enhances the cryopreservation of cells in suspension.[14–16] It is hypothesized that the activity of PVA is related to its regularly spaced hydroxyl groups.[17] Matsumura and Hyon have developed polyampholytes[18] that are cryoprotective but have moderate IRI activity.[19,20] Wang and co-workers have demonstrated the significant IRI activity of graphene oxide.[21] Ben and co-workers have developed low-molecular-weight surfactants that also inhibit ice growth.[22] A major setback is that the above synthetic IRIs are neither biodegradable nor bioresorbable and have not been applied to the significant challenge of cell monolayer storage.

There are no crystal structures for AFGPs but solution-state NMR and circular dichroism (CD) spectroscopy suggest a polyproline II (PP II)-type helix.[23] Polyproline is unique amongst the canonical amino acids in that it has no amide N–H, meaning that it cannot form intramolecular hydrogen bonds. Therefore, it is water-soluble and quite hydrophobic at the same time, as is the case for AFP I, which contains 70% alanine (a hydrophobic amino acid). We thus hypothesised that polyproline could be a minimal AF(G)P mimic owing to its amphiphilicity.[24] Homopolypeptides are appealing targets compared to vinyl polymers as they can be prepared by solid-phase synthesis,[25] solution-phase polymerisation,[26] or recombinant methods,[27] proving vast (bio)synthetic space.

Herein, we introduce polyproline as a minimum (bio)-synthetic antifreeze protein mimic. We demonstrate that polyproline has ice recrystallisation inhibition activity, which is linked to its amphipathic PP II helix structure. Polyproline was found to improve the post-cryopreservation recovery of cell monolayers compared to DMSO alone, demonstrating

[*] B. Graham, T. L. Bailey, Prof. M. I. Gibson
Department of Chemistry, University of Warwick
Gibbet Hill Road, Coventry, CV47 AL (UK)
E-mail: m.i.gibson@warwick.ac.uk
J. R. J. Healey, Prof. M. I. Gibson
Warwick Medical School, University of Warwick
Coventry, CV4 7AL (UK)
Dr. M. Marcellini, Dr. S. Deville
Ceramics Synthesis and Functionalization Lab
UMR3080 CNRS/Saint-Gobain
550 Avenue Alphonse Jauffret, 84306 Cavaillon (France)
Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.201706703.



Scheme 1. Condensation polymerisation of proline. The materials were used in stereopure form but both the L- and D-isomers were used, hence no stereocentres are shown.

a new macromolecular approach for the storage of complex cells to enable next-generation therapies.

L-, D-, and (racemic) D/L-polyproline were synthesised by condensation polymerisation using 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC, Scheme 1), alongside several commercial samples. Following dialysis, the polymers were characterised by size exclusion chromatography (SEC; Table 1). The polymers were less disperse than expected owing to fractionation during dialysis.

Table 1: Polyproline characterisation.

	M_n [g mol^{-1}]	$Đ^{SEC}$[a]	DP	Secondary structure
$PPro_{11}$	1300[a]	1.03	11	
$PPro_{15}$	1700[a]	2.12	15	PP II
$PPro_{19}$	2100[a]	1.50	19	
$P(D\text{-}Pro)_{15}$	1700[a]	1.01	15	enantiomeric PP II
$P(D/L\text{-}Pro)_{21}$	2400[a]	1.01	21	random coil
$PPro_{10\text{-}100}$	1–10 000[b]	–	10–100	PP II[e]
$PPro_{10}$	900[c]	[d]	10	PP II[e]
$PPro_{10\text{-}25}$	1–3000	1.01–1.03	10–25	PP II[e]
$PPro_{20}$	2000[c]	[d]	20	PP II[e]

[a] Determined by SEC. [b] Value from supplier. [c] Determined by mass spectrometry. [d] Single species. [e] From Ref. [28–30].

CD spectroscopy confirmed that $PPro_{15}$ adopted a PP II helix (Figure 1A; see also the Supporting Information, Figure S1)[31] with characteristic signals present at 207 and 228 nm, whilst a random coil would exhibit slight peak shifting, with signals absent in the 220 nm region.[32] $P(D\text{-}Pro)_{15}$ gave the mirror spectrum whilst the D/L racemic mixture showed no secondary structure. This series of peptides were subsequently tested for IRI activity using a splat assay.[33] This involved seeding a large number of small ice crystals, which were annealed for 30 min at −8 °C before being photographed. The average crystal size was measured relative to a PBS control, with smaller values indicating more IRI activity (Figure 1B,C).

All polyproline variants were found to display dose-dependent activity but weak molecular-weight dependence in the range tested (Figure 1B). The shortest peptide ($PPro_{10}$) lost activity below 10 mg mL^{-1}, but the longer ones retained activity at 5 mg mL^{-1}. The magnitude of activity was significantly smaller than for AF(G)Ps, which function at concentrations as low as 0.14 μg mL^{-1},[34] but comparable to that of polyampholytes.[19,20] Knight and co-workers have observed that poly(hydroxyproline) has IRI activity, which was assumed to be due to the regularly spaced hydroxyl groups along the backbone.[35] However, the observations made here suggest that the PP II helix, rather than (or in addition to) the

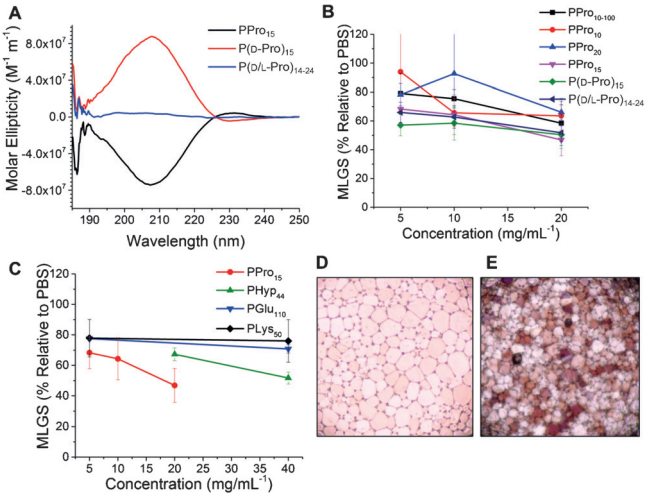


Figure 1. A) Circular dichroism spectra. B) IRI activity of the polyproline series. C) IRI activity compared to other homopolypeptides. D) Cryomicrograph of a PBS negative control. E) Cryomicrograph of 20 mg mL^{-1} polyproline. Photographs taken after 30 min at −8 °C. Error bars represent ± standard deviation from a minimum of three replicates. Images shown are 1.2 mm across. MLGS = mean largest grain size.

hydroxyl groups, gives rise to the observed activity. Figure 1C compares the IRI activity of poly(hydroxyproline) with those of $PPro_{15}$ and two α-helical poly(amino acid)s.[36] Polylysine ($PLys_{50}$) and poly(glutamic acid) ($PGlu_{110}$) showed no IRI activity. $PPro_{15}$ was found to be more active than poly(hydroxyproline) of higher molecular weight. This finding confirmed that hydroxyl groups are not essential for activity in IRI-active compounds. $P(D\text{-}Pro_{15})$ and $P(D/L\text{-}Pro_{21})$ had statistically identical activity to $PPro_{15}$, suggesting that local rather than long-range order is crucial for activity.

We hypothesise that IRI activity requires segregated hydrophilic and hydrophobic domains (amphipathy).[37,22,24] $PPro_{10}$ was compared to a non-glycosylated type I sculpin AFP[38] and also to $PGlu_{10}$ by mapping their hydrophobic/hydrophilic domains (Figure 2). The type I sculpin AFP (Figure 2A) possesses "patches" of hydrophobic/hydrophilic groups. $PPro_{10}$ (Figure 2B) also possesses this facial amphi-

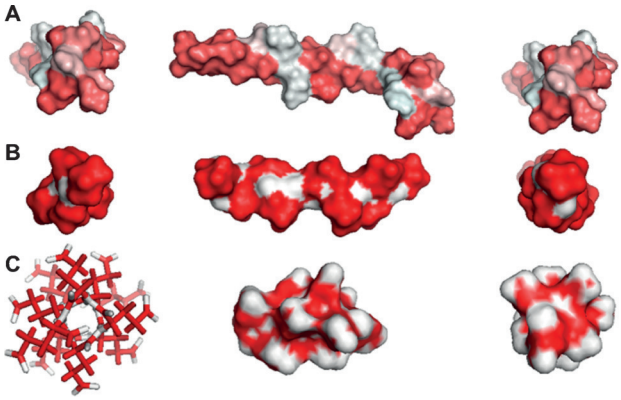


Figure 2. Hydrophobic surface mapping of A) recombinant type I sculpin AFP, B) $PPro_{10}$, and C) $PGlu_{10}$, showing charged hydrophilic surfaces. Hydrophobic regions (red), hydrophilic regions (white).

philicity. In comparison, $PGlu_{10}$ (no IRI activity) has charged groups around the core of the helix, which prevents the presentation of hydrophobic domains. This agrees with our previous study on nisin A, which has IRI activity associated with segregated domains,[37] and also the results obtained with amphiphiles developed by Ben et al., which only function below the critical micelle concentration.[22]

Aside from IRI activity, AF(G)Ps display unwanted ice shaping, which promotes the formation of needle-like ice crystals, which damage cell membranes.[12] Cryo-confocal microcapillary microscopy has emerged as a tool for monitoring ice crystal shaping,[39] and was also employed here (Figure 3). A non-IRI-active dye, sulforhodamine B, provided

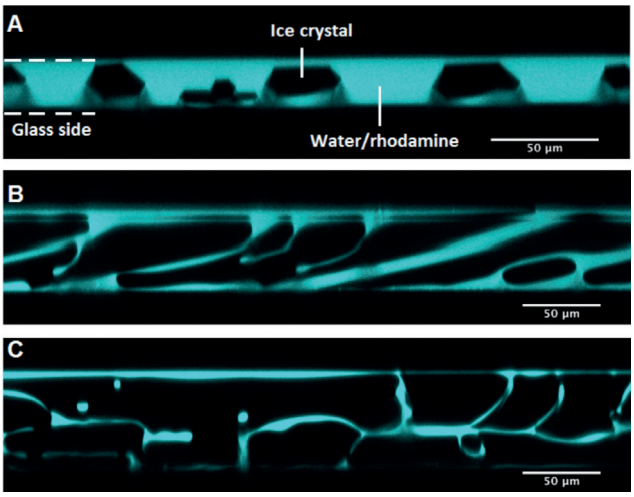


Figure 3. Cross-section of ice crystals perpendicular to the temperature gradient: A) ZrAc (positive control), B) $PPro_{19}$, C) PBS (negative control). The ice crystals expel the dye while growing, appearing in black, while the remaining liquid fluoresces.

contrast against the ice (which appears dark). A control using pure PBS showed no shaping whilst zirconium acetate (ZrAc), which is a strong ice shaper, produced hexagonal crystals.[39] $PPro_{19}$ did not induce shaping, supporting the concept that polyproline inhibits ice crystal growth without inhibiting the formation of a specific plane of ice; however, as these are relatively weak IRIs, the concentrations needed for ice shaping would be very high.

To explore polyproline as a macromolecular cryopreservative, A549 cells were employed as a prototypical adherent cell line.[40] The protective osmolyte proline (which has no IRI activity; see the Supporting Information) was used as a secondary cryoprotectant. A549 cells were incubated with 200 mM (23 $mg\,mL^{-1}$) proline (blue bars; Figure 4) or medium alone (red bars; Figure 4) for 24 h. The medium was then removed and replaced with a medium containing 10% DMSO with varying concentrations of $PPro_{11}$ (1250 $g\,mol^{-1}$, $Đ = 1.03$). After 10 min exposure to this solution, all excess solvent was removed, and the cells were subjected to controlled-rate freezing at 1 °C min^{-1} to −80 °C. Following storage at −80 °C, the cells were thawed by addition of warm medium (37 °C), and the total number of viable cells was determined by trypan blue staining 24 h after thawing.

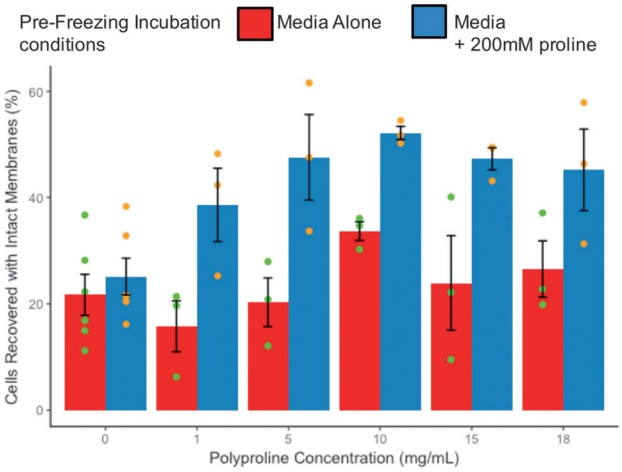


Figure 4. A549 cryopreservation. Cell recovery determined by trypan blue assays. Cells were first incubated either in the medium alone or with 200 mM proline for 24 h. They were subsequently cryopreserved by addition of 10% DMSO with the indicated $PPro_{11}$ concentration. Error bars ± S.E.M. from $n = 3$ with two nested replicates. # $P < 0.05$ compared to 10% DMSO treatment; * $P < 0.05$ compared to 200 mM proline exposure with 10% DMSO treatment.

Figure 4 shows that the use of DMSO alone led to 27% cell recovery. Addition of polyproline alone to 10% DMSO failed to give any additional protection. However, for cells that had been preconditioned with 200 mM proline for 24 h before treatment with 10 $mg\,mL^{-1}$ $PPro_{11}$/10% DMSO, the cell recovery doubled to 53%. Increasing the concentration of polyproline beyond 10 $mg\,mL^{-1}$ did not increase recovery further, suggesting that the additive benefits plateau at 10 $mg\,mL^{-1}$.[14] It should be highlighted that the cell viability assays measure intact cells, and that detailed functional analysis will be needed in the future for demonstration of complex function. For comparison with other macromolecular cryopreservatives, Matsumura and co-workers have reported poly(ampholyte)-enhanced monolayer storage using vitrification solutions, giving near-quantitative cell recovery.[41] However, this required very high DMSO concentrations of 6.5 M (> 500 $mg\,mL^{-1}$) plus 10 wt% (ca. 100 $mg\,mL^{-1}$) of the polymer, and there was a reduction in the post-thaw proliferation rate associated with the large solvent volumes, which may limit practical applications. In our PPro system introduced here, the total recovery levels were less, but far lower concentrations of DMSO were employed (10 wt%/ca. 100 $mg\,mL^{-1}$), and the total exposure time to this potentially toxic component was only 10 min. To critically compare PPro, another batch ($PPro_{10-25}$) was synthesised and tested for cytotoxicity and heamocompatibility. A549 monolayers were exposed to PPro for 24 h, and the cell viability was assessed (see the Supporting Information). This extended exposure period led to a reduction in alamar blue to 60% for 5 $mg\,mL^{-1}$ PPro, suggesting some cytotoxicity if exposed to elevated concentrations for long periods of time. It is important to note that in this cryopreservation procedure, PPro is only in contact with the cells for 10 min before the excess is removed and the cells are frozen. Red blood cell heamolysis experiments (see the Supporting Information)

showed this was not due to any inherent membrane activity of the (amphipathic) PPro.

In summary, we have demonstrated that polyproline is a potent additive for cell-monolayer cryopreservation when appropriate freezing conditions are employed. Polyproline has moderate ice recrystallisation inhibition activity, which was hypothesised to be due to its "patchy" amphipathic structure associated with its PP II helix. Addition of polyproline to adherent cell cultures led to an increase from 20% to >50% in total cell recovery post-cryopreservation, which is significantly better than for the use of DMSO alone. This increase in recovery is thought to be associated with the inhibition of ice recrystallisation. Short exposure times of just 10 min to the polyproline/DMSO solution, followed by removal of the excess solvent, reduced the cytotoxicity associated with long-term (24 h) exposure to elevated levels of polyproline. The minimal solvent exposure times may give benefits in downstream processing and biomedical applications compared to current high-solvent-concentration methods using vitrification. Polyproline is appealing compared to other macromolecular cryoprotectants as it only comprises native amino acids and can be obtained by chemical and biochemical methods.

Acknowledgements

This study has received funding from ERC grants (CRYOMAT 638661, 278004 FreeCo), the BBSRC (BB/F011199/1), and the Royal Society. The University of Warwick WCPRS partially supports T.L.B. J.R.J.H. thanks the EPSRC for funding via MOAC DTC EP/F500378/1. M. Menze is thanked for providing the Biocision CoolCell to enable controlled-rate freezing.

Conflict of interest

The authors declare no conflict of interest.

Keywords: biomaterials · cryopreservation · ice recrystallization inhibitors · monolayers · polymers

How to cite: *Angew. Chem. Int. Ed.* **2017**, *56*, 15941–15944
Angew. Chem. **2017**, *129*, 16157–16160

[1] A. Fowler, M. Toner, *Ann. N. Y. Acad. Sci.* **2005**, *1066*, 119–135.
[2] G. Walsh, *Nat. Biotechnol.* **2014**, *32*, 992–1000.
[3] G. Seth, *Methods* **2012**, *56*, 424–431.
[4] K. Brockbank, M. Taylor, *Adv. Biopreserv.* **2007**, *5*, 157–196.
[5] M. Iwatani, K. Ikegami, Y. Kremenska, N. Hattori, S. Tanaka, S. Yagi, K. Shiota, *Stem Cells* **2006**, *24*, 2549–2556.
[6] K. Kawai, Y.-S. Li, M.-F. Song, H. Kasai, *Bioorg. Med. Chem. Lett.* **2010**, *20*, 260–265.
[7] P. Mazur, *Science* **1970**, *168*, 939–949.
[8] P. Mazur, J. Farrant, S. P. Leibo, E. H. Chu, *Cryobiology* **1969**, *6*, 1–9.
[9] X. Stéphenne, M. Najimi, E. M. Sokal, *World J. Gastroenterol.* **2010**, *16*, 1–14.
[10] C. H. Boon, P. Y. Chao, H. Liu, S. T. Wei, A. J. Rufaihah, Z. Yang, H. B. Boon, Z. Ge, W. O. Hog, H. L. Eng, T. Cao, *J. Biomed. Sci.* **2006**, *13*, 433–445.
[11] Q. Xu, W. J. Brecht, K. H. Weisgraber, R. W. Mahley, Y. Huang, *J. Biol. Chem.* **2004**, *279*, 25511–25516.
[12] H. Chao, P. L. Davies, J. F. Carpenter, *J. Exp. Biol.* **1996**, *199*, 2071–2076.
[13] M. I. Gibson, *Polym. Chem.* **2010**, *1*, 1141–1152.
[14] R. C. Deller, M. Vatish, D. A. Mitchell, M. I. Gibson, *Nat. Commun.* **2014**, *5*, 3244.
[15] B. Wowk, E. Leitl, C. M. Rasch, N. Mesbah-Karimi, S. B. Harris, G. M. Fahy, *Cryobiology* **2000**, *40*, 228–236.
[16] R. C. Deller, J. E. Pessin, M. Vatish, D. A. Mitchell, M. I. Gibson, *Biomater. Sci.* **2016**, *47*, 935–945.
[17] C. Budke, T. Koop, *ChemPhysChem* **2006**, *7*, 2601–2606.
[18] K. Matsumura, S. H. Hyon, *Biomaterials* **2009**, *30*, 4842–4849.
[19] D. E. Mitchell, M. Lilliman, S. G. Spain, M. I. Gibson, *Biomater. Sci.* **2014**, *2*, 1787–1795.
[20] D. E. Mitchell, N. R. Cameron, M. I. Gibson, *Chem. Commun.* **2015**, *51*, 12977–12980.
[21] H. Geng, X. Liu, G. Shi, G. Bai, J. Ma, J. Chen, Z. Wu, Y. Song, H. Fang, J. Wang, *Angew. Chem. Int. Ed.* **2017**, *56*, 997–1001; *Angew. Chem.* **2017**, *129*, 1017–1021.
[22] C. J. Capicciotti, M. Leclere, F. A. Perras, D. L. Bryce, H. Paulin, J. Harden, Y. Liu, R. N. Ben, *Chem. Sci.* **2012**, *3*, 1408–1416.
[23] D. H. Nguyen, M. E. Colvin, Y. Yeh, R. E. Feeney, W. H. Fink, *Biophys. J.* **2002**, *82*, 2892–2905.
[24] D. E. Mitchell, G. Clarkson, D. J. Fox, R. A. Vipond, P. Scott, M. I. Gibson, *J. Am. Chem. Soc.* **2017**, *139*, 9835–9838.
[25] R. B. Merrifield, *J. Am. Chem. Soc.* **1963**, *85*, 2149.
[26] M. I. Gibson, N. R. Cameron, *J. Polym. Sci. Part A* **2009**, *47*, 2882–2891.
[27] E. Gutierrez, B. S. Shin, C. J. Woolstenhulme, J. R. Kim, P. Saini, A. R. Buskirk, T. E. Dever, *Mol. Cell* **2013**, *51*, 35–45.
[28] A. A. Adzhubei, M. J. E. Sternberg, A. A. Makarov, *J. Mol. Biol.* **2013**, *425*, 2100–2132.
[29] P. Wilhelm, B. Lewandowski, N. Trapp, H. Wennemers, *J. Am. Chem. Soc.* **2014**, *136*, 15829–15832.
[30] A. V. Mikhonin, N. S. Myshakina, S. V. Bykov, S. A. Asher, V. Pennsyl, *J. Am. Chem. Soc.* **2005**, *127*, 7712–7720.
[31] Protein Circular Dichroism Data Bank **2016**, pCD0004553000.
[32] J. L. S. Lopes, A. J. Miles, L. Whitmore, B. A. Wallace, *Protein Sci.* **2014**, *23*, 1765–1772.
[33] T. Congdon, R. Notman, M. I. Gibson, *Biomacromolecules* **2013**, *14*, 1578–1586.
[34] S. Lui, W. Wang, E. von Moos, J. Jackman, G. Mealing, R. Monette, R. N. Ben, *Biomacromolecules* **2007**, *8*, 1456–1462.
[35] C. A. Knight, D. Wen, R. A. Laursen, *Cryobiology* **1995**, *32*, 23–34.
[36] M. I. Gibson, C. A. Barker, S. G. Spain, L. Albertin, N. R. Cameron, *Biomacromolecules* **2009**, *10*, 328–333.
[37] D. E. Mitchell, M. I. Gibson, *Biomacromolecules* **2015**, *16*, 3411–3416.
[38] A. H. Kwan, K. Fairley, P. I. Anderberg, C. W. Liew, M. M. Harding, J. P. Mackay, *Biochemistry* **2005**, *44*, 1980–1988.
[39] M. Marcellini, C. Noirjean, D. Dedovets, J. Maria, S. Deville, *ACS Omega* **2016**, *1*, 1019–1026.
[40] B. Stokich, Q. Osgood, D. Grimm, S. Moorthy, N. Chakraborty, M. A. Menze, *Cryobiology* **2014**, *69*, 281–290.
[41] K. Matsumura, K. Kawamoto, M. Takeuchi, S. Yoshimura, D. Tanaka, S.-H. Hyon, *ACS Biomater. Sci. Eng.* **2016**, *2*, 1023–1029.

Manuscript received: July 4, 2017
Revised manuscript received: September 27, 2017
Accepted manuscript online: October 18, 2017
Version of record online: November 22, 2017

Document check over: Change captions to small caps as per fig 1.5 in intro

Document check over: ensure cite citations are being included as with citep

Document check over: go back and change citep to cite where the paper has been referred to directly in the text to get rid of unecessary brackets

Document check over: ensure all tables (particularly in Methods) have references if needed

Revisit captions, particularly in phylogenetics chapter to make them more descriptive.

Check formatting of all references, ensure no duplicates (e.g. Smith2000, and Smith2000a relating to the same article), and ensure all in-text citations resolve fully (including captions)

Document check over: Complete appendices including: All HHpred hits/Scripts/annotate multiple seq alignments

Document check over: finish declaration and abstract

Document check over: ensure consistent capitalisation of headings

Epigraphs for all chapters

Ensure all hyperreferences resolve correctly

Replace gray!10 everywhere so that it reproduces on paper better

To Do