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 Protein Delivery Mechanism**
DATE OF DEPOSIT:	**September, 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:

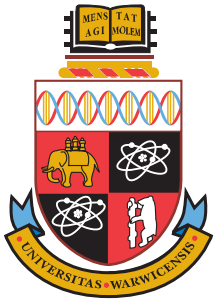
> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information 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 Protein Delivery Mechanism

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

Acknowledgements

This was a multidisciplinary PhD, and consequently I have many, many people to thank for making this thesis a reality, both formal and informal.

Formally, I'd like to acknowledge the Molecular Organisation and Assembly in Cells (MOAC) Centre for Doctoral Training and the Engineering and Physical Sciences Research Council for giving me the environment and funding to complete my studentship under grant EP/F500378/1. MOAC has been a thoroughly enjoyable experience, and simultaneously, quite possibly the hardest thing I've ever done. I cannot recommend the CDT model strongly enough for anyone that wants to undertake a PhD. On a more individual basis, thanks go to Prof. Alison Rodger, Dr. Hugo van Den Berg and Dr. Nikola Chmel for giving me the opportunity to join, quite frankly, one hell of a community. Thank you also to the indispensable Mrs. Naomi Grew, for putting up with my (and no doubt everyone else's) incessant requirements!

Thanks also go to everyone within the CDT, especially the 2013 cohort, for making the last 4 years so enjoyable. We've had some excellent times at away days and residential courses, and an abundance of stories to tell, whether that be from walking for an hour and a half across London on our way back from a night out in Soho, to watching one another fall in the river Severn whilst canoeing, and a great many more that are probably best left out of a public discourse!

Next, I am of course, indebted to my supervisors, Dr. Nick Waterfield and Prof. Matt Gibson. It has been an absolute pleasure to work with two of the most enthusiastic and insightful scientists I've ever met. Every meeting left another dozen PhDs worth of ideas and work to be done, and this enthusiasm was infectious. This attitude to discovery and science in general, is everything I think of when I picture an exemplary scientist. More

specifically, thanks to Nick for putting together the project for a lowly Masters student who turned up his office one day, and allowing me to 'own' the project from the outset - and for letting me run off around the world at various points attempting to turn his ideas in to a company! On the topic of turning up out of the blue, thanks also go to Matt for agreeing to supervise an unusual project that was somewhat outside the scope of the lab's research at the time, and for really throwing him self into the project. I hope I get to continue this project with you both far in to the future in some form or another.[1]

Additionally I'd like to acknowledge everyone in both of my labs, past and present for providing advice, help, and the camaraderie required to get through a PhD! Particularly, thanks go to Dr. Alexia Hapeshi for all her help throughout my Masters and PhD, a constant source of excellent advice and support in the lab.

Though it doesn't get much 'air-time' in this thesis, a journey as exciting as my PhD unfolded along side it, which was my foray in to the world of biotechnology and commercialisation of research. For their encouragement, advice, and support in this process, I have to thank Dr. Laura Lane and Dr. James Lapworth from Warwick Ventures. I hope we get to continue our world tour of jazz bars for a good while to come yet!

In no particular order, I'd like to acknowledge Dr. Andy Millard for his support with bioinformatics tasks, and for agreeing to participate in my advisory panel. Similarly, thanks to Dr. Chris Corre for being on the advisory panel and helping out with submission of projects to the Warwick Synthetic Biology centre; and once again to Dr. Nikola Chmel for also participating in the advisory panel and helping out with Circular Dichroism experiments. I would like to acknowledge the Medical Research Council's Cloud Infrastructure for Microbial Bioinformatics, particularly Marius Bakke, for allowing me access to computer resources that 'proper' computational scientists would kill for! I also wish to thank Chris Thoroughgood, Dr. Avinash Punekar and Prof. David Roper for help and access to equipment for protein purification and crystallography experiments.

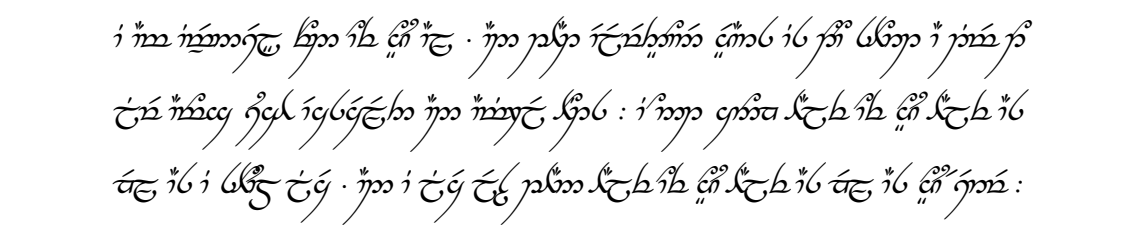
Last but by no means least, I'd like to thank three of my mentors from the past. I feel like this doesn't happen as often as it should in many acknowledgement sections, and I intend to rectify that - in fact, before this thesis was so much as a scribble on a page or a byte on my laptop, I knew this section would be written. I can categorically

[1]Not to mention, they both make excellent drinking companions!

say, and without any sense of exaggeration, that I would not have gotten this far without their influences, and they are entirely the reason I chose to pursue the biological sciences. Firstly, Mrs. Elena Segalini-Bower, my GCSE and A Level Biology teacher encouraged me to apply to university and set me on the path of the biological sciences in the very beginning. Without her, I would never have even made it to undergraduate! She is a fabulous teacher (and in my opinion was absolutely robbed of the Royal Society of Biology teacher of the year award in July 2015!), and has no doubt influenced many students since me to come this far! She challenged me at every opportunity as a student, and even if I was less than thrilled about it at the time (I wanted easy homework too!), it has been the absolute making of me as a student and scientist. I really can't express my gratitude enough in words, so I won't labour the point further!

For all the same reasons, Dr. Justin Pachebat and Prof. Luis Mur, my undergraduate tutor/supervisors were instrumental in my understanding and appreciation of the biological sciences. If Mrs. Bower is the reason I made it to undergraduate, there is no doubt in my mind that Justin and Luis are the reason I made it to postgraduate. They pushed me academically, and provided me with incredible freedom and opportunities for my dissertation project which were unequivocally the catalyst for me becoming hooked on experimental biological research, and they supported me every step of the way.

Thank you once again to everyone who has influenced my journey, I wish I had the space to name everyone individually, but I think 3 pages is quite enough!



Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It is an entirely novel work of the authors own creation, 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 was conducted by the author except in the cases outlined below.

Assistance with protein purification (including running FPLC-IMAC and Gel Filtration processes), as well as Crystallography screening and diffraction testing, was kindly performed with Dr. Avinash Punekar and Dr. Chris Thoroughgood. Thanks go to Dr. Alexia Hapeshi for all the NGS data generated in sequencing constructs and for new genomes. A Masters student, Katie Smart, created the promoter constructs utilised in Chapter 6 on page 234.

Publications arising from this candidature (see Chapter A on page 330):

- Graham, B., Bailey, T. L., Healey, J. R. J., Marcellini, M., Deville, S. and Gibson, M. I., **(2017)**. "Polyproline is a minimal antifreeze protein mimetic and enhances the cryopreservation of cell monolayers." **129**, 16157-16160 *Angewandte Chemie International Edition* DOI: 10.1002/ange.201706703

Author Contribution: Creation of simulated structural and modelling data for exemplifying amphipathy of antifreeze protein mimetics.

"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. Feynman was substantially ahead of his time in predicting not only what could be achieved with biology, but also making astute predictions about how they may manifest. Not bad for a physicist!

Abstract

The "*Photorhabdus* Virulence Cassette" (PVC) is an elaborate macromolecular protein complex evolved to specifically and potently deliver effector molecules to the interior of target cells. A PVC can be thought of as a "headless bacteriophage", with a protein cargo rather than a nucleic acid one. This thesis sheds light on some of the 'dark matter' in the content and operon structure of PVCs, though mysteries remain. The Introduction guides the reader through the universe of similar protein complexes. The first 2 results chapters 'pave the way' for lab experiments by examining the PVCs in a number of computational workflows. In Chapter 3, the disconnect between sequence and structure for proteins is examined - and what this means for PVCs. Through sensitive methods, less dependent on sequence, new, informative homologies are detected, and we get a 'first look' at the likely structure for many elements of a PVC. Chapter 4 explores the operon structure of PVCs, identifying proteins which may be non-essential to PVC function, and demonstrating the 'microevolution' in, and variability between, operons. In the first of the 'wet lab' chapters, Chapter 5 examines the enigmatic proposed 'tail fibres' of the PVCs, identified as hypervariable in Chapter 4. This chapter represents the first experimental study of a potentially unique chimeric fibre protein, and provides the first experimental data confirming their true nature as *bona fide* tail fibres. Finally, Chapter 6 details the efforts made to heterologously clone controllable PVCs - a non-trivial task it transpires; and to understand the regulation and population dynamics of how PVCs are deployed by *Photorhabdus* naturally, with preliminary observations implicating a role for RfaH-like transcriptional regulation proteins and anti-termination/operon polarity suppression.

Abbreviations

μF	MicroFarads
1°	Primary
2°	Secondary
AAA+ ATPase	ATPases Associated with diverse cellular Activities
ATCC	American Type Culture Collection
AWC	Adjusted Wallace Coefficient
Afp/AfpX	Antifeeding prophage
BLAST	Basic Local Alignment Search Tool
C-score	Confidence Score
CAR	Coxsackie Adenovirus Receptor
CDS	Coding DNA Sequence
DMSO	DiMethylSulfOxide
DTT	DiThioThreitol
EDTA	EthyleneDiamineTetraacetic Acid
ELISA	Enzyme-Linked ImmunoSorbant Assay
EM	Electron Microscopy/Micrograph
EMDB	Electron Microscopy DataBank
F	Forward
FITC	Fluorescein IsoThioCyanate
FPLC	Fast Protein Liquid Chromatography
FRT	Flippase Recognition Target

GFP	Green Fluorescent Protein
HEPES	4-(2-HydroxyEthyl)-1-PiperazineEthaneSulfonic acid
HPLC	High Performace Liquid Chromatography
HSI	Hue/Saturation/Intensity colourspace
IJ	Infective Juvenile nematode
IMAC	Immobilised Metal ion Affinity Chromatography
INDEL	INsertion/DELetion in a sequence
LD_{50}	Lethal Dose at 50% mortality
LPS	LipoPolySaccharide
MAC	Metamorphosis Associated Complex
MSA	Multiple Sequence Alignment
Mbp	Mega-basepairs
NEB	New England Biolabs
NMR	Nuclear Magnetic Resonance
NRMSD	Normalised Root Mean Square Deviation
NTA	NitriloTriacetic Acid
PAAR	Proline-Alanine-Alanine Repeat protein
PCR	Polymerase Chain Reaction
PDB	Protein DataBank
PLTS	Phage-Like Translocation System
PVC	*Photorhabdus* Virulence Cassette
PVDF	Polyvinylidene Fluoride
R	Reverse
RCF	Relative Centrifugal Force
RMSD	Root Mean Square Deviation
RVA	Rapid Virulence Annotation
SDS-PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis

SOC	Super Optimal media with Catabolise Repression
T6SS	Type 6 Secretion System
TAE	Tris-Acetate-EDTA
TBC	Tube-Baseplate Complex
TEMED	TetraMethylEthylene
TES	Transcript Elongation Complex
TM-score	Template Modelling Score
TVISS	see T6SS
TXSS	Type *x* Secretion System
Tm	Melting/Annealing Temperature
Tss	Type Six Subunit
UV	UltraViolet
Vgr	Valine-glycine repeat protein
bp	basepairs
cif	Cell-cycle Inhibition Factor
cnf	Cytonecrosis Factor
dsDNA	double-stranded DNA
gDNA	genomic DeoxyriboNucleic Acid
gp##	Phage T4 gene product ##
hcp	hemolysin co-regulated protein
kDa	kilo-Daltons
kV	kilo-Volts
kbp	kilo-basepairs
lopt	*Photorhabdus* YopT-like toxin
lumt	Large Unknown Mosaic Toxin
nRF	Normalised Robinson-Foulds distance
ops	Operon Polarity Suppressor

pADAP	Amber Disease Associated Plasmid
pnf	*Photorhabdus* Necrosis Factor
ppGpp	Guanosine pentaphosphate
qRT-PCR	quantitative Reverse-Transcriptase Polymerase Chain Reaction
ssRNA	single-stranded RNA

Note to the reader

Welcome to the future! This thesis contains elements which can be viewed in augmented reality! (I know right?!). If you want to view these features, you'll need a smart phone or tablet device (the more powerful the better as some figures are complex), and you'll need to download the "Augment" app. It can be found in the the Google Play Store or Apple App Store (`www.augment.com`) - don't worry, it's free!

Any time you see the symbol below on a page (bottom left footer), you can scan that page with the app, and see the model spring to life on the page! (Note, that due to some rendering limitations of mobile devices 'horsepower' and reduced resolution, the augmented reality image may not entirely resemble the image on the page, and in cases where the render would have been too large, the image might have been replaced with a related one for simplicity). Experiment with moving the 'camera' (your mobile device) as well as the physical page itself!

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, nano-scale, protein secretion and delivery system 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 two ways. Firstly, the PVCs are of interest from a fundamental biology standpoint, given their unusual nature and proposed purpose. 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, though not dwelt on in this thesis particularly, is that the PVCs represent a potentially interesting new mechanism for delivery of protein cargos to cells of interest, and therefore a 'next-generation' drug delivery tool. At the same time as all the work that will be presented in this thesis was ongoing, efforts to develop the idea commercially were also undertaken.

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 three 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 one (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, for instance the "HIT" and "JUN" European isolates (Mulley *et al.*, 2015).

In fact, underscoring the point made in the first paragraph, even after this section was first written, a large polyphasic study (combining sequence typing, proteomics, DNA hybridisation, determinative bacteriology etc.) by Machado *et al.* (2018) was released which has proposed that the diversity within the genus should promote a number of subspecies to species in their own right, as well as recognise a number of new subspecies.

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); while 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 preliminary work (unpublished) has demonstrated that the core genome of the clade may comprise some 1625 chromosomal genes (for approximately 35 available genomes) 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 change as more genomes are studied). Unsurprisingly, these stark genetic differences can manifest in substantial phenotypic differences. As mentioned, certain *P. asymbiotica* strains can infect human hosts (and possibly other mammals), 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 on the following page below shows, in a schematic manner, the host and temperature restrictions of some exemplar strains from each species.

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 translational 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

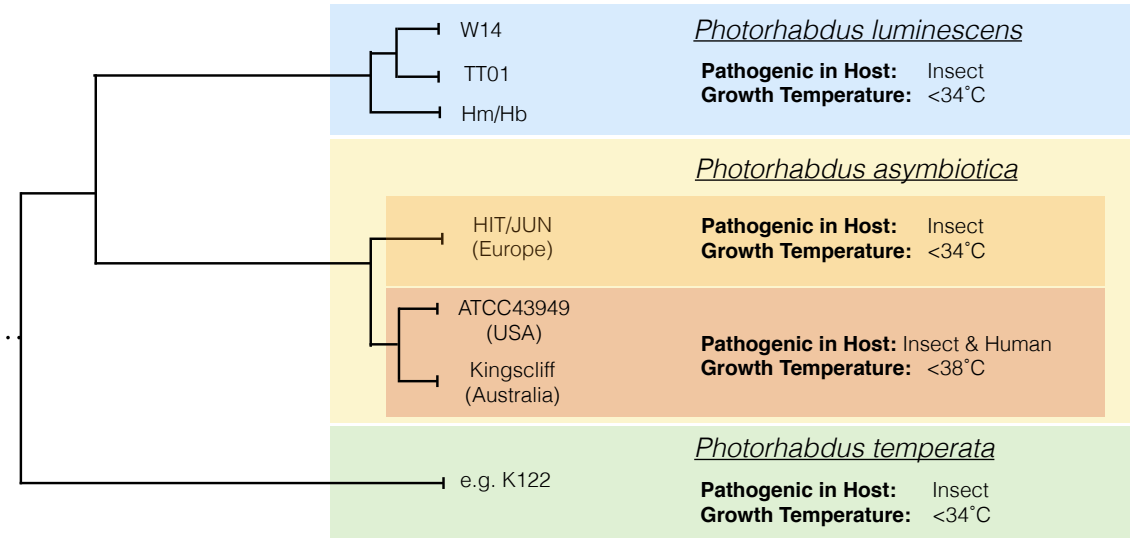


Figure 1.1 | A SCHEMATIC DIAGRAM DEPICTING THE SUBCLADES WITHIN THE *Photorhabdus* GENUS. Not drawn to scale. The schematic shows the host ranges and thermotolerance of the archetypical species within the *Photorhabdus* genus, with some exemplar strains. Adapted from (Waterfield *et al.*, 2009), and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

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 reduced genome size (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 its 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*, which, for a long while, *Photorhabdus* was thought to be unique

in being the only non-Plant organism known to produce it (Joyce *et al.*, 2008) - more recently it has also been detected as a *Bacillus* metabolite too (Kumar *et al.*, 2013). 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 its 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) (which is developmentally equivalent to the

"Dauer juvenile" of *Caenorhabditis elegans*) stage of the nematode host, which is free living and actively seeking insects to infect/parasitise.

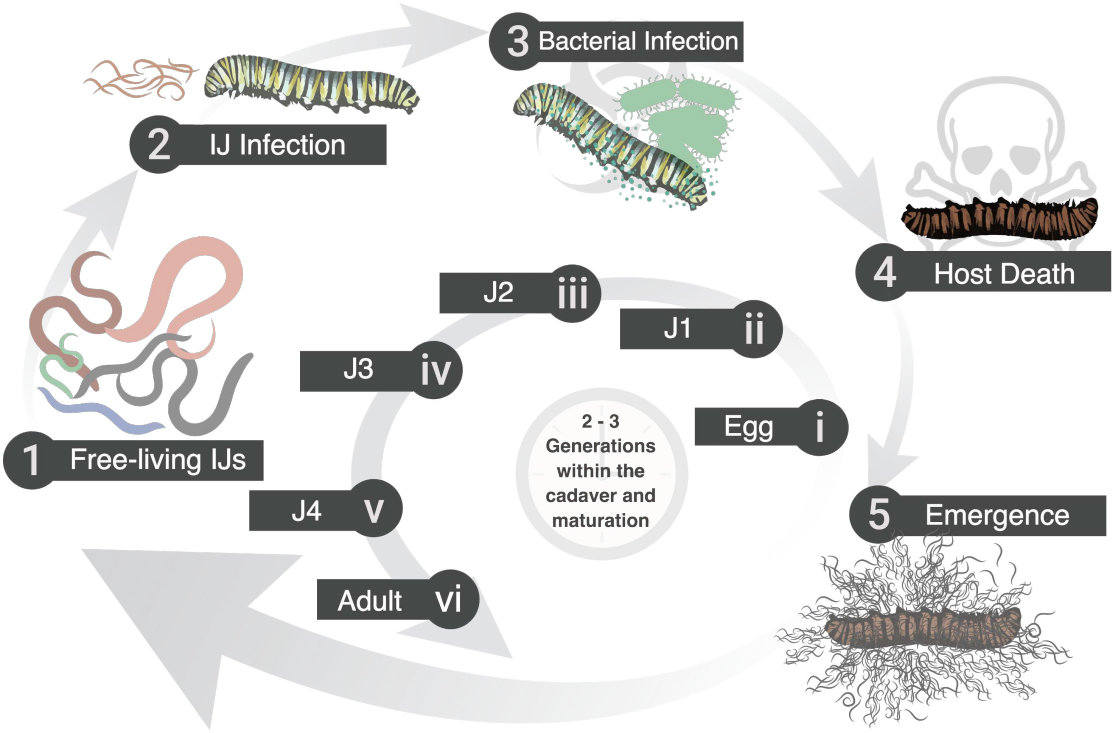


Figure 1.2 | The infection lifecycle of the Entomopathogenic Nematode complex. **(1)** The free living "Infective Juveniles" (IJs) in the soil seek out a new insect host to prey on. **(2)** IJs ingress in to the insect prey either through natural openings or boring through the cuticle. **(3)** *Photorhabdus* 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).

As Figure 1.2 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 (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 depleted, the EPN complexes vacate the site and go off in search of fresh prey, repeating the cycle.

There are a number of theories suggesting a basis 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 were defective in some way (e.g. truncations, deleted 5' regions and so on), 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 234.

The PVCs were able to be 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

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.

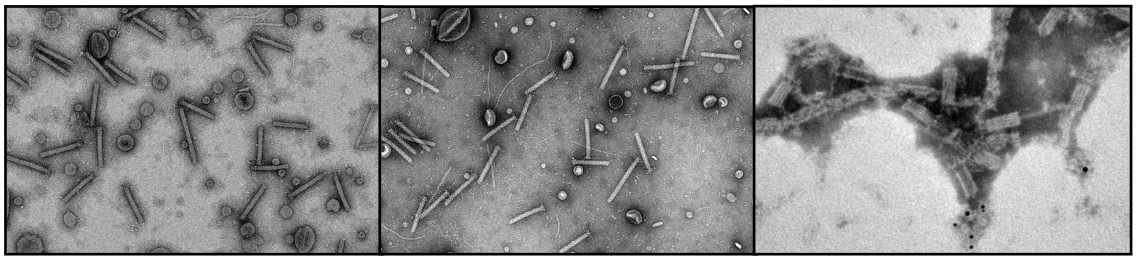


Figure 1.3 | A selection of electron micrographs of the PVCs.
A selection of micrographs are shown here revealing their caudate structure and the resemblance to other contractile tail systems. The left and centre panels show more recent, but unpublished data from an ongoing collaboration with the Max 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 (Yang *et al.*, 2006).

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 six distinct PVC operons, each with one 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 four 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*, being in fact, primarily orthologous to the *P. luminescens* TT01 "Unit 4".

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 one 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.

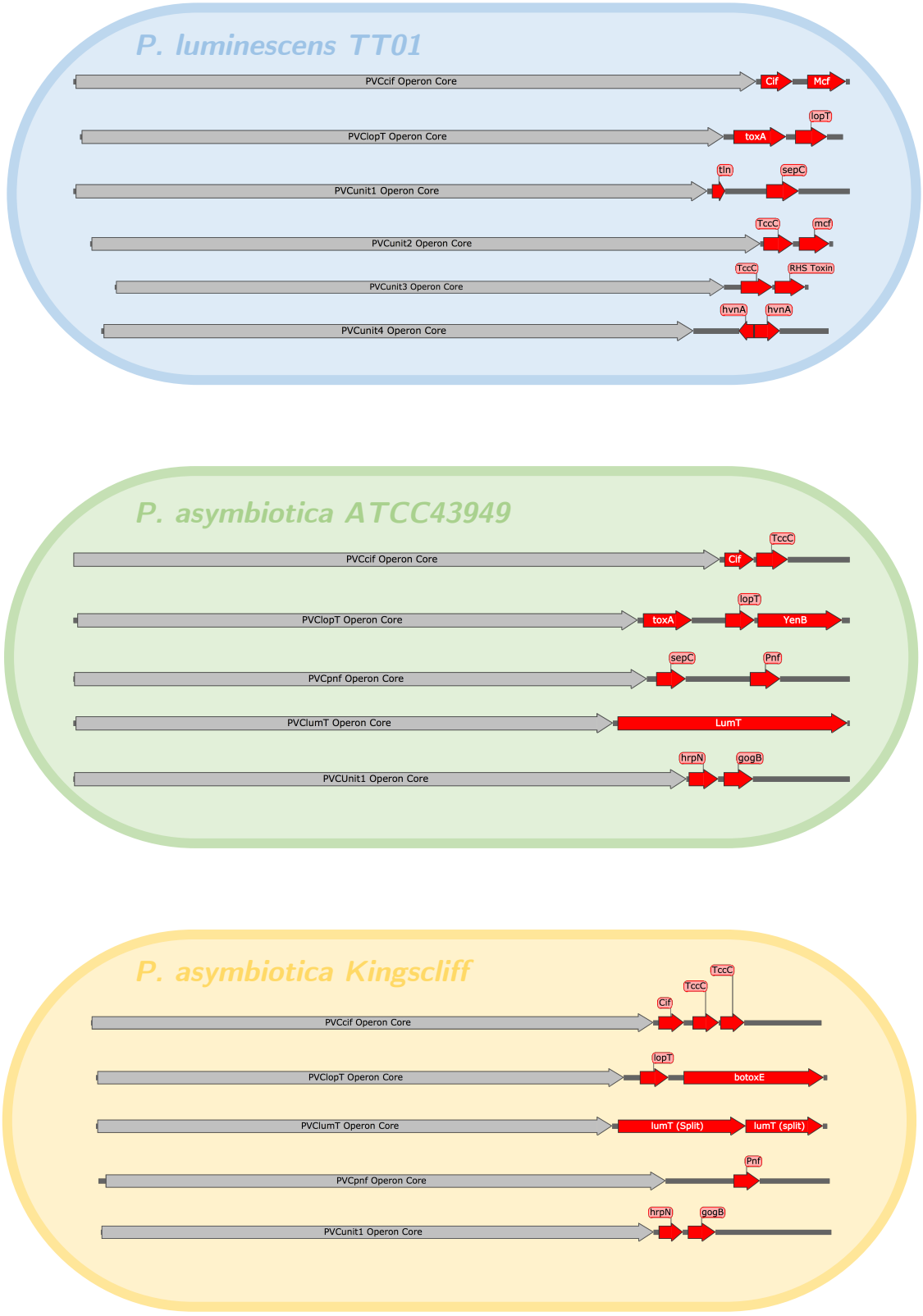


Figure 1.4 | The PVC variants found in 3 *Photorhabdus* genomes.

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

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.*, 2015), the Antifeeding prophage (Heymann *et al.*, 2013), and other prophage sequences, which have had their structures resolved or electron microscopy (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, 2015; 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 three excellent reviews from both a structural and genetic perspective (Kube and Wendler, 2015; 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.*, 2016; 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.

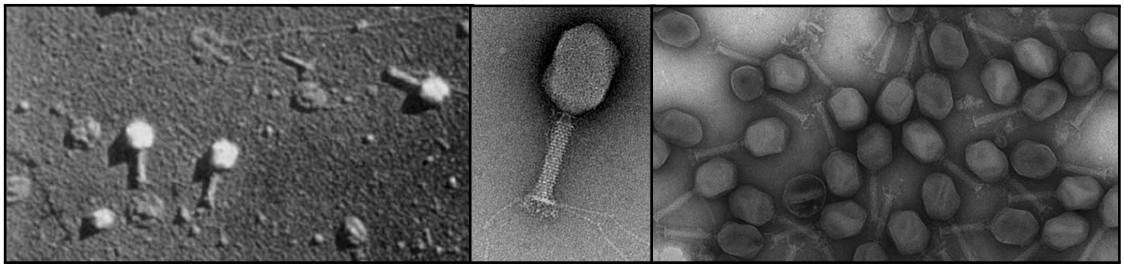


Figure 1.5 | Electron Micrographs of T4 bacteriophage virions.
The left panel shows an early T4 phage micrograph from 1958, reproduced from `https://www.molbio.unige.ch/eng/about/history`. The centre panel shows a close up of the T4 phage, revealing the tail fibres, baseplate, capsid and tube in detail, 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 ≥20 proteins, 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 ("gene product") 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 two 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 hexameric, helical, 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 subunit offset is approximately 17.2°, and in the contracted state, this twist increases to 32.9° (Kube and Wendler, 2015; 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 three 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; Katsura, 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

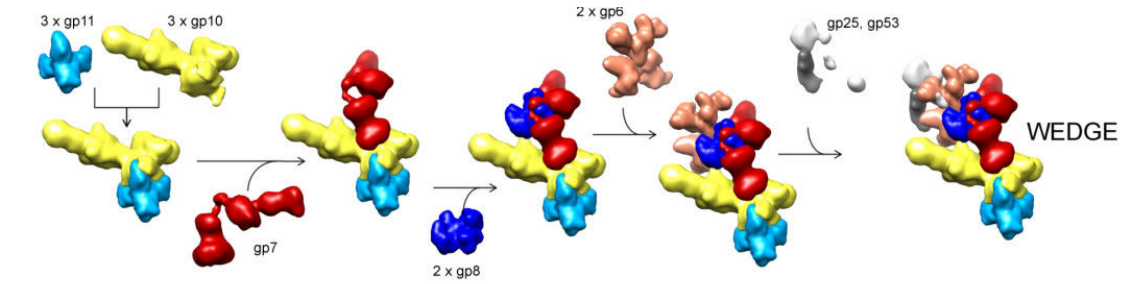
T4 baseplate complex is exceedingly intricate, comprising some 18 different protein types (including the baseplate spike/hub, and the tail fibres), and roughly 57 separate protein molecules (some of which are, themselves, made up from multiple chains). As can be seen in Figure 1.6 on page 20 from Leiman *et al.* (2004), six "wedge" complexes are formed from a gp6-gp7-gp8-gp10-gp11-gp25-gp53 complex. Each of these six wedges then come together around the baseplate hub spike complex (gp27-gp5), itself comprised of three different proteins and at least seven distinct chains. A further 12 proteins are added (six 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, 2015). 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.

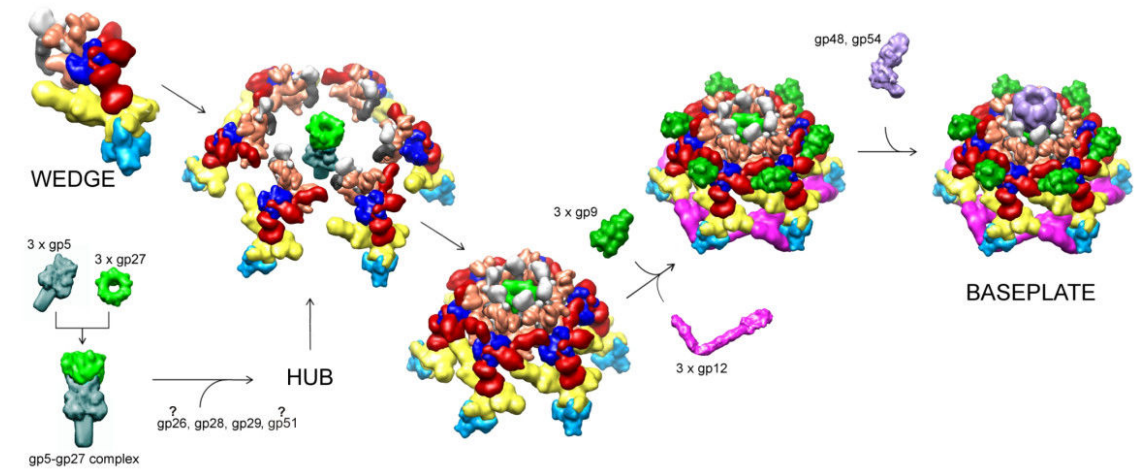
Finally, the PVCs are thought to contain putative tail-fibre like genes, proposed to be for cell binding in the same manner as T4. So far, there appears to be no evidence of both long and short fibres as is the case in T4 however (Bartual *et al.*, 2010; Thomassen *et al.*, 2003). Again, the PVCs seem to elaborate a much simpler version of these analogous structures. The long tail fibres of T4 are comprised of four 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 six trimer complexes of the gp9 protein (Figure 1.6B on the following page). At the outermost edge of the dome, the six 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 187, where a more detailed introduction to the tail fibre structure and proposed function can also be found.

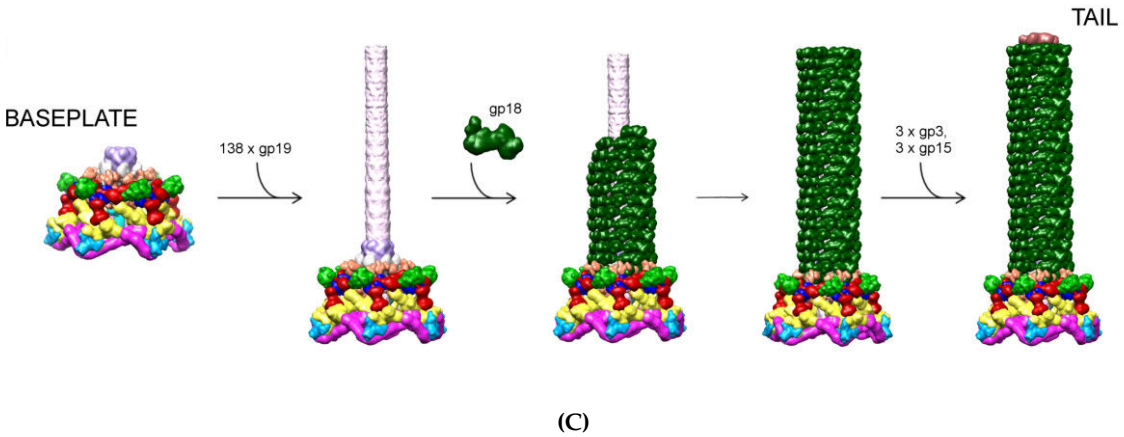
This review will not consider the assembly of the T4 capsid, as the PVCs do not contain analogous structure, but for an excellent all-round review of the full assembly of T4, see Yap and Rossmann (2014) and Leiman *et al.* (2010).



(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 (2014).

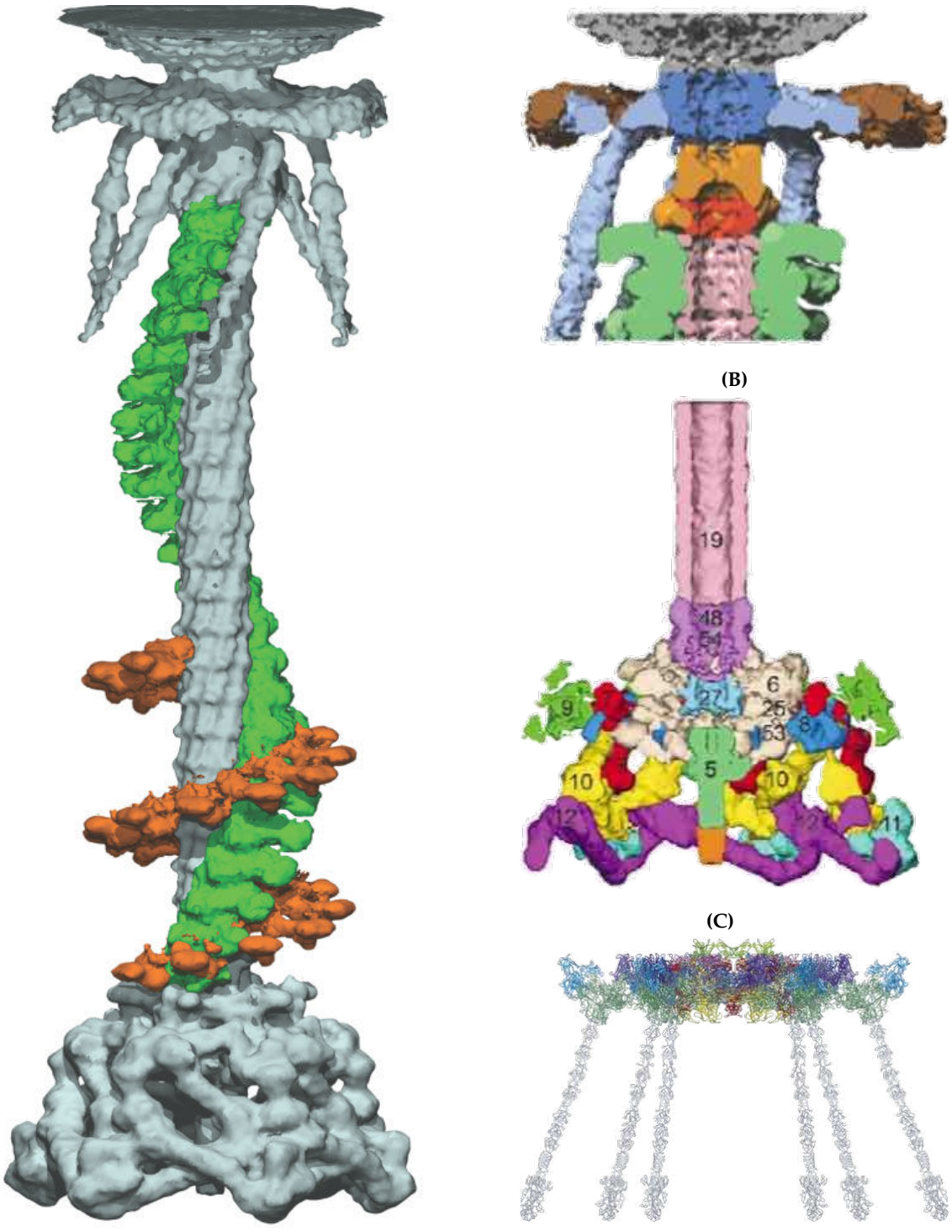


Figure 1.7 | A selection of the resolved structural components of Bacteriophage T4.
(A) The T4 EM density reproduced from Kostyuchenko *et al.* (2005), the helical outersheath protofilaments are shown in the extended (green) and contracted (orange) conformations. **(B)** The architecture of the 'neck'/'collar' region of the T4 phage, showing the top of the tube. Adapted and reproduced from Kostyuchenko *et al.* **(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.* (2016). The structure of the genome-containing capsid has been omitted, as the similarity to the tail and baseplate is more relevant. Augmented reality structures coloured according to cylinder radius (red (≤100Å) to blue (≤20Å).

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.*, 2015) (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.

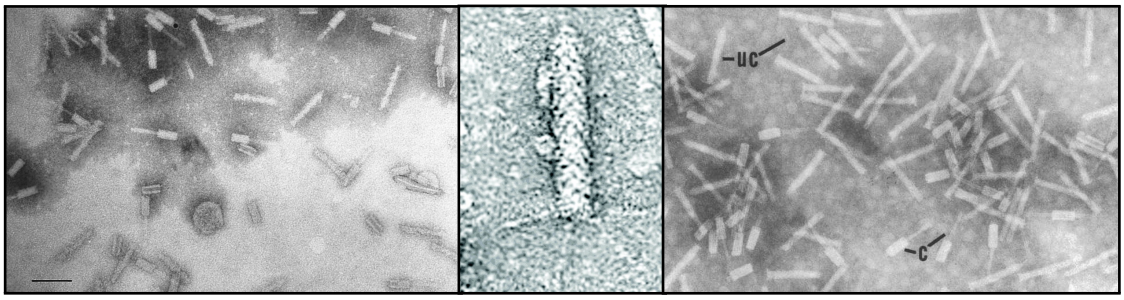


Figure 1.8 | Electron Micrographs of *Pseudomonas* R-type pyocin particles
The left panel shows a number of R-type pyocin molecules purified. The contracted nature of several of the particles reveals clearly the size difference between the inner and outer sheaths. Reproduced from Lee *et al.* (1999). 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 structurally resemble something of a 'halfway house' between phage and PVC-like systems; they have 'streamlined' genetics, by way of removal of the capsid genes and the associated replicative machinery, though the evolutionary relationships between them remain unknown. 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* is 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 a 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 Paragraph 1.2.3.5.1 on page 47 and Paragraph 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 28).

The seminal paper which finally resolved the intricacies of the structure of the R-type pyocins was that of Ge *et al.* (2015). 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.

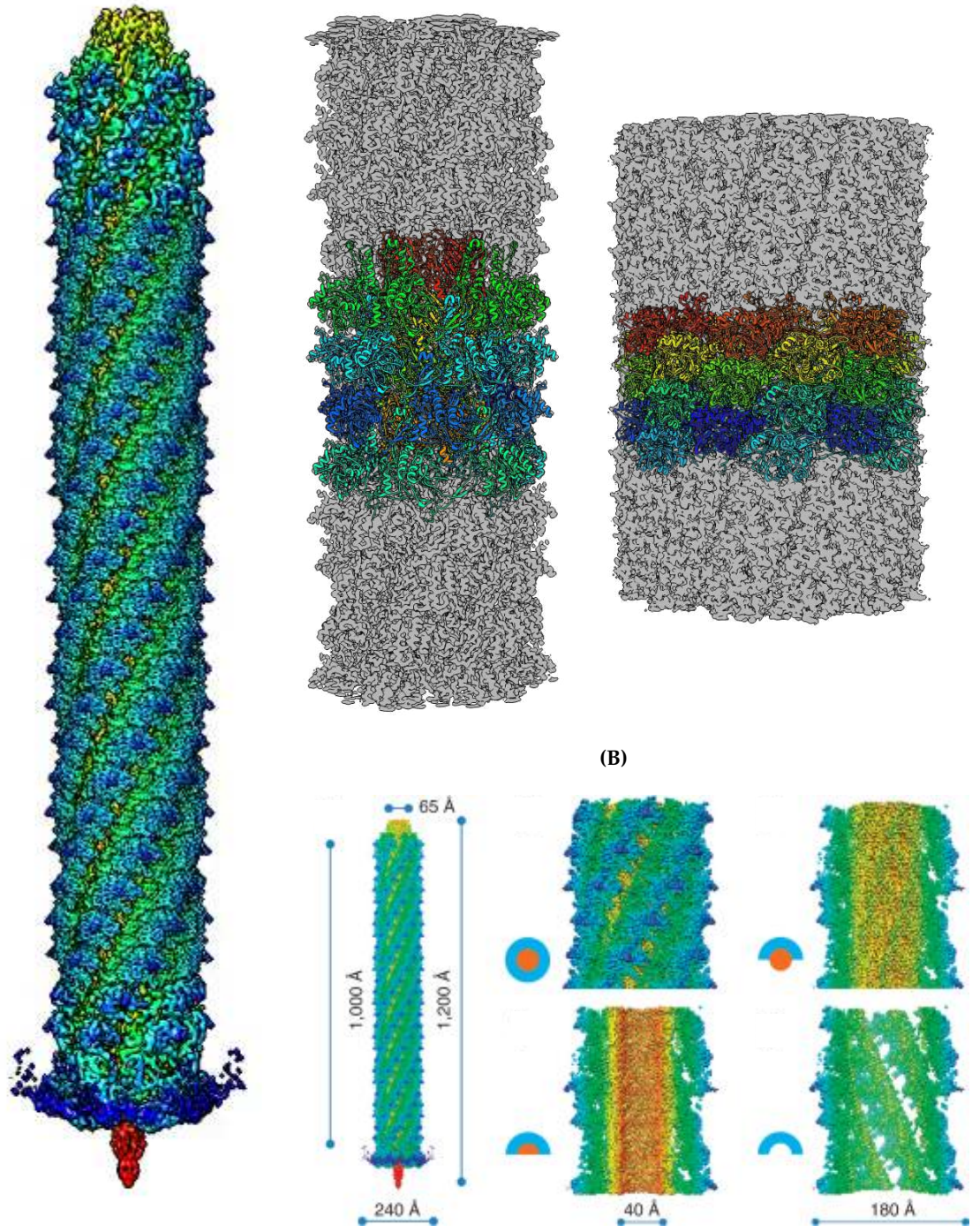


Figure 1.9 | The structures of the R-type pyocin from Ge *et al.* (2015).
(A) The reconstructed pyocin EM density reproduced from the Supplementary Video 1 of the Ge *et al.* (2015) paper as a snapshot. The structure is coloured according to its distance from the central axis (colder colours are further 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). The augmented structure to accompany this page shows the resolved densities for the pre- and post-contraction sheaths (as in **(B)**).

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.*, 2015; Kube and Wendler, 2015). 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 ≈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, 2015). From the atomic reconstruction, it was shown by Ge and colleagues that each protomer of the outer sheath interacts with the adjacent two 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.*, 2015). 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.*, 2015). The inner sheath monomers of the pyocin consists of two 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 two sets of four 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 six 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.*, 2015).

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 two '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 two 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 provide, nor speculate on, 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, 2015; 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 the nearest cousins of the PVCs - the so-called "Antifeeding Prophages" of *Serratia*. The Afp was, until the advent of the Ge *et al.* (2015) 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 giveni* (formerly *Costelytra zealandica* (Coca-Abia and Romero-Samper, 2016)) 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:

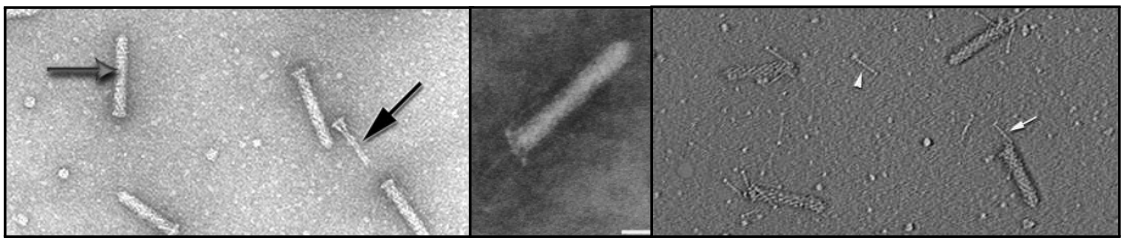


Figure 1.10 | Electron Micrographs of the Antifeeding Prophage of *S. entomophila*.
The left panel shows EMs of the Afps, revealing their "bullet like" shape, and similarity to PVCs. The grey arrow denotes a mature, fully intact Afp particle. The black arrow highlights a "Tube-Baseplate Complex". Adapted 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.* (2007). 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. giveni*. The plasmid has been shown to contain other virulence factors such as the *sep* toxins (homologues of the well characterised *Photorhabdus* "Toxin Complex" genes, and thought to be similarly mobile (Dodd *et al.*, 2006)), 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) (Hurst *et al.*, 2000). It was observed in the *sep* studies that another large locus on the plasmid caused a cessation of feeding effect one to three 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, four primary papers were published which steadily elucidated the genetic components and pathology (Hurst

et al., 2004), the regulation (Hurst *et al.*, 2007) 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, 2015). 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, four 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). More recently, another orthologue of the Afp, termed AfpX has been found in a further strain of *Serratia*, *S. proteamaculans*, once again located on a plasmid, is distinct from the 'original' Afp in a number of ways which will be discussed in upcoming sections (Hurst *et al.*, 2018).

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 two 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.*, 2015; 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. Unusually, the AfpX identified in *S. proteamaculans* has two paralogues of the tape measure protein, which vary in length by 36 amino acids. Accordingly, when particles where identified micrographically, a wider distribution of sizes was observed than in the original Afp. The significance of this is not yet understood.

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).

Despite this extensive study, the EM map that was obtained, displayed in Figure 1.11 on the following page, 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-

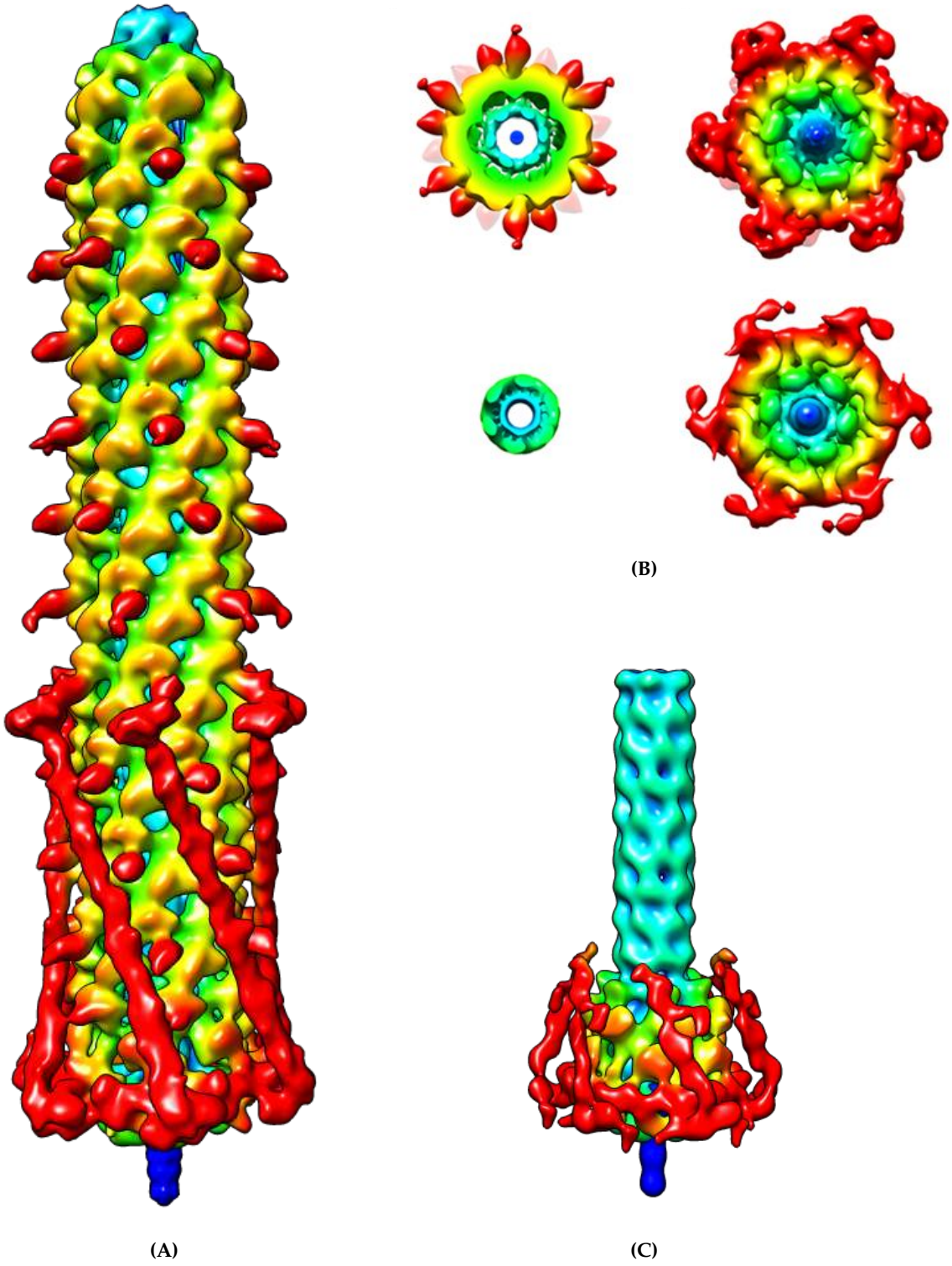


Figure 1.11 | The structures of the Antifeeding prophage from Heymann *et al.* (2013).
(A) The reconstructed 20 Å electron density map for the *S. entomophila* Antifeeding prophage, based upon Heymann *et al.* (2013), and reproduced independently from the deposited data under EMDB-2419. All panels in this image are coloured by the distance in Ångstroms from the centre of rotational symmetry (blue ≤ 20Å to red ≤100 Å). Some features of note include the dark red 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.

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.* (2015). 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 and PVCs are extremely similar, which is perhaps not unsurprising given their host's similar lifestyles as insect pathogens. 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). A compelling explanation for significantly different selection pressures is the fact that Afps have been demonstrated to be extremely selectively toxic to only the New Zealand grass grub. The geographic isolation of New Zealand, known for its unusual flora and fauna, may point to a long co-evolution of *S. entomophila* and *C. giveni* which limits the host range. In the recent paper by Hurst *et al.* (2018), AfpX was demonstrated to be toxic to the larvae of the Manuka beetle (*Pyronata festiva*), another organism which is endemic to New Zealand, and has yet to be found elsewhere. *Photorhabdus*, by contrast, has demonstrated wide ranging lethality to insects, and is found throughout the world. Lastly, the PVCs are present in various forms, scattered throughout *Photorhabdus*

genomes, and this alone has potentially lead to enormously different selection pressures (due to their paralogy), and potentially morphology - whereas *Serratia* is limited to only two examples (and still only a single operon per genome).

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 nine "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 three 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 four 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 two 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 five 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 five T7SSs, as *Photorhabdus* habours up to six 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, mediate symbiosis, 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 two 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, 2015), 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, 2015).

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

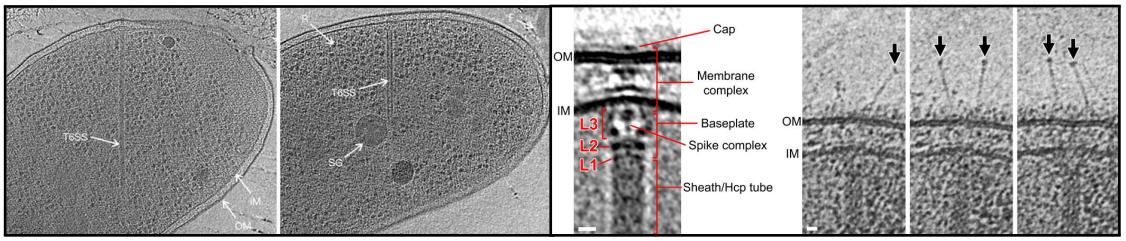


Figure 1.12 | Electron Micrographs of the Type VI Secretion System.
The left panel shows two EMs of the T6SS, and displays the enormous magnitude and variability in size of the tube complex, in the left most inset, the T6SS is labelled. In the right inset, the T6SS is labelled again, along with putative 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).

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 and in synergism. The diversity of competition 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).

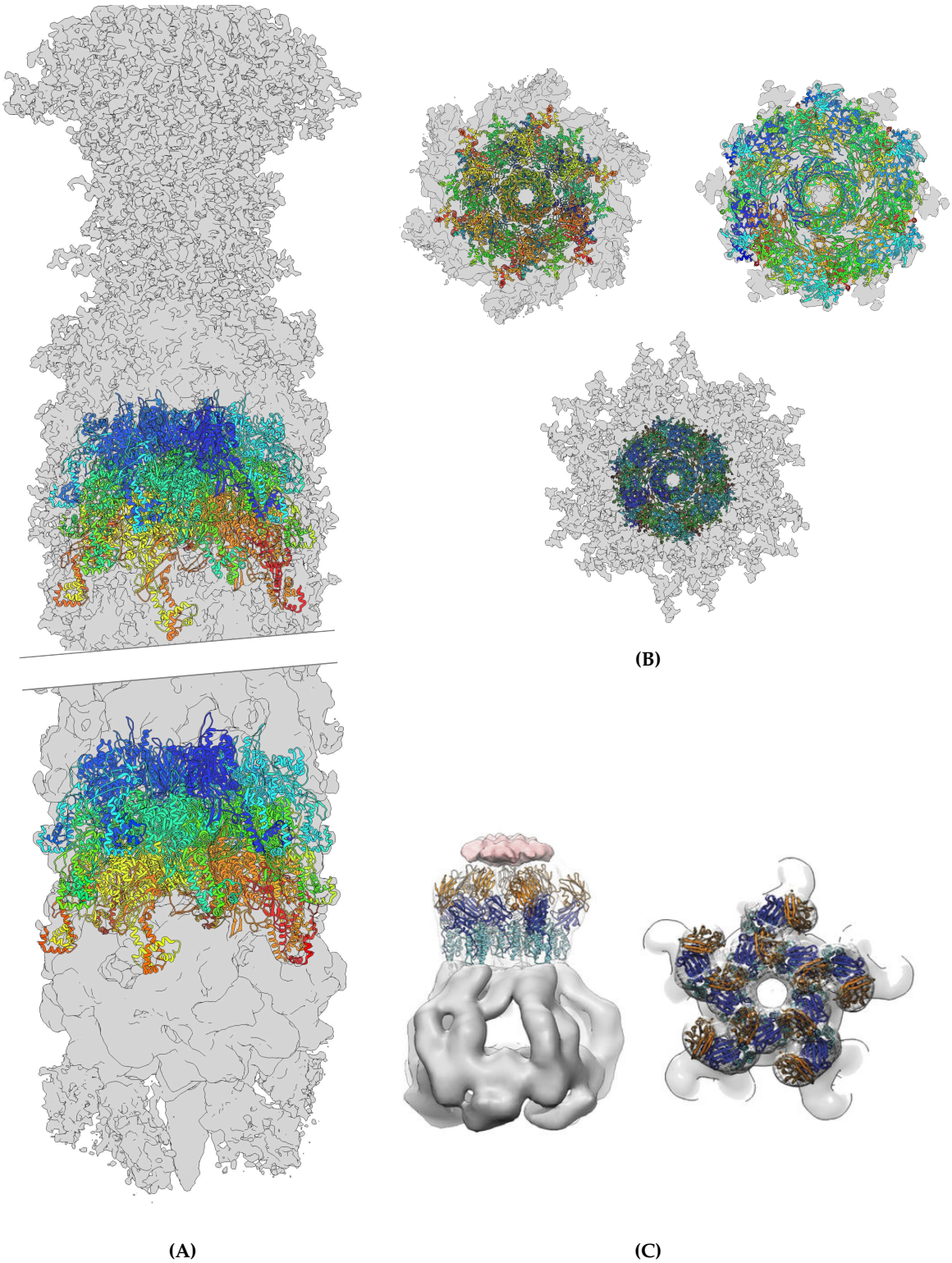


Figure 1.13 | The structures of resolved components of the Type VI Secretion System.
(A) The EM density for the proximal and distal ends of the pre-contraction Type VI secretion tube. Figures were reproduced independently from the deposited data under EMDB-3878 and 3879 from Nazarov *et al.* (2017). 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 four structural complexes - a contractile tube (one could make the case that the spike complex forms a fifth 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 two different proteins: TssB and TssC. Type 6 secretion proteins have been given alphabetic nomenclature preceded by "Tss" (Type six subunits), as such, the outer two 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.*, 2017; 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.*, 2017). 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, 2015). 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 two 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 five '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 aside from just virulence/lethality. 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 eight.

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 five 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.

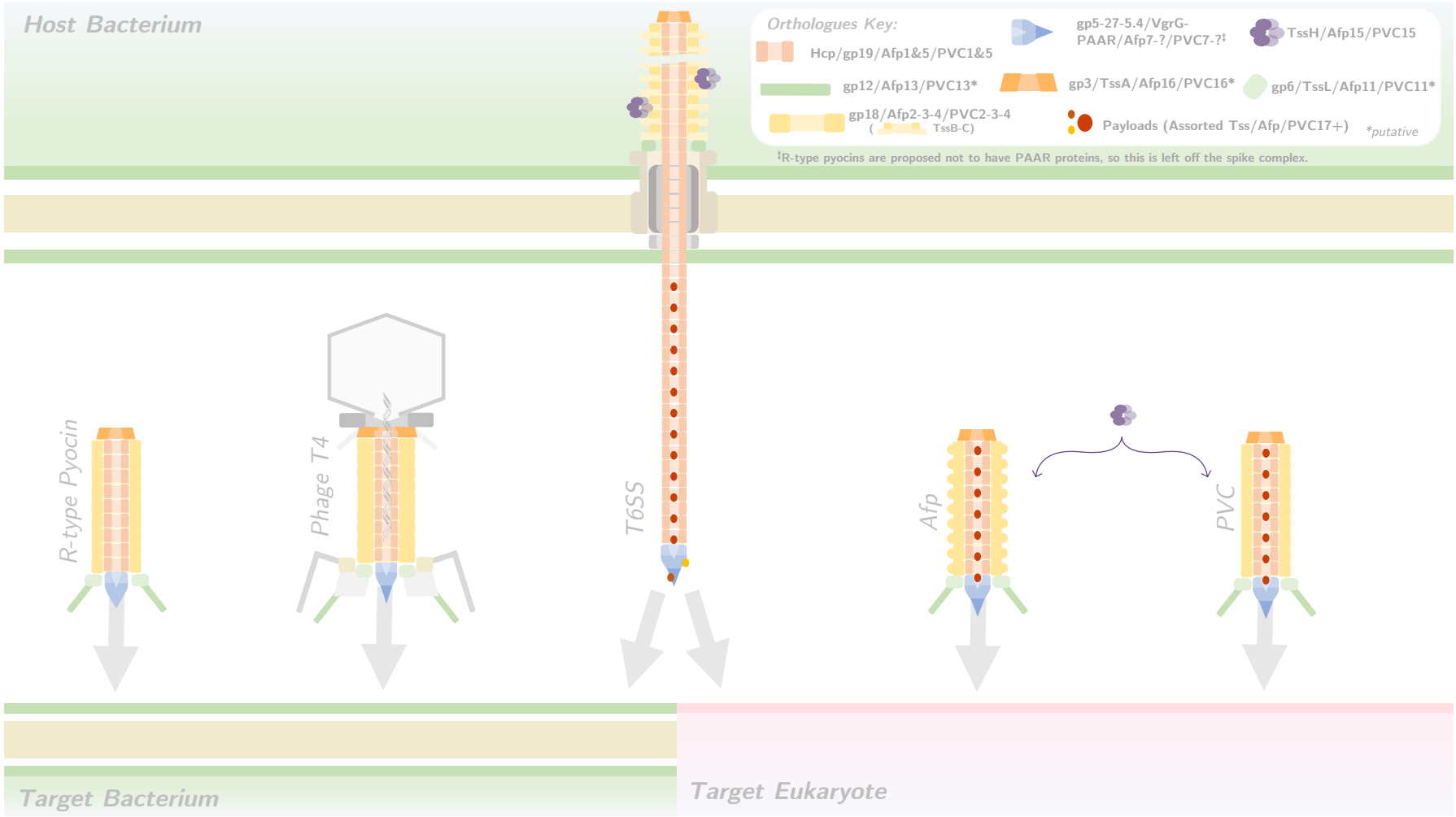


Figure 1.14 | Schematic of conserved architecture in various caudate structures.

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, 2015; 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 one 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, 2015).

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, 2015). This leads to some very impressive statistics. Wang *et al.* (2017) 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.* (2017) 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 *P. luminescens* 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 three operons. This raises questions about why the PVCs maintain two inner sheath and up to three 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 *P. asymbiotica* 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 187.

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, 2015). 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 distinct 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 Hyper-variable 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 one 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, which

once again highlights the similarity between them and PVCs. 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 *obvious* 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, 2015). 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 they 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 (2015), 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 152. 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 187 explores the tail fibres in more detail, and represents possibly the first experimental studies of naturally occurring chimeric tail fibres.

- Chapter 6 on page 234 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 promoters, 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 and pLac promoter systems, 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 2.2 shows concentrations of compounds used.

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

Purpose	Supplement	Working Concentration
Antibiotic Selection	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)
Induction	Arabinose	0.2% (w/v)
	IPTG	2 mM
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/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 following page details all the existing plasmids used in this study that have been previously constructed and/or published. Table 2.5 on page 67 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 following 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.	(Cárcamo-Oyarce *et al.*, 2015)
pAGAG	Promoterless GFP bearing plasmid, without GFP start codon for promoter fusion reporter construct creation. Conjugative plasmid requiring *λpir E. coli* for propagation.	This study.
	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)
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
		Recombineering Constructs	
pJETBAD-FRT-Kan	*araBAD* promotor system and terminators	pJET-FRT-Kan	Recombineering helper plasmid derivatised from pJET-FRT-Kan by addition of the pBAD30 promoter system adjacent to the Kanomycin resistance cassette.
pJETBAD-FRT-Cm	*araBAD* promotor system and terminators	pJET-FRT-Cm	Recombineering helper plasmid derivatised from pJET-FRT-Cm by addition of the pBAD30 promoter system adjacent to the Chloramphenicol resistance cassette.
		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*Unit 4* operon promoter fused to GFP
pAGAG_TT01PVClopT	*P. luminescens* TT01 PVC*lopT* promoter	pAGAG	*P. luminescens* strain TT01 PVC*LopT* operon promoter fused to GFP
pAGAG_TT01PVCcif	*P. luminescens* TT01 PVC*cif* promoter	pAGAG	*P. luminescens* strain TT01 PVC*cif* operon promoter fused to GFP
pAGAG_TT01PVCU1	*P. luminescens* TT01 PVC*Unit1* promoter	pAGAG	*P. luminescens* strain TT01 PVC*Unit 1* 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		GATGGAGTAAAT**GGTACC**TTGGG	55.1	
hyfC_Junction_F	Recombineering sequencing confirmation	CCCTCATTACTGTTGCTGTTAC	50.0	1847
hyfC_Junction_R		GCAGCCGCCTGTAATTTC	50	
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]	

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 - Protein Purification		
PVCpnf13-NdeI_F	*pnf* Tail Fibre	GAGTTA**CATATG**AACGAAACTCGTTATAATGC	[67]	1548
PVCpnf13-BamHI_R		TTTTCA**GGATCC**TTAAAGCTTTATGATGAAAGC		
PVCpnf13-KpnI_R		TTTTCA**GGTACC**AAAAGCTTTATGATGAAAGC		
PVClumt13-NdeI_F	*lumt* Tail Fibre	GCCGGA**CATATG**GACAACAAAAATAAC	[67]	675
PVClumt13-BamHI_R		TTACTT**GGATCC**TTACACAACCTTAATCATATAG		
PVClumt13-KpnI_R		TTACTT**GGTACC**AACACAACCTTAATCATATAG		
		Classical Cloning - Promoter Fusions		
TT01U4-BamHI_F	TT01 PVCU4 Promoter	ATT**GGATCC**TCGCTGTTCTCTCTTTTCACC	[59]	≈500
TT01U4-KpnI_R		ATT**GGTACC**TGGAGTTGTAGACATGATTTTTTCC		
TT01U1-BamHI_F	TT01 PVCU1 Promoter	ATA**GGATCC**CTACTGGGATGTGTATTCAAACA	[59]	≈500
TT01U1-KpnI_R		ATT**GGTACC**TGGAGTTATAGCCATGATTCTTC		
TT01Cif-BamHI_F	TT01 PVCCif Promoter	ATA**GGATCC**CGAGCAATAATGCTGTGAAT	[59]	≈500
TT01Cif-KpnI_R		ATA**GGTACC**GACAGTTGTAGACATTGTTATTTCC		
TT01LopT-BamHI_F	TT01 PVCLopt Promoter	ATA**GGATCC**CGCTGTAGTTTGTTTTAAAAAGG	[59]	≈500
TT01LopT-KpnI_R		ATT**GGTACC**TGTAGTTACGGACATAGTTTATTTCC		
PB68.1Pnf-BamHI_F	PB68.1 PVCPnf Promoter	ATA**GGATCC**ATCCCAACGTATCTTGTCC	[58]	≈500
PB68.1Pnf-KpnI_R		ATT**GGTACC**TGTACTTGTAGACATAAAAGCCC		
PB68.1LopT-BamHI_F	PB68.1 PVCLopT Promoter	ATA**GGATCC**CCAATAACCTGACATATTAAACCG	[58]	≈500
PB68.1LopT-KpnI_R		ATT**GGTACC**TGTGGTTGTAGTCATAATTATTTCCT		
PB68.1Cif-BamHI_F	PB68.1 PVCCif Promoter	ATA**GGATCC**GCATGTTATTTTCCTGCCTATTAT	[59]	≈500
PB68.1Cif-KpnI_R		ATA**GGTACC**GGCAGTTGTAGACATCGTTA		
PB68.1U1-BamHI_F	PB68.1Cif PVCU1 Promoter	ATA**GGATCC**CAATTTTAACTATTTACTGGACTTCG	[57]	≈500
PB68.1U1-KpnI_R		ATT**GGTACC**TGGAGTTGTAGACATAATGTTTCC		

Gibson Assembly				
pBAD30frag_F	pBAD30	ATGTAATTAATTCAACCATCACGGAGAGTTTATCAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCC	[72]	4791
pBAD30frag_R		CTTGTAGACATAAAAGCCCCTTTTTAGACAAAAAATAGCCCAAAAAAACGGGTATGGAGAAACAGTAGAG		
PNFfrag1_F	PVC*pnf* 1-8	CTCTACTGTTTCTCCATACCCGTTTTTTTGGGCTATTTTTTGTCTAAAAAGGGGCTTTTATGTCTACAAG	[72]	7181
PNFfrag1_R		GTGTCAGTATTTGATTTTCCATTCATCGTCACCTTTCATTGGGTAAGATTAATTTTTGCGCCTTTGATTT		
PNFfrag2_F	PVC*pnf* 9-12	AAATCAAAGGCGCAAAAATTAATCTTACCCAATGAAAGGTGACGATGAATGGAAAATCAAATACTGACAC	[72]	6039
PNFfrag2_R		TCTTGTACAGTTGCATTATAACGAGTTTCGTTCATGATTAACTCCAGAAAACATATTTAATTCAACATCA		
PNFfrag3_F	PVC*pnf* 13-15	TGATGTTGAATTAAATATGTTTTCTGGAGTTAATCATGAACGAAACTCGTTATAATGCAACTGTACAAGA	[72]	5514
PNFfrag3_R		TTATTGACATCAATAATAGTTTGCGTGTTTAACATAAAAAACCTCTCTTAAATTATATCGTGATAACTTT		
PNFfrag4_F	PVC*pnf* 16-18	AAAGTTATCACGATATAATTTAAGAGAGGTTTTTTATGTTAAACACGCAAACTATTATTGATGTCAATAA	[72]	4416
PNFfrag4_R		GGGGAGACCCCACACTACCATCGGCGCTACGGCGTTGATAAACTCTCCGTGATGGTTGAATTAATTACAT		
Recombineering Primers				
ψendA_no1_F	*endA* Deletion site	AAACAGCTTTCGCTACGTTGCTGGCTCGTTTTAACACGGAGTAAGTGATGCGCACTTCCAGACCCAGATC	[72]	1100
ψendA_no2_R		GTTAACAAAAAGAATCCCGCTAGTGTAGGTTAGCTCTTTCGCGCCTGGCAGATGGAGTAAAT**GGTACC**TTGGG		
speB_no1_F	*speB*	GTTTTACCCGTGCGCATCGCATCTGGTGCTTACTCGCCCTTTTTCGCCGCCGCACTTCCAGACCCAGATC	[72]	1100
speB_no2_R		GACGCGGAAGGGTTTTTTTATATCGACTTTGTAATAGGAGTCCATCCATGGATGGAGTAAAT**GGTACC**TTGGG		
hyfC_no1_F	*hyfC*	GTTTTACCCGTGCGCATCGCATCTGGTGCTTACTCGCCCTTTTTCGCCGCCGCACTTCCAGACCCAGATC	[72]	1100
hyfC_no2_R		GACGCGGAAGGGTTTTTTTATATCGACTTTGTAATAGGAGTCCATCCATGGATGGAGTAAAT**GGTACC**TTGGG		

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 68.

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 Section 2.2.1.2 on page 64.

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 72. 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 68, 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 Section 2.2.1.2 on page 64 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 79).

2.2.7.3 Electrocompetent Cells

For recombineering protocols (see Section 6.2.1.1 on page 242), 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 69. 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 72). 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 (see Table 2.4 on page 66 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. 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 76.

Electroporation was conducted with the previously detailed parameters for *E. coli*, using ≈ 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 78 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 68.

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 generally be assembled in to single contiguous sequences after discarding contaminating host reads - see Section 2.5.2 on page 87. Libraries were prepared according to the manufacturers specifications, using the paired-end 2x250bp Nextera XT kit. Reads were downsampled if assembly issues were

encountered.

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 187). 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 × 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 187.

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 Aktä 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 upon.

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

2.3.6.1 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.

2.3.6.2 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.

2.3.7 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.

2.4 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.

2.4.1.1 Image Normalisation and Consistency

Images were taken under equivalent conditions, with an exposure determined empirically at the time of imaging, in order to ensure minimal GFP bleaching, but sufficient signal to visualise. Heterogeneity in sample preparation on the GeneFrames meant that some images appeared darker than others for equivalent exposures. To make images more consistent, `ImageMagick v7.0.8-2 Q16` was used to normalise the images to a single reference image that was considered sufficiently bright for clarity. This was done using the Hue/Saturation/Intensity (HSI) colourspace of the `matchimages.sh` script which wraps a number of ImageMagick functions[3]

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

[3] `http://www.fmwconcepts.com/imagemagick/matchimage/index.php`

on page 187 due to it requiring some empirical experimentation, and results are presented there.

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. All scripts created for this thesis are available online at GitHub[4]. 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 91).

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

[4] `https://github.com/jrjhealey`

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 two 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[5] (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 two 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 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 two methods of choice. In this case, the clusters would be two 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

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

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.

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

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 relaxed parameters in an effort to gather even low quality hits and was run against the latest version of the PDB70 database available from the HHSuite database site[6].

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).

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

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.

RMSD is calculated as follows:

$$\text{RMSD(v, w)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz}))^2} \tag{2.5}$$

Root Mean Square Deviation of protein model superpositions

where a set of n points, v and w, correspond to the x, y, and z atomic positions of the atoms respectively (as per their subscripts). Thus the function returns the average Euclidean distance in any direction between 2 points or sets of points. The lower this value is, the closer 2 points are and this functionally translates to a better superimposition of 2 protein structures (for instance).

Similarly, TM-score, another measure of structural similarity, is calculated as follows:

$$\text{TM-score} = max\left[\frac{1}{L_{target}}\sum_{i}^{L_{aligned}}\frac{1}{1+\left(\frac{d_i}{1.24\sqrt[3]{L_{target}-15}-1.8}\right)^2}\right] \tag{2.6}$$

Template Modelling Score of protein model superpositions

where L_{target} and $L_{aligned}$ are the lengths of the target protein and aligned region respectively, and d_i is the distance between the ith residues. The cube root is a scaling term which normalises the distances.

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`[7]. The most frequent or longest set of repeats were retained after a default number of iterations.

2.5.12 RNA structure analysis

Hairpin structures were probed using the `Vienna RNAfold v2.4.6` program (Lorenz *et al.*, 2011)[8], and images were rendered using the `Forna` webserver (Kerpedjiev *et al.*, 2015)[9]

2.5.13 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[10] 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`[11]

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).

General figures such as schematics and diagrams were typically prepared with Microsoft Powerpoint, or BioRender[12].

[7] `https://www.ebi.ac.uk/Tools/pfa/radar/`
[8] `https://www.tbi.univie.ac.at/RNA/RNAfold.1.html`
[9] `http://rna.tbi.univie.ac.at/forna/`
[10] GSL Biotech LLC
[11] `http://tree.bio.ed.ac.uk/software/figtree/`
[12] `https://biorender.io/`

Part II

Computational Results

Chapter 3

Structural Bioinformatics of PVC Proteins

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

Richard P. Feynman

3.1 Introduction

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 project. 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, such as homology modelling, which is becoming increasingly popular as the rate of high quality genome sequences vastly outstrips the rate of experimental protein structure resolution (Rodriguez *et al.*, 1998). 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', through the application of techniques which have not been attempted to date.

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, and also leverages the benefits of a number of new genomes. By re-running these analyses at various points, new putative functions and homologies may be discovered as databases improve.

This chapter is also intended to 'pull double duty' somewhat; to serve both as an exploration of new roles and information regarding PVC proteins, but also as an extended, introduction or 'guided tour'. Hopefully, therefore, this chapter will provide the reader with an understanding of the PVCs with regard to what is already known, and what has been discovered, on essentially a gene-by-gene basis.

3.1.1 The Sequence Identity Problem

Despite progress in the elucidation of many related structures, as shown in Chapter 1, much is still unclear with regard to PVCs. Many PVC proteins in the existing *Photorhabdus* annotations are listed as 'hypothetical proteins' with no useful information whatsoever - in some cases, this can be the entire operon (Duchaud *et al.*, 2003). The significant diversity that can be seen amongst PVC operons (see Chapter 4 on page 152 for a more in-depth discussion), and the frequent lack of quality sequence analogy that can be detected via BLAST and annotation tools, suggested that structural studies might offer additional/better information.

It is now a widely accepted phenomenon that sequence identity cannot always provide sufficient structural understanding of proteins on its own, as the evolutionary rules that constrain structures are somewhat different from those that constrain sequences. Consequently, protein structures are known to evolve slower than the amino acid sequence, and slower still than the nucleotide sequence. In fact, this rate of change has even been quantified, and is estimated to be between three and ten times as slow (Illergård *et al.*, 2009). Moreover, proteins with entirely unrelated sequences can give rise to functionally equivalent proteins (through convergent evolution for example), meaning this analogy

between proteins would be completely missed through sequence studies alone. Now, of course this does not completely devalue the position of the sequence in determining protein structure and function, however, as Illergård *et al.* (2009) and Chothia and Lesk (1986) explain, the relationship between structural similarity (as quantified by Root Mean Square Deviation (RMSD)), and sequence identity is not linear. One of the earliest papers to explore this disconnect defined a so-called 'twilight zone' of structural similarity, observing that once sequence identity dropped to around 25% and below, false positive hits begin to predominate (Rost, 1999). Just to labor the point a little further, Holm and Sander (1996) explain how structures are able to remain much the same, "even when all sequence memory appears to have been lost". This means that for proteins whose common ancestors are extremely far back in time, it effectively becomes meaningless to compare the DNA or amino acid sequences, and only the 3D structure matters. The authors summarise this quite nicely with the analogy:

> "Comparing protein shapes rather than protein sequences is like using a bigger telescope that looks farther into the universe, and thus farther back in time, opening the door to detecting the most remote and most fascinating evolutionary relations."

To make matters worse, even proteins which *do* share significant sequence identity, do not necessarily have the same structure, and therefore may differ in function. A really good example of this is epitomised by the "Paracelsus challenge". Posed by Rose and Creamer (1994), the challenge was to convert one protein's conformation to that of another protein by altering less than 50% of the sequence. This was achieved first by Dalal *et al.* (1997) (winning them a $1,000 wager in the process), and though they did have to change roughly 44% of the protein and include an additional 7 amino acids for solubility (bringing to mind a somewhat 'Ship-of-Theseus' argument), they were able to convert an almost entirely β-sheet protein in to an entirely α-helical one. Subsequently, their effort has been bested by the work of Alexander *et al.* (2007), and thoroughly hammers home the message of "sequence $\neq$ structure". They were able to convert domains from *Streptococcus* G proteins to retain 88% sequence identity, yet to change their fold structures, all the while keeping their ligand binding activities.

In short, there is no substitute for being able to *see* the tertiary structure and individual folds for each and every protein. This is far from a solved problem however, as it currently requires exhaustive computation and experimental efforts to generate this kind of data. Protein structure simulation has advanced significantly, but is not without its flaws, and of course, the lottery that is protein structure determination is an incredibly intensive process with a high failure rate. New advancements in ultraresolution (cryo)Electron Microscopy are improving this outlook, however it is still a slow process (Kühlbrandt, 2014).

Finally, to underscore this idea with some relevant examples, Leiman *et al.* (2010) observed that "evolutionary relationships cannot be detected in their [tailed phages] amino acid sequences", but many structural proteins of phage origin share tertiary structure. One rationale for this is that given the high turn over rate of phage genomes, their evolution will occur rapidly, and thus they will explore the 'chemical space' rapidly also. This is combined with the fact that phages experience a multitude of evolutionary pressures, and being proteinaceous entities only, they have to be particularly 'inventive' when it comes to protein structure robustness. A further example from the same paper concerns the structures of the VgrG spike proteins. They are often structurally very well conserved which is usually immediately apparent on visual inspection, and yet, the sequences of certain orthologues may share <15% amino acid similarity, and due to redundancy, even lower nucleotide identity. Table 1 in Browning *et al.* (2012) highlights this, using HHPred to identify the same OB-fold in VgrG and T4 spike complexes, despite as little as 16% amino acid similarity in that particular case. This effectively means that to try and study sequences such as these using sequence data alone - which falls well within the threshold for false positives to predominate, as explained by Rost (1999) - would be almost indistinguishable from random noise. Silverman *et al.* (2012) make the same observation for the tube proteins (Hcp/TssBC and the VgrG spike).

A further complication with purely sequence based studies, for long, multicomponent operons such as the PVCs and other caudate structures, is that it seems synteny and gene copy may also be significant, potentially playing a role in how the assembly is choreographed (phage early versus late proteins for instance). Papers like that of Sarris

et al. (2014), note that while many of the gene products are identifiable between orthologous operons, the gene copy number varies, as does the syntenic arrangement. In the case of the MAC complex discussed in the first chapter for instance (Paragraph 1.2.3.5.1 on page 47), the whole complex isn't even encoded in a single location in the genome (Shikuma *et al.*, 2014).

To summarise, this chapter looks to address the question of 'conservation of structure not sequence' among the PVCs, with the aim of identifying new roles for proteins for which sequence identity-based studies have failed to produce useful information; and to infer from these structures in such a way as to provide lab-testable hypotheses.

Chapter Aims

- Complete a thorough and sensitive functional annotation of as many PVC proteins as possible.
- Generate structural data and compare it to known proteins.
- Examine quality simulations for physical characteristics of the PVCs.

3.2 Experimental Procedures

Given the lack of informative annotations from existing genomes available, the first step was a deeper dive in to the orthologies for all the CDSs from 16 PVC operons. Any information that can be obtained which improves on annotations of 'hypothetical protein', even if weak, might provide insights which can be used to plan experiments in future.

3.2.1 Methods used for Probing the Structures of PVC proteins

3.2.1.1 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 database version. This was performed using `HHsearch` from the HHsuite of tools (Remmert *et al.*, 2011), and was queried against the Protein DataBank each time.

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 according to their likelihood (Eddy, 2004). This arises from the fact that not all amino acids are equally likely to appear adjacent to one another - a contrived example might be, 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 through probabilities, HMMs capture domain information very well. They are especially suited to the task of identifying distant and very variable homologies, in fact, to quote directly from the HHSuite manual: "HHsearch and HHblits [different algorithm implementations] can detect homologous relationships far beyond the twilight zone, i.e., below 20% sequence identity. Sequence identity is therefore not an appropriate measure of relatedness anymore."

Individual homologies for different sections of the PVC operons will be discussed in the coming sections. In overall terms, there are 2 loose distributions to the hits. As the histogram in Figure 3.1 on page 102 shows, of the roughly 350 sequences that were

profiled, about half are extremely well matched, with probability scores of approximately 85% or higher. The probability score from HHSuite is essentially a measure of how good a match is to a query in terms of their likelihoods of being orthologues. It functions similarly to the more common E-Value, but also takes in to consideration the secondary structures of the proteins matched, which is partly why it is able to reliably identify more distant matches than sequence identity alone. Since all the PVCs carry different effector molecules, and they are reasonably well characterised to date, this chapter will focus on the 'dark matter' of the structural region of the operons - not to mention, effector biology of the PVCs could be a PhD thesis in itself. Since any orthology might provide more insight than annotations of 'hypothetical proteins', match scores were examined even if very weak. Furthermore, the appropriateness of any given cutoff can vary by protein domain and the host taxon, so no hard cutoffs were applied *per se*[1].

At the outset of this work, in the ≈ 350 studied PVC CDSs, from 16 individual operons, 249 were annotated as hypothetical proteins. As a result of further probing through HMMs, all loci have garnered a hit/annotation of some sort, though there are still some unreliable hits. The least well profiled genes (by score) primarily consist of:

- PVC14 - a protein inferred to be a tube tape measure protein from the studies of Rybakova *et al.* (2013).
- PVC7 - an unknown protein which does, however, attract consistent annotation in almost all operons to a SIR2 metalloprotein deacetylase (Shore, 2000), but with poor hit quality scores.
- Other assorted unique genes which disrupt the general naming convention applied to the operons. These include several from the PVC "lumt" operon, which in many ways appears to be a particularly unusual operon. As unique genes, it is even more difficult to speculate on their role as one cannot appeal to the other operons for clues.
- Further unusual genes such as an additional locus in the PVCUnit2 locus of *P. luminescens* TT01, between the canonical PVC13 and 14.

[1]See the following URL for an interesting discussion about benchmarking HMMs by taxon and protein type: `http://csbl.bmb.uga.edu/dbCAN/download/readme.txt`

Unsurprisingly, the proteins with the highest confidence hits matched those with the best annotations before the project began, namely the sheath proteins and payload enzymes. That said, many of the initial annotations for the tube proteins provided little more information than that they shared similarity with phage tail "major" and "minor" proteins. Through HMM searching, these annotations have begun to point to more specific matches such as gp19 orthologues and the sheaths of R-type pyocins. Moreover, subtle differences in the 'conserved' sheath proteins reveal that not all operon's sheath proteins match the same orthologues.

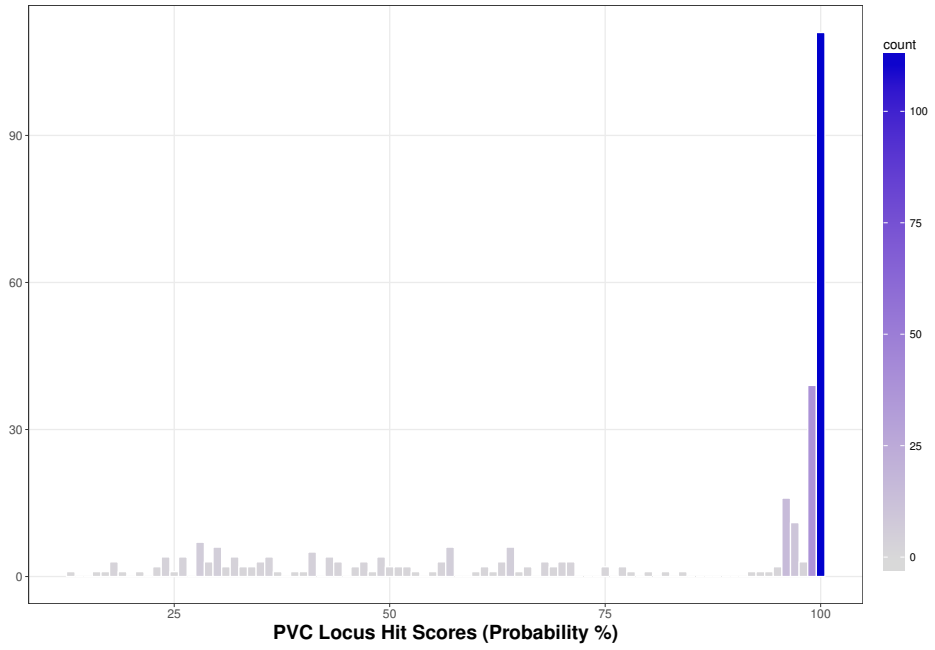


Figure 3.1 | Distribution of match probabilities for PVC proteins.
A histogram of the "Probability" score for approximately 350 PVC protein sequences. Bars are coloured according to the bin density. Around half of all the proteins are well profiled in this manner, with Probability scores over 80%. The remaining proteins are distributed widely with a number of examples of proteins which remain with low scores and little to no informative matches.

3.2.1.2 Homology Modelling, Threading and Structural Refinement

Protein structure modelling was performed using a local installation of the I-Tasser pipeline, on 12 virtual machines, utilising over 300 CPUs (Yang *et al.*, 2014; Roy *et al.*, 2010; Zhang, 2008). I-Tasser was chosen as it generates full length models (whereas some tools only produce models for regions adequately represented by sequence homology), and

because it takes a somewhat hybrid approach to modelling, using threading templates, but also *ab initio* molecular dynamics steps for modelling regions (or whole sequences) with no reliable sequence templates, and for final model refinement. I-Tasser is consistently ranked as one of, if not the best, modelling servers/algorithms in the international Critical Assessment of Structure Prediction competitions (Moult *et al.*, 2015).

In most cases, I-Tasser returns 5 models (though sometimes only 1 if the models converge well). Where multiple models were produced, in all upcoming analyses and images the model with the lowest RMSD to the PDB entry identified as closest via HHPred is used. Figure 3.2 shows the distribution of RMSD values obtained from the simulations matched to these structures. The majority of the simulations were able to score well against deposited structures, with RMSDs less than ≈10 Å.

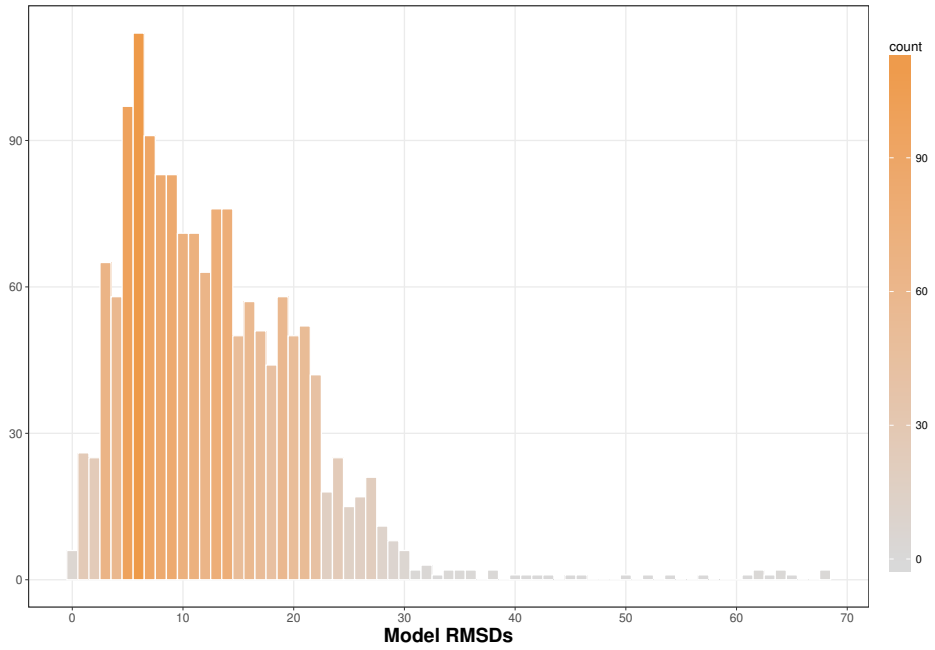


Figure 3.2 | Distribution of atomic deviations for all simulated models.
A histogram of the Root Mean Square Deviation (in Ångstroms), between a modelled sequence and its closest structural match, as determined by HHSuite. RMSDs were calculated using the MatchMaker function within UCSF Chimera. Reasonable/useful RMSDs of less than ≈10 Å are obtained for a significant proportion of the loci simulated. Bars are coloured according to the bin density.

That said, despite being the most widely used metric, RMSD is not a perfect measure of structure similarity, as it is prone to large RMSD values (indicative of poor fits) if just a few localised regions are not well matched, even if the protein overall shares similar

topology. To this end, I-Tasser also produces calculations of "C-score" (confidence score) and "TM-score" (template modelling score) (Zhang and Skolnick, 2005). I-Tasser provides additional data to assess the best models produced during its molecular dynamics steps, such as the cluster density, which represents the number of times a model passes through a region of 3D modelled space during its molecular dynamics trajectories. Thus the 3D space which is sampled most often is likely to represent the conformation of the most favourable models.

There are a large number of models with TM-scores of just over 0.2, (TM-score is in the interval $[0, 1]$), which is considered the lower bound on a plausible match between structures (Zhang and Skolnick, 2005), suggesting that these protein structures may be of somewhat dubious quality. Roughly half of all modelled proteins have TM scores greater than 0.5 however, and these structures are therefore likely well modelled. This correlates well, as expected, with the C-scores, where higher scores are indicative of better models. C-score is particularly useful when distance based metrics like TM-score and RMSD cannot be used, namely, in situations where a template protein could not be identified at all.

It is important to mention however, that despite all these metrics, whether or not a model is useful depends heavily on the question at hand, since models which diverge from the reference structures may still have important/useful information in. Moreover, not all of the metrics correspond perfectly with one another (for instance, some models which have poorer C-scores, actually have good RMSD values). As the epigraph at the start of this chapter suggests, there is often no substitute for visually inspecting the models.

Eleven of the sequences attempted failed to simulate. Firstly, 8 sequences from the *P. asymbiotica* Kingscliff genome failed to simulate due to the presence of ambiguous amino acids as a result of an older, lower quality, genome assembly. Three proteins were rejected for exceeding the upper amino acid length limit; I-Tasser's algorithms have a hard cut-off of 1500 amino acids.

For the purposes of this chapter, structures are going to be assumed to be predominantly correct (positions of certain loops and small discrepancies notwithstanding). In

cases where the simulations are likely spurious they will be noted in the discussion. That is to say, where good templates have been identified and plausible structures have been obtained, the electrostatics, general fold, and any inferences based on these will be assumed to be correct in the absence of truly resolved structures which could reveal unexpected differences. This is so that the chapter can focus on what might be learned from the models, rather than a lengthy discussion of whether the simulation quality is optimal.

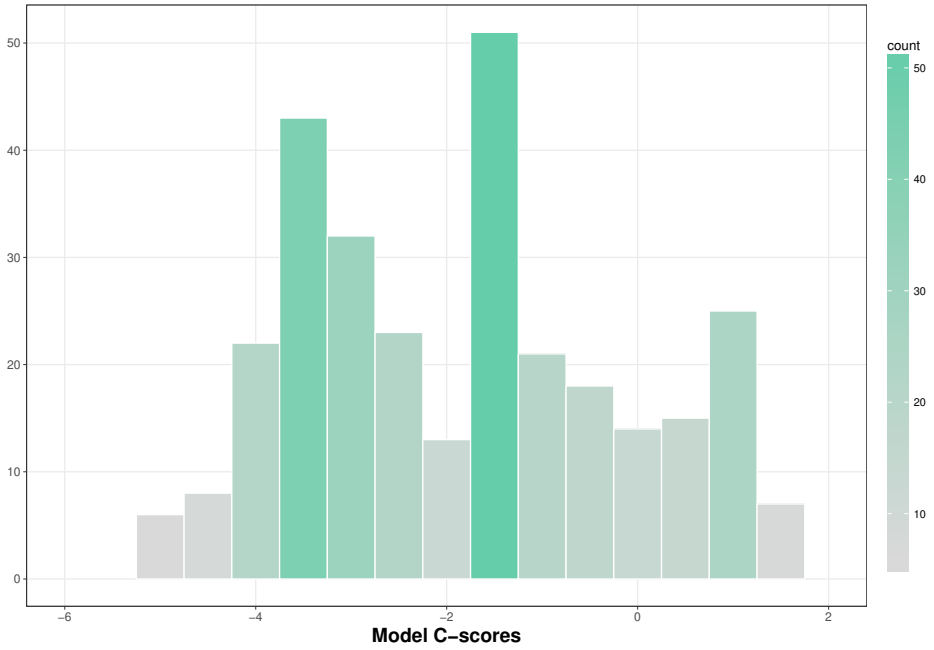


Figure 3.3 | Distribution of the "C-scores" for all simulated models.

A histogram of the "C-score" from I-Tasser. The C-score is a confidence score calculated internally by I-Tasser based on its confidence in the threading template alignments and convergence of the models. To a rough first estimation, C-scores of greater than approximately -2 indicate reasonable confidence. Approximately half of the models produced therefore have acceptable scores. C-score is in the interval $[-5, 2]$.

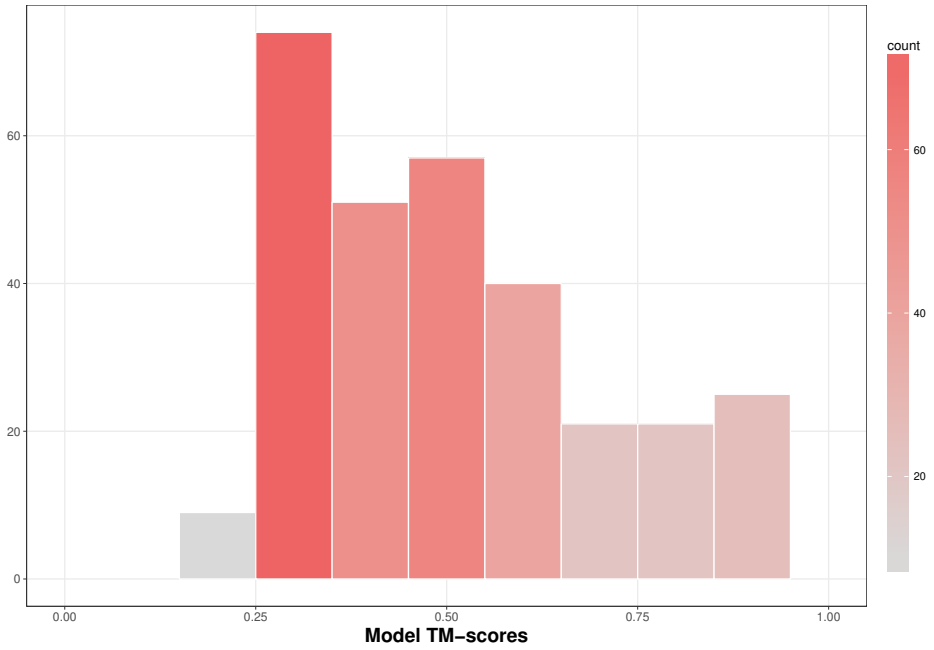


Figure 3.4 | Distribution of Template Modelling score for all simulated models.

A histogram of the TM-score, between a modelled sequence and its closest structural match. TM-scores reflect how well 2 structures match globally, with less susceptibility to local deviations affecting overall score. The vast majority of protein models have TM-scores of greater than 0.25, an acceptable if slightly generous lower bound. TM-score is in the interval $[0, 1]$.

3.2.2 Exploration of the Structure of PVCs by 'Functional Unit'

3.2.2.1 The PVC Tube

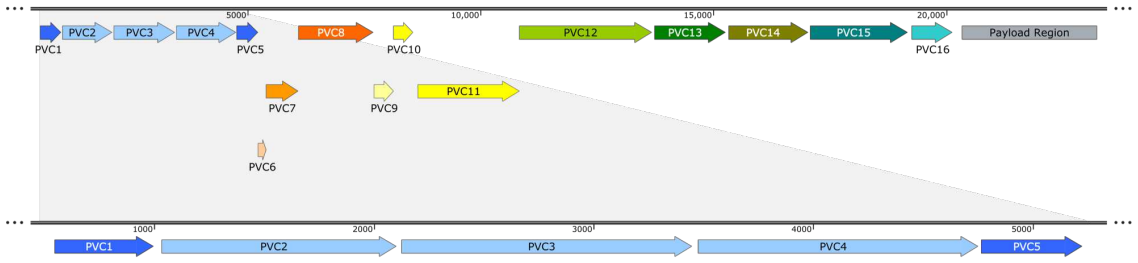


Figure 3.5 | The five loci that comprise the tube structure of a PVC.
(Top) The Pnf operon is used as an exemplar of the numbering of loci within an archetypal PVC operon, and three colourings are used to demarcate the 'functional units' of an operon: blue - tube proteins, orange/yellow - spike complex proteins, green/cyan - the operon 'core'. **(Bottom)** The six loci that give rise to the functional unit of the spike complex itself are blown up below as an aid to understanding the operon organisation and the proteins under discussion.

Among the better annotated genes at the outset of this work, the first five loci of the PVCs, are predicted to match phage tail tube proteins of various sorts, though the existing annotations were not much more informative than this (and for the rest of the operon the vast majority of which were "hypothetical proteins"). On further inspection, these genes have become 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 (see Table 3.5 on page 134). In some previous rounds of annotation, homologies to the Type 6 Secretion Systems of *Edwardsiella tarda*, *Vibrio cholerae*, and *Burkholderia pseudomallei* were also detected for the inner sheath proteins (PVC1 & 5), which underscores the original notion that PVCs seem to resemble a sort of hybrid, somewhere between a T6SS and a phage in sequence similarity.

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.*, 2016; Zheng *et al.*, 2017)), and the outer sheath of PDB ID 3J9Q which corresponds to the resolved pyocin tube structure (Ge *et al.*, 2015). 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.

Table 3.5 on page 134 shows a summarised selection of the orthologues which HHPred has picked up over time. The full list of the most recent orthologies detected is not shown for brevity.

Table 3.1 | HHPred orthology summary for the tail tube proteins.
A summary of homology matches via HHPred for the first 5 PVC loci. The hits have varying degrees of confidence but have Probabilities greater than 60% and E-Values of less than 1×10^{-5}. They represent a 'collapsed' set of common hits from all the variants for each locus. Hit scores for the inner sheath proteins (PVC1 and 5) are consistently lower than for the outer sheath proteins (PVC2-4).

Locus	PDB ID Hit	Structure	Component
PVC1 & 5*	5IV5	Bacteriophage T4	Baseplate wedge protein (gp6)
	5W5F	Bacteriophage T4	Tail tube protein (gp19)
	3EAA	T6SS	Inner sheath protein (Hcp/TssD)
	4TV4	T6SS	Inner sheath protein (Hcp/TssD)
PVC2, 3, & 4	3J9Q	R-type Pyocin	Outer sheath protein

* PVC5 attracts the same three homology matches as PVC1, plus 4TV4.

Figure 3.6 on page 111 and Figure 3.7 on page 112 showcase the similarities and differences in structure between the first five PVC loci and their nearest structural homologs either as determined via HHPred or previously shown in the literature. In addition, they also show a simulated structure from this study for the Antifeeding Prophage equivalent locus. The Afp is the closest orthologue when only DNA or amino acid sequence is considered, but no resolved structure exists (and therefore cannot be picked up by HHPred when using HMMs of the PDB database). Figure 3.7 on page 112 replicates this for an example of the outer sheath proteins.

Rather than show all the PVC loci compared to their homologs, where the PVC loci are similar, one example is shown (as in Figure 3.6 on page 111) and then all the PVC variants are compared with one another in the subsequent Figure 3.8 on page 114.

The inner sheath proteins from all these structures form a conserved pair of anti-parallel β-sheets which are sandwiched together approximately perpendicular to one another (with the exception of the T4 baseplate protein which appears to be an erroneous homology/annotation). They all exhibit a lower right hand extension positioned almost exactly in the centre of the amino acid chain (cyan/aquamarine) which would overlap the monomer adjacent to it within the tube hexamer. In the case of the T4 gp19 tube protein (chain F from PDB ID 5IV5), this is quite extended, along with a much longer protruding

C-terminus (dark blue).

The matched PDB ID 5IV5 is the entire baseplate complex of T4, including part of the tube (and encapsulates PDB ID 5W5F, so there is a slightly redundant match), from the T4 phage. It seems unusual that many of the proteins match (at the sequence and secondary structure level) to chains within the model which are *not* the tube, instead forming part of the baseplate complex, primarily gp6. It is clear from looking at the structures visually that these proteins are definitively homologues of the inner sheath proteins. On further inspection of the specific sequence matches detected by HHPred, it appears that what may have transpired is an error in the annotation of gp6 being attributed to some of the gp19 chains within the enormous 5IV5 PDB complex.

Inset in each figure is a pairwise distance matrix for alignments of the tube orthologues. Interestingly, though perhaps unsurprisingly, the sequence identity between the PVC and Afp is significantly better (a lower distance) than it is between either of them and the T6SS, T4 or Pyocin, despite the structures being extremely good matches, further reinforcing the message of this chapter that the sequence alone is insufficient to fully capture the subtleties of the evolution of these structures. Excluding the erroneous gp6 orthology, the remaining structures can be superimposed to an RMSD of 1.2 Å or less.

For the outer sheath proteins, of which there are three in most operons, there is no ambiguity in their sequence matches, all having orthology to the R-type pyocin outer sheath under the PDB ID 3J9Q. The structures of the PVC2 locus from the "Pnf" operon, and equivalent outer sheath proteins from other structures are also shown as in Figure 3.6 on page 111, though they are less compelling matches than the pyocin at the sequence/HMM level (with the exception of the Afp, which is not detected by HHPred as the structures are unresolved), again highlighting that the sequence is not the be-all and end-all.

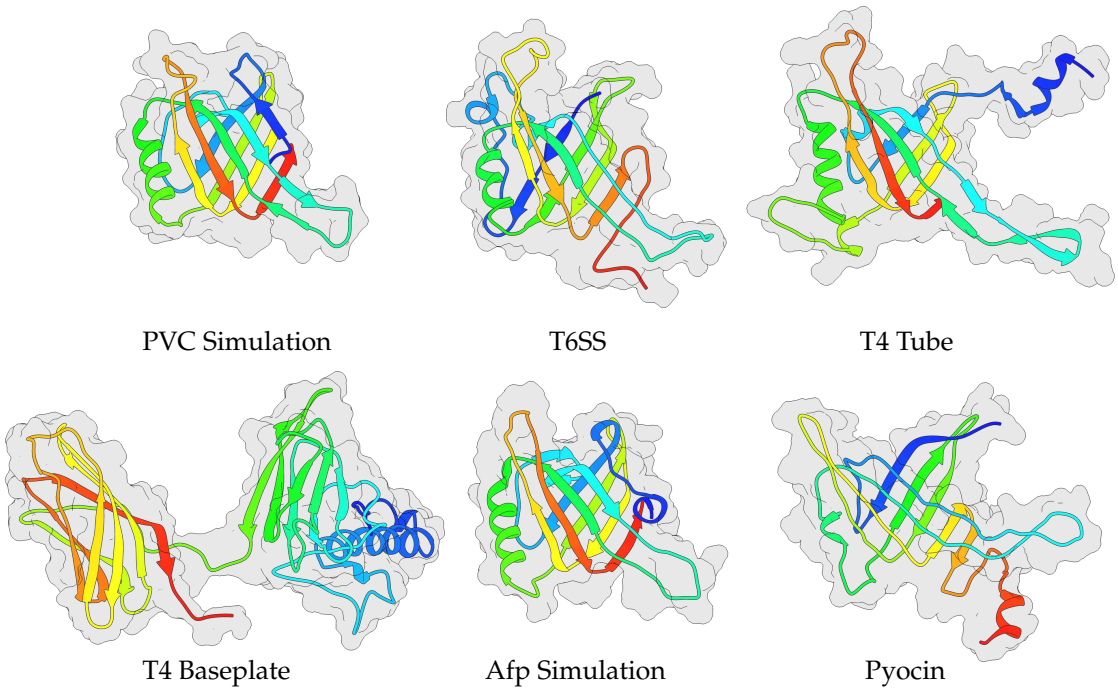
The outer sheath proteins are characterised by two outstretched termini which interlace with the adjacent monomers to form the contractile mesh over the rigid inner tube. In the case of the T4 tube, its C-terminus actually forms an additional domain known as the "protease resistant fragment", which the other structures lack (Aksyuk *et al.*, 2009a). The Type 6 Secretion System is unique amongst T4-like caudate structures, having a second outer sheath protein (TssB and TssC - known as VipA and VipB in the case of *V. cholerae*),

making the equivalent outer tube a heterododecamer. All the structures contain a pair of helices extending vertically which neighbour a pair of anti-parallel strands/sheets, that form the attachment interface to the inner tube electrostatically (Ge *et al.*, 2015).

Curiously, the 20 Å Afp EM map depicted in Figure 1.11 on page 32 seems to show protrusions of the sheath monomers radially, though the simulated sheath proteins very closely match that of the R-type pyocin, as do the PVCs, and the R-type pyocin map lacks any such protrusions. Similarly, preliminary cryoEM data for the PVCs suggests a lack of any protrusions too. This might be indicative of an unknown third protein augmenting the exterior of the sheath, somewhat like the T6SS dodecameric arrangement, but is more likely just an artefact of a comparatively low resolution EM map.

The outer sheaths of all the non-T4, released/secreted, caudate structures therefore appear to be much simpler proteins, representing perhaps a minimal functional unit capable of exerting the contraction, which is then decorated with domains such as the protease resistant fragment of T4 to confer further features to the sheath.

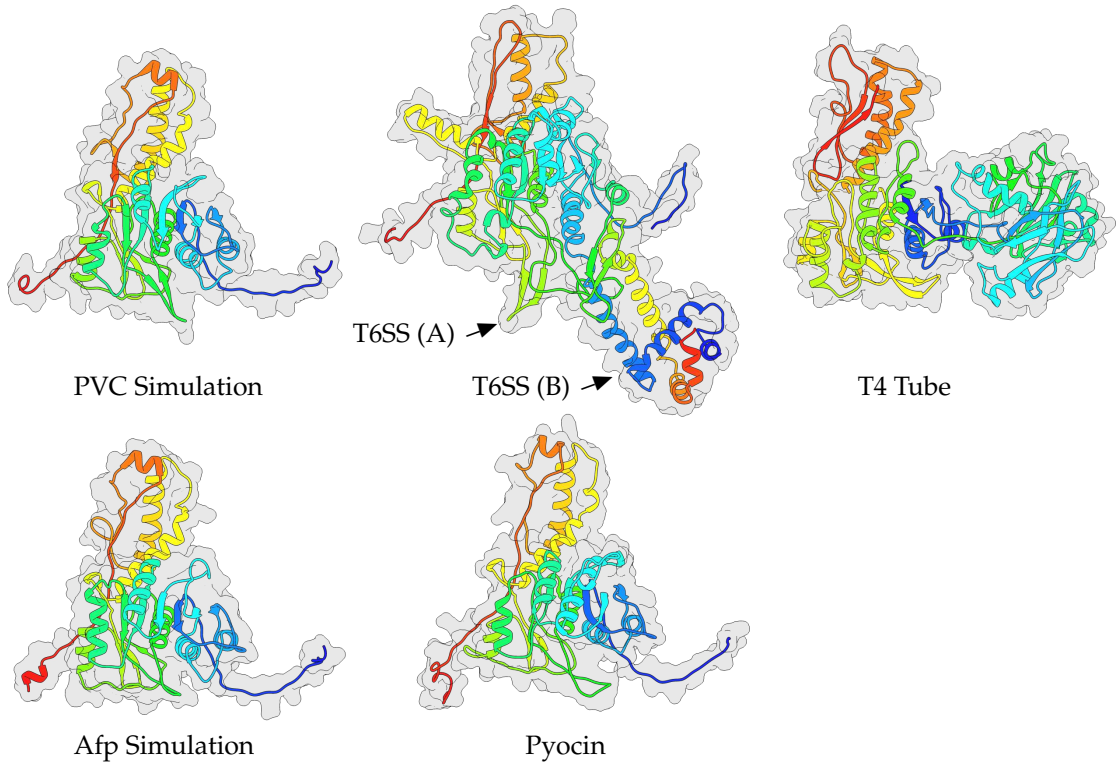
The pairwise sequence distances between the outer sheaths tell a similar story to the inner sheaths, with the PVC and Afp being the closest by some margin, compared to the remaining structures. However, the sequence distance is greater between the PVCs/Afps outer sheaths than it was for the inner sheaths. This will be due in part to the fact that the proteins are longer, but it seems that the surfaces of caudate apparatuses may be open to modification, possibly to cope with different environmental stresses, or at the very least have some sequence redundancy. All proteins can be superimposed to an RMSD of less than 1.2 Å, despite significant sequence differences.



Distance Matrix

Model	Pairwise Distances					
PVC	0.000000					
T6SS	0.872483	0.000000				
T4 Tube	0.885906	0.895706	0.000000			
T4 Baseplate	0.865772	0.876471	0.846626	0.000000		
Afp	0.255034	0.865772	0.879195	0.852349	0.000000	
Pyocin	0.892617	0.892857	0.877301	0.880952	0.899329	0.000000

Figure 3.6 | Comparisons of the structure of PVC locus 1 to orthologous proteins.
An example model of PVC locus 1 from the PVCPnf operon of *P. asymbiotica* ATCC43949. Here the structure is compared to (from top left to bottom right) the Hcp protein (5OJQ chain W) of the *V. cholerae* T6SS, gp19 (5W5F chain V) from the T4 bacteriophage, gp6 also from T4 (5IV5 chain F), a simulated Afp1 structure from *S. entomophila*, and an interior sheath monomer from the R-type pyocin (3J9Q chain V) from *P. aeruginosa*. Models are shown 'rainbow' coloured, from one terminus to the other, with a molecular surface 'ghost'. It is evident that the 2 T4 proteins, despite being frequent sequence based matches to PVC1 (and 5), are not the closest matches in terms of structural conformation. The 3D model to accompany this page shows superimposed α-carbon chain traces for all the structures (though without the T4 baseplate as its superposition is poor) - the PVC protein is rendered in thicker red wires/pipes, and the homologs in various monochrome shades. The inset table shows the all-vs-all pairwise alignment distances for the amino acid chains depicted



Distance Matrix

Model	Pairwise Distances					
PVC	0.000000					
T6SS VipA	0.890141	0.000000				
T6SS VipB	0.851613	0.864516	0.000000			
T4	0.856338	0.883721	0.845161	0.000000		
Afp	0.539548	0.892655	0.851613	0.870056	0.000000	
Pyocin	0.845070	0.898701	0.864516	0.875325	0.838983	0.000000

Figure 3.7 | Comparisons of the structure of PVC locus 2 to orthologous proteins.
An example model of PVC locus 2 from the PVCPnf operon of *P. asymbiotica* ATCC43949. Here the structure is compared to (from top left to bottom right) the VipA/VipB proteins (5OJQ chains c and m) of the *V. cholerae* T6SS, gp18 (3J2N chain U) from the T4 bacteriophage, a simulated Afp2 structure from *S. entomophila*, and an outer sheath monomer from the R-type pyocin (3J9Q chain A) from *P. aeruginosa*. The T6SS structure uniquely comprises 2 distinct chains, along with some sequence extensions to the main chain. The T4 gp18 contains a considerable augmentation to its C-terminus consistent with protrusions that can be seen in Figure 1.7 on page 21. Models are shown 'rainbow' coloured, from one terminus to the other, with a molecular surface 'ghost'. The 3D model to accompany this page shows superimposed α-carbon chain traces for all the structures - the PVC protein is rendered in thicker red wires/pipes, and the homologues in various monochrome shades. The inset table shows the all-vs-all pairwise alignment distances for the amino acid chains depicted.

3.2.2.1.1 Structural comparisons of multiplied tube proteins

Why *Photorhabdus* needs, and has managed, to maintain so many copies of what appear to be highly paralogous proteins in the tube region is an unanswered question. Not only are the proteins paralogous within a single operon, but multiplied up by 5 or 6 times to account for the other operons, and somehow proteins with as many as 10 or more copies are somehow well preserved, when it might be expected that they would be open to drift.

Figure 3.8 on the next page demonstrates the structural and sequence similarities between all 5 inner and outer sheath loci from the Pnf operon in *P. asymbiotica* ATCC43949. There are some subtle structural differences (spans of loop structure etc.) between the loci, though they generally agree very well and this may be a result of the stochasticity in the molecular dynamics trajectories calculated by I-Tasser. PVC locus 1 does exhibit greater sequence variability compared to its 'compatriot' PVC locus 5, though its overall fold is the same.

To explore this further, a comparison is made in Figure 3.9 on page 115 between the least similar PVC paralogues, as determined by their distances in the guide tree that accompanies the alignments, to show the most extreme differences (alignments can be found in Appendix Chapter B on page 335). In both cases (PVC1 and PVC5) the "Lumt" operon from *P. asymbiotica* Kingscliff and ATCC43949 respectively, present one of the pair of least similar sheath proteins. For the counterpart PVC1, the least similar paralogue is the inner sheath from the "Cif" operon of *P. asymbiotica* ATCC43949, and for PVC5 it is the tube protein of the "Pnf" operon from the same genome, highlighting the diversity present in PVC operons.

Figure 3.8 | Comparisons of sequence conservation in inner and outer sheath paralogues.

Models of PVC1 to PVC5 for the Pnf operon are shown as front and rear rotations, coloured according to their per-column sequence conservation (as per the sequence alignments in Appendix B); rendered from high (green), to low (red), as per the inset scalebar. Structural differences are minimal, and sequence conservation is generally high, though PVC1 is slightly more variable (more blue sites) than its paralogue PVC5. Of the outer sheath proteins, the interior aspect is better conserved than the exterior, and all 3 loci exhibit increased sequence diversity.

Figure 3.9 | Comparisons of the most dissimilar inner and outer sheath paralogues.

The most dissimilar homologues of the tube proteins (PVC1 to PVC5) and their structural/sequence differences are shown here. The paralogues on the top versus the bottom represent the farthest separated tips of the phylogeny and therefore the most dissimilar gene variants. This panel highlights that, despite these being the least similar variants of the same locus, the structure is still well conserved, and thus the sequences are evidently at liberty to drift without compromising the PVC structure significantly. The variability in exterior sheath proteins, particularly PVC2 is highlighted.

The interior facing side of the outer sheath is generally well conserved (middle three structures in the top row of Figure 3.8 on page 114), as might be expected since it will need to interface with the similarly well conserved inner tube proteins. The outer side of the sheath however, which represents the bulk of the structure which would be 'exposed to the elements', shows a great deal of diversity in all three loci.[2]

A further observation can be made in Figure 3.9 on page 115. It has been known for some time that the PVC "Lumt" operon in *P. asymbiotica* ATCC43949 has deleted one of the structural loci (as have both examples of the PVC "LopT" operon in *P. asymbiotica* strains) - namely PVC3. However, in the Kingscliff "Lumt" operon, the remaining PVC2 locus (which can be seen in the upper row of Figure 3.9 on page 115), appears to have lost the crucial twin helices that protrude upwards. The PVC4 locus for this operon is intact. Figure 3.10 on the following page shows the remaining genes, and compares the sequence to the same operon in *P. asymbiotica* ATCC43949. It appears that a similar deletion has occurred in both the "Lumt" operons from *P. asymbiotica* ATCC43949, and Kingscliff, however the Kingscliff PVC2 has also undergone a gene split, resulting in three CDSs. This presents an interesting conclusion, as, if the genes are assumed to still be functional, it would mean that the outer sheath can be made up of two chains, similar to the Type 6 Secretion System. If the split genes *are not* functional however, but the PVC itself is, this would prove that the remaining PVC4 is capable of 'cis-complementing' the defunct genes, and therefore the existence of multiple gene copies may not simply be due to maintaining expression stoichiometry.

3.2.2.1.2 Electrostatic comparisons amongst tube proteins

There do not appear to be obvious structural or fold differences that seem likely to lead to functional differences between these paralogues accounting for their maintenance. Perhaps then, the proteins may largely preserve their structure, but the differences brought about by diverse sequence manifest in electrostatic potentials which are in some way key to the structure of the PVCs. Given that the PVCs are proposed to carry different payload proteins with variable biophysical characteristics, there would stand to be two hypotheses:

[2]In Figure 3.9 on page 115 The most distant PVC2 locus according to the guide tree is from the PVCPnf operon of *P. asymbiotica* Kingscliff, but it failed to simulate due to ambiguous amino acids, so the next most dissimilar locus has been used instead - PVCUnit2 from *P. luminescens* TT01.

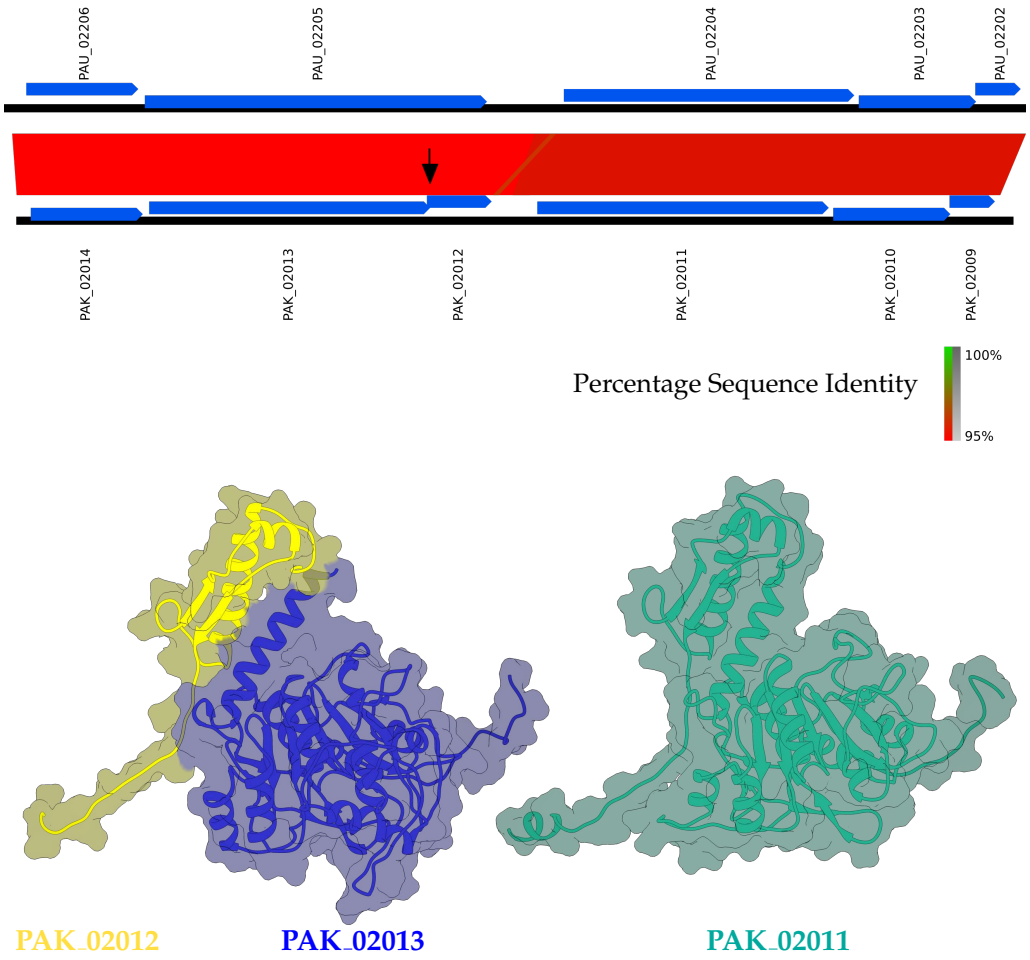


Figure 3.10 | Outer sheath proteins of PVC "Lumt" showing a domain split.
The outer sheath proteins of PVC "Lumt" from the *P. asymbiotica* Kingscliff operon are shown. The cyan/aquamarine model corresponds to PVC4, an intact outer sheath protein. The dark blue model is the same model as depicted in Figure 3.9 on page 115 in the top row for PVC2. The 2 helices that protrude from the top in other proteins are missing, instead being present as a gene split, depicted as a separate chain in yellow. The arrow in the upper alignment panel shows the gene split in in the Kingscliff "Lumt" operon, and is compared by sequence identity to the equivalent operon in the ATCC43949 strain (just the first 6 proteins of the operon).

one, that the tube lacks any specialisation at all, allowing for many different payload to pass through, which appears to be the observed phenomenon for the T6SS (Ge *et al.*, 2015), or alternatively, each PVC is adapted in a particular way to its cognate payloads.

In Ge *et al.* (2015), comparisons are made regarding the electrostatics of the tube interiors, where they reason that the R-type pyocin, as a proton conducting apparatus, has a largely negative charge in order to convey such ions. Similarly, they remark that the inner sheath of the T6SS is mixed/neutral overall to cope with conveyance of diverse cargoes, and conversely that the interiors of the PS17 and λ phages are mostly positive to convey negatively charged DNA. Curiously for the T4 proteins, the opposite appears to

be true, with an substantial negative charge. Figure 3.11 replicates a similar analysis for both inner tube proteins (PVC1/5), and the orthologues identified earlier (without the gp6 monomer and with PVC5 in its place however). Some potential structural differences are more apparent between PVC1 and 5 in this hexameric arrangement, appearing to have a wider, more funnel-shaped conformation. As these differences are largely attributed to assorted loops however, it is probably unwise to speculate too far.

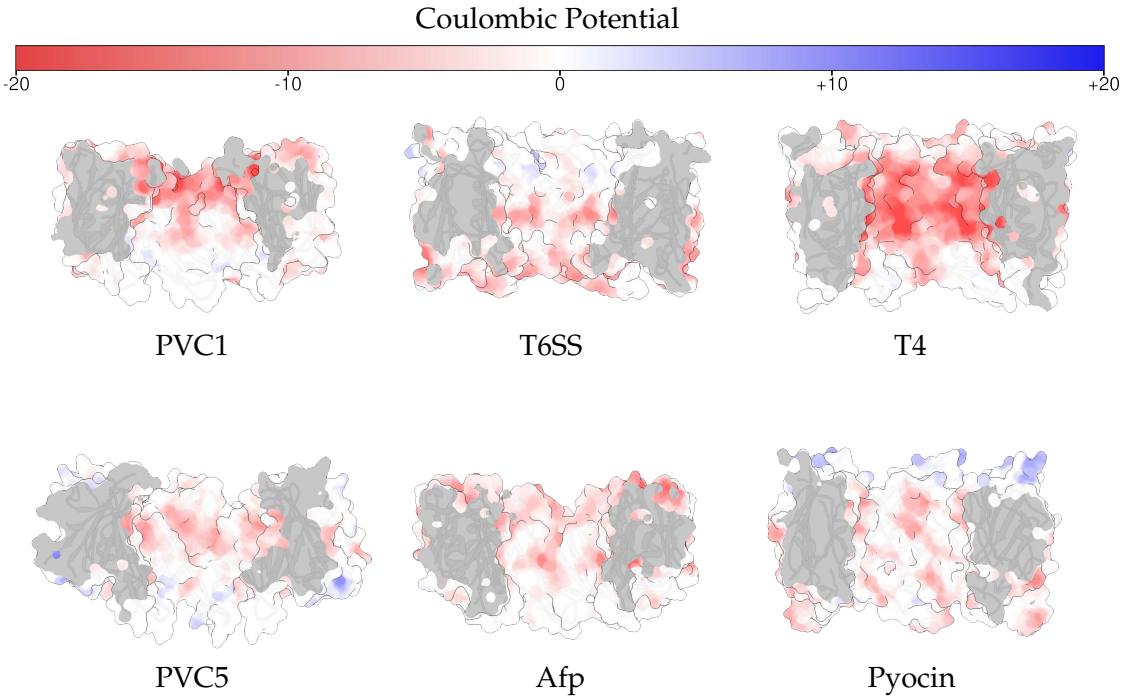


Figure 3.11 | The Coulombic potentials of caudate structure inner sheaths.
The various inner sheath protein orthologues are rendered by their Coulombic electrostatic potential (i.e. charge), from strongly negative (dark red) through neutral (white with transparency) to strongly positive (dark blue). In consensus with the study of Ge *et al.* (2015), the R-type pyocin has a neutral to moderately negative charge in the interior. The sheath hexamers are shown as cut-aways (translucent grey faces show the 'cut' path). The T6SS has mostly neutral charges (and a mixture of positives and negatives). The T4 tube appears to be overwhelmingly negatively charged, c.f. Ge et al.'s findings for the λ and PS17 phages which were seemingly positively charged. The PVC (in both loci) and Afp, have similarly neutral-to-negative charges. In the case of PVC1, a halo of greater negative charge appears around the top, which is discussed further subsequently. The augmented reality model to accompany this figure shows the PVC1 hexamer in full 3D to give an idea of the all round electrostatics and structure.

Amongst all PVC1s there does not appear to be a great discrepancy in the surface potentials. All are largely negatively/neutrally charged. There are a few cases where some more prominent positive charges appear. In the case of PVC1, the exterior of the inner sheath is highly negatively charged. However, there is a prominent difference with PVC5, as it maintains a neutral and even positively charged exterior by comparison. This

is surprising as it would be expected that if each monomer has to interact with an exterior sheath protein, and the sequences are largely conserved as shown previously, that both loci would carry similar charge profiles on this face.

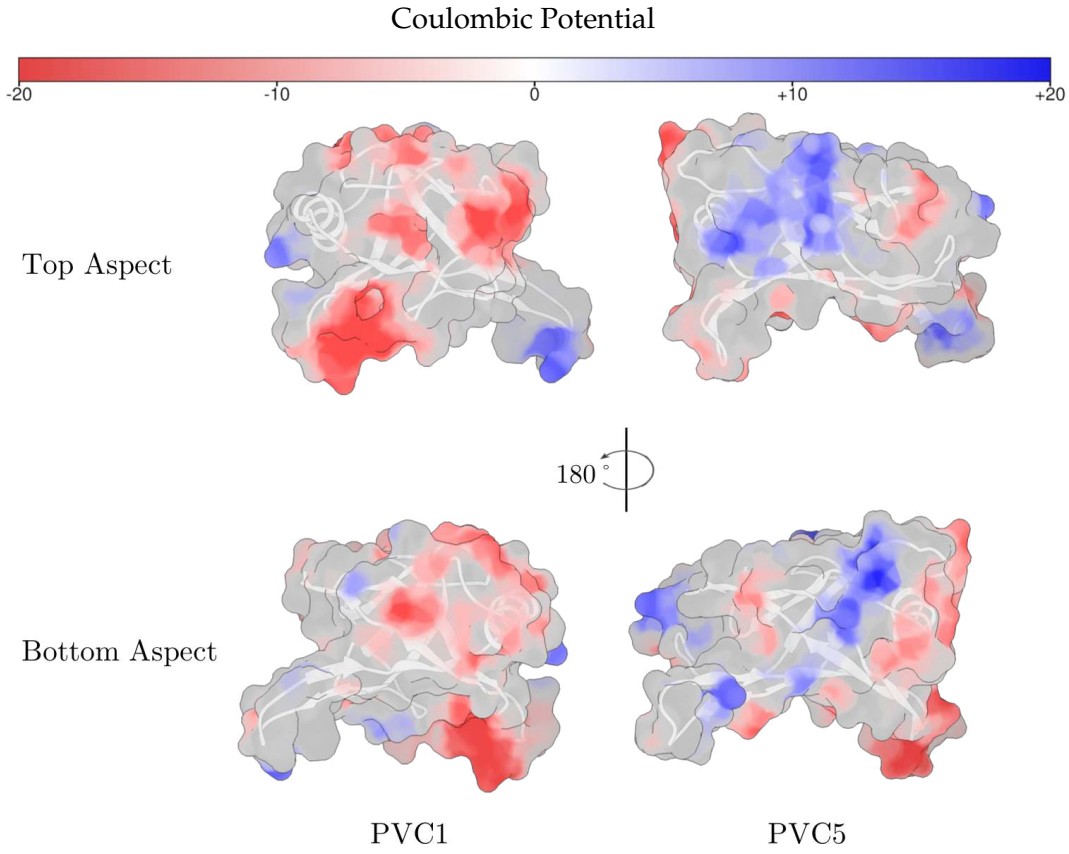


Figure 3.12 | Polarities of the top and bottom faces of inner sheath monomers.
The electrostatic potentials of exemplar inner sheath monomers from PVC "Pnf", loci 1 and 5. PVC1s exhibit predominantly positive charges on the 'top' and 'bottom' faces (the faces between disks of the tube), whilst PVC5s exhibit predominantly negative charges. This suggests that the 2 loci may form alternating hexameric stacks to assemble the tube in a directed manner like so: `(++)(--)(++)(--)`, where `(+-)` represents a monomer and its charges at either end.

There may be further subtlety to the structural assembly of the PVCs incorporating these proteins, as the other predominant electrostatic pattern shows that PVC1 has predominantly negative charges at its 'top' and 'bottom', whilst PVC5 has predominantly neutral/positive charges, as demonstrated in Figure 3.12. Therefore, one hypothesis is that the PVC is actually built from alternating stacks of PVC1-PVC5-PVC1-PVC5 with the negative and positive faces interacting. This is in contrast to, for example, the Pyocin tube, which has a prominent negative face, and a positive face within the same hexameric disk (visible in the bottom right of Figure 3.11 on page 118, akin to bar magnets placed

end-to-end, thus forms directionally controlled spontaneously self-assembling tube. As per the rest of the chapter so far Figure 3.12 on page 119 shows the case for just one of the inner sheath pairs (PVC1 and 5 from PVC "Pnf" again), but this largely holds for all 32 proteins. The side aspects of the proteins (not shown) demonstrate predominantly neutral charges, suggesting there is little significant within-disk ordering.

The outer sheath proteins remain something of a mystery. Not only is there often an additional copy of the gene, with very little obvious significant structural difference, but there do not appear to be drastic differences in charge between them. They all have predominantly neutral/positively charged interiors, but overwhelmingly negatively charged surfaces. Even with the sequence variability demonstrated in Figure 3.8 on page 114 and 3.9, there appears to have been a drive to maintain a negative overall surface charge. As an experimental indication of the models validity, an anion exchange chromatography protocol has been developed in the lab which relies on positively charged resins to retain negatively charged molecules and has worked well for purification of several PVCs to date.

As their inter- and intra-protomer interactions are more complex, Figure 3.13 on the next page shows a half-sheath where 3 of the 6 helical protomers have been scaffolded against the R-type pyocin's structure (a rise of 38.4 Å, and a twist of 18.3 Å). Each protomer is made of one of the 3 paralogues from the "Pnf" operon of ATCC43949. Ge *et al.* (2015) suggest that the majority of interactions in the outer sheath are within a helical protomer, and this manifests in one of the outstretched 'arms' of the outer monomers, carrying a negative charge which interacts with a positively charged groove at the neck of the monomer's two upstretched helices, below and to the left of it (as the helix is left handed).

The most compelling alternative hypothesis, certainly for the inner sheaths, is that one of the pair is specialised as an adapter or collar. Given that PVC5 is translationally coupled to PVC6, another putative structural protein discussed in the next section, it might be likely that this is the case for PVC5. In Figure 3.12 on page 119, some greater structural differences do seem to manifest when viewed as a sliced hexamer. PVC5 has a slightly 'squashed' confirmation, being somewhat wider than the PVC1 hexamer, and seems to have a slightly more restricted lumen suggesting that it is perhaps less suited

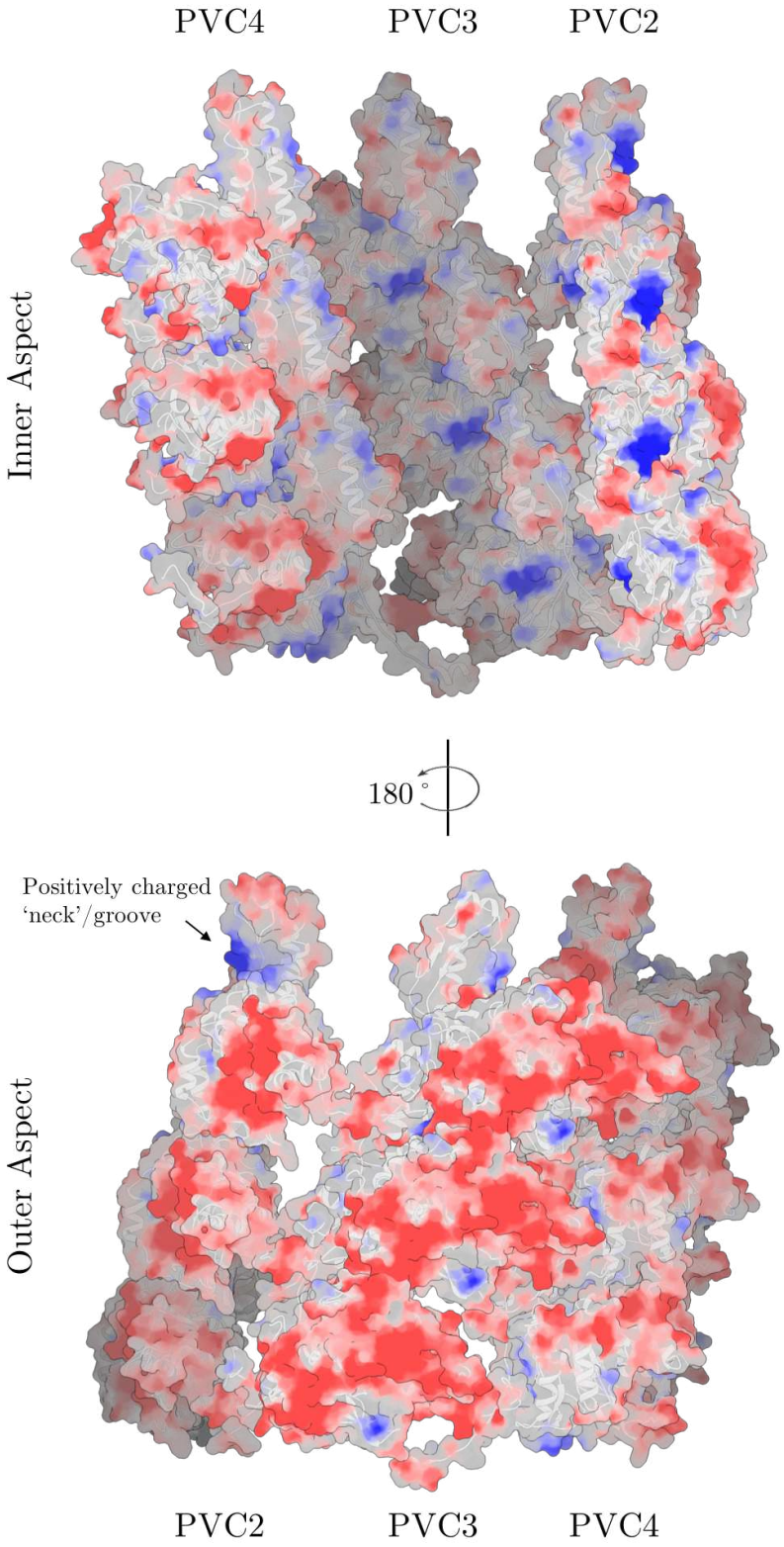


Figure 3.13 | Interior and exterior aspects of outer sheath protomer electrostatics.
The interior and exterior of the outer sheath proteins in biologically relevant positions/conformations are shown, scaffolded positionally against the R-type pyocin tube structure (PDB ID 3J9Q). All three paralogues share similar electrostatic profiles, with the tube interface as neutral/positive, but the exterior overwhelmingly negatively charged in all cases.

to comprising the bulk of the tube. This fits with legacy 2D gel proteomic analysis which found PVC1 in abundance but lower levels of PVC5, and would also fit with the Coulombic patters noticed above in that PVC5 did not demonstrate the same outer aspect potentials as PVC1 (not shown explicitly), instead being much more neutral in charge. This might make sense if the hexamer doesn't have to interact with the outer sheath proteins like PVC1 might. Figure 3.14 on the next page shows a colour coded, stripped down, T4 baseplate complex, demonstrating that there are 2 additional proteins, gp48 and gp54 (red/orange), which form a transition between the inner tube and the spike, and resemble subtly modified inner sheath proteins. Interestingly, HHPred fails to attribute this orthology, instead matching primarily to the tube proteins themselves which may also have influenced the homology modelling process. The figure shows the bulk of the baseplate hub in monochrome shades, the tube in yellow, and 2 fitted PVC proteins as white surface models. Outer sheath proteins are not shown.

Better RMSDs are obtained when matching PVC5 to gp54, than to gp48, so it is unclear whether there is another protein which fulfils the role of gp48 or whether PVCs have a simplified baseplate; certainly, their baseplates lack the vast majority of peripheral T4 baseplate proteins (which are omitted from the figure but the extent can be seen in the EM map). To complicate matters further, a recent paper by Renault *et al.* (2018) has demonstrated that, in the T6SS at least, the Hcp hexamers (equivalent to PVC1/gp19) interact directly with the face of gp27 in the spike complex explored in the next section. This may suggest that in this particular aspect, the PVCs are more similar to T4 than the T6SS, alternatively, PVC5 may form an adapted inner tube at the distal end of the tube, interfacing with terminator proteins to cap off the tube (similar to gp3-13-14-15 in T4 (Fokine *et al.*, 2013), though no orthology to those proteins has ever been detected).

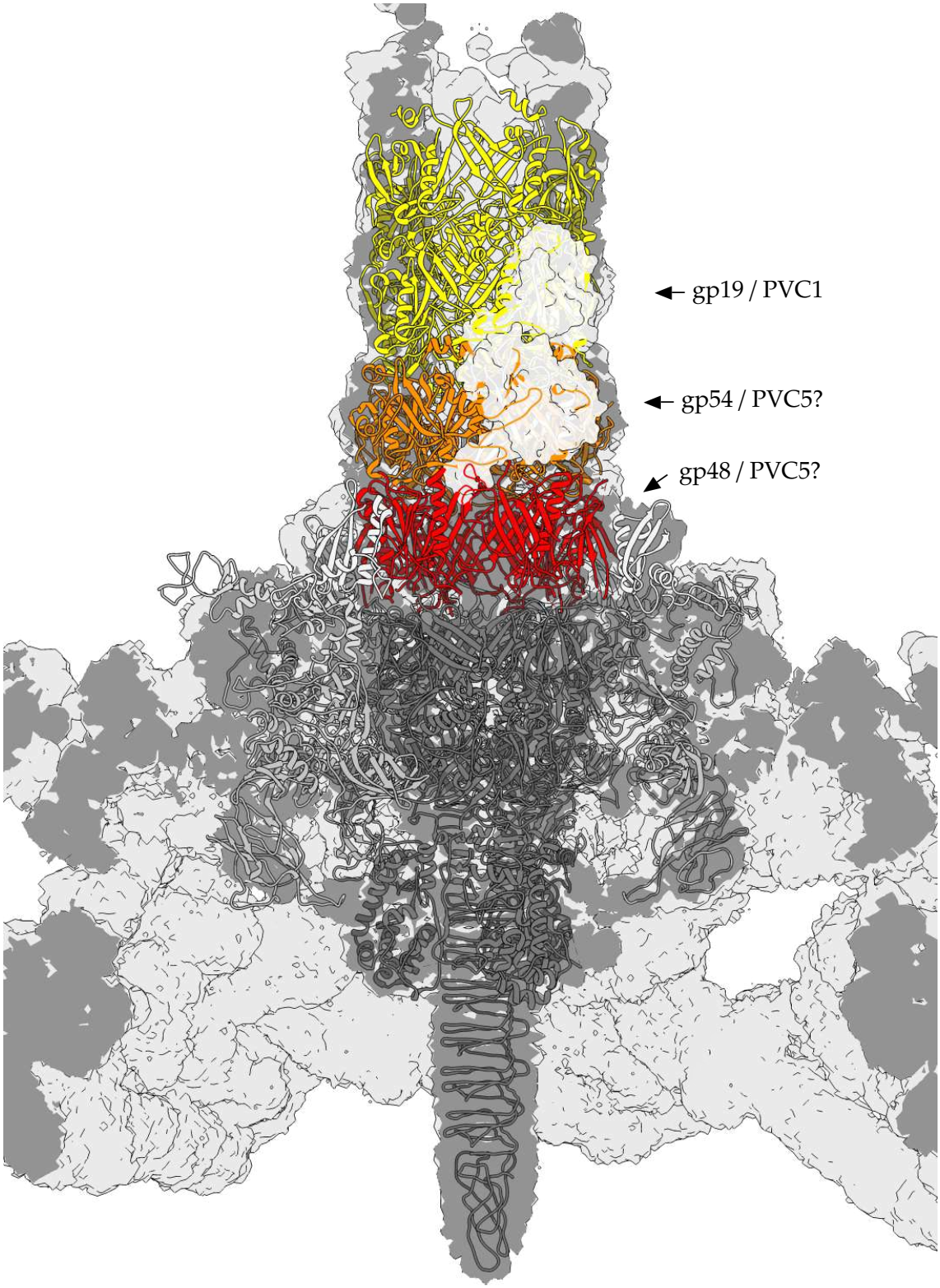


Figure 3.14 | Colour coded T4 baseplate collar proteins compared to PVC subunits.
The baseplate hub and collar of the inner tube for the T4 bacteriophage (PDB ID 5IV5 and EMDB 3374) are shown colour coded to identify the adapter proteins that interface the spike complex (monochrome) with the tube (yellow). This chapter proposes that PVC5 is not actually a tube monomer, but instead may form one of these 2 strata of adapters in a simplified complex. The accompanying 3D model for this figure omits the surfaces and EM map as the geometry count is too high and transparency is not fully supported.

3.2.2.2 The Spike Complex

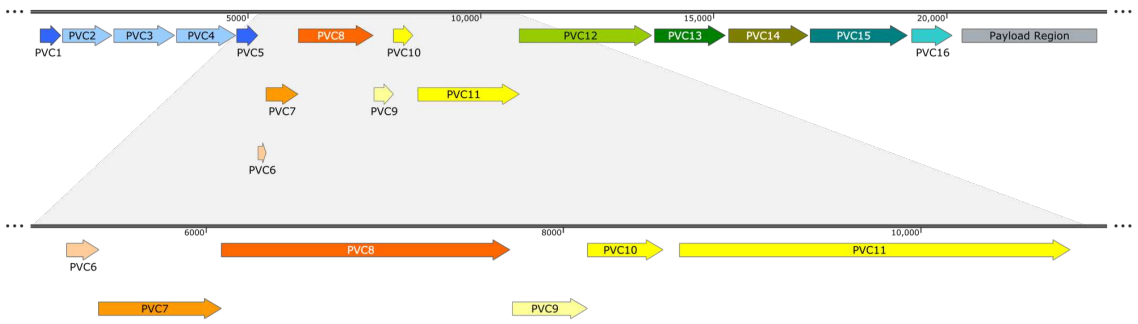


Figure 3.15 | The six loci predicted to comprise the spike and baseplate complex of a PVC.

(Top) The Pnf operon is used as an exemplar of the numbering of loci within an archetypal PVC operon, and three colourings are used to demarcate the 'functional units' of an operon: blue - tube proteins, orange/yellow - spike complex proteins, green/cyan - the operon 'core'. **(Bottom)** The five loci that give rise to the functional unit of the spike complex itself are blown up below as an aid to understanding the operon organisation and the proteins under discussion.

The next five genes are a mixture of well resolved structures and complete enigmas. The most prominent gene in this cluster is PVC8, a VgrG/gp27-5 orthologue which is proposed to form the membrane puncturing spike at the tip of the tube. A complex of miscellaneous baseplate proteins and the spike itself are proposed to be a kind of 'nucleation' site for the initial assembly of the tube. Given their syntenic position and the fact that several of the genes are translationally coupled, it seems likely that they may all have roles in the structural basis of the tube, likely within a collective collar/baseplate. PVC6 is actually translationally coupled to PVC5 in the 'tube unit' of the operon, but without any good orthologies to suggest possible functions, the best hypothesis is that it may form an adapter or collar protein which interfaces the inner tube with the spike protein, since there is a transition at this interface from the 6-fold symmetry of the tube, to 3-fold in the spike complex.

A number of the (extremely weak) orthologies detected for PVC loci 6 and 7 seem to suggest enzymatic domains so it may be the case that these proteins are involved in the assembly of the PVC, but not in the structure itself.

Table 3.4 | HHPred orthology summary for the putative baseplate and spike complex.
A summary of homology matches via HHPred for PVC loci 6-10. They represent a 'collapsed' set of common or plausible hits from all the variants for each locus. Many of the loci in this section of the operon have poor orthologies detected. PVC8 and 9 are the only proteins with high scoring orthologies detected.

Locus	PDB ID Hit	Structure	Component
PVC6	Various	Various	Assorted enzymes/kinases (?)
PVC7	5A3A	n/a	SIR2 Family metalloprotease (?)
PVC8	2P5Z	T6SS	VgrG
PVC9	2IA7	T4	Tail lysozyme
PVC10	4JIV	T6SS	PAAR-spike tip (?)
PVC11	3H2T	T4	Baseplate structural protein gp6

(?) denotes common or potentially plausible hits with very weak scores.

3.2.2.2.1 Enigmatic putative collar/baseplate proteins

Proceeding in turn, PVC6 demonstrates no reliable orthologies whatsoever. Even with the more sensitive approach of using HMMs, the lowest E-value obtained for a hit is >1. The only potentially informative information available is that the proteins are all quite short, at only ≈61 amino acids, and the structural simulations resulted, almost universally, in a pseudo-L-shaped helix-turn-helix structure, with a degenerate position forming the start of the turn (all structures not shown for brevity). While this doesn't provide information much to go on, it may suggest that this protein forms an augmentation to another structural protein, or perhaps forms a kind of lynchpin to lock elements of the structure together, given its size and potentially simple shape. One might expect that the protein would be larger if it were likely to have an enzymatic role or similar.

In the case of PVC7, the hits are mixed, but dominated by matches (albeit weak ones) to an SIR2 family ADP-Ribsosyltransferase from *Streptococcus pyogenes* (Shore, 2000), with the next most plausible hit being to an EssC-like chaperone of the *Pseudomonas* Type 3 Secretion System (Vogelaar *et al.*, 2010). While again, drawing too many conclusions from such weak similarities is unadvisable, it may be the case that these hits indicate that PVC7 has a role in assembly rather than the structure itself, potentially though some enzymatic or chaperone-like mechanism. Since the structural proteins are typically identifiable through structural homology, it is possible that PVC7 does not contribute directly to the structure proper. It is a much larger protein therefore the models simulated are likely

among the more spurious, and they do indeed have some of the larger RMSDs to even the closest HMM matches (on the order of ≈20 Å). Both PVC6 and 7 are chief among the 'unsolved mysteries' of the operon.

3.2.2.2.2 The putative spike complex and PVC baseplate hub is more reminiscent of the T6SS

This chapter proposes the idea that this section of the operon is made up of proteins which contribute to a baseplate and spike complex in some form or other. This may not be the case for the enigmatic PVC6 and 7, but almost certainly is for PVC8, 9, and potentially 10. Firstly, PVC8 has long been identifiable as a structural orthologue of the VgrG class of trimeric 'spikes' which sit atop the inner tubes of every caudate structure identified to date. In T4, this spike is comprised of gp27 which forms a sort of trimeric adapter which interfaces with the hexameric tube, bridging the difference in symmetry. As mentioned in Chapter 1, VgrG's can be decorated with many additional functional domains (also forming toxins in their own right in some cases). The models obtained here suggest that PVCs employ a simplified spike complex, closer in structure to a T6SS like that of uropathogenic *E. coli* (Leiman *et al.*, 2009), than like T4.

Particularly prominent appears to be the loss (in the non-T4 complexes) of the 'wings' that the T4 spike exhibits, flanking the extended "β-prism". Kanamaru *et al.* (2002) identified these as tail spike lysozyme domains of the gp5 protein, with which T4 is proposed to 'chew through' the peptidoglycan cell wall of target bacterial cells. With no cell wall to negotiate in eukaryotic targets, it stands to reason that the T6SS, PVCs and Afps have lost this (though the T6SS does mediate some prokaryotic interactions as well, so may have other mechanisms it can substitute in to deal with peptidoglycan). In T4, the structure is comprised of two separate proteins, gp27, which forms the 'OB-fold' at the base of the structure and gp5 which contains the lyosozyme domain and forms the prominent β-prism at the very apex of the spike. In the simulations, this upper prism area is less well defined, and is even unresolved in the crystallised VgrG from the uropathogenic *E. coli*. Based purely on protein chain length however, it does appear that the stretch of prism in the PVCs and Afps might be shorter. As a further example of

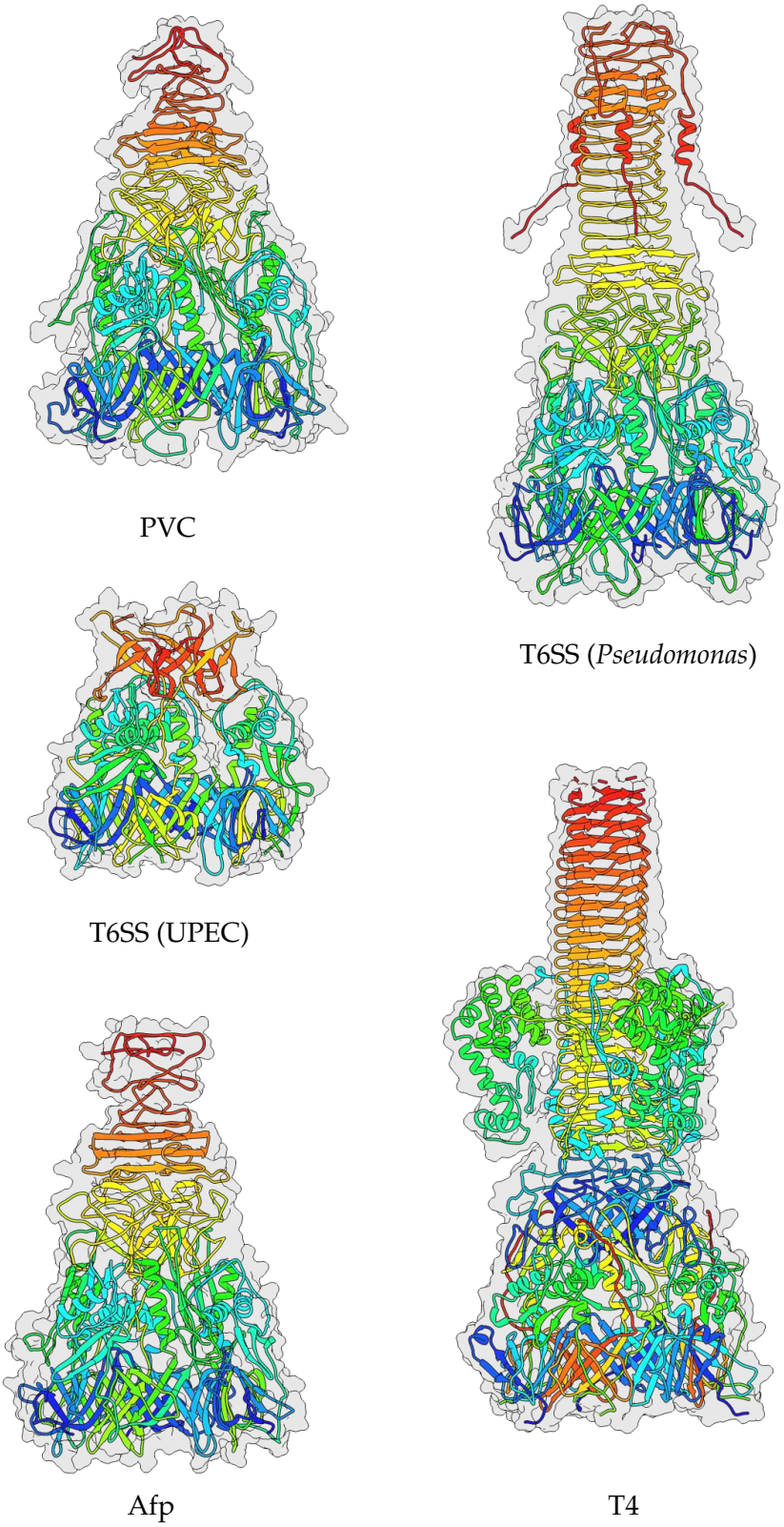


Figure 3.16 | The PVC8 spike complex compared with homologues from T4, T6SS and Afp. As one of the most distinctive proteins in the operon, annotations for this protein have proven reliable, however VrgG-like spikes are well known in the literature to be variable and share little sequence similarity. PVC spikes appear more like a simplified T6SS than like T4. The accompanying 3D model for this page shows a chain trace.

the sequence variability and apparent potential for drift or adaptation of the VgrG-like spikes, the distance between the PVC and Afp orthologues is comparatively higher than for other loci studied so far, even with very similar tertiary structures among orthologues. This is consistent with published observations of high VgrG variability among caudate structures.

3.2.2.2.3 A new suggested role for PVC9 in tube initiation

PVC9 has been identified by sequence searches and annotation, as an orthologue of the PDB ID 2IA7 - a putative "tail lysozyme" from *Geobacter sulfurreducens*. This has always been a point of some confusion though, as it is not clear what role a lysozyme would have in the biology of PVCs, since they have never demonstrated any antimicrobial activity. With the loss of the gp5-like lysozyme domain in the PVC8/VgrG/gp27-5 orthology locus, it seemed that this might present a domain split, and that this domain had moved to a locus of its own. Nevertheless, with no cell wall to act upon in anti-eukaryotic virulence, the only other hypothesis proposed was that this may be a mechanism of release for the PVCs, by degrading the host bacterium cell wall. This also seemed some what unlikely, since the locus is carried squarely within many other putatively structural proteins.

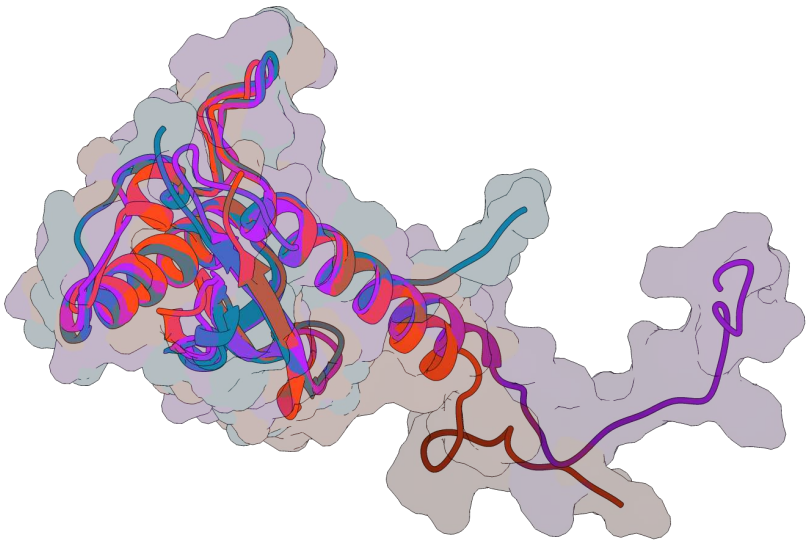


Figure 3.17 | PVC9 COMPARED TO A PUTATIVE LYSOZYME AND TUBE INITIATOR.
PVC9 (purple), structurally compared to PDB ID 2IA7, a putative lysozyme (blue), and the T4 tube initiator protein gp25 (orange). Based on their structure similarity, PVC9 is likely a tube initiator protein, and this was not previously identified due to a spurious annotation of lysozyme on the better sequence match for PDB 2IA7.

On inspecting the structures, PVC9 candidates do indeed closely resemble the best HMM hit, PDB ID 2IA7. However, they also closely structurally match the gp25 baseplate initiator protein of the T4 bacteriophage (PDB ID 5IW9), even though the E-value for this match (whilst still good) is three orders of magnitude greater (worse). Figure 3.17 on page 128 demonstrates the similarity between an example PVC9, the *G. sulfurreducens* structure, and gp25 of T4 in purple, blue and orange respectively. The RMSD between the putative lysozyme and the example PVC9 used is 1.2Å, whilst between PVC9 and the gp25 PDB ID 5IW9 is less than 0.3Å.

Since the role of a tube initiator protein is relevant to the structure and assembly of all caudate structures (as far as is known presently), and no other compelling candidates have been identified so far, PVC9 is therefore likely fulfilling this role. The descriptor ascribed to the PDB ID 2IA7 of a putative lysozyme is almost certainly incorrect.

3.2.2.2.4 Identifying subtle hallmarks of potential spike "PAAR" proteins

PVC10 has been another enigmatic protein for some time and still recruits matches to a number of different potential orthologues including proteins such as RNA polymerase subunits and Amyloid proteins. In just a couple of cases when querying the various PVC orthologues of this protein, homology was noted to "PAAR"-repeat motif proteins, though only ever for the same one or two sequences. "PAAR", or Proline-Alanine-Alanine-Repeat proteins are integral components of most caudate structures studied to date, but are extremely variable and difficult to detect. The T6SS alone employs a bewildering diversity of these proteins (Shneider *et al.*, 2013). Curiously however, on first inspecting the sequences, there are few repeats to speak of when analysed by RADAR, and none of the orthologues even contain the "PAAR" subsequence itself.

PAAR-repeat proteins sit at the apex of the extended β-prism of the spike complex, sharpening the otherwise slightly blunt face, down to a single atom's diameter potentially. Shneider *et al.* (2013) describe PAAR proteins, in the case of model PDB ID 4JIV, as being comprised of nine short β strands, three of which form a base interacting with the rest of the spike complex following the β-prism conformation, and the remainder forming β hairpins of different length.

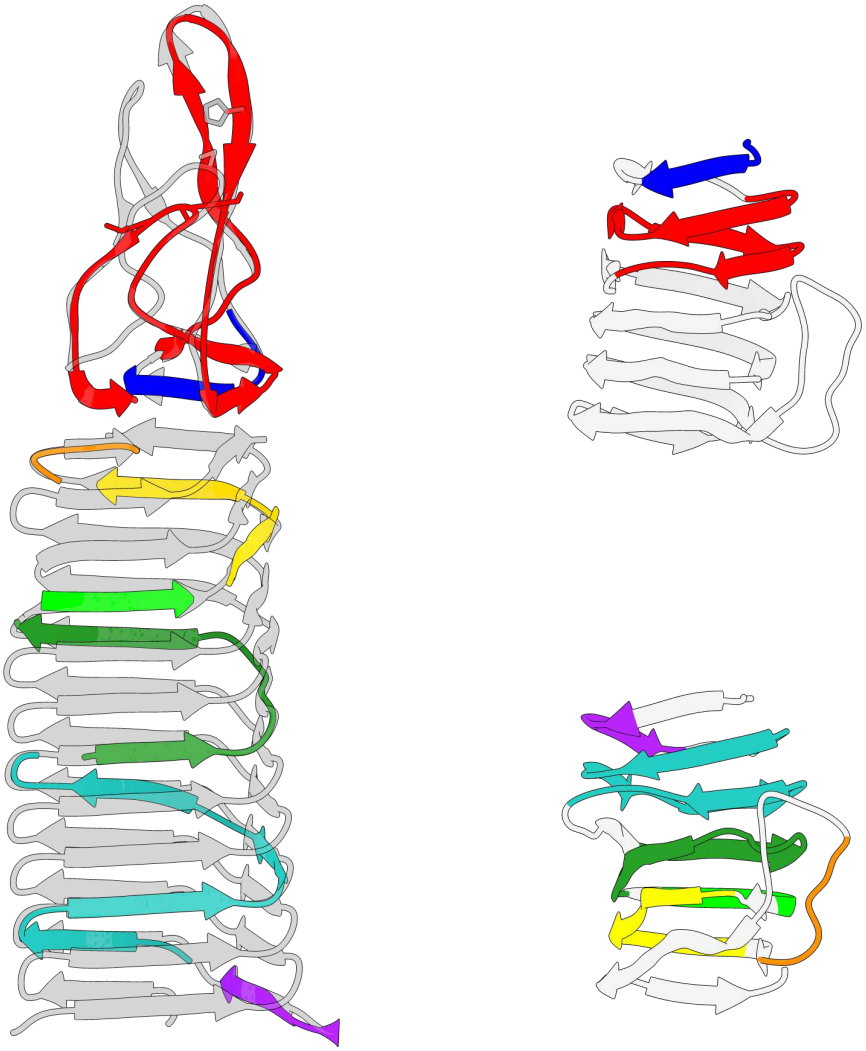


Figure 3.18 | Exemplar PVC10 model with conformations mapped to the T6SS spike.
An example of a PVC10 locus obtained model is shown (twice) on the right hand side. These two models have had their conformations mapped on to the spike and β-prism domain from a T6SS on the left (PDB ID 4JIV). Unmapped stretches are shown in white and the reference structure is shown in grey. Colours indicate contiguously mapped stretches of sequence. Side chains are shown for metal coordinating amino acids which are relevant to Figure 3.19 on page 132.

The simulated models obtained are fairly poor facsimiles of *bona fide* PAAR proteins of resolved structure, however, all of the models share some characteristics which support the hypothesis. The majority of models adopted a β-sheet dominated structure, which run head to tail, reminiscent of the β-prism conformation seen at the base of PAAR protein spikes, suggesting they augment the prism, if not form the spike itself. As these proteins are slightly longer than canonical PAAR proteins, it may be the case that

they incorporate more of the β-prism in to the spike itself, which could account for the slightly shorter prisms seen in the PVC8 models, even taking in to consideration limitations in the simulation quality. This gives the models a similar triangular cross section to the PAAR proteins. Where the structures are not β-stranded they produce turn regions, which would provide the alternating strand-turn-strand conformation needed to enable β-hairpins like those seen in the resolved structures. These proteins possibly present a limitation for the simulation software, lacking good tertiary structure models and therefore resorting to largely molecular dynamics based approaches, which has given an energy minimised 'collapsed' structure. However, Figure 3.18 on page 130 shows how the secondary structure and sequence of one of the putative PAAR protein model compared to the β-prism and PAAR protein from a *Vibrio* T6SS. The conformations of the simulated structure are morphed on to the spike complex (in grey, PDB ID 4JIV). The largest span of amino acid/secondary structure similarity forms two of the major loops of the spike (shown in red on the upper monomer). There is some split structure which matches an additional β-strand region in dark blue (domains are split when an alignment gap is created). As mentioned, the PVC10 loci tend to be longer than canonical PAAR loci, and in this case, a significant proportion of the structure remains un-morphed on to the reference structure. Since one of the three major loops is missing, this suggests that part of the structure is too low similarity to be matched, but would mean that at least part of the unmatched structure (shown in white), could potentially contribute to the spike. Nonetheless, the increased length of the protein, the dominant β-sheet, and roughly triangular cross section might suggest that the proteins contribute to the spike prism 'stem' as well as the spike itself.

Despite the fact that long stretches which can align to the prism or the spike are identifiable (the red region of the spike and the cyan region of the prism map to the same two secondary structure spans in the models), the spike contains the longest matched stretch. Moreover, in order to form the prism, three copies of the protein are needed. If the locus forms the spike as well as the prism, this would result in three spike domains which is unlikely. Moreover, to map the sequence on to the prism domain reference, many more sequence breaks were required, though a greater total span of PVC10 sequence could

be morphed on to the structure.

Further evidence is present in the simulations for this loci's role as a spike tip; though they are certainly not fully correctly positioned, as the process of morphing conformations introduces strained geometries. The reference structure uses two histidine, and one cysteine residues to coordinate a metal atom at the tip of the spike, which holds the six strands of the three loops in place. In this model shown in Figure 3.19, one of the loops is not well modelled as mentioned, however in the two that were, one of the histidines is replaced by an arginine and the cysteine position can be superimposed over by a lysine (shown as ball-and-stick-models). This is consistent with the proposals of Shneider *et al.* (2013) who note that these residues may be replaced with "similar or complementary metal-binding residues (arginines, lysines and glutamines) in more distant homologues suggesting that they also carry a metal ion roughly at the same position".

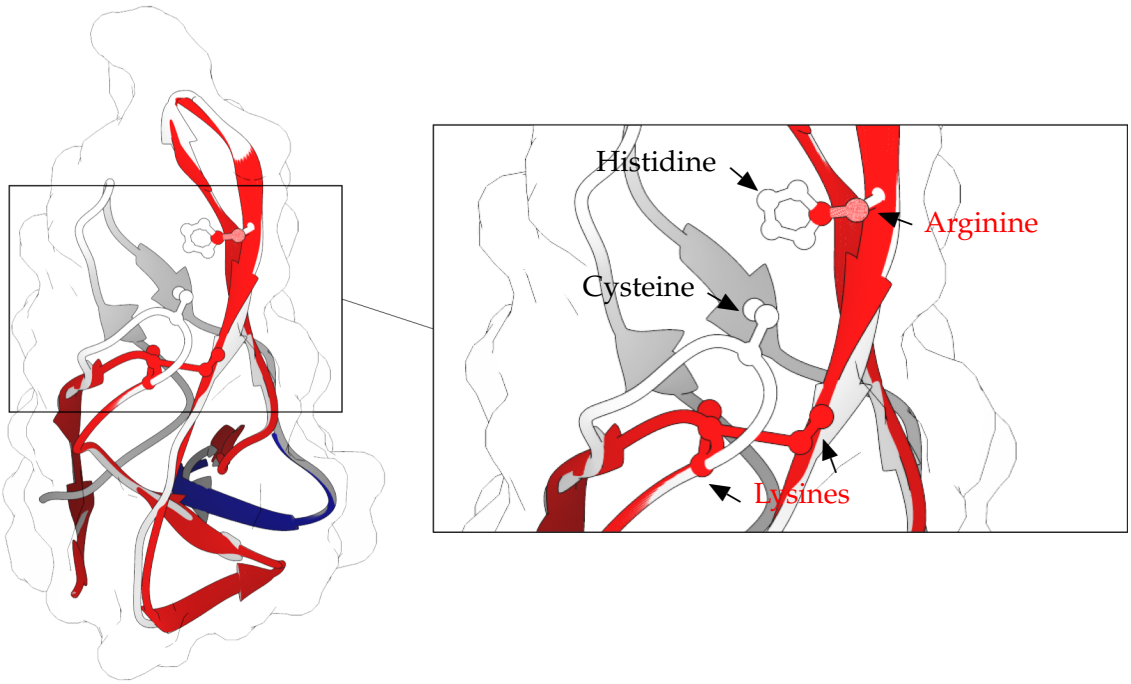


Figure 3.19 | Evidence for potential conservation of metal binding in spike proteins.
The superimposed spike proteins preserve the localised positioning of metal coordinating residues (metal atom not shown). The reference T6SS structure is shown in white, the putative PVC spike protein in red. The positioning of an arginine in place of a histidine in the top most loop is very similar. The lower loop has some strained conformations so will not completely reflect the true structure, but conserves a lysine capable of metal coordination in place of a cysteine, and an arginine in place of a histidine, in the reference structure consistent with the hypothesis of Shneider *et al.* (2013).

3.2.2.2.5 PVC11 as the major baseplate structural component

There is not much to be said for the last gene in this cluster, PVC11. It has had high quality orthology matches to the gp6 major baseplate hub protein of the T4 phage for some time. As a large gene, it likely contributes the bulk of the PVC baseplate which is certain to be drastically simpler than that of the T4. In T4, gp6 is critical to the conduction of the contraction signal, which likely accounts for why PVCs retain an orthologue. The only significant feature of these proteins appears to be that they have conserved N and C-termini, as seen in the alignments in Appendix Chapter B on page 335, but 3 of the 16 proteins studied (all belonging to the "LopT" operon) retain an additional lengthy expanse of sequence. Interestingly, the "LopT" operons are one of the few exceptions in terms of conservation of their 'operon cores' which are explored in the next section. It appears that they do not harbour receptor binding 'tail fibres', so perhaps the baseplate itself has some additional augmentation to replace them for allowing for binding/contraction, perhaps in a non-specific manner, as generalist delivery mechanism.

3.2.2.3 The Operon Core

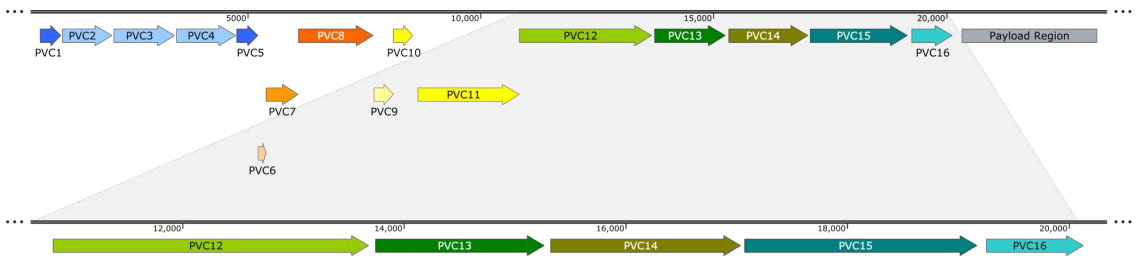


Figure 3.20 | THE FIVE LOCI COMPRISING THE 'CORE' OF A PVC OPERON WITH VARIOUS FUNCTIONS. **(Top)** The Pnf operon is used as an exemplar of the numbering of loci within an archetypal PVC operon, and three colourings are used to demarcate the 'functional units' of an operon: blue - tube proteins, orange/yellow - spike complex proteins, green/cyan - the operon 'core'. **(Bottom)** The five loci that give rise to the operon core are blown up below as an aid to understanding the operon organisation and the proteins under discussion. The 'core' is a collection of conserved proteins present in almost all PVC operons, typically comprised of larger, single copy genes, some which are proposed to be structural, but also the proteins responsible for cell surface adhesion, and an ATPase which potentially loads payloads or recycles the tube apparatus.

The final 'module' of a PVC operon that this chapter will discuss is what is being termed here as the 'operon core' (the chapter will ignore the payload region of the PVCs as that likely presents a thesis in itself). The operon core is typically reasonably well conserved, though poorly characterised - almost always containing 5 generally larger, single copy, proteins.

Table 3.5 | HHPRED ORTHOLOGY SUMMARY FOR THE PUTATIVE BASEPLATE AND SPIKE COMPLEX. A summary of homology matches via HHPred for PVC loci 6-10. They represent a 'collapsed' set of common or plausible hits from all the variants for each locus. Many of the loci in this section of the operon have poor orthologies detected. PVC8 and 9 are the only proteins with high scoring orthologies detected.

Locus	PDB ID Hit	Structure	Component
PVC12	Various	Various	Assorted nucleotide binding (?)
PVC13	1V1H / 2FKK	T4	Tail-fibre/baseplate
PVC14	Various	n/a	Assorted
PVC15	5E7P	n/a	AAA+ ATPase
PVC16	Various	n/a	Assorted

(?) denotes common or potentially plausible hits with very weak scores.

3.2.2.3.1 PVC12, an enigmatic nucleotide binding protein?

PVC12, typically the largest protein in a PVC operon, has remained uncharacterised.

Previous sequence annotations attributed weak similarity to so called "GGDEF domain" proteins. These proteins are ubiquitous bacterial regulators or signal sensors, and are proposed to bind cyclic-di-GMP, with the domain linked to other enzymatic protein domains, with known roles in processes such as exopolysaccharide synthesis (Ryjenkov *et al.*, 2005). Given its localisation in the operon it is tempting to assume it likely contributes another structural protein which is evidently quite unique, with no equivalent structures in the better characterised T4 or T6SS, however it was not seen in legacy 2D gels of particle isolations. With little to no informative orthologies detected nor any consistent structure between obtained models, PVC12 remains a mystery for the time being.

3.2.2.3.2 The putative tail fibres equivalents of PVC operons

PVC13 has long been of interest in the biology of PVCs, though the entirety of Chapter 5 on page 187 is dedicated to their experimental study, so they won't be covered in depth here. The tail fibres are the most diverse gene in the entire operon (also discussed in greater detail in the upcoming Chapter 4), and are missing from the "LopT" operons altogether. They appear to have split domain structures, and have attracted unusual 'anti-eukaryotic' homologies, such as Adenovirus, appearing to be fused to 'anti-prokaryotic' phage domains. In fact, a PDB ID hit which has come to predominate the matches is 1V1H, an artificial fusion between Adenovirus fibre heads, and T4 phage short fibres from the van Raaij lab (Papanikolopoulou *et al.*, 2004b).

The simulations are of mixed reliability/quality, likely due in large part to their diversity resulting in the selection of various templates for the threading stages of the I-TASSER pipeline. Figure 3.21 on the following page shows a selection of the lowest RMSD models obtained. They have been symmetry modelled as trimeric proteins, since tail fibres are known to be trimeric only. While the finer details of the models is likely incorrect due to the sequence divergence, the gross architecture looks promising, and most interestingly of all, the hypervariable region of the alignment (annotated on Figure 5.6 on page 199 in Chapter 5) can be seen prominently at one end of the structure (in red). At the opposite end, a largely conserved region can be seen which suggests that generally similar 'mounting hardware' might be used to fasten the fibres to the tube baseplate.

Figure 3.21 | Conservation mapping of tail fibres reveals a hypervariable binding region and conserved mounting region.

The conservation of amino acid positions for the extremely variable putative tail fibres of the PVCs. Tail fibres are responsible for target recognition and conveyance of the contraction signal to the contractile sheaths via the baseplate. PVC tail fibre models resemble the receptor binding tip (PDB ID 1PDI) of the short fibre of bacteriophage T4. The hypervariable domains are proposed to have homology to adenoviral motifs for eukaryotic targeting, and the high degree of diversity seen in the potential binding region speaks to the diversification of PVC variants against numerous target cell/tissue types. Inset is the T4 structure itself, a key to amino acid conservation and a small diagram for orientation since the perspective view of the tail fibres does not reproduce depth well in 2D. The accompanying augmented reality model for this page shows a space-filling surface of the fibres rendered as above.

3.2.2.3.3 PVC14 as a putative 'tape measure' protein

The work of Rybakova *et al.* (2015) showed that Afp14, ostensibly the equivalent locus to PVC14, is involved in determining the length of the tube structures. There is little known of the structural conformation of these proteins in general, compounded by the fact that they are proposed to span the length of a viral tail tube *in vivo*. This would mean that efforts to heterologously express and crystallise the proteins would likely not reveal their true, biologically-relevant or correct conformations. As discussed in the Introduction, these proteins are thought to exhibit a disproportionately high degree of helicity, as well as having distributed and broadly conserved hydrophobicities. This can be seen in resolved putative tape measure proteins such as that of PDB ID 5A20 (Chaban *et al.*, 2015).

Inspecting the HHPred results for these proteins reveals no consistent hits, belying their variability, but many of them are to proteins such as a 3-helix bundle of *S. aureus*, a heavily α-helical nucleoporin of *Saccharomyces* and typically helical membrane associated structures such as Photosystem II of *Arabidopsis*. It is unusual that the homologies do not include hits to other tape measure proteins such as that of the phage SPP1 protein resolved by Chaban *et al.* (2015), further reinforcing the idea that sequence homology is often an inadequate tool for these types of proteins.

Taking a selection of models obtained shown in Figure 3.22 on the following page, it can be seen that there is little to no consistency in final globular structure (as would likely be the case with efforts to heterologously crystallise the proteins away from the tube), however, they are utterly dominated by alpha helical structures (the figures are coloured by secondary structure, with helix in magenta, β-sheets/strands in cyan, and coils in gray). Also rendered on the models is a hydrophobicity surface, which shows a homogenous distribution of hydrophobic/philic residues (orange/blue) across the structure, which is consistent with previously proposed orthologues.

Another indicative factor is that these proteins vary substantially in overall length, anywhere from ≈450 to ≈650 amino acids. As was also explored in the Introduction, Rybakova *et al.* (2015) and Pedulla *et al.* (2003) note that there is a strong linear correlation between final tube length and tape measure protein length. To a quick 'back of the

envelope' approximation using the trend identified, this would suggest that PVC tape measure proteins would be between ≈80 nm to ≈110 nm, based purely on sequence length - with the upper values roughly the same as the Afp itself. Estimates of the full particle length from electron micrographs suggest they are closer to 150-200 nm though, so it may be the case that this trend is not entirely universal, or perhaps fails to account for baseplate/spike complex contributions. The natural conformation for these proteins *in vivo*, therefore, is likely extended helices, lying along the length of a growing tube, connected by short unstructured linker spans, somewhat like a length of rope with 'α-helical knots' along it.

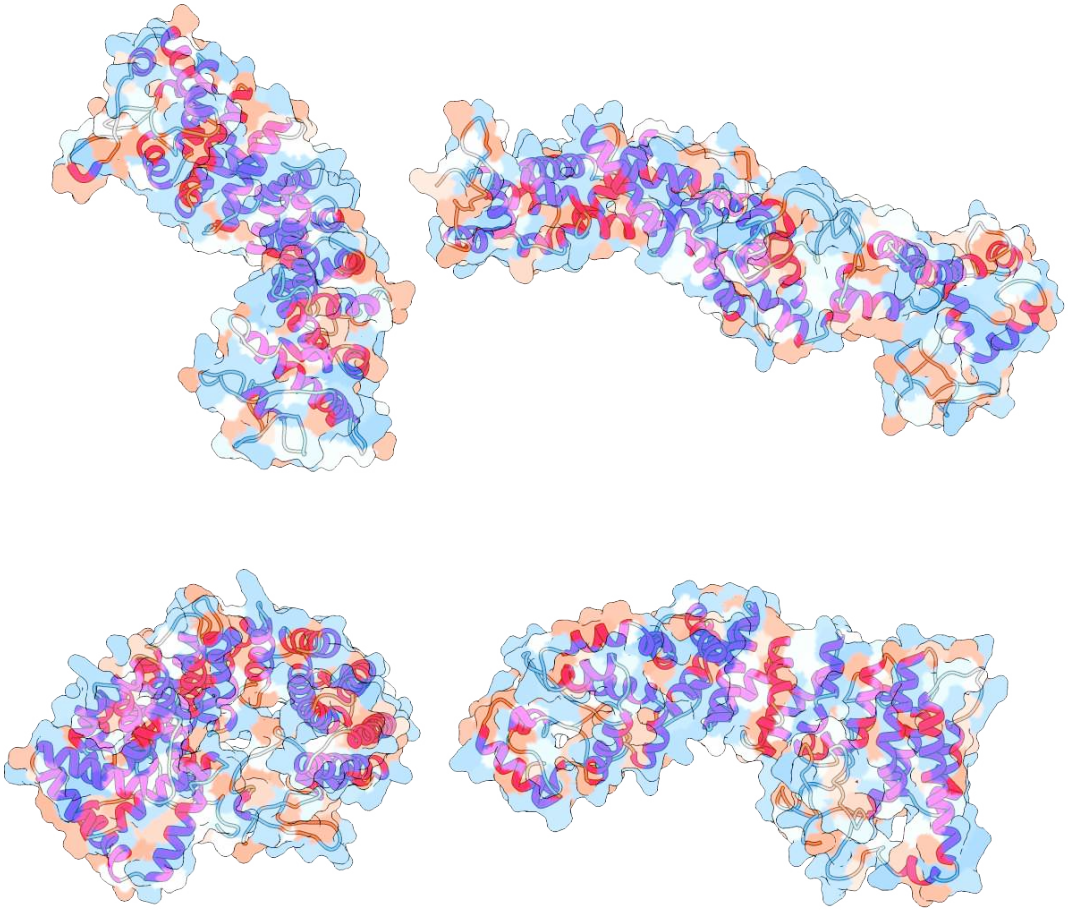


Figure 3.22 | Characteristic α-helicity and hydropathy of 'tape measure proteins'.
The simulated models for a selection of putative PVC14 tape measure proteins, revealing structure dominated, almost if not entirely, by α-helix (magenta ribbons). Also rendered is a molecular surface coloured according to hydrophobicity/hydrophilicity, with the most hydrophobic residues in orange, and the most hydrophilic in blue. Tape measure proteins are proposed to have a largely conserved and homogeneously distributed pattern of residues with different hydropathy.

3.2.2.3.4 The conserved, characteristic, ATPase of PVC operons

The ATPase carried by the PVCs and related structures are well conserved. In the original genome annotations, they are the only CDS which is given a 'proper' gene name, rather than simply a locus tag. Namely, as an FtsH orthologue, the ATPases contain a characteristic protease domain at the C-terminus, and a comparatively variable N-terminus. The role the ATPase plays is still unknown, though in the T6SS the similar ClpV ATPase is responsible for recycling the tube structure. Since the proteins are conserved, but have unknown mechanistics within a PVC operon, they won't be covered in significant detail here.

Scrutinising the models obtained unsurprisingly reveals consistent structures between operons (not shown) (since the sequences are also highly similar). The active site of ATPase is preserved when compared to other known AAA+ ATPases, though the non-ATPase active site domain differs fairly substantially from structural and sequence orthologues such as PDB ID 5E7P for instance. The only remaining clue as to their role in PVC mechanistics, is that Hurst *et al.* demonstrated that the ATPase is required for disease causing phenotypes in the Antifeeding prophage (Hurst *et al.*, 2004; Rybakova *et al.*, 2013), and its deletion led to lack of assembly or misassembly of the complex.

3.2.2.3.5 PVC16 as a tube terminator or cap protein

Hurst and colleagues further demonstrated a role for the sixteenth gene of the operon in potentially capping the growing tube, leading to its growing length termination. Deletions and truncations of these proteins lead to aberrant forms of the structures. Similar to the putative tape measure proteins, no existing reliable structural or sequence orthologies are readily identifiable, with no E-values less than 1 that could be detected. The only consistent features of the obtained models are the presence of ≈3-5 helices, with the remainder of the structure being comprised of disordered turns. β-sheets were only observed in one or two models. If PVC16 is indeed a cap protein, it appears to be significantly structurally and sequentially different to any known functionally equivalent orthologues with resolved structures, and clues to its *in vivo* conformation remain sparse.

3.3 Discussion

The intention of this chapter as a results chapter was to explore new hypotheses for roles of as-yet not fully understood loci within PVC operons, whilst also acting as an extended introduction, explaining the existing hypotheses and orthologies and orienting the reader for the specifics of the chapters to come. To be clear, this chapter does not claim to have *solved* any structures; merely the intention was to exploit a rare opportunity to use high performance computing resources with impunity, in the hopes of providing data that might be used to inform practical experiments further down the line, such as targeted mutagenesis with the benefit of at least some structural understanding. Furthermore, as we continue to probe the structure experimentally in collaboration with the Max Planck Institute, these models may prove invaluable for characterising the PVCs in a high throughput manner, using the experimentally resolved electron density of a single PVC, to 'scaffold' other PVC structures against, without the need to experimentally resolve all of them.

PVCs represent a hybrid between the T4 phage, R-type pyocins and T6SS like structures. More specifically however, they appear to resemble the different structures in different parts of the overall apparatus, highlighting a genetic modularity different selection pressures. For example, the outer sheath proteins of PVCs appear more like the T4 phage/Pyocin, than they do like the T6SS for instance, whilst the opposite is true for the spike complexes. This could perhaps speak to the 'intracellular life' of a T6SS apparatus, having to cope with conditions inside the cell primarily, before the contraction is triggered; on the other hand, phage, pyocins, and PVCs would spend much more of their time in extracellular environments. Likewise, the PVCs carry the characteristic AAA+ ATPase which is also a typical synthesis hallmark of the T6SS, but is not the case for T4 phages (excluding the DNA packaging motor).

Hopefully this chapter has also proven the statement made in its introduction, that sequence-based orthology studies are insufficient, particularly when dealing with protein folds which probably first originated in deep time (with the first phages). These fundamental fold structures which give rise to the inner and outer tubes for example, appear robust to reuse in all manner of translocation applications (DNA/protein/ions etc.). The

co-opting bacteria exploit the redundancy and robustness of these structures to hone the surface properties such as hydropathy and charge to adapt to varied niches, all the while preserving an incredibly complex and elegant structure.

Table 3.6 summarises the new and existing proposed roles for the proteins within the PVC operons, though despite all the advancements made experimentally, and even hypotheses proposed here, there remain enigmatic proteins with few if any clues as to even a plausible role.

Table 3.6 | Summary of putative loci functions for PVC structural proteins.
A summary of the proposed primary roles for PVC structural loci. Loci with asterisks indicate new functions proposed as a result of this work, or new evidence which contributes to existing proposed roles.

Locus	Proposed structure/role in the PVC
PVC1	Inner tube
PVC2	Outer sheath
PVC3	Outer sheath (frequently deleted)
PVC4	Outer sheath
PVC5*	Possible collar adapter
PVC6	Unknown (augment/lynchpin?)
PVC7	Unknown (Possible enzyme/chaperone?)
PVC8	Spike trimer
PVC9*	Tube initiator
PVC10*	PAAR-like spike augment
PVC11	Major baseplate protein
PVC12	Unknown
PVC13	Host binding tail-fibre chimeras
PVC14	* Tube 'tape measure protein'
PVC15	Loading/Assembly ATPase
PVC16	* Tube terminator (?)

For the bulk of the PVC tube, it is interesting that the inner and outer tube orthologies do not share the same provenance. It might be expected that they should all match to the next closest structure (e.g. all matching to the *P. aeruginosa* R-type pyocin) with the operon effectively 'linking' the evolution of all the proteins, but instead the inner sheath proteins

retain T4 hits. This potentially indicates that the inner and outer sheaths may be subject to different evolutionary pressures, and that the inner sheaths require more substantial 'alteration' to their new protein translocating function, from the ancestral phage-like 'blueprint'. Additionally, interior sheath proteins are smaller and better conserved. This could indicate that *Photorhabdus* is honing the exterior of the sheath toward different environments. An appealing hypothesis, especially since legacy transcriptomic data has shown that different PVCs are deployed in very particular inducing conditions, and results obtained in the upcoming Chapter 6 on page 234 concur. Alternatively, its 'increased paralogy' (at three copies instead of two), may have allowed the structures to drift further.

This difference in provenance for the inner sheath proteins compared to the outer sheath proteins is consistent with published observations that, for instance, the R-type pyocin is generally negatively charged to convey protons for depolarisation, and that the T6SS is comparatively neutral for conveyance of proteins of various properties (Ge *et al.*, 2015). This is therefore indicative of the inner and outer sheath proteins potentially being under different selection pressures, reflecting their deployment environment (in the case of the outer sheath), their cargo (in the case of the inner sheath), and their ultimate role.

Why PVCs carry multiple copies of tube proteins has been a curiosity for some time. The original speculative hypothesis for this was that, as the bulk of the structure, they would need multiple copies to provide stoichiometric ratios to the rest of the operon components (i.e. two proteins would be produced 'per transcript'). There are flaws in this idea though. Firstly, the inner and outer sheaths would be, to the best of our current knowledge, present in roughly equal parts, as hexameric concentric disks in an approximately 1:1 ratio. Yet, there are (normally) three outer sheath loci, and two inner sheath loci. Plus, the fact that at least one of these loci can be deleted with (as far as is known about their functional state) no consequence, suggests the stoichiometry is not that sensitive to disruption. In the case of the "Lumt" operon from *P. asymbiotica* Kingscliff, the domain split observed in one of the remaining outer sheath proteins would mean, if the operon still elaborates functional PVCs, that one outer sheath CDS is sufficient for synthesis.

From comparing the sequential and structural conservation, and biophysical proper-

ties of the tube proteins, the following alternative hypotheses for their maintenance are proposed. For the inner sheath proteins, there are prominent differences in the electrostatic nature of the subunits and subtle differences in structure. This suggests that the tube is either made of alternating stacks of the different monomers, providing a different mechanism for conferring directionality to the tube, or increasingly plausibly, that PVC5 actually fulfils a more specific role. This role may be as an orthologue of gp48/gp54 in T4, as a collar/spike adapter, or possibly gp3 in the tail termination complex at the opposite, distal, tube end (though no crystal structures yet exist to compare to). There are slight contradictions here too however, as in the T6SS, a recent paper has shown that no such adapter/collar proteins are needed (Renault *et al.*, 2018), and the inner tube can interface directly with the VgrG spike. Given that the tube appears more reminiscent of a phage than a T6SS, but the spike is conversely seemingly T6SS-like, only experimental structural resolution will yield the final answer. Preliminary data and EM maps from the ongoing structural resolution collaboration with the Max Planck Institute in Dortmund (not shown) have suggested that the baseplate region might be more substantial than has been seen in either the Afp or R-type pyocin so far, which could indicate that it is more similar to a T4 phage. In which case, the likelihood of a collar/adapter baseplate component protein remaining is somewhat higher. Ge *et al.* (2015) were unable to obtain high enough resolutions in their Pyocin structure to reveal the architecture of its hub and baseplate, so some guesswork remains.

The lower conservation of outer sheath outer aspects suggest that they are open to adaptation to their deployment niches. Phage or T6SS-like exteriors, are prone to significant immunogenicity in a higher eukaryote infection, which runs counter to *Photorhabdus*' extensive efforts to suppress immune responses during infection (a theory proposed to at least partially explain the enormous repertoire of toxins that it carries (Eleftherianos *et al.*, 2010)). An appealing hypothesis therefore, is perhaps these nano-scale delivery systems are adapted for use in these infections and have naturally 'functionalised' surfaces to improve their stability and survivability *in vivo* in the host (Del Tordello *et al.*, 2016; Kaur *et al.*, 2012). The pronounced negative surface charges noted with all of the loci might also speak to a similar phenomenon, as previous publications have identified marked

reduction in trypsin-based proteolysis of negatively charged particles compared with positive ones (Liu and Huang, 1992). Similarly, positively charged molecules have been shown to be preferentially endocytosed at greater rates (Chung *et al.*, 2007), something the PVCs would likely try to avoid so as not to be destroyed. Though admittedly, neither of these studies focussed on proteinaceous entities, so the rules may not be exactly the same. Hurst *et al.* (2004) demonstrated that the outer sheath protein Afp2 is required for formation of a toxic Afp. In several PVC operons a deletion in one of the 3 proteins which match to outer sheaths are observed. Further inspection by sequence similarity indicated that the deletions in these outer sheath proteins is always PVC3 rather than 2 or 4. Unpublished communication from the Hurst lab has proposed that PVC4 may actually be a slightly modified variant of the outer tube. This may be such that it functions as an adapter protein at one end of the tube, possibly as a collar that interfaces with the baseplate (and therefore maybe analogous in role to PVC5, but for the outer sheath instead). It is plausible that such adapter proteins would be needed, as a protein which is capable of transducing the contraction signal 'wavefront' from the baseplate (which in turn received such a signal from the tail fibres), to the start of the outer sheath would be needed; since the prevailing hypothesis in the literature Kube and Wendler (2015), is that contraction resembles an action potential, emanating from the proximal spike-baseplate complex, to the distal terminator at the top of the tube - rather than a concomitant contraction of the whole sheath.

Stepping through the operon, PVC6 and 7, the first putatively 'non-tube' loci remain poorly characterised as no experimental work has been yet conducted which offers functional information, nor are there any published protein structures which offer up any reliable insights based either on sequence identity or structural similarity. PVC7 matches very loosely to certain metalloprotease orthology, perhaps suggesting that it may have an enzymatic role rather than a structural role, but these hits are much too weak to speculate on to any real extent.

Moving on to PVC8, the exact conformation of the gp27-5/VgrG-like spike may have significance to the apparatus' overall role. In the case of the T6SS, the spikes are known to frequently be augmented with other enzymatic activities and other toxins, functioning

like a poison arrow rather than a syringe in some cases. The T6SS is both anti-eukaryotic and anti-prokaryotic, and the T4 spike is purely anti-prokaryotic. The PVC spikes are (as far as has been observed to date), purely anti-eukaryotic. This implies different evolutionary pressures on the protein, which may be manifested in subtle differences in the structures, since the membranes/walls and internal cellular conditions of targets will vary. For instance the pH within the target cell has been shown to be a cue for the spike complex to dissociate (Kumar Sarkar *et al.*, 2006), 'uncapping' the tube, and so the electrostatics of the structure and affinities between the monomers will potentially be subtly sensitive to the environments the complexes are deployed in, which may manifest in their sequences, substituting amino acids of particular electrostatic properties in accordance. As mentioned, the spike complex appears to be one of the areas of the operon where the PVCs are more similar to the T6SS than to the T4 phage, being somewhat simpler for lacking the lysozyme domain, but also being homotrimeric, unlike the T4 which is heterohexameric (gp27 and gp5) (Kumar Sarkar *et al.*, 2006; Arisaka *et al.*, 2003). This may be accounted for by the fact that they need only to negotiate a much simpler cellular barrier (with no cell wall or multiple membrane layers).

A major remaining mystery within the operons is that of PVC12. Weak orthologies to nucleotide binding proteins does not provide much information to speculate upon. Even if the nucleotide binding potential is valid (which is far from certain), its role within the operon is still unknown. Given its syntenic position, the default assumption is that it too contributes to the structure. As typically the largest protein in the operon, a role in the overall structure would likely be fairly pronounced and one might expect it to be easier to identify. Similar proteins have not been seen in other caudate structures to date. Probably the best current hypothesis for its presence, and actually fairly considerable conservation, is as a regulator of the operon in some manner. The GGDEF domain which has been weakly identified within the protein previously is typically implicated in regulatory mechanisms in prokaryotes (Paul *et al.*, 2004; Ryjenkov *et al.*, 2005), but for the PVCs this too remains hugely speculative and weakly supported. A curious observation is that GGDEF domain proteins have also been implicated in the production of exopolysaccharides and other components destined for the extracellular environment,

as have RfaH proteins and *ops* sequences which are explored further in Section 6.2.3 on page 265.

Extensive discussion of PVC13 can be found in Chapter 5 on page 187. It has been proposed that these proteins fulfil the role of the tail fibres within the structure, but little more is known. A subset of the models obtained here appear to offer some useful insight, disambiguating the regions of the structure which could be responsible for host specificity. The true structure will no doubt differ in some respects, and the aforementioned chapter delves in to garnering the first experimental validation of their potential structure and role.

There is likely little useful or reliable information in the tertiary structures obtained for PVC14, however the secondary structure signal appears pronounced, and a first look at the possible hydropathy of these proteins offers some supporting evidence for their role as tape measure proteins, controlling the polymerisation of the PVCs. Consistent with the observations of Rybakova *et al.* (2015), Mahony *et al.* (2016), and Katsura and Hendrix (1984), among others (albeit among the λ phages instead), the high degree of α-helicity is considered a hallmark. Similarly, according to Mahony *et al.* (2016), conserved N- and C-termini are also required (which makes logical sense, since there are likely conserved binding domains at each end of the chain, somewhat reminiscent of a bolas), and this can be seen in the multiple sequence alignment in Appendix Chapter B on page 335. Lastly, given the relative similarity of Afps and PVCs, the preservation of these genes in terms of synteny offers further evidence that PVC14 is indeed likely to be a tape measure protein.

While reasonable quality models were obtained for PVC15, the putative ATPase, its as-yet-unknown role means that it is difficult to ask meaningful questions of the structures. The ATP-binding active site of the enzyme appears to be intact, so, as is the case with the Afp, its likely that the enzymes are still functional within the PVC operons, and not simply a vestige. Since AAA+ ATPases are often known to have a secondary proteolytic domain which appears to be less conserved in the case of the PVCs, it is possible that the ATPase domain has been maintained, but is now fused to a domain of new function. Given the closeness in sequence of the PVCs and Afps however, Occam's razor would suggest that the protein is likely functioning in a similar way, with a cryptic role in the

(dis)assembly choreography, as evidenced by the aberrant forms obtained by Hurst *et al.* (2004) when it was deleted.

Lastly, PVC16 remains enigmatic. The work of Rybakova *et al.* (2015) remains the only clue as to its function to date, having an effect on tail length when deleted, implicating it as a cap or terminator of sorts. Unfortunately, at present, both HMM searching and homology modelling fail to offer any further useful insight.

To briefly make note of the limitations of this approach, obviously the simulations presented in this chapter are of assorted qualities, and care must be taken when concluding anything from theoretical data alone. There is however, a sufficient body of related proteins for many PVC loci, such that comparative modelling can provide, and has provided, hypotheses and potential answers for a number of questions which are not answerable without resolving the structures either experimentally or otherwise. Homology modelling is a standard tool that is often reached for in many publications when the alternative is to have to laboriously resolve the structure (if it is even possible), but as George Box's famous aphorism goes: "all models are wrong, but some are useful."

Some of the primary limitations of this approach are discussed below. I-TASSER is still a semi-template-based method, though it breaks down structures in to essentially meaningful chunks, rather than relying on a single overall structure too much. This is conceptually similar to finding 'structural k-mers' - amino acid 'words' which 'spell out' a particular helix structure or span of β-sheet etc. Such methods may not be ideal for 2 opposing reasons which relate to the 'morals' of this chapter. Firstly, sequence matches which are *too good* will result in over-fitting of sorts. This will then provide a model of high confidence, which mirrors an existing structure, but that does not guarantee than *in vivo* in the host organism, that that protein actually folds in the same way. As the example in the introduction (the protein G from *Streptococcus* (Alexander *et al.*, 2007)) makes plain, 2 proteins with extremely similar sequences can have entirely different folds in practice. This is even before additional biological considerations are factored in, such as chaperones, whereby proteins of very similar sequence, but some with chaperons and some without, will likely lead to substantially different structures. Consequently a threading/templating tool is essentially coercing the sequence to be modelled in to the

conformation of existing proteins to too great a degree, and so some subtleties of the real structure may be lost. I-TASSER (and other tools) do ameliorate this somewhat by performing refinements via molecular dynamics simulations. This enables strained regions of the modelled structures to relax in to more natural conformations, however, by this point they will still be somewhat constrained by the bulk of the structure which has been templated, meaning the effects are largely dominant in the turns/strands/disordered regions which are freely mobile, so the degree to which they approximate the real structure may vary.

These molecular dynamics processes are somewhat of a double-edged sword however. In cases where few/no reliable templating structures can be detected for a query sequence, the bulk of the structure will be modelled dynamically. Most of the time this leads to a 'collapsed' globular structure resembling a tangled ball of thread, with no meaningful structure whatsoever. Consequently, there is still a need for visual inspection of obtained structures, whether they have favourable metrics or not, as well as a little scientific intuition on the part of the researcher as to whether a structure looks plausible for a proposed role - essentially echoing the epigraph of this chapter once again.

Finally, a word on the limitations of the HMM based methods used. HMMs are extremely sensitive tools for querying sequences, but are not magic, and still have a lower bound of utility. This results in many hits of assorted quality which are likely not meaningful, but match as they share some spans of secondary structure or sequence, even if only very short. In this chapter, this was not a significant problem, as the E-values of these matches can often be readily identifiable as spurious, and to some degree all new homologies detected were taken in to consideration, since they improve on 'hypothetical proteins'. Instead, the main stumbling block that was encountered during this work was incorrectly named hits in databases (e.g. in the case of the match for PVC9 against a "Lysozyme", and PVC1 being matched to the obviously erroneous gp6).

3.3.1 Summary and Future Work

This chapter has tried to focus on just one or two particularly interesting aspects of the different loci in the PVCs, largely in the interests of brevity. To perform every analysis - conservation mapping, electrostatic mapping, density occupation, structure superimpo-

sition, and so on - for every single protein/locus, would fill volumes in figures alone (not to mention may not provide any particularly useful information in some cases). However, all of the data and models produced are now available as a 'database' that can be referred back to in future. Naturally this means that there is much future work which could be done comparing and contrasting these structures, however, as the intention of this chapter was to propose hypotheses and ideas which could be tested in the lab in future, below are suggestions of future *experimental* work which can build on ideas from this dataset.

Of course, the most obvious task is to experimentally resolve the structure of at least one PVC. This is in progress at the Max Planck Institute, but is still a slow process. While finally resolving a PVC structure will somewhat render the efforts of this chapter redundant, the diversity of PVCs means that homology modelling remains the most practical and feasible approach for comparing all of their structures - and the differences between the various PVC operons is likely the most interesting aspect of their biology. With the advent, hopefully in the not too distant future, of the first atomistically resolved PVC, a return to this homology modelling process might be worthwhile. More accurate models could be obtained, having a better template to start from, with the benefit of experimental proof, but also with the advantage of being able to rapidly compare the structural differences between operons for biological significance.

This chapter proposed that the role of PVC5 should potentially be amended from the currently accepted general tube protein; instead potentially forming a part of a collar or terminator complex. Some existing 2D electrophoresis and mass spectrometry data suggests that this may be true, since the ratios of PVC1:PVC5 suggested much less PVC5 was detected in PVC samples. It would be worthwhile repeating the experiment and proceeding directly to quantitative peptide enumeration through mass spectrometry to confirm this finding. Similarly, getting a good grasp of the proteins which are present in the structure of a mature PVC might shed some light on whether all of the 16 loci which are currently proposed to be structural actually contribute to the structure itself, or whether some of them are responsible for assembly instead. Similarly, PVC2-4, for which there is still no compelling explanation for their paralogy, may be incorporated in to the

final assembly in different proportions, and could offer information about their potential role.

PVC6 and 7 are entirely enigmatic proteins at this stage. The aforementioned mass spectrometry would be informative to discovering whether they too are structural or not. In the case of PVC7, very weak enzymatic orthologies were detected, so it may be the case that it has a synthetic role instead of a structural one. These proteins are also prime candidates for mutagenesis, particularly because they have been assumed to be structural for so long and thus there is no existing experimental data concerning them, unlike PVC14-16 from the Hurst group for example.

PVC9 is another good candidate for mutagenesis. If the theory proposed here is correct, its deletion should impact the polymerisation of the PVC tube - presumably preventing it from occurring, though assembly of the central spike hub and inner sheath may still occur, since the existing literature suggests that these are spontaneous and electrostatically driven processes.

For PVC10, the putative spike tip, sequence based approaches are unlikely to unambiguously identify them in all cases as they are simply too diverse. In the absence of structural resolution, studies in to protein-protein interactions between it and PVC8, the major spike protein would provide valuable information if interactions can be detected.

The large CDS in position 12 with weak homology to GGDEF-domain bearing proteins is another prime candidate for deletion from the operon. It is not anticipated that this homology is particularly meaningful, and it may yet transpire that the protein does form part of the structure, though such a large protein raises questions about what its role might be. A fairly straightforward mass spectrometry, imaging, and deletion experiment might shed light on its role structurally. If it truly is a regulator of the operon, a transcriptomic study of deletion mutants might be warranted.

Experimental study of the putative tail fibre proteins is well under way, and discussed in Chapter 5 on page 187.

It is unlikely that PVC14 will be structurally resolvable as part of the total complex, and experimental structural resolution through crystallographic methods will also probably yield a tertiary structure of questionable utility. The protein has been seen in past 2D gel

electrophoresis experiments of particle preparations, confirming that it is present in the structures, and at surprisingly high levels. The best approach to confirm its role within the PVCs is likely simply to recapitulate similar existing studies, with artificial truncations and extensions of the proteins, to see if the PVC itself concomitantly shrinks or extends (Rybakova *et al.*, 2015). A simple experiment to confirm its secondary structure profile at least, might be to heterologously express it and perform circular dichroism studies, though this is still indirect proof, and there is no guarantee of the stability of the protein under heterologous expression.

For PVC15 and 16, experimental structural resolution is probably feasible, through crystallography for example. Understanding the non-ATP-binding domain structure for the ATPase will likely be key to proposing a function and role. It might be possible to perform protein-protein interaction studies following similar experiments like those done for the T6SS (Douzi *et al.*, 2016), to determine if the ATPase interacts with sheath proteins, indicating (dis)assembly. If PVC16 does form a terminator/cap for the helically 6-fold symmetric tube, its quite likely that it too will form hexameric structures, and therefore its crystal conformation and fold structure could be extremely determinative. The distal end of many caudate structures are comparatively poorly characterised, so this is likely a worthwhile endeavour. An experiment in the lab is planned to attempt to GFP tag the protein and examine its localisation. A similar approach worked for the baseplates of the MAC complex, though the effect of fusing a protein to a structural component may well lead to unintended effects on the PVC structure itself.

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, 2014) for a good review), R-type pyocins/tailocins (Ge *et al.*, 2015; 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, 2015) (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 five to six 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 Experimental Procedures

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.

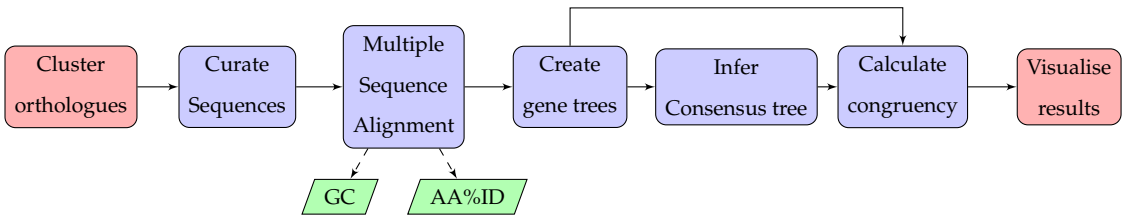


Figure 4.1 | A flowchart to demonstrate, at high level, the steps involved in the process of the congruency analysis presented in this chapter.

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. Section B.1 on page 335 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 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, 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.

4.2.2 Curation of Sequences

There is legacy sequencing data published in NCBI for the three 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 2.5.4 on page 88, 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 B.1 on page 336.

As a further note on the existing confusing nomenclature; two 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 four 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 4.4 to 4.19 on pages 159–166 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 the supplementary Appendix Chapter B on page 335 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.

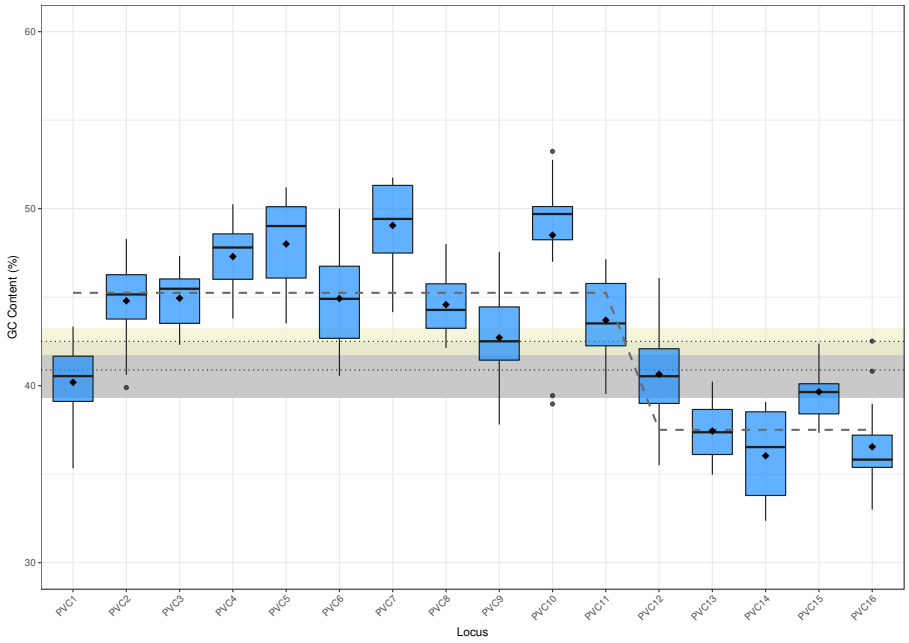


Figure 4.2 | The GC content (%) distribution of PVC loci.

There is a trend toward significantly lower GC content at the 3′ end of the operon. ♦ denotes the mean, • denotes extreme outliers outside 1.5× the interquartile range for the sample, and the black line is the median. 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.

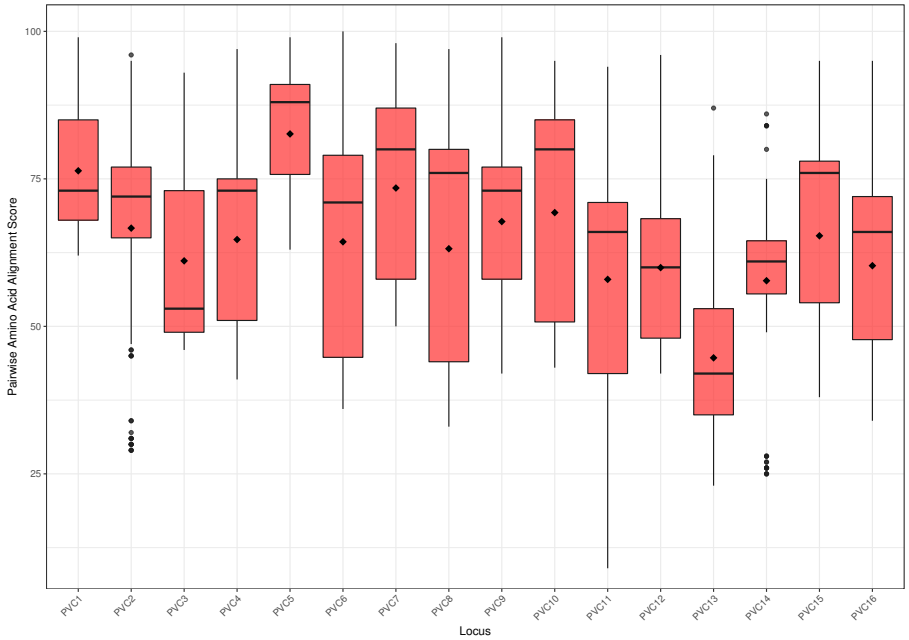


Figure 4.3 | Pairwise amino acid similarity of all-vs-all genes for each PVC locus

Pairwise alignment similarity scores of a multiple sequence alignment of all the sequences within a given syntenic position. This demonstrates the distribution of similarity within a locus, and highlights that there are significantly different PVC 'alleles' due to their position as outliers. ♦ denotes the mean, • denotes extreme outliers outside 1.5× the interquartile range for the sample, and the black line denotes the median.

4.2.4 Gene Trees

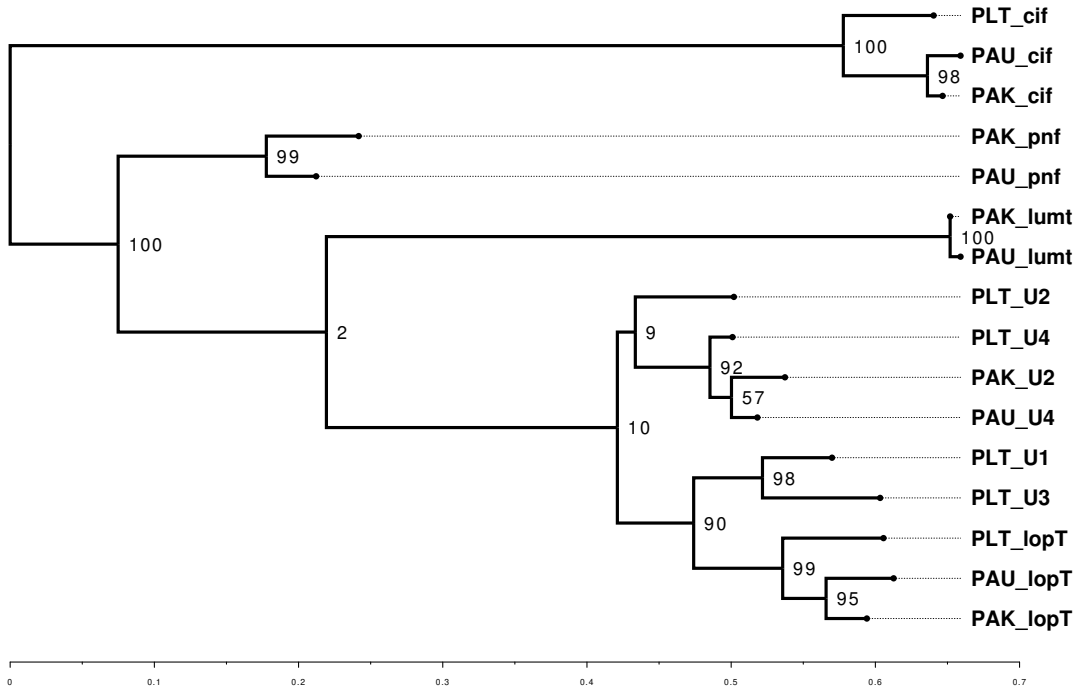


Figure 4.4 | Maximum-likelihood tree of the locus position (PVC1) from each operon.

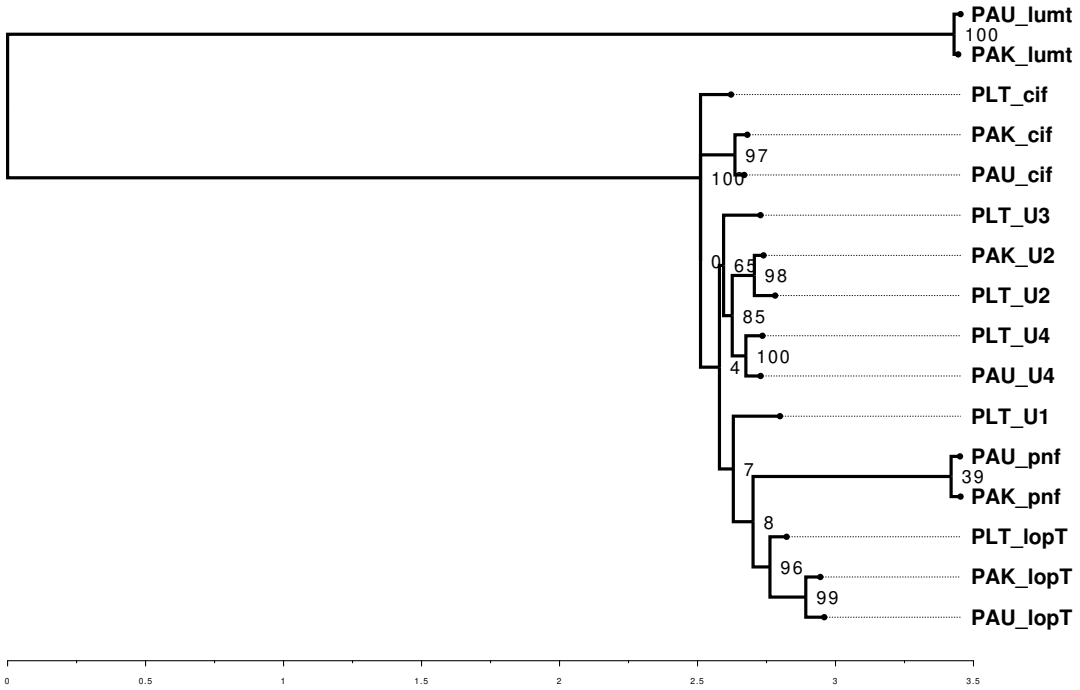


Figure 4.5 | Maximum-likelihood tree of the locus position (PVC2) from each operon.

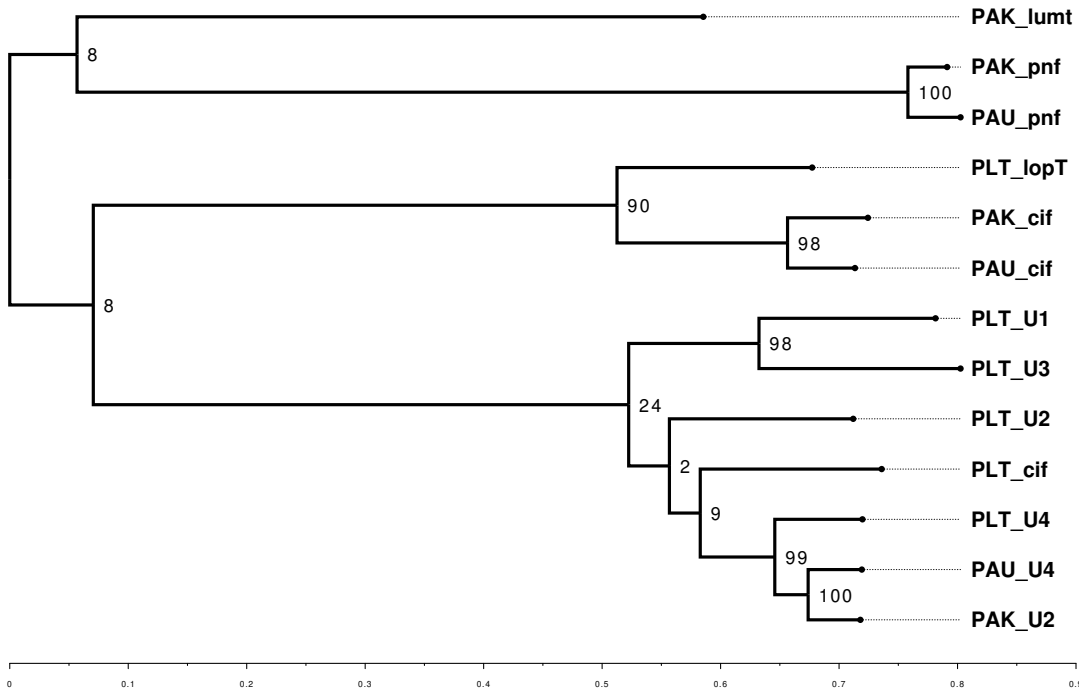


Figure 4.6 | Maximum-likelihood tree of the locus position (PVC3) from each operon.

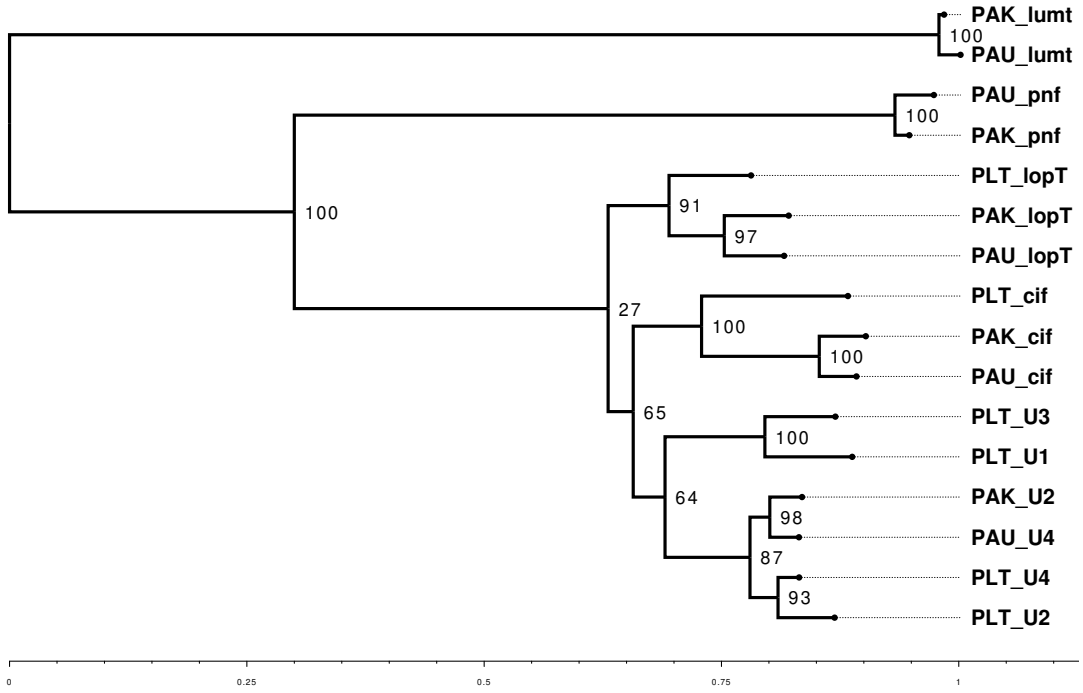


Figure 4.7 | Maximum-likelihood tree of the locus position (PVC4) from each operon.

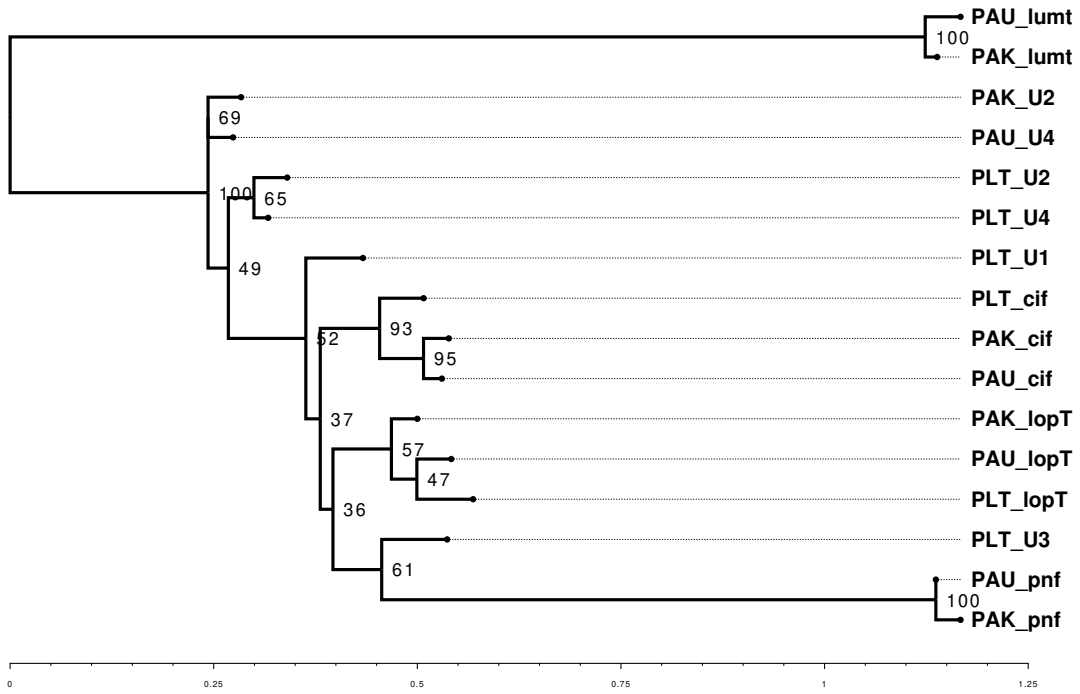


Figure 4.8 | Maximum-likelihood tree of the locus position (PVC5) from each operon.

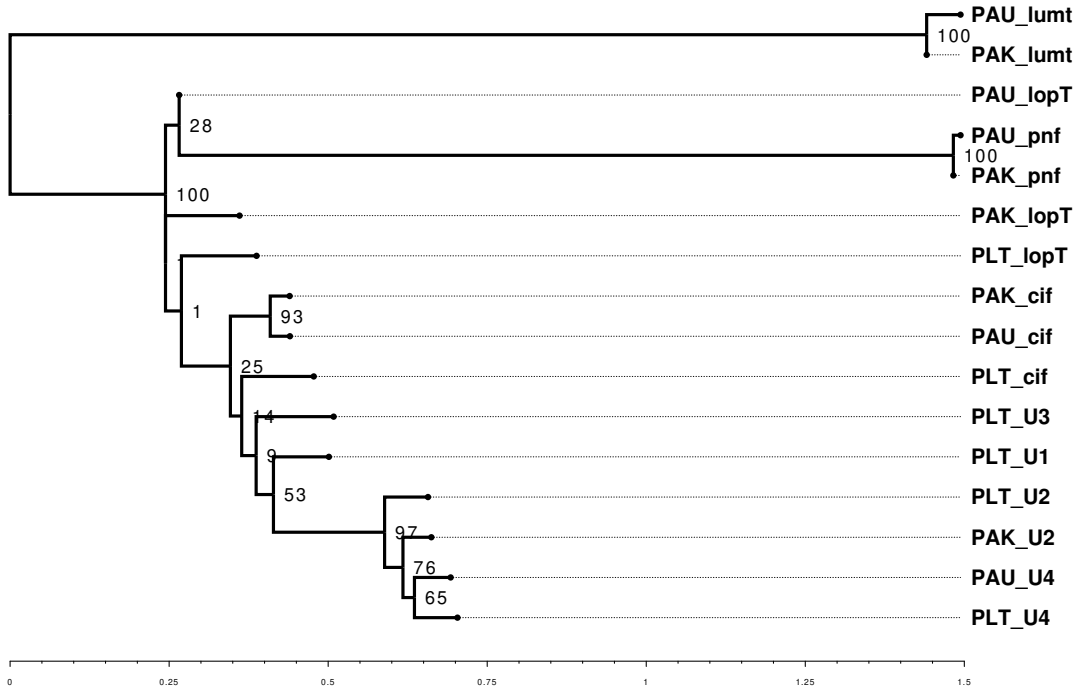


Figure 4.9 | Maximum-likelihood tree of the locus position (PVC6) from each operon.

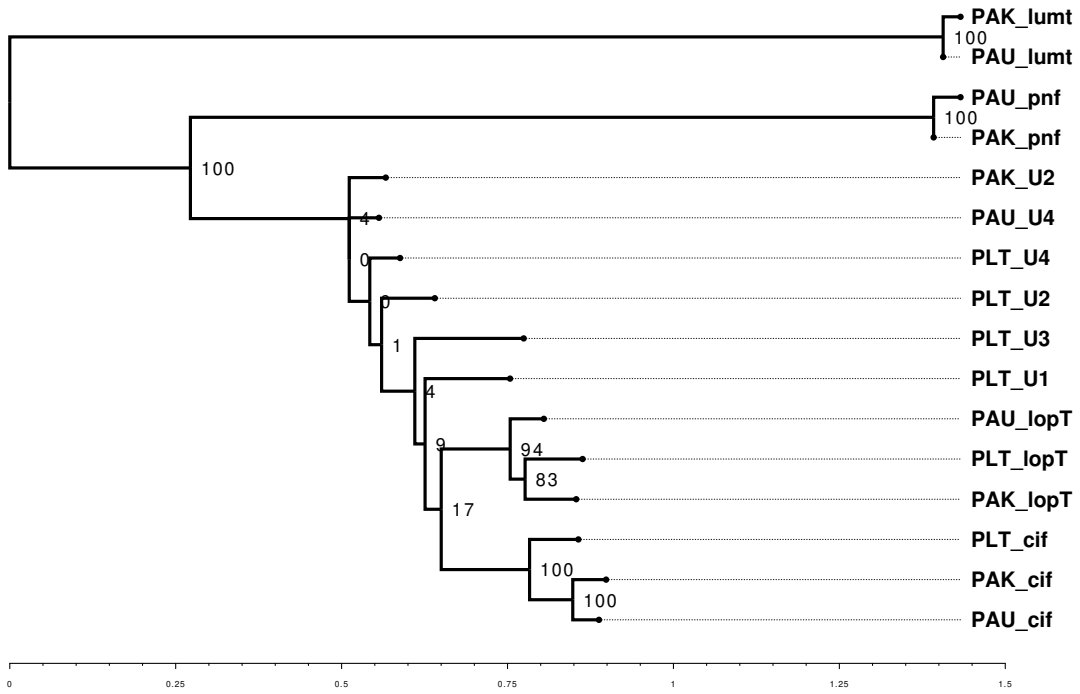


Figure 4.10 | Maximum-likelihood tree of the locus position (PVC7) from each operon.

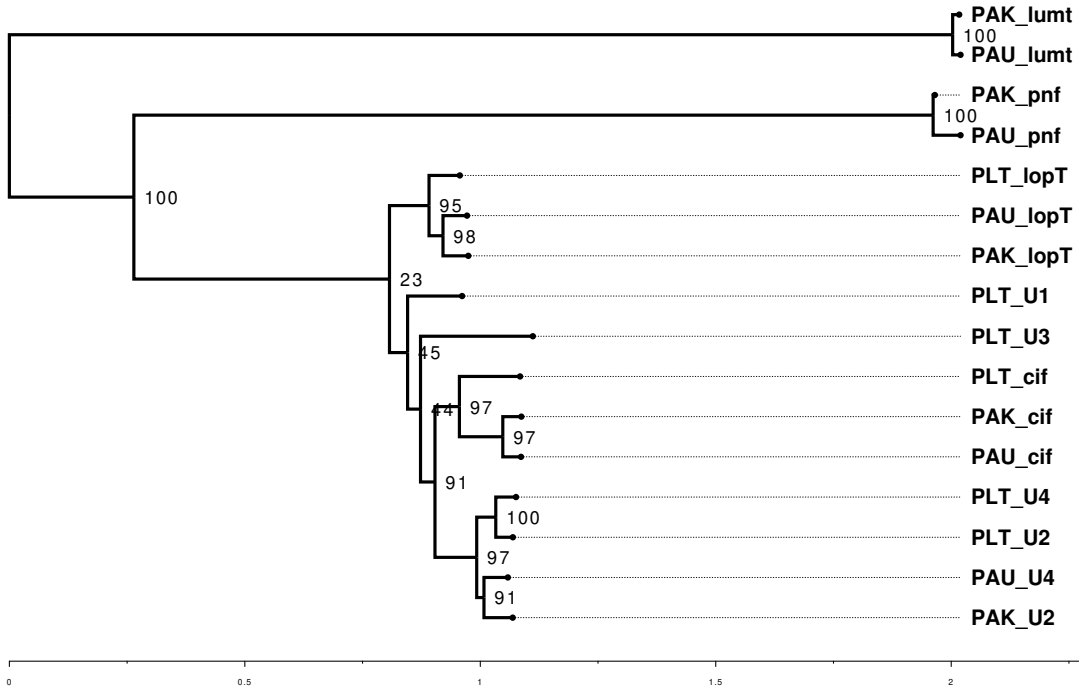


Figure 4.11 | Maximum-likelihood tree of the locus position (PVC8) from each operon.

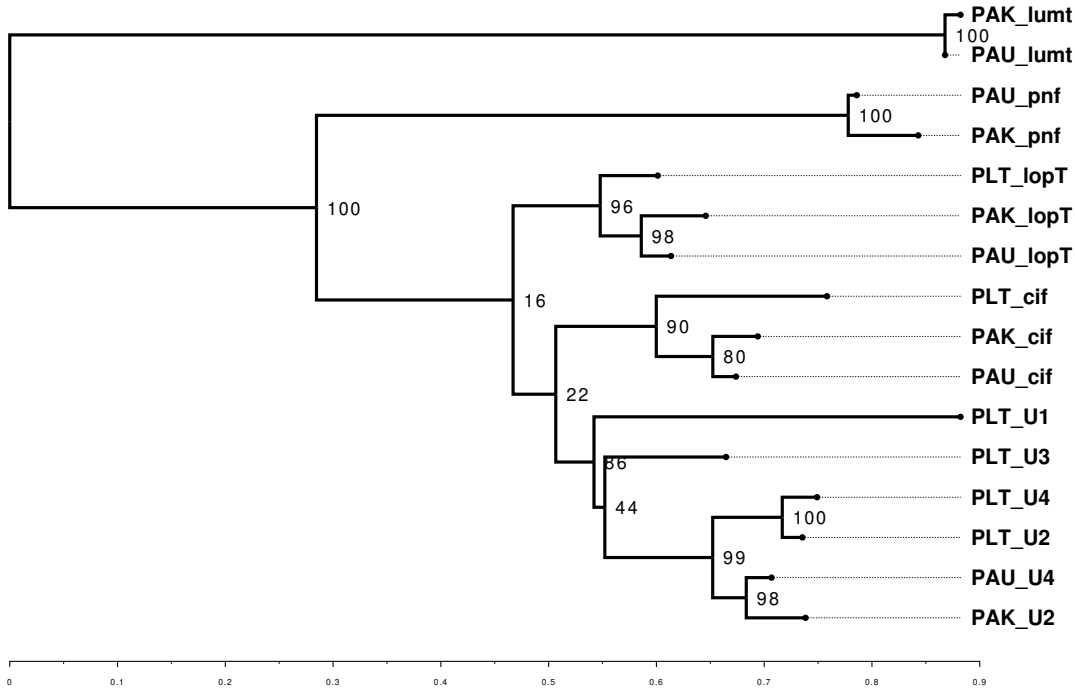


Figure 4.12 | Maximum-likelihood tree of the locus position (PVC9) from each operon.

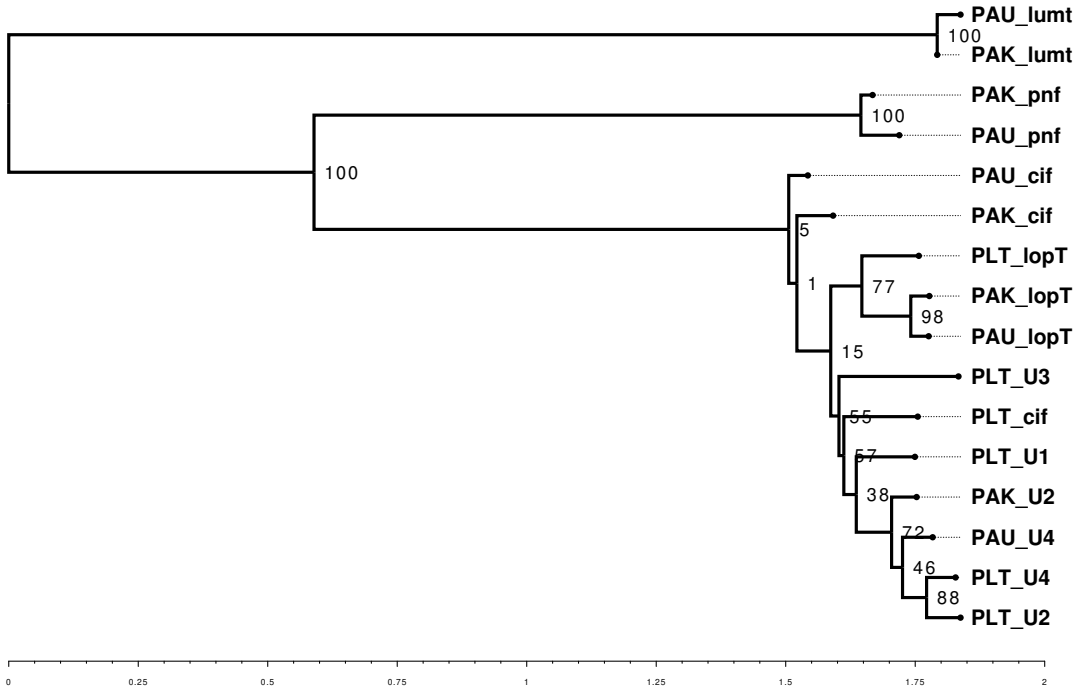


Figure 4.13 | Maximum-likelihood tree of the locus position (PVC10) from each operon.

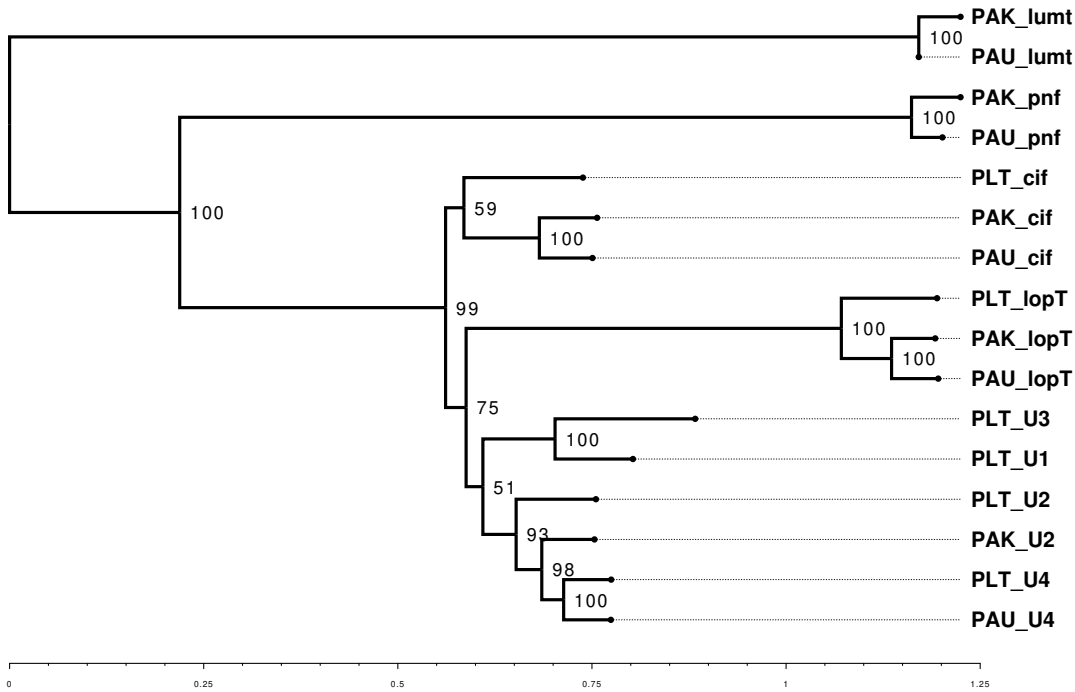


Figure 4.14 | Maximum-likelihood tree of the locus position (PVC11) from each operon.

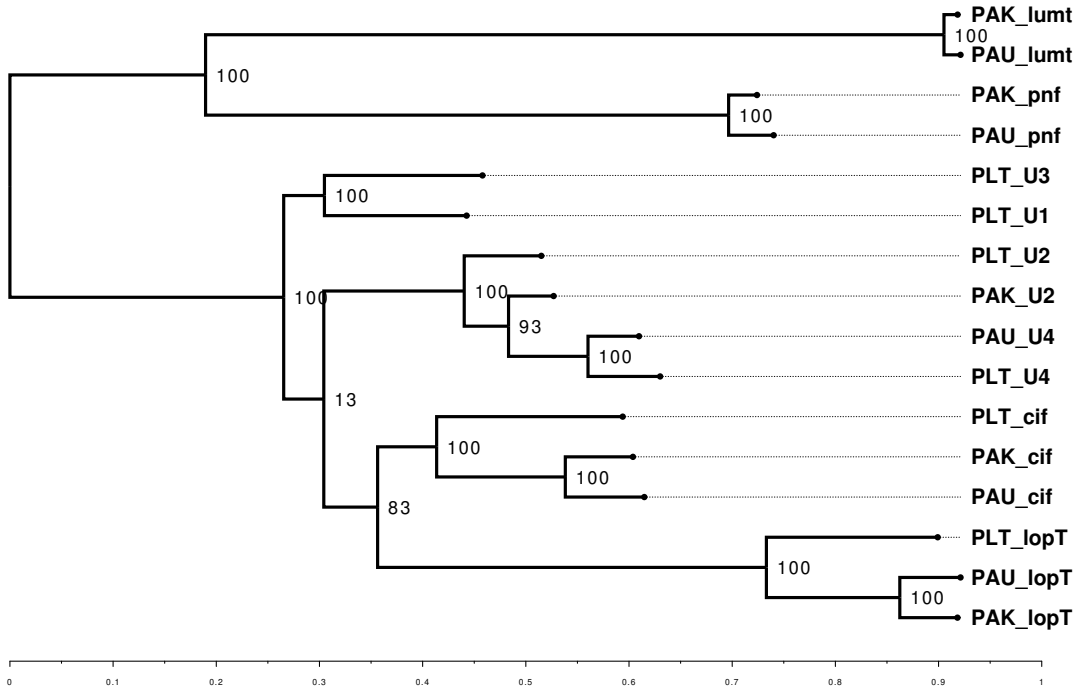


Figure 4.15 | Maximum-likelihood tree of the locus position (PVC12) from each operon.

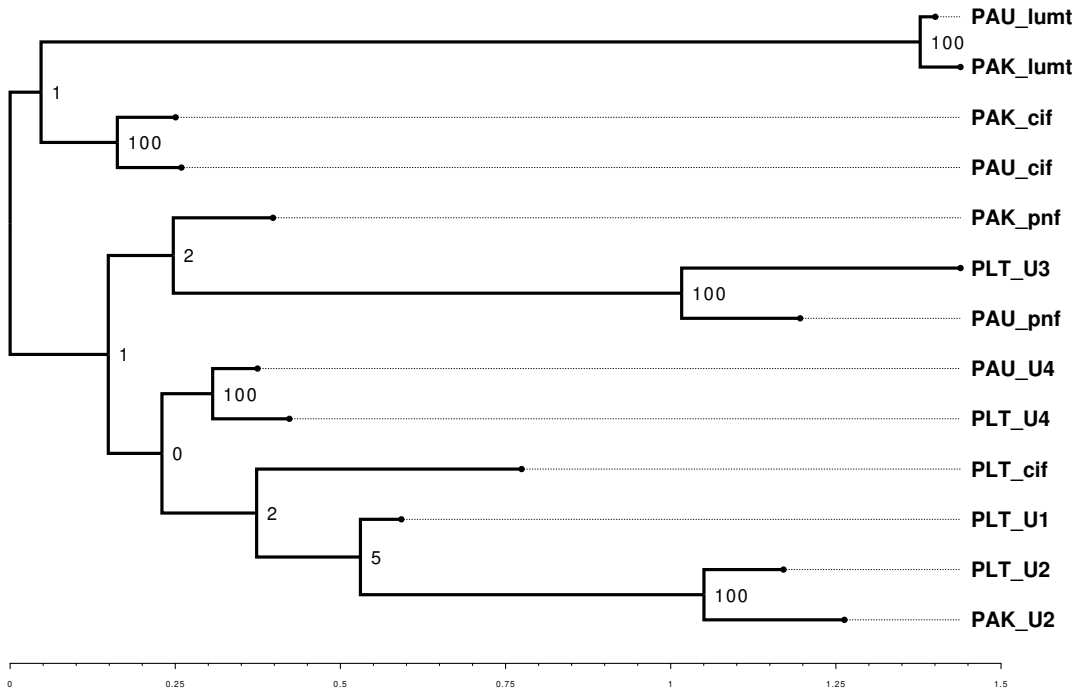


Figure 4.16 | Maximum-likelihood tree of the locus position (PVC13) from each operon.

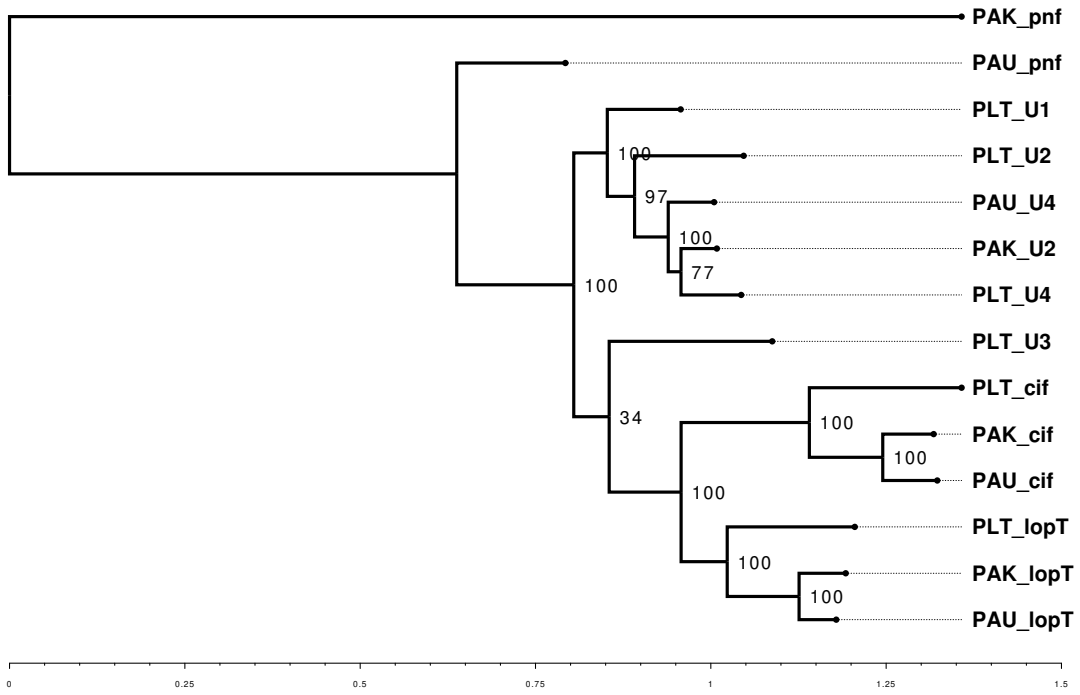


Figure 4.17 | Maximum-likelihood tree of the locus position (PVC14) from each operon.

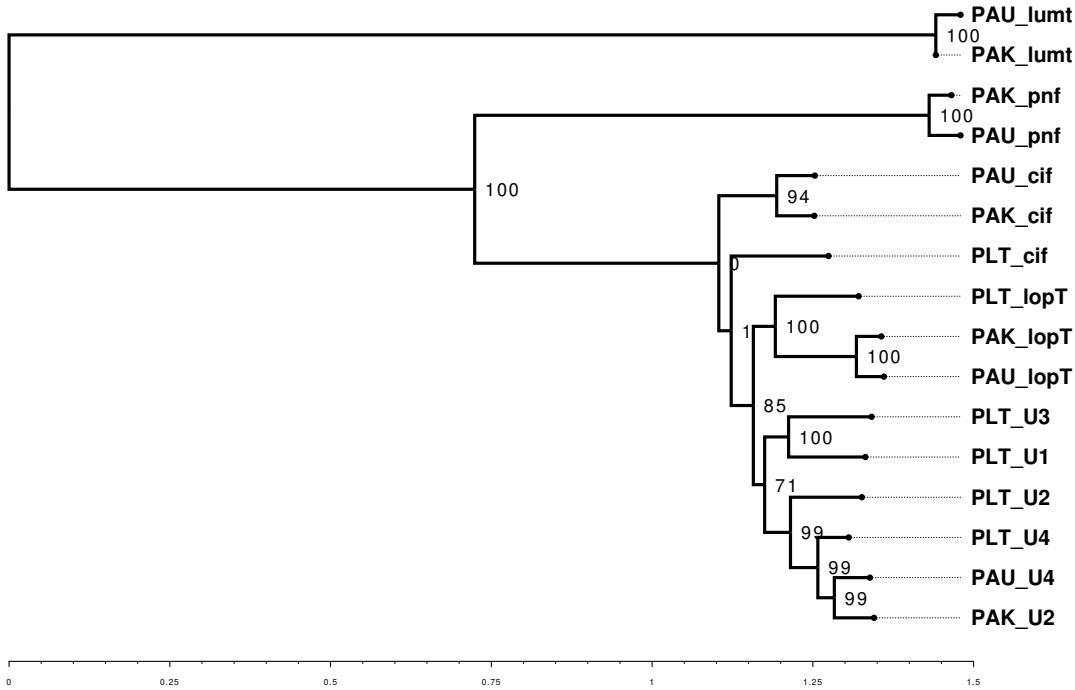


Figure 4.18 | Maximum-likelihood tree of the locus position (PVC15) from each operon.

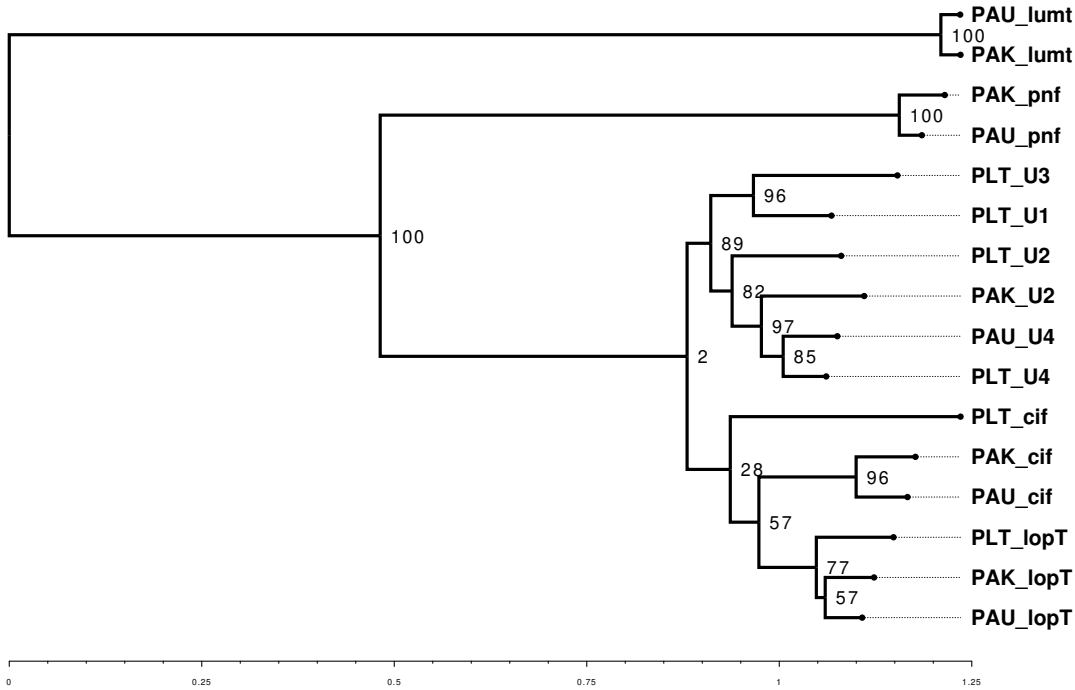


Figure 4.19 | Maximum-likelihood tree of the locus position (PVC16) from each operon.

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.

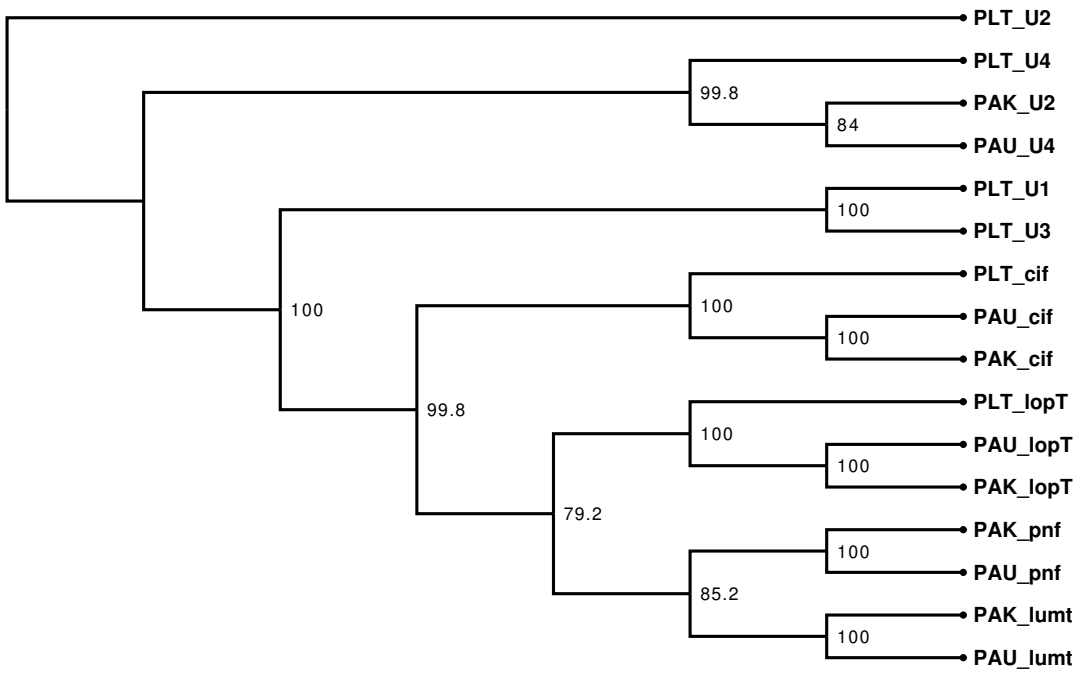


Figure 4.20 | Inferred consensus tree from ASTRAL-II.
The inferred consensus tree from the genetrees PVC1-16. Branch lengths are arbitrary for this tree and therefore it has been depicted as an cladogram with no scale.

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

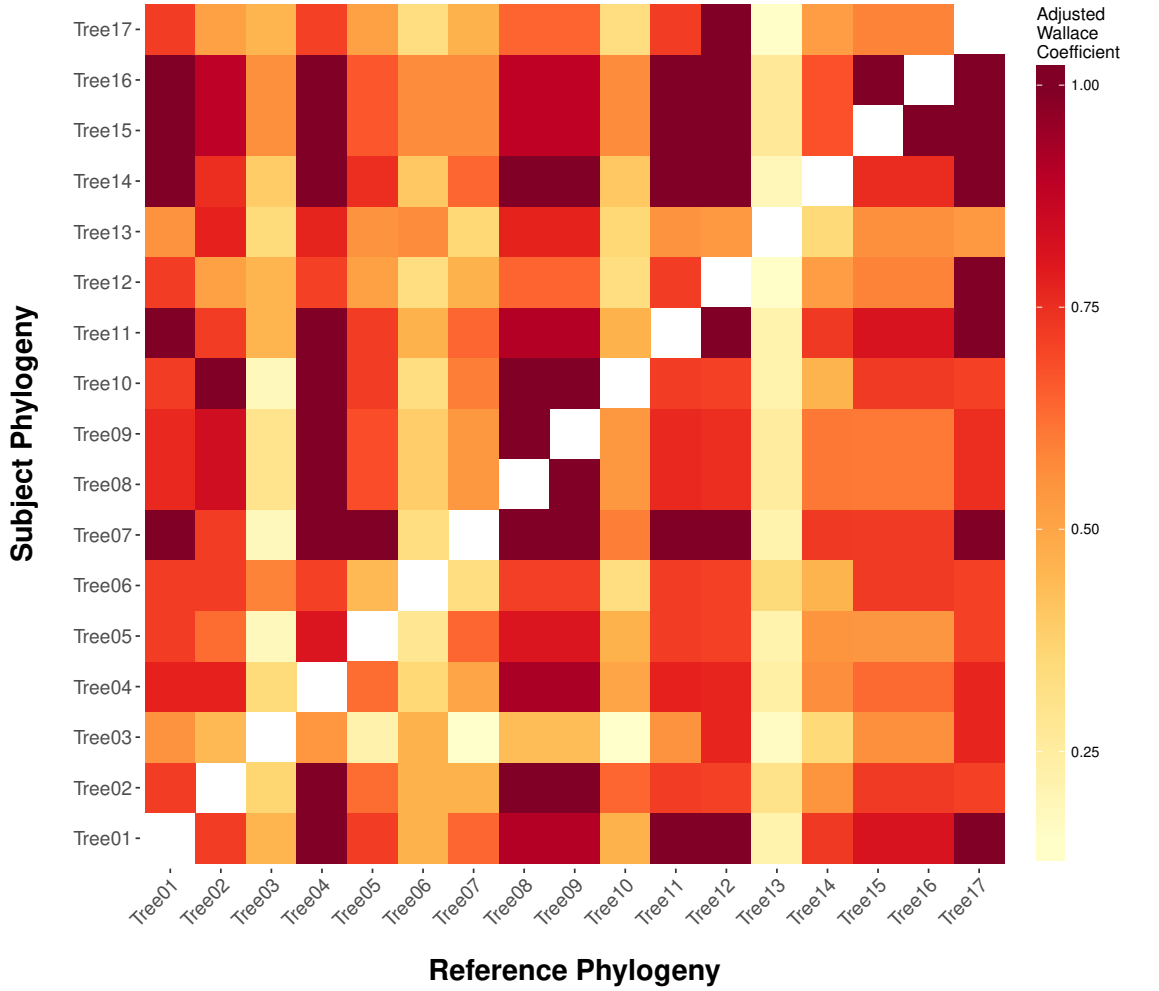
With each gene tree output from RAxML and the 17th tree as the inferred tree from ASTRAL-II, the pairwise congruency between all trees was calculated. Evaluation of congruency metrics ultimately means one is able to put a single number on to a subject tree with respect to a reference, to say how similar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

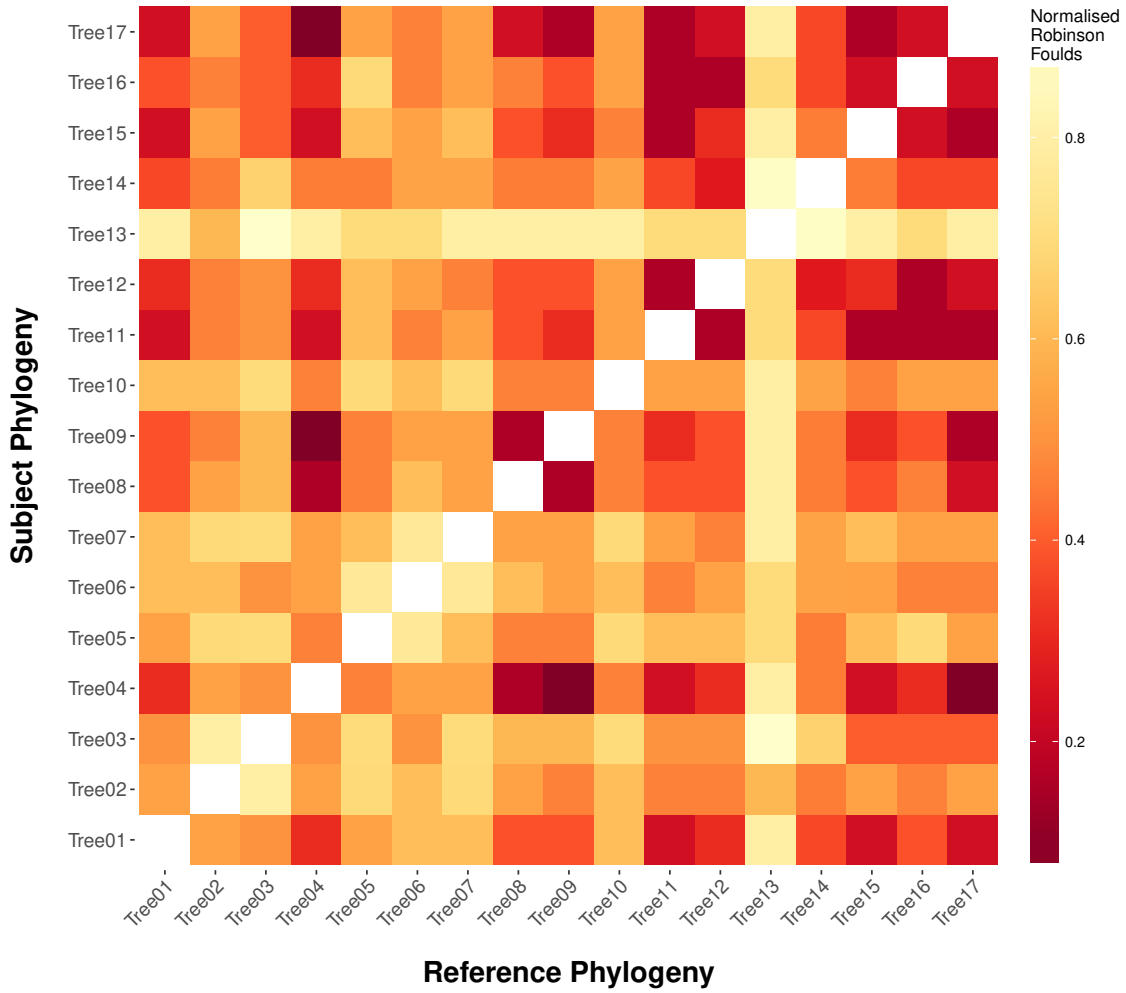
4.2.6.2 Normalised Robinson-Foulds

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 | Visualised congruency between trees (Adjusted Wallace Coefficient).
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 | Visualised congruency between trees (Robinson-Foulds).
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.

4.3 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 Appendix Chapter B on page 335, 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 two genes in the operon (with the exception of "PVC0" that was discussed in Section 4.2.1 on page 154). As can be seen from Figure 4.3 on page 158, 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) Chapter B on page 335, 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, potentially altering the target organisms' immune response to the complex, while maintaining the contractile functionality.

PVC3 scores generally lower (more apparent in Figure 4.21 on page 169, than Figure 4.22 on page 170), and this too is unsurprising as this gene is missing from three 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. 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 may be 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 (six 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 paralogues 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 two genes would follow similar evolutionary patterns, though the average pairwise amino acid similarity scores shown in Figure 4.3 on page 158, shows PVC4 to be better conserved (fewer extremes) overall than PVC2, despite having similar mean and median values. Hurst *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 paralogue 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 158). 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.

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 its 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 $\approx$ 250 residues generally aligning better than the beginning of the sequence with a much higher number of 100% identical residues at any given position (see Appendix Chapter B on page 335). 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 its partners to form a triangular prism-like shape, as shown in Chapter 3 on page 95. 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.

PVC9 shows a similar pattern of congruence as PVC8, with similarities to tail lysozyme domains as shown in Chapter 3 on page 95 (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. Based on identification of a better structural match to tube initiator proteins which have to interface with the spike complex, their proximity and potentially intertwined roles mean that a similar pattern of congruence is 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 155, the operons actually harbour a second PVC11 paralogue, though only one 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 ≈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 158 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 Chapter 3 on page 95. In this workflow PVC13s are clearly the least congruent genes within the operon with Figure 4.16 on page 165 not clustering the PVCs well, with low confidence nodes and long branch lengths. These proteins have low overall identity (Figure 4.3 on page 158), and could therefore be responding to very specific, and potentially very strong selection pressures. The current hypothesis is that the co-opting/convergent evolution of eukaryotic viral molecular patterns has allowed *Photorhabdus* to repurpose these structures as a toxin system for use during infection of higher organisms, or for manipulating the nematode partner. 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 187).

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 is 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 two 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.*, 2015). 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 genes to identify for some time, and is 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 are 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. One exception to this might be as a means of recycling the subunits of PVCs that could

be produced prematurely/aberrantly. By doing so, the cells would reduce the deficit to the 'cellular economy' of building so many large and energetically demanding structures. Similarities between PVC15s and Afp15 have been demonstrated previously but there is currently no known role for the analogous Afp15 from *Serratia* either, other than it is known to be required for assembly/activity (Hurst *et al.*, 2004, 2018). 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), however evidence remains scant.

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 154, 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 4.4 on page 159, 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 three 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 its 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*, 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 161) 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 158 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 three 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 two, and this suggests that the operons which have a gene deletion, are actually lacking PVC3, retaining the two 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 two 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 Figure 4.5 on page 159, 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 160 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). As explored in Chapter 3 on page 95, this appears to be a gene split, where PVC3 is actually fully deleted and PVC2 has undergone a mutation to introduce a new start codon. PVC3, therefore, is seemingly 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 Appendix Chapter B on page 335), 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, and preliminary evidence in the lab has implicated at least *P. luminescens* TT01 Unit 4 in manipulating the nematode via induction of *endotokia matricida*. In the *P. luminescens* TT01 genome, all the "Unit#" operons are located tandem to one another, likely integrated from the same mobile element. This would essentially have coupled the operon's evolutionary histories, making it small wonder that they group well, though this does not offer much to explain the occurrence of orthologues of just a couple of these sequences within *P. asymbiotica*.

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 Pnf operons (for instance) are more alike across the genera, than any two PVCs within a genome are like

one another.

4.3.2 Identifying the PVC 'Blueprint' Elsewhere

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.
  - One or more copies of an outer sheath protein, with an additional variant paralogue (PVC4) which will likely match to the same structural homologues, 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 one or more unknown loci at PVC5 and 6, immediately followed by two 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/baseplate 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. Identifiable orthology to eukaryotic viral motifs/domains may well be a useful characteristic for their identification however.
  - PVC14 is not a well characterised gne, 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 as obvious. 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 effectors 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 158. 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, and is certainly also the case for the Afp (Hurst *et al.*, 2004). In which case, the GC skew is not a unique feature of PVCs, but may be somewhat characteristic of caudate structures, or at least protein translocating ones. An intriguing hypothesis to explain this may be that the GC content at 3' ends of long operons such as these can have comparatively low %GC content, such that fewer hydrogen bonds hold the strand together, promoting strand separation and therefore potentially easier transcription, in a manner similar to that which promotes strand separation for replicon origins (Artsimovitch and Landick, 2002).

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.

4.3.3 Summary and Future Work

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 Chapter 3 on page 95, 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.

In future, it would be good to extend this workflow to a greater number of PVC operons from more genomes, after identifying them based on some or all of the criteria defined here. In particular, it would be better to reimplement this whole workflow in an automated manner. Two particular weaknesses of the approach used here are the subjective process of identifying PVC orthologues (particularly since the operon contains paralogues/deletions/rearrangements/unique genes etc.), and the subjective clustering of trees when creating the input for the Adjusted Wallace calculations. There are methods for clustering trees objectively, though not always accurately, so test cases would have to be explored where accuracy versus subjectivity is assessed. Clustering the operon orthologues is a little more tricky, as it requires not only an assessment of protein sequence similarity or orthology, but also needs to encapsulate synteny, such that, for example, PVC 3 can be demarcated from PVC 4. This is important, as this chapter has shown that, despite PVCs 2, 3 and 4 all demonstrating orthology to phage/pyocin outer sheath proteins, only PVC3 is ever deleted, and this points to a role for PVC4 which the sequence annotations have not yet sufficiently untangled (such as a baseplate adaptor protein).

Part III

Experimental Results

Chapter 5

Structure and Function of PVC Tail Fibre-like Genes

"Consider the possibility that we too can make a thing very small which does what we want - that we can manufacture an object that maneuvers at that level!"

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 Chapter 1 on page 2 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. 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, the cell surface target is the OmpA protein or lipopolysaccharides of *E. coli* (Granell *et al.*, 2014; Taylor *et al.*, 2016; 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.*, 2016), and that the fibres are responsible for the initial conveyance of a conformational change to the peripheral baseplate proteins. Taylor *et al.* (2016) 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 190 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.*, 2014). 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 152, 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

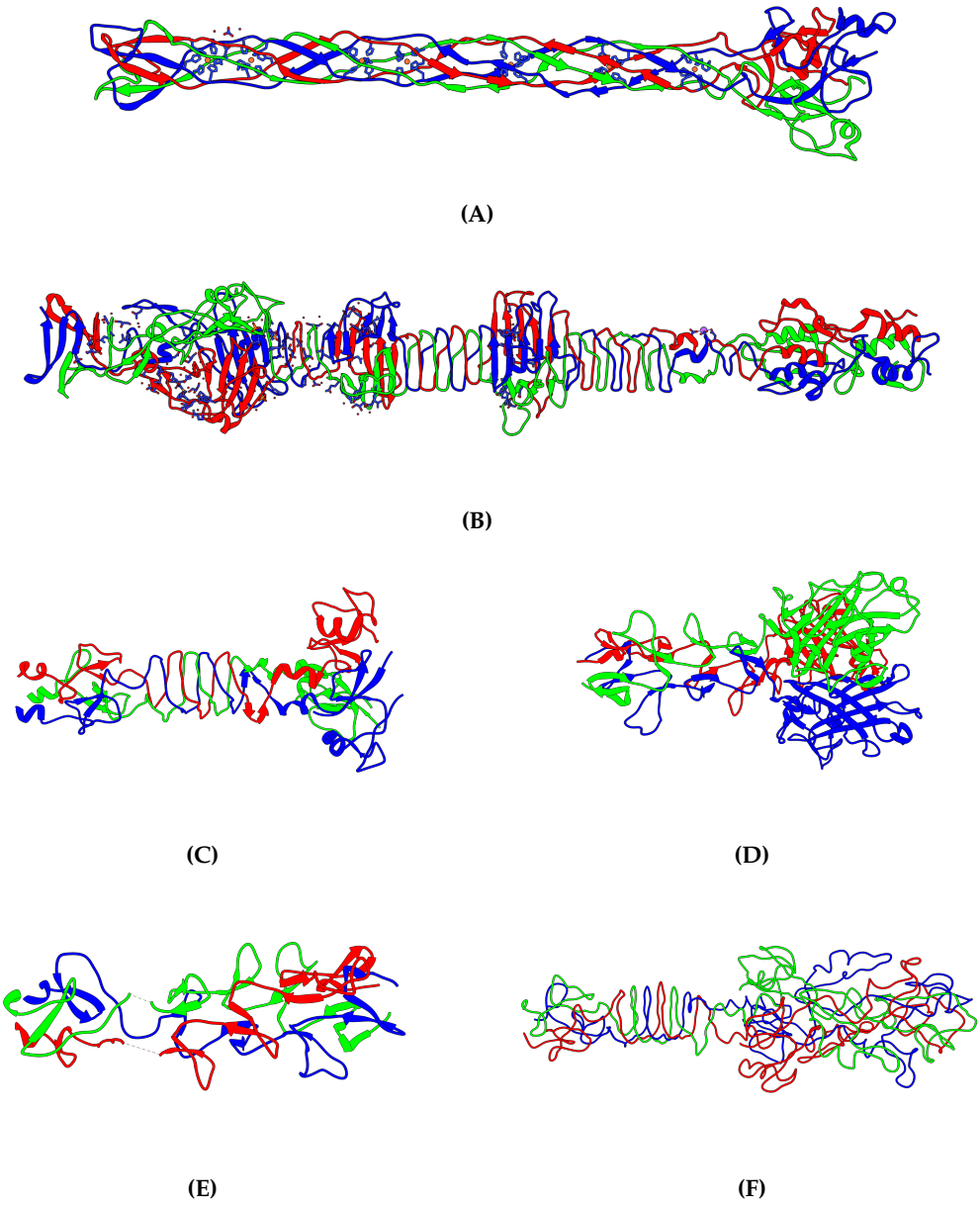


Figure 5.1 | Crystal structures for resolved tail fibre-like proteins.

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.* (2014). **(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).

overlap, and are averaged out, as in the case in the Ge *et al.* (2015) paper for instance. In Heymann *et al.* (2013) (see Figure 1.11 on page 32), there appear to be tail fibre like structures laid back along the length of the Afp 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 Chapter 3 on page 95 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 190), 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 effector delivery systems, suggests that their tail fibre proteins may be of particular interest. Understanding the existing target binding spectrum may elucidate the role of the PVCs in virulence and pathogenesis during *Photorhabdus*' infection cycle in the various hosts it exploits. 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 rather than prokaryotic.

In summary, this chapter examines tail fibres in isolation from the difficult-to-manipulate PVCs; in order to scrutinse 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* Profiling 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.

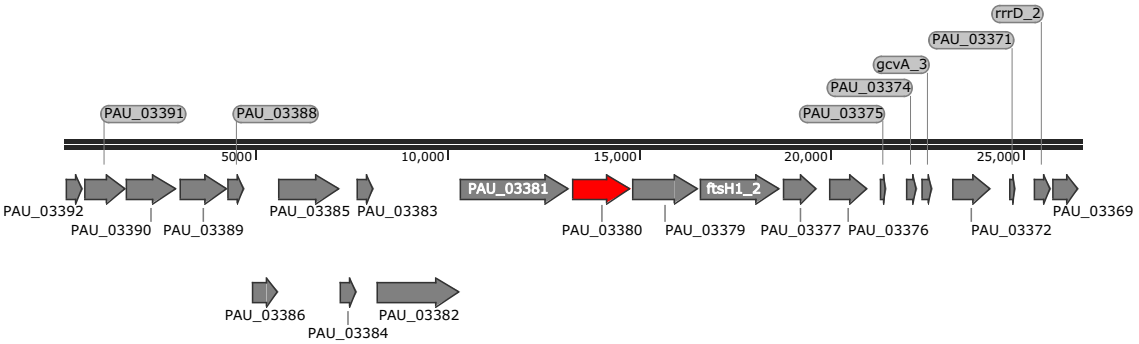


Figure 5.2 | The "pnf" operon from *P. asymbiotica* ATCC43949.
The PVC*pnf* operon, with the putative tail fibre gene (PVCpnf13) that was cloned in red. Locus tags correspond to the most recent genome annotation used throughout this study.

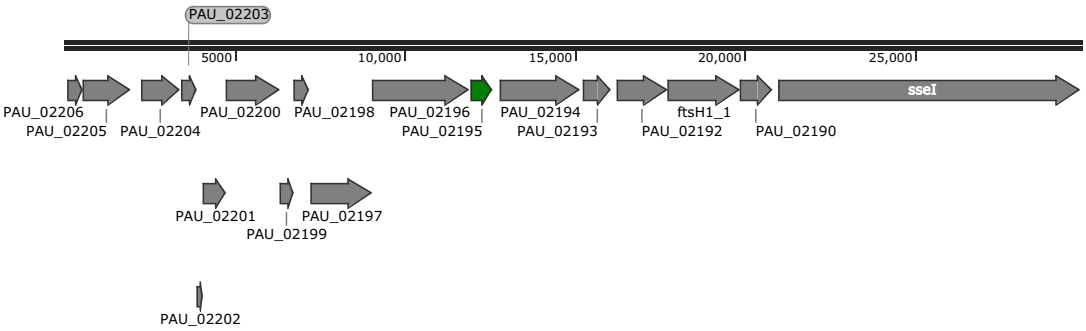


Figure 5.3 | The "lumt" operon from *P. asymbiotica* ATCC43949.
The PVC*lumt* operon, with the putative tail fibre gene (PVClumt13) that was cloned in green. Locus tags correspond to the most recent genome annotation used throughout this study.

These two 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 167, it can be seen from Figure 5.2 and Figure 5.3 that the two

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 two 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 two proteins (and tail fibres from PVC operons in general) is their apparent chimerism in domain structure as demonstrated in the HHPred results from Section 3.2.1.1 on page 100. Table 5.1 on the following page summarises the HHPred hits for these two tail fibres. They match distinct PDB entries in different regions, typically from either T4 bacteriophages or human Adenovirus fibre proteins.

This is suggestive of *Photorhabdus* perhaps 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 two 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 two 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}

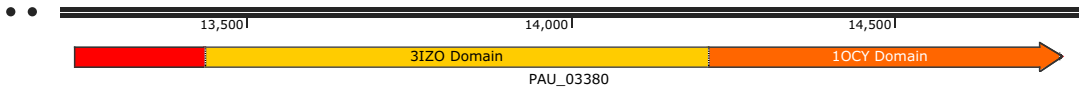


Figure 5.4 | The PVCpnf13 tail fibre with dominant domain homologies depicted
A map of the dominant protein domain splits within the PVCpnf13 tail fibre protein according to HHPred. Toward the 3′ end there is a region which matches well to the PDB ID 3IZO, a fibre protein-penton base from the human Adenovirus 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).

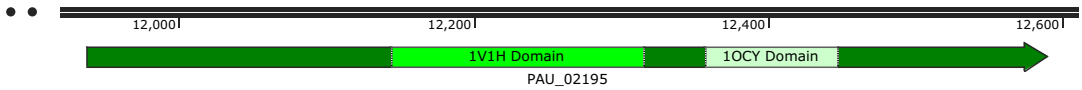


Figure 5.5 | The PVClumt13 tail fibre with dominant domain homologies depicted
A map of the dominant protein domain splits within the PVClumt13 tail fibre protein according to HHPred. Toward the 3′ end there is a region which matches to the PDB ID 1V1H, an artificial fibre fusion protein from the human 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 Thomassen *et al.* (2003)) 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 195, 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 190. 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 seven 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 two proteins being studied here, the

results are displayed in Tables 5.2 to 5.3 on the next page. Repeats are identifiable in all but one of the tail fibre proteins, and typically runs of four to six 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 two amino acids (see Table 5.2 on the following 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 two 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[1]. The significance that this might have is not yet understood however. Interestingly, in the recent paper by Hurst *et al.* (2018), the tail fibres of the newly identified AfpX demonstrated a tail fibre repeat architecture very similar to that of the PVCpnf13, with an even greater number of tandem 15-residue repeats, that were also quite rich in valine, serine and glycine. Since the tail fibres which are seen in the EM density map in Figure 1.11A on page 32 appear to be very long (substantially longer than some very preliminary EM densities obtained for a PVC via collaboration with the Max Planck institute at Dortmund (data not shown)), it makes sense that the sequences for the PVC tail fibres contain fewer repeats and thus potentially shorter shaft regions.

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 199. 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.

[1]Data for other tail fibres not shown

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

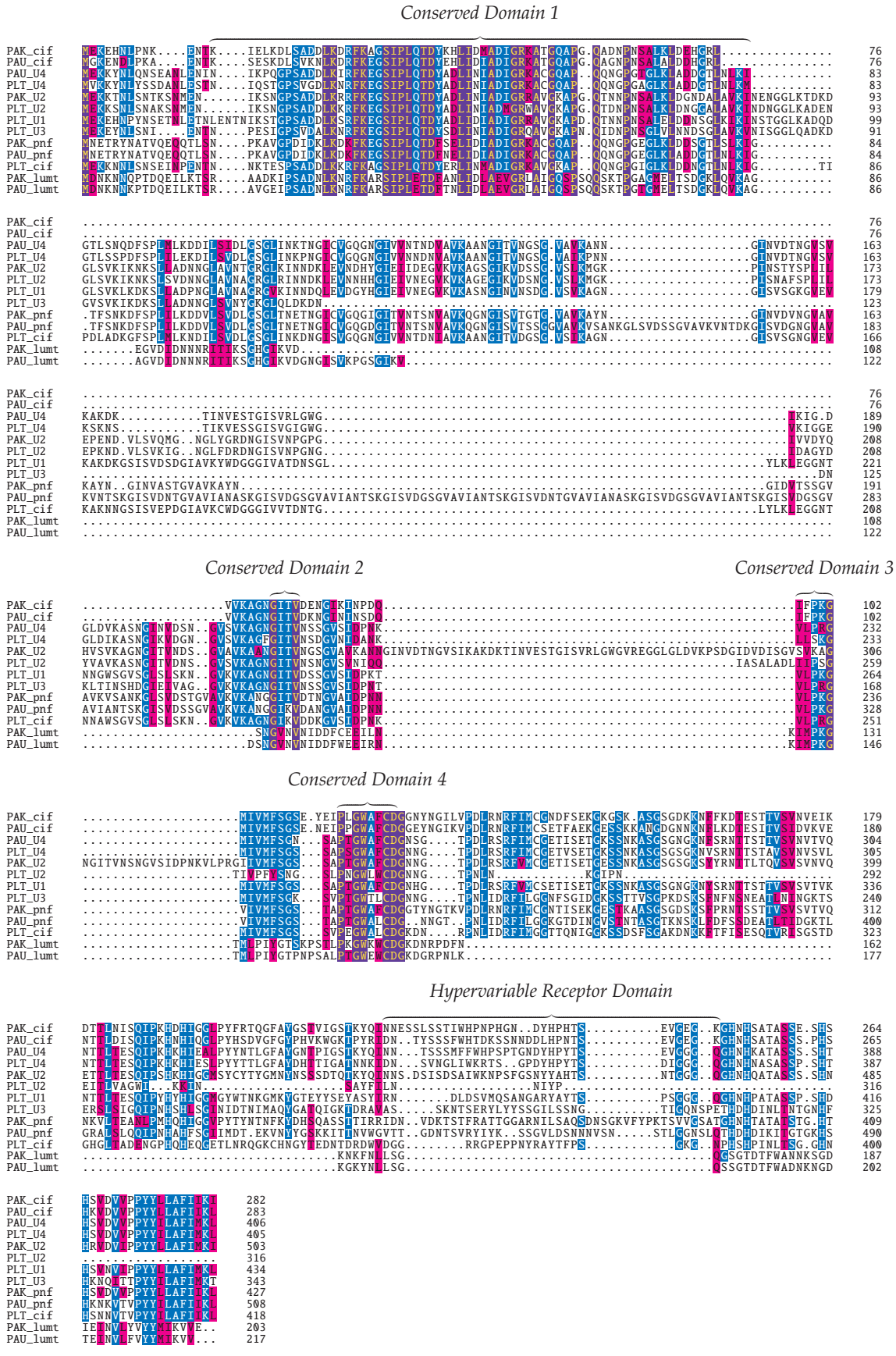
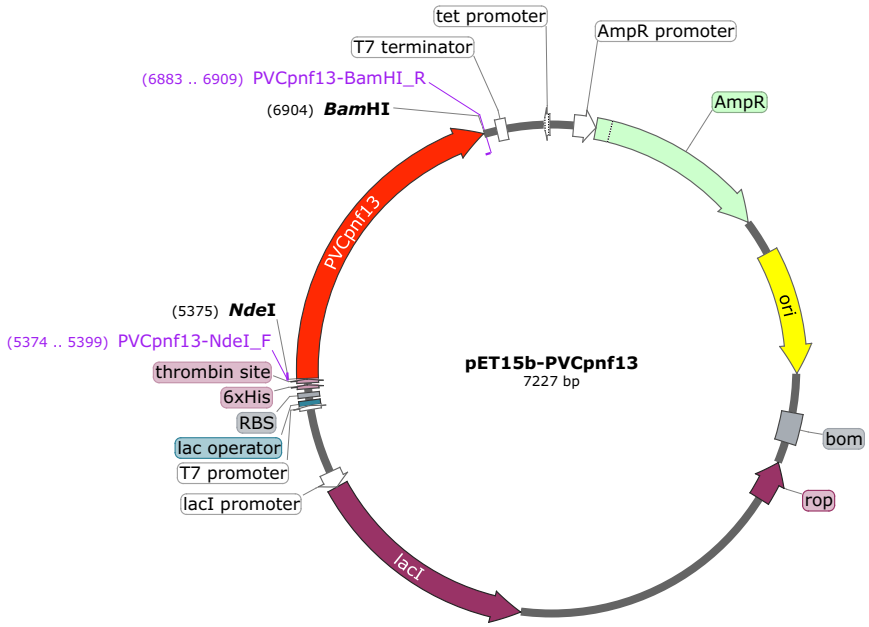


Figure 5.6 | Annotated multiple sequence alignment for putative PVC tail fibres

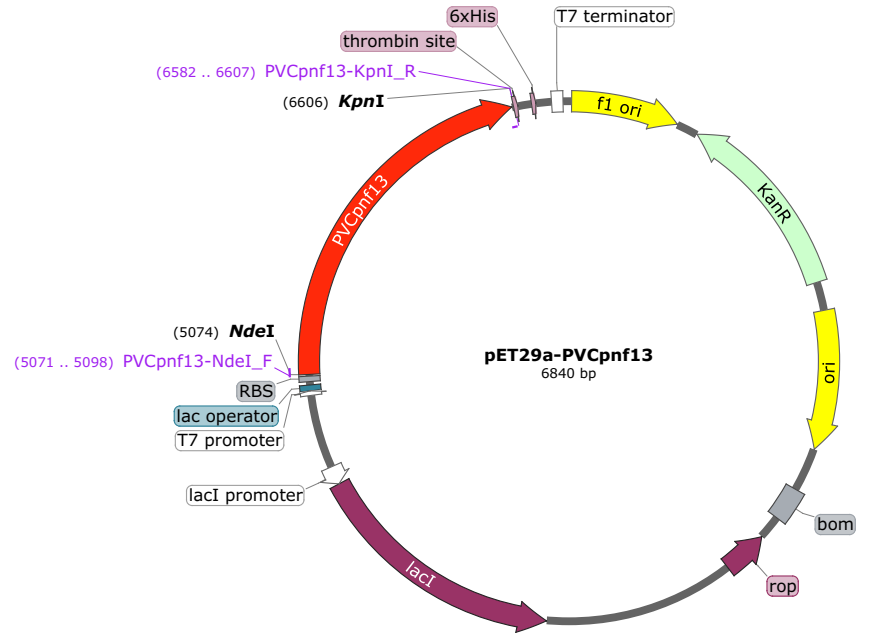
Alignment reproduced from Appendix Chapter B on page 335. Residues are colour coded by residue similarity, and identifiable domains of interest are annotated.

5.2.1.3 "*in silico* Cloning"

Even with the hypervariable sequence identified and some promising 3D models, it was not possible to tell categorically where the region of the tail fibres responsible for binding to the host would be in the final folded protein. 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 two 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 67 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 Figures 5.7 to 5.8 on pages 201–202.



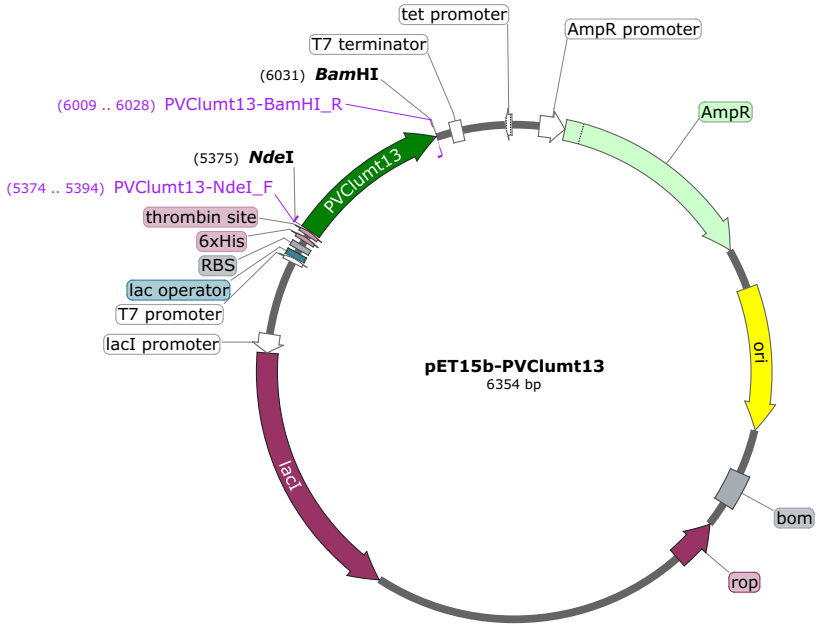
(A)



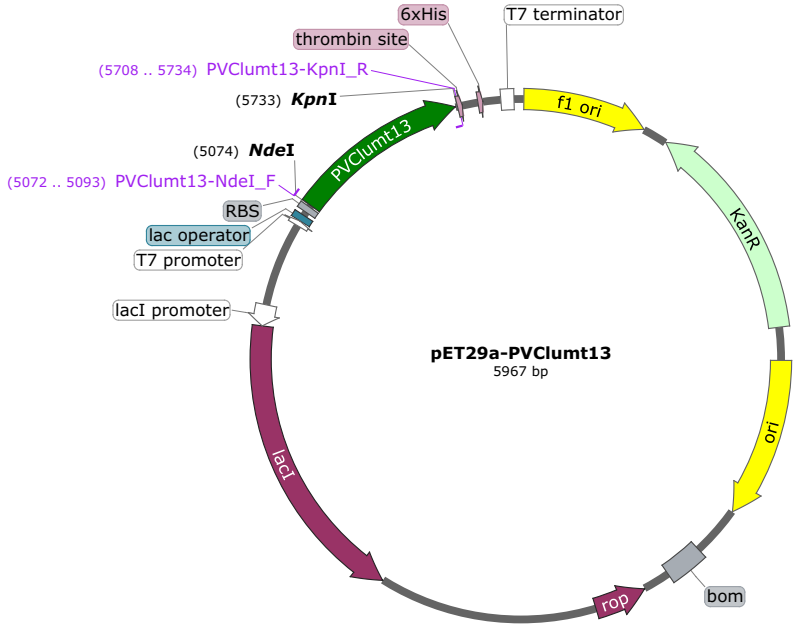
(B)

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 69, and the PCR conditions outlined in Table 2.10 on page 72, 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 $OD_{600\ nm}$ of 0.4-0.6. Figures 5.9 to 5.10 on the next 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 two litres of bacterial culture. Purification was performed using an Äkta FPLC system and a

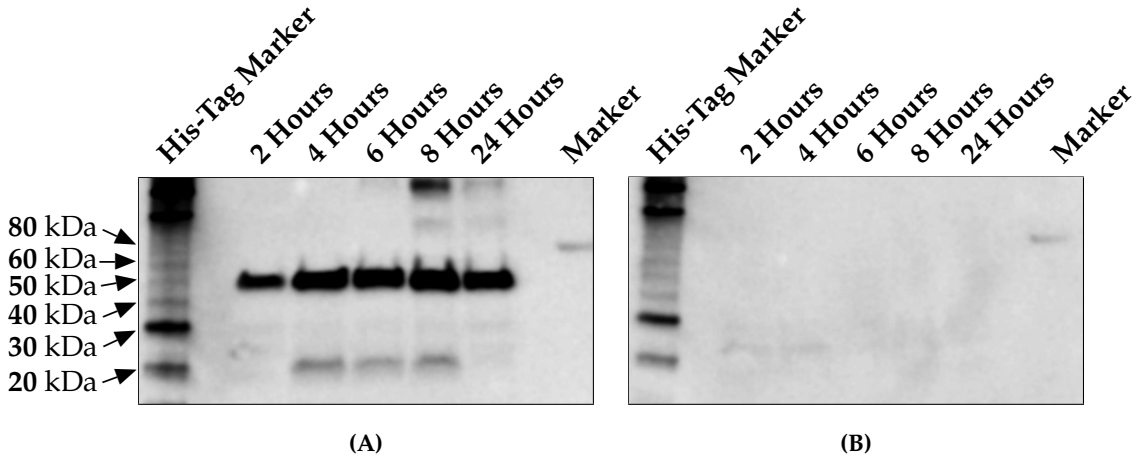


Figure 5.9 | A Western blot of pnf13 expression in BL23(DE3) cells after induction. **(A)** Western blot of inductions of pnf13 from the pET15b vector, with an N-terminal Hexahistidine tag. **(B)** Western blot of inductions of pnf13 from the pET29a vector, with a C-terminal Hexahistidine tag. Good yields 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.

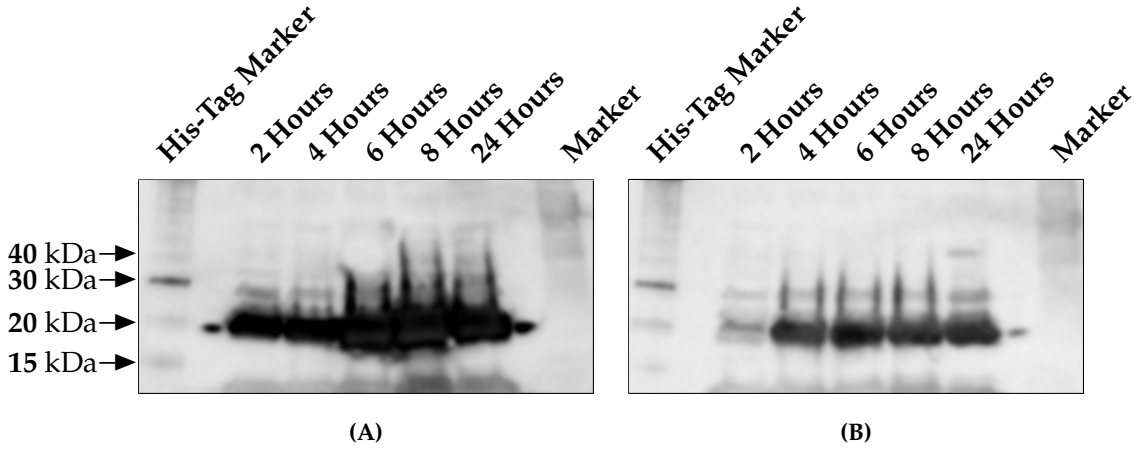


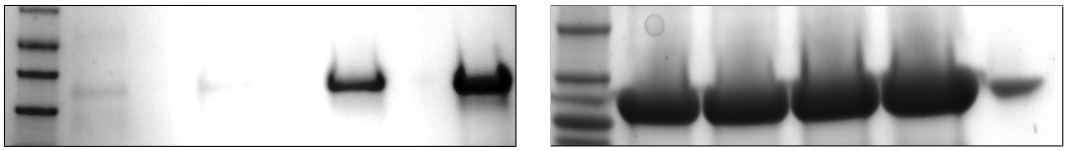
Figure 5.10 | A Western blot of lumt13 expression in BL23(DE3) cells after induction. **(A)** Western blot of inductions of lumt13 from the pET15b vector, with an N-terminal Hexahistidine tag. **(B)** Western blot of inductions of lumt13 from the pET29a vector, with a C-terminal Hexahistidine tag. Good 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, three hexahistidine tags would be present per final protein structure. A potential explanation for the unusual chromatograms could, therefore be, that stochastically, some proteins manage to bind one, two, or three histidine tags, resulting in differences in binding strength. It might be expected, in this case, that three 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 themselves may be quite 'sticky' and are therefore interacting with the column or other binding partners as yet unknown.

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 following 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)



(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.

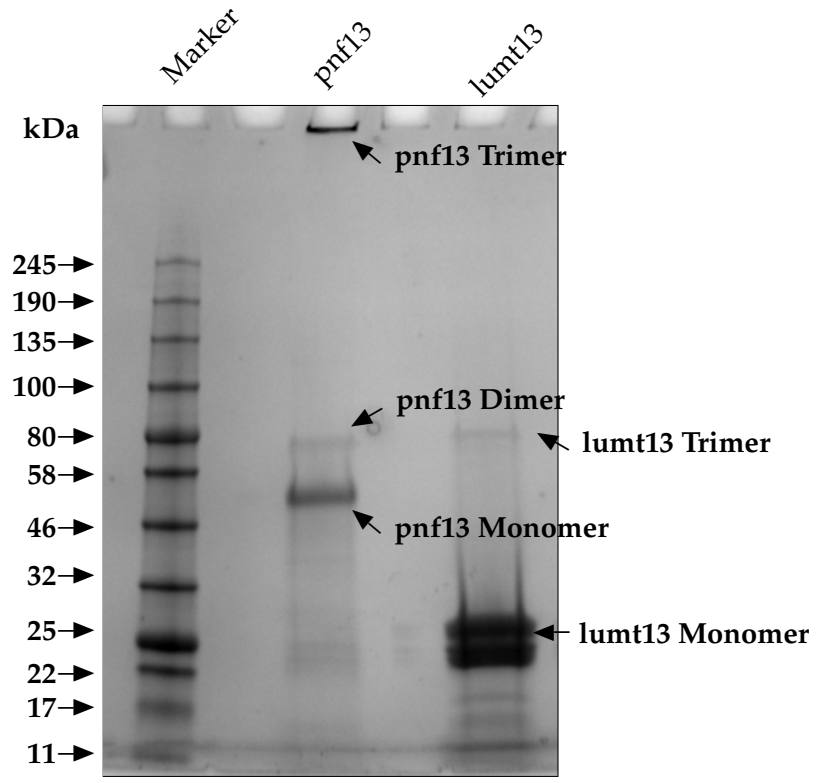


Figure 5.12 | Trimerism of PVC tail fibres revealed via SDS-PAGE

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 circular dichroism 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 six times (technical replicates) in each run, and for each protein, three 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 two-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 two 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

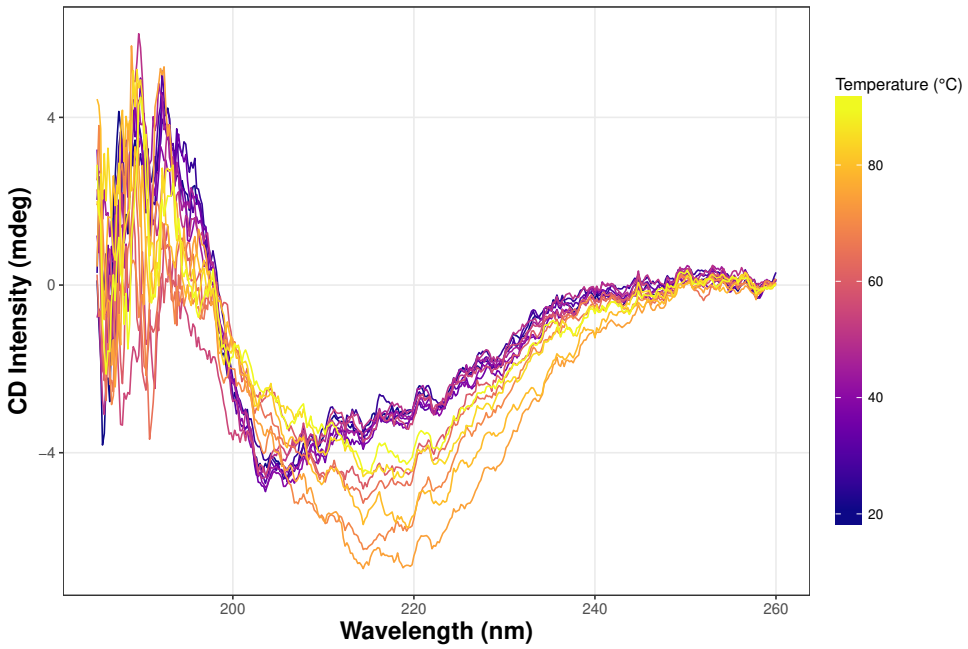


Figure 5.13 | CD TEMPERATURE RAMPING SPECTRA FOR PNF13

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.

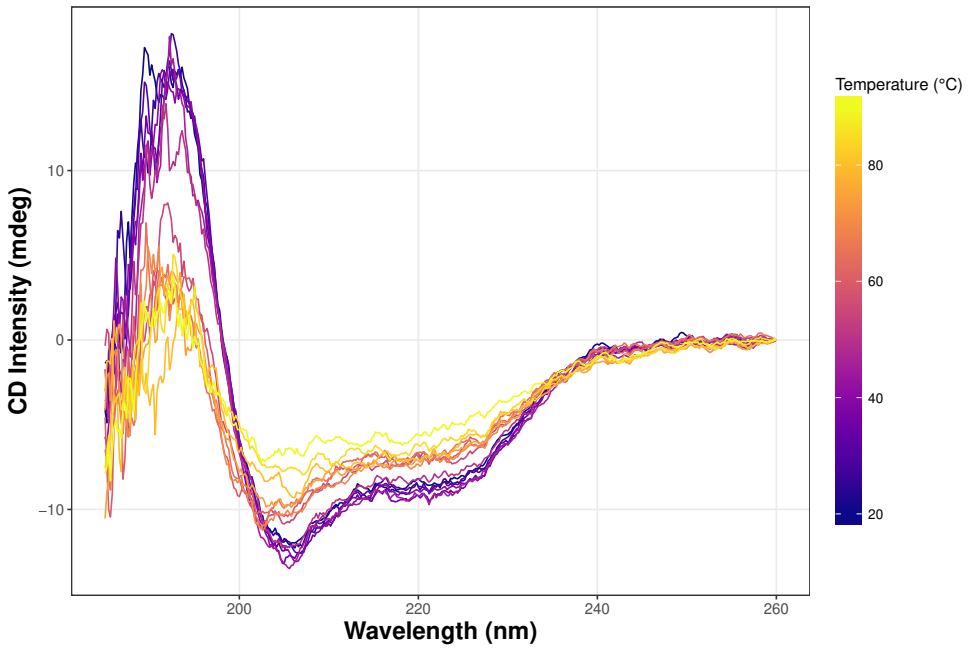


Figure 5.14 | CD TEMPERATURE RAMPING SPECTRA FOR LUMT13

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 three 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 208). 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 Normalised Root-Mean-Square-Deviation (NRMSD) values by a significant margin (see Figure 5.15B on page 211), 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[2]. 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

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

considered unwise to analyse beyond 190 nm.

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 next page for one of the input spectra tested under several models. Dichroweb further provides an NRMSD 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 212 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.

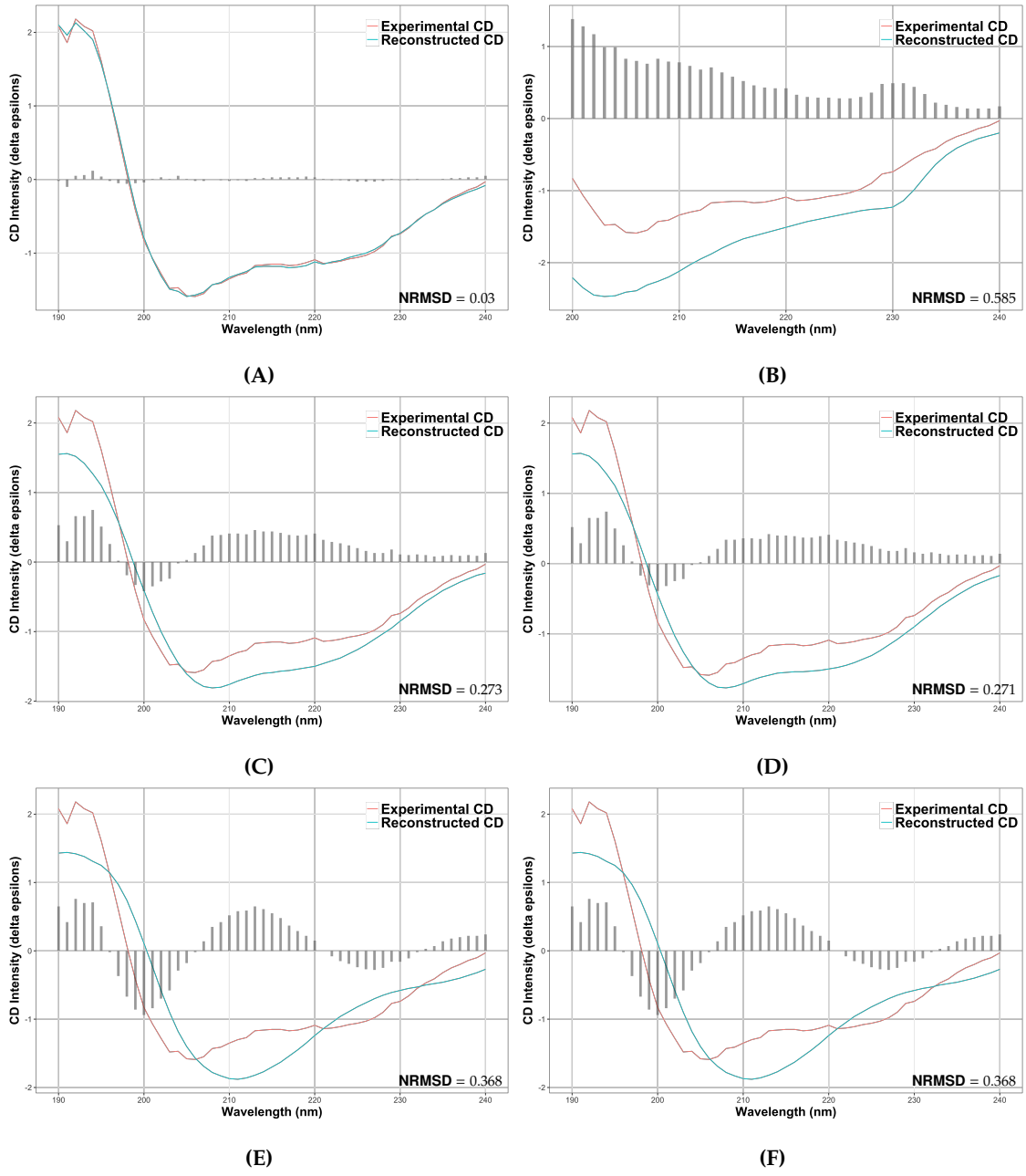


Figure 5.15 | Comparisons of optimal Dichroweb algorithms and reference sets.
Each of these charts shows a comparison of a different algorithm and reference set for identifying the optimal parameters for estimation of secondary structure proportions from the acquired spectra. As an example, each spectra 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.

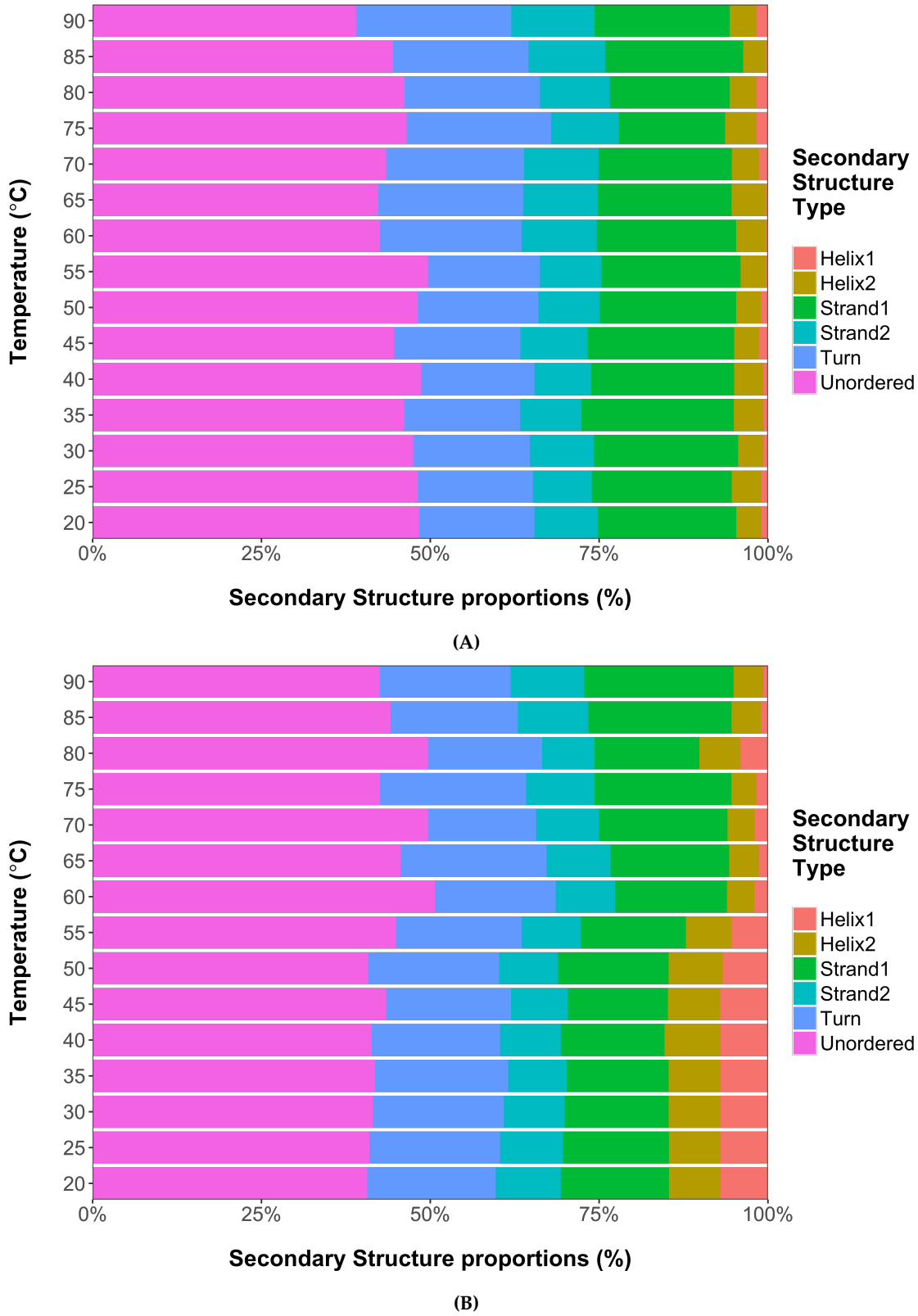


Figure 5.16 | CD MELT SECONDARY STRUCTURE PROPORTIONS FOR TAIL FIBRE PROTEINS.
Stacked bar charts showing the proportions of secondary structure as estimated by Dichroweb, for averages of the 3 replicate spectra, for each protein at 15 different temperatures in the melting gradient experiment. **(A)** 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 three identical monomers which each have a functioning 'lone' structure, and 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 190, and for these tail fibres is reinforced by Figure 5.12 on page 206. 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 two 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"[3], 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 190, and domain homologies detected by HHPred are reproduced in Table 5.4 on the following page. Note, these results are extracted from DSSP assignments,

[3] `http://pdb2cd.cryst.bbk.ac.uk/`

however DSSP only recognises three 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 190 and in Table 5.1 on page 195, 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 187.

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 217 shows the buffer conditions for which crystals could be seen. Figure 5.17 on page 218 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 diffrac-

tion 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 218). 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 removes 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 two 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 following 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

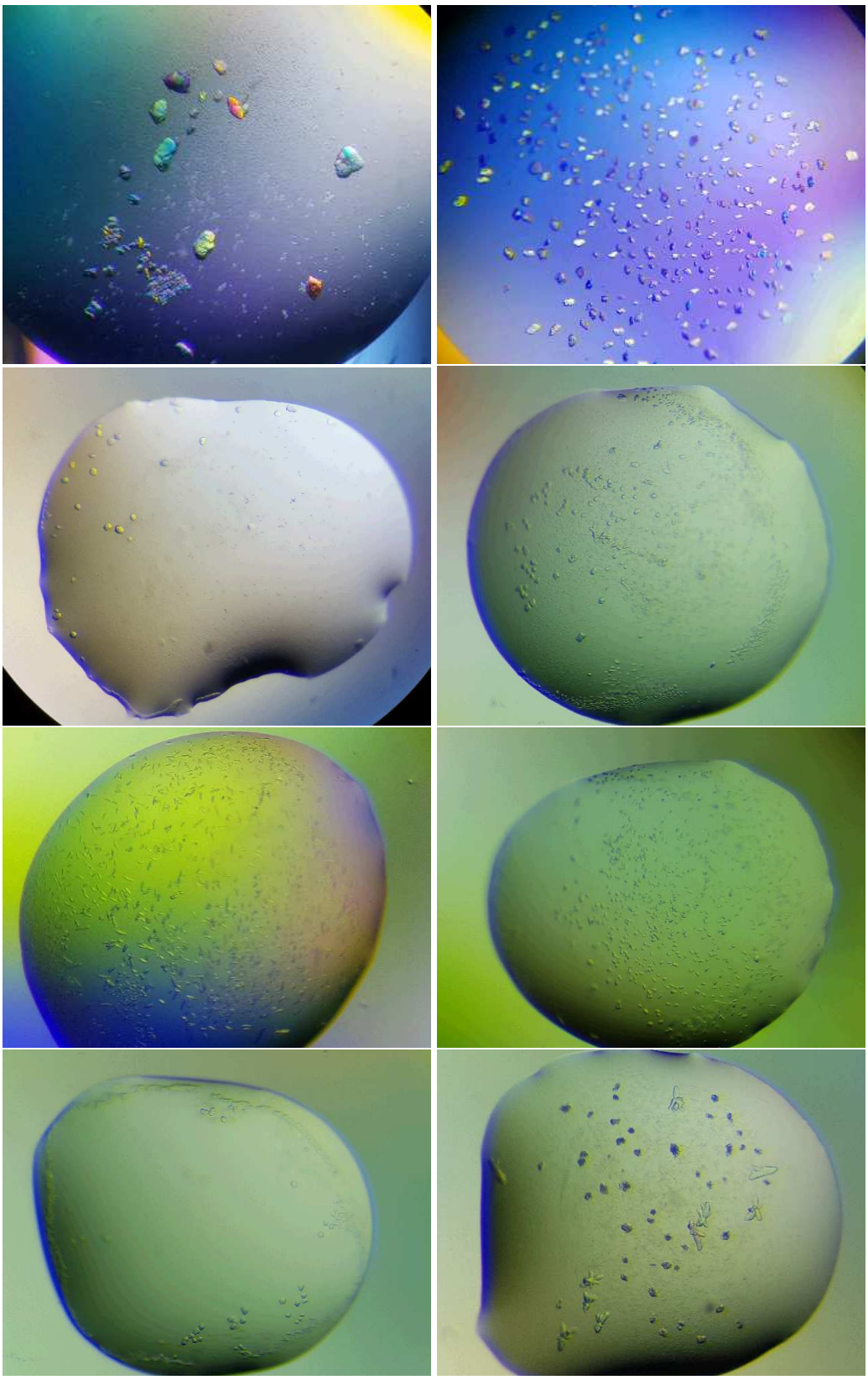


Figure 5.17 | A selection of lumt13 crystal morphologies.
A selection of the crystal morphologies obtained from crystal screening with lumt13 and the Mosquito robot, corresponding to some of the conditions in Table 5.6 on page 217.

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 following 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 next 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.

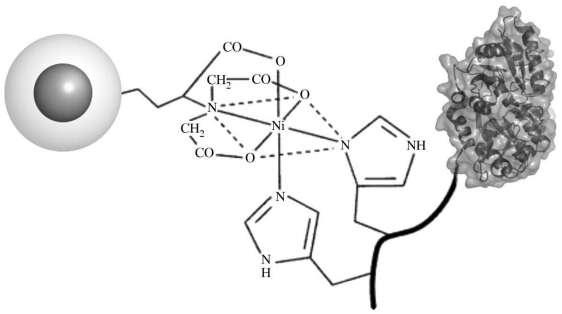


Figure 5.18 | Nanoparticle-Ni-NTA interactions with polyhistidine tracts.
A structural and illustrative diagram of the interaction between metal nanoparticles, Nickel nitrolotriacetic acid chelating coordination groups, and the polyhistidine tracts of a target protein. Image adapted and reproduced from 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

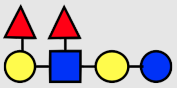
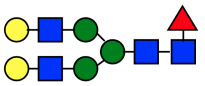
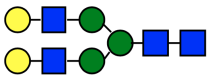
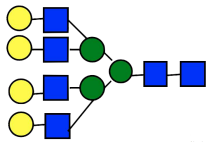
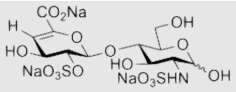
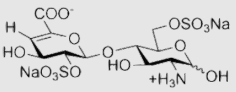
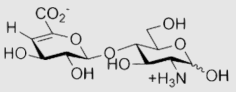
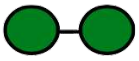
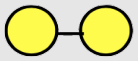
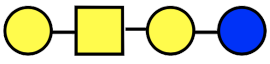
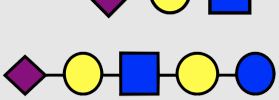
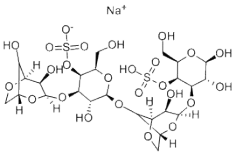
Since the previous two approaches primarily aimed at identifying cell/tissue types and proteins which were preferentially binding the tail fibres, an additional assay utilising glycan arrays was designed to screen for sugar binding. The main motivation for this is that there is significant precedence in the literature for other tail fibre proteins binding surface sugars. The T7 phage short fibre has been proposed 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, three array tests were run, and spots with a fluorescence intensity fold change of at least one between control and sample were counted. Table 5.8 on the following 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 204). 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 206)), 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 purification to purification. 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 213). 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 218 and Table 5.6 on page 217), 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 two additional chaperones, gp38 and gp57 in order to ensure correct structural formation (Granell *et al.*, 2014; 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 194, 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.*, 2004b,a). 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 195. 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 Tibres

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 efficacy. 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 four 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[4] 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 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

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

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-Galactosamine 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 152. 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 additional 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 Expression

"...cells are very tiny, but they are very active; they manufacture various substances; they walk around; they wiggle; and they do all kinds of marvellous things..."

Richard P. Feynman

6.1 Introduction

PVCs were originally identified due to their toxic effects in a screening process known as "Rapid Virulence Annotation" (RVA) (Waterfield *et al.*, 2008; Yang *et al.*, 2006). The "RVA" screen examined a cosmid library comprised of *Photorhabdus* sequences which were heterologously expressed in ordinary lab *E. coli*, and tested against insects, nematodes, amoebae, and macrophages for activity. Should the cosmid clone contain a genomic region which encoded a toxin or virulence factor, assorted lethalities or morbidities could be detected in the screened cell or organism types.

PVC-bearing cosmids were obtained in various states of completeness, often with the 5′ regions deleted, though enough functional cosmids were recovered to produce the first work on the PVCs (Yang *et al.*, 2006). The frequency of deletion of regions of the

PVC bearing cosmids suggested that intact cosmids were not easily recovered in *E. coli* and were unstable and potentially toxic to the host cells. This was borne out by steadily decreasing viability of certain cosmid clones when routinely cultured for experiments.

The precise basis of this toxicity is still unknown. One hypothesis was that the PVCs (being somewhat phage-like), were potentially lysing the cells that produced them (some of the operons are also found in close register to holin-lysin pairs). Since the PVCs were obtained in the cosmids with their own promoters and regulatory signals, their expression was likely aberrantly controlled, in a non-*Photorhabdus* host, even though *Photorhabdus* and *E. coli* share similarities as members of the Enterobacteriaceae

However, there were a number of PVC operons, that despite being cloned in their entirety, did not cause such pronounced cell toxicity. Inferring from RNAseq data for the PVCs in *Photorhabdus*, many PVC operons are silent under normal culture conditions, or seemingly expressed at lower levels. These silent operons, while propagable, were of limited use for further study due to not producing active PVCs. Thus, one of the earliest aims in this project was to explore methods for more reliable and controllable heterologous expression of PVCs on demand. This is also desirable as the cosmid clones often contain captured genomic sequence flanking the PVCs, which may contain genes/proteins which could confound assays.

In addition to genetically engineering control of PVC sequences, a better understanding of the natural expression patterns and regulation of the PVCs is also desirable. Why, for example, are some PVCs expressed in standard culture conditions, whilst others are entirely silent?

The wide variety of PVC loci suggests they may be responding to a plethora of environmental cues to trigger their deployment, and thus their expression patterns could be very prescriptive. Furthermore, due in large part to its esoteric lifestyle, *Photorhabdus* is known to exhibit a significant degree of population heterogeneity in many other phenotypes (Langer *et al.*, 2017; Heinrich *et al.*, 2016). This suggests that not only might the PVCs expression/deployment depend on environmental cues, but that it may also depend on culture conditions (such as density), growth phase, and phenotypic state (*Photorhabdus* exists as 'primary' and 'secondary' cells (Boemare and Akhurst, 1988)). Its

primary/secondary variability has been shown to manifest in various phenotypes. For instance, the bulk of the bioluminescence produced by *Photorhabdus* cultures is attributable to the primary (1°) cells (Boemare and Akhurst, 1988; Akhurst, 1980), the majority of anthraquinone-based pigmentation occurs in 1° cells with a particular metabolic suppression in 2° whereby the same transcription factor *AntJ* has diametrically opposite effects in primaries versus secondaries (Heinrich *et al.*, 2016; Langer *et al.*, 2017). Other examples include the differential deployment of a variety of proteolytic enzymes (Marokházi *et al.*, 2004), a global shift in fundamental biochemical processes such as respiration (Smigielski *et al.*, 1994) and production of secondary metabolites (Turlin *et al.*, 2006), and, perhaps most prominently and significantly of all, 2° cells are incapable of association with the nematode vector. Thus, a second aim of this chapter is to asses the PVC operons for any heterogeneity in expression among a bacterial population.

Little is precisely known about the choreography of PVC (and, more broadly, many caudate structure's) expression and assembly. The T4 phage is the best understood, as evidenced in Figure 1.6 on page 20, however, T4 genetics are quite far removed from that of simple caudate structures, with a much larger genome and both "early" and "late" expression profiles (O'Farrell and Gold, 1973). The best indication for the regulatory system in the PVCs comes, once again, from the work of Hurst *et al.* (2007, 2004) on the Afp. In the original studies that attempted to elucidate the structure, expression of the Afp cluster was only achievable when an additional protein from the pADAP plasmid was also heterologously expressed. Namely, *anfA1*, the protein was identified upstream of the Afp cluster as part of a pair of genes (the other is termed *anfA2*), which contained a NusG-like domain with some sequence similarity to the RfaH transcript elongation factor, and over-expression resulted in a concomitant over-expression of the Afp itself. To slightly confound these findings however, Hurst *et al.* (2018) have identified another Afp locus in *Serratia proteamaculans* strain AGR96X, termed AfpX, which has a disparate 5' region, and responds to mitomycin C induction (unlike the Afp), suggesting it has somewhat different regulatory mechanisms. It does, however, retain the 2 *anfA* loci upstream, and associated *ops* binding site (discussed in the next paragraph).

RfaH, a specialised version of NusG (a more generalist transcription factor, which curi-

ously has roles in both termination *and* anti-termination) has been shown to be important in the expression of long operons (Bailey *et al.*, 1996), by prolonging the transcripts and preventing the polymerase from prematurely dropping off the strand (increasing processivity), ordinarily brought about by Rho-dependent termination. RfaH is recruited to the RNA polymerase transcript elongation complex (TES) at a specific DNA motif called the *ops* (operon polarity suppression) site (which in *E. coli* has the sequence `5'-GGCGGTAGnnTG-3'`). Operon polarity refers to the mechanism by which under 'ordinary' expression circumstances, Rho proteins bind to exposed mRNA regions brought about by ribosome halting, but continued RNA polymerase processing. In prokaryotes, translation occurs directly on the nascent 5'-end of the transcript as it is produced from the RNA polymerase, meaning the two complexes are in close register the majority of the time. However, when a stop codon in the nascent RNA is met by the ribosome - it halts - meanwhile, the RNA polymerase continues to process, and therefore a small gap of exposed RNA can form. The Rho proteins bind to this region of exposed RNA, track along it, and disrupt the RNA polymerase, preventing/reducing the 'unintended' transcription of downstream genes. This reduction in expression of downstream regions is what is known as operon polarity (Santangelo *et al.*, 2008; Adhya *et al.*, 1974; Banerjee *et al.*, 2006; Nudler and Gottesman, 2002). The *ops* site therefore earns its name by mediating a counteracting ("suppressing") effect to this 'premature' transcription termination for long operons, by recruiting additional transcription factors (such as RfaH) to the TES, ensuring that the RNA polymerase stays attached to the template and continues to process the mRNA for a full length operon. Transcription is enhanced for roughly 20 kb downstream (Artsimovitch and Landick, 2002; Leeds and Welch, 1996, 1997), which is the approximate minimal length of the structural components of a PVC operon. Effectors for PVCs often appear with their own promoters and can be transcribed independently of the rest of the operon, as seen in previous transcriptomic studies of *Photorhabdus* - suggesting they may not require the influence of transcript elongation factors, and that it is primarily implicated in elaboration of the structural complex only.

Gene clusters which have been demonstrated to be dependent on RfaH, include the LPS core, exopolysaccharides (Wilkinson *et al.*, 1972), haemolysin toxins (Landraud *et al.*,

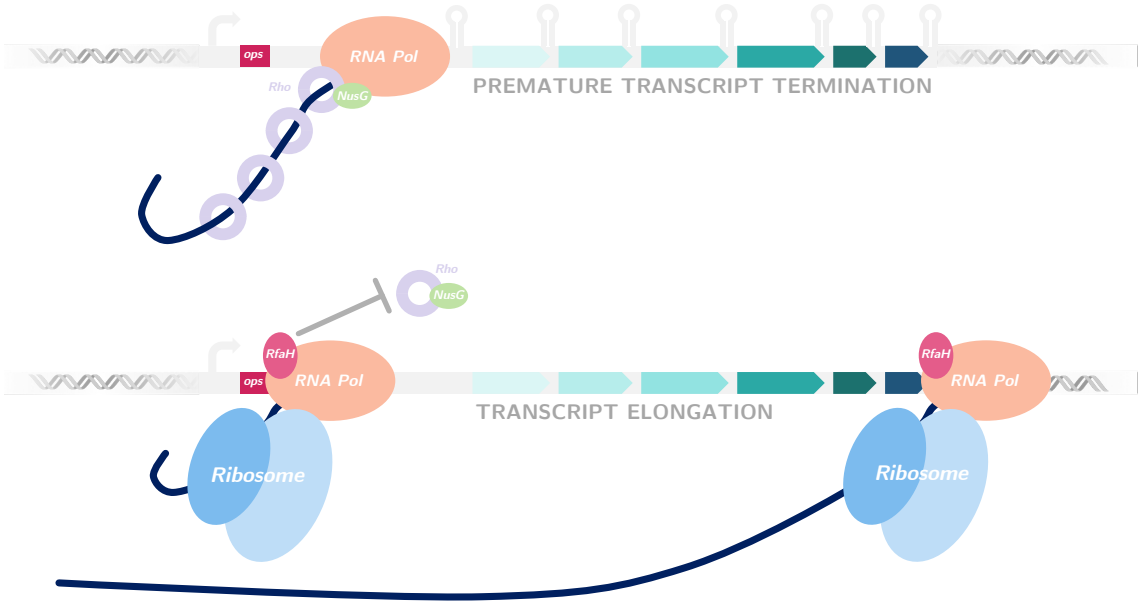


Figure 6.1 | A schematic diagram depicting elongation mechanism of RfaH.
A schematic diagram to demonstrate the transcript elongation effect of RfaH. In the upper panel, in the absence of RfaH, the transcript terminates early (thus a high degree of polarity) when encountering terminators in the RNA structure. NusG and Rho proteins complex with the polymerase and the nascent strand, blocking any translation. In the lower panel, with RfaH present and recruited to the polymerase at the *ops* site, and the protein blocks the activity of Rho and NusG proteins, and in complex with the RNA polymerase and ribosome, increases the processivity of the polymerase, traversing through terminators. RfaH is believed to have some activity in direct loading of the ribosome on to the RNA strand. Diagram based on Figure 1 from Hu and Artsimovitch (2017).

2003; Leeds and Welch, 1996, 1997) and the F-type conjugation pilus. The *ops* site which RfaH obligately depends upon has also been found in *Shigella flexneri*, *Yersinia enterocolitica*, *Vibrio cholerae* and *Klebsiella pneumoniae* polysaccharide synthesis clusters, as well as the RP4 fertility operon of *Pseudomonas aeruginosa* (Bailey *et al.*, 1997) - and of course the *S. entomophila* Afp (Hurst *et al.*, 2007). The *ops* site is therefore widespread in the *Proteobacteria*, and possibly further. Deletion of either the *ops* site, or RfaH itself results in equivalent reduction in transcription (Bailey *et al.*, 1997). Interestingly, the *ops* site has also been identified as part of a larger 39 bp sequence known as the "JUMPStart" sequence, which consists of a conserved but somewhat degenerate GC-rich hairpin structure immediately prior to the *ops* sequence. The JUMPStart sequence has been detected in most of the examples listed previously (Wang *et al.*, 1998). It is not, however, found in the *tra* operon that gives rise to the F pilus, nor in the typical *hyl* haemolysin operon (Nieto *et al.*, 1996). However, Leeds and Welch (1997) showed the *hyl* operon of the *E. coli* J96 strain *does* use a full JUMPStart motif). Why the JUMPStart sequence, which can be thought

of as an extended *ops* sequence, is necessary, is not clear, since the *ops* site is evidently sufficient for RfaH recruitment and long operon expression in many cases.

Interestingly, literature such as the papers by Bailey *et al.* (1996) and Santangelo and Roberts (2002) note that RfaH regulation seems to also be a hallmark of long operons whose products are destined for the extracellular environment/membrane (e.g. LPS, pili, secreted enzymes). Though it is far from clear if this is a strict dependence/relationship, and there is certainly no understood mechanism; it is tempting to speculate that this might be the case for the PVCs and Afps which are also destined for the extracellular environment, however.

Finally, therefore, another aim for this chapter was to investigate the role of RfaH (or RfaH-like proteins), the *ops* or JUMPStart sequences, and antitermination in the mechanistics of PVC operons.

Chapter Aims:

- Engineer controllable PVC expression constructs.
- Examine the natural regulation and expression patterns of PVCs in *Photorhabdus* populations.
- Investigate the putative role of RfaH-like proteins in PVC regulation.

6.2 Experimental procedures

Figure 6.2 depicts a high level overview of the experimental threads running through this chapter as it contains areas of different focus. The chapter is broadly divided in to efforts to reconstitute the PVCs heterologously in *E. coli* through a number of methods (two blue threads), and similarly, efforts to understand the activity of the natural expression system in several ways (two orange threads).

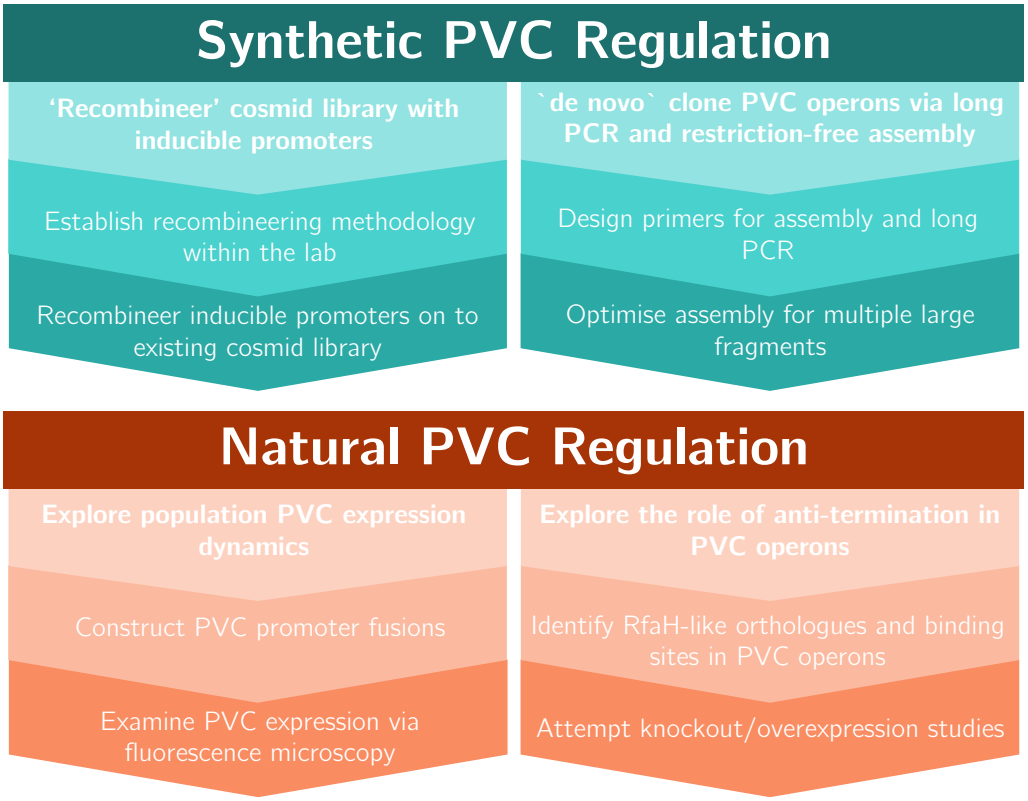


Figure 6.2 | A high level flowchart of the threads for studying PVC regulation.
A high level flowchart showing the experimental threads for this chapter. The upper blue chart shows the efforts towards making controllable synthetic PVC constructs through recombineering and synthetic fragment assembly approaches. The lower orange chart depicts the 2 primary packages of work undertaken to probe the natural regulatory system of the PVCs, and the 'sub-threads' within this. Namely, construction of promoter fusions allowing for studies of the population heterogeneity (or lack thereof) in PVC expression, and an examination of AnfA1 orthologue control of PVCs within *Photorhabdus*.

6.2.1 Cloning and engineering PVC operons

As discussed in the introduction to this chapter, the only existing heterologous expression system that was available was the 'raw' cosmid clones from the initial identification of

the PVCs. The most bioactive of these, PVCpnf from the *P. asymbiotica* ATCC43949 strain, was also the most genetically unstable, and over time freezer stocks had dwindled since it could not be easily propagated. This section will explore the experimental efforts undertaken to try and reconstruct these long operons in *E. coli*.

All of the techniques that follow have specifically avoided attempts at classically cloning PVCs, instead opting for techniques that rely on homologous recombination. Classical restriction based cloning was ruled out early on, as the operons are very long and rich in restriction sites that rendered a lot of the reliable 'standard' lab enzymes unsuitable, especially as the operons would probably need to be cloned in several chunks. Ideally, a robust technique was desired that would allow any of the PVC operons to be cloned with minimal protocol changes. Therefore, a dependence on particular restriction sites might preclude certain operons from being cloned. Additionally, it was unknown whether there was internal structure to the operons such as internal promoters. Certain proteins are certainly transcriptionally coupled for example, so the incorporation of restriction sites and any intervening spaces could impact the expression and assembly of the PVCs. However, a previous post-doc of the group did achieve the cloning of a single operon in this manner separately, contemporary with the efforts here. While this demonstrates in principle that classical cloning is a viable approach for constructing this particular operon, a previous attempt by the same lab at an alternative operon had lead to various assembly errors internal to the sequence, and rendered the construct unusable.

Moreover, much of the existing literature for cloning of long operons pointed to homologous recombination based techniques as a promising approach for subcloning recalcitrant or difficult targets (for example, Wang *et al.* (2016); García *et al.* (2004), both of whom present novel DNA capture techniques). Despite the interest in bioprospecting or 'genome mining' (combing genomes for biosynthetic gene clusters of interest) of late (Ziemert *et al.*, 2016; Van Lanen and Shen, 2006; Netta *et al.*, 2009; Lautru *et al.*, 2005; Bergmann *et al.*, 2007; Wenzel and Müller, 2009; Charlop-Powers *et al.*, 2014), due in no small part to the explosion of interest and improved capabilities in (meta)genomics, it is still non-trivial to manipulate sequences of this size.

6.2.1.1 Recombineering

Since the original cosmids containing segments of the captured genome were available, it was decided that a promising approach might be to try and engineer control in to the replicons that were already obtained, attributing the reductions in cell viability simply to uncontrolled expression via natural promoters.

Recombineering is a long standing technique for 'recombination-mediated genetic engineering' (hence the name). It relies on the use of three phage λ proteins, named Gam, Beta, and exo, to facilitate the incorporation of segments of linear DNA with 5' and 3' homologous stretches to an insert site of choice. Perhaps its most 'famous' use is the technique of choice for the generation of the Keio collection of *E. coli* knockouts, which first defined a minimal essential genome (Baba *et al.*, 2006). Two major alternative techniques exist, using two different enzyme sets: RecET from the Rac prophage and "Lambda-Red", which is the mechanism discussed here. Exo is an exonuclease responsible for creating 3' overhangs (it has 5' $\rightarrow$ 3' exonuclease activity, hence its name); Beta is a single strand binding protein which stabilises these overhangs so that they aren't degraded by host enzymes and recombination can occur more efficiently. To further improve the recombination efficiency, the Gam protein inhibits the RecBCD nuclease complex (Yu *et al.*, 2000). Figure 6.4 on page 244 shows a schematic of how the process works.

6.2.1.1.1 Chromosomal engineering

Since recombineering was a new technique in the lab, a working protocol was first devised by attempting to knock out chromosomal targets. Three chromosomal *E. coli* targets were chosen as trials. Initially the *att* site used for phage insertion was chosen as it was proposed that this should be non-essential. All attempts to engineer this site failed however, for reasons not fully understood. Instead, the three sites chosen to replace it were the *hyfC* locus, which encodes part of the hydrogenase complex, since in the Keio collection data this was scored as the least essential gene (and thus should be easy to knock out) (Yu *et al.*, 2000). Moreover, the hydrogenase is only transcriptionally active under anaerobic conditions (Skibinski *et al.*, 2002), which suggested that the region would be experiencing less 'cellular traffic', such as RNA polymerases, thus making recombination as simple

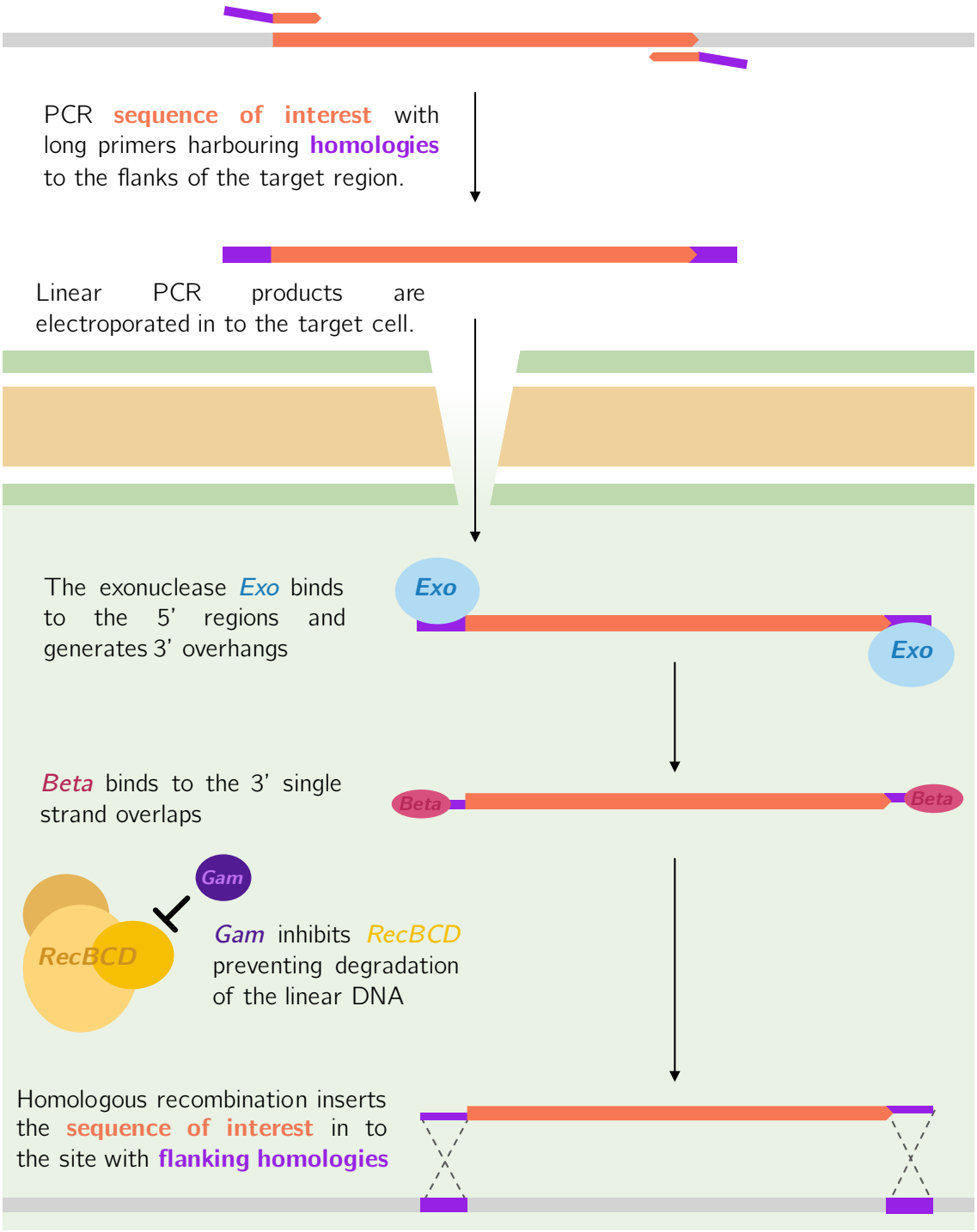


Figure 6.3 | The mechanism of action for "λ Red"-based recombineering.

"λ Red"-mediated engineering is performed by amplifying an insert sequence of interest using primers with overhanging sequences (up to ≈50 bp), which are homologous to the flanking regions of the intended insert site. The linear DNA is electoporated in to the desired cells. Phage λ proteins are induced in the cells prior to making them electrocompetent, and thus the exonuclease Exo is readily available to begin digesting the linear DNA. As it does so, single stranded binding proteins, Beta, bind the overhangs to stabilise the fragment and prevent degradation. The Gam protein which is also expressed with Beta and Exo, inhibits the Rec nuclease complex to preserve the linear DNA and improve efficiency. Finally, the sequence of interest recombines with the target site as determined by the flanking regions.

as possible. Next, using the same approach *speB* was identified as the next-least essential protein. *speB* encodes an 'agmatinase' enzyme, which is responsible for producing putrescine and urea from the precursor agmatine (Szumanski and Boyle, 1990). Lastly, the endonuclease encoded by the gene *endA* was targeted for knockout. The primers in Table 2.8 on page 69, show the sequences used to amplify antibiotic resistance cassettes encoded on the helper plasmids pJET-FRT-Kan and pJET-FRT-Cat. These plasmids carry an antibiotic cassette flanked by conserved primers that can be used to amplify either, and Flp-flippase recombination sites which can be used to optionally 'pop out' the introduced marker to create (semi)clean deletions. Note, *endA* is denoted as a 'pseudo' *endA* in the table with a ψ, since the protein is actually mutated in laboratory *E. coli* already, to reduce its activity, however the locus is still largely sequence-intact.

Recombineering was performed as outlined in Section 6.2.1.1 on page 242. Successful colonies were screened with the primers used to amplify the cassette, and a number of successful colonies were observed for all 3 inserts. To be certain that the insert was correct, flanking primers to the *hyfC* locus were designed ≈200 bp up and downstream of the insertion site, and used to confirm the correct insertion via Sanger sequencing.

With the technique successfully applied in the test case, the protocol was adapted and applied to recombineering the cosmids.

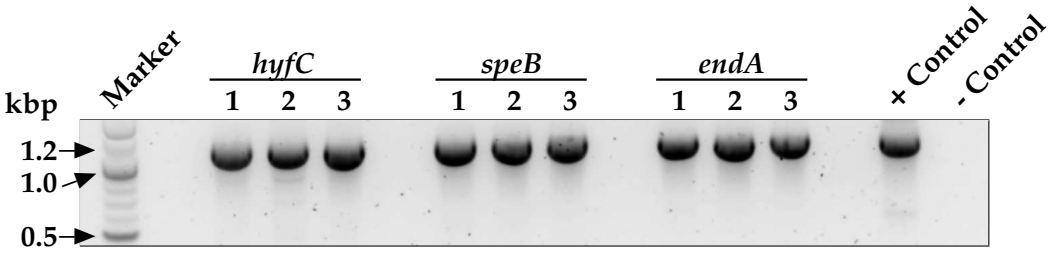


Figure 6.4 | Validation of recombineering techniques for chromosomal targets.
This gel shows the first indication that the recombineering protocol had finally been sufficiently optimised and was working satisfactorily to test against PVC elements in subsequent study. The gel shows the result of colony PCR against DH10β, demonstrating insertion of a kanamycin cassette in place of the genes *hyfC*, *speB* and *endA*. The helper plasmid which bears the kanamycin cassette is used as a positive control. Expected band size was 1168 bp.

6.2.1.1.2 Cosmid engineering

Unlike the chromosomal targets, which were simply 'overwritten' with a resistance cassette, the objective of engineering the cosmids is to incorporate an inducible promoter upstream of the operon, thus providing control potentially without having to reconstruct the operons. The first step was to generate helper plasmids which would bear the template for incorporation. To this end, the helper plasmids pJET-FRT-Kan and pJET-FRT-Cm were derivatised with the *araBAD* promoter system from pBAD30, this provided a selectable marker and inducible promoter adjacent to one another which could then be PCR'd for use as the electroporated linear oligonucleotide. The resultant plasmids maps are shown in Figure 6.5 on the next page, with the primer sites used for cassette amplification which incorporate the selection marker, promoter system, and FRT ("Flippase Recognition Target") sites, so that the marker can be subsequently removed if desired.

With a functional protocol for recombineering established to good efficiency in the chromosome, the next step was to apply it to the cosmids. Since the cosmids contained extraneous genomic sequence, their own replication origins, and multiple resistance markers, a more complicated engineering process had to be designed. While the recombination process is as simple as the chromosomal protocol, the extra complications arise from these resistance markers and the origins would limit the plasmids that could be used in future in transcomplementation etc.

The first complication is that the pKD46 plasmid which was used for chromosomal recombination could not be used with many of the cosmids as they both harbour Ampicillin resistance markers and thus the selection would not maintain each of them. To circumvent this problem, the DY380 strain was procured which features the same enzymes integrated in to the chromosome, and has a tetracycline marker. This also alleviates the need for an additional incompatibility group between plasmid origins (fortunately, the 101ts origin that pKD plasmids use is compatible with ColE1 origins). This allows the cosmids to be propagated within the strain with the prerequisite enzymes present.

Next, the template cassette from the two helper plasmids had to be chosen according to the library sequence. PVC operons were present on either cosmids with Ampicilin/Kanamycin/Neomycin resistance, or on fosmids with only Chloramphenicol resistance,

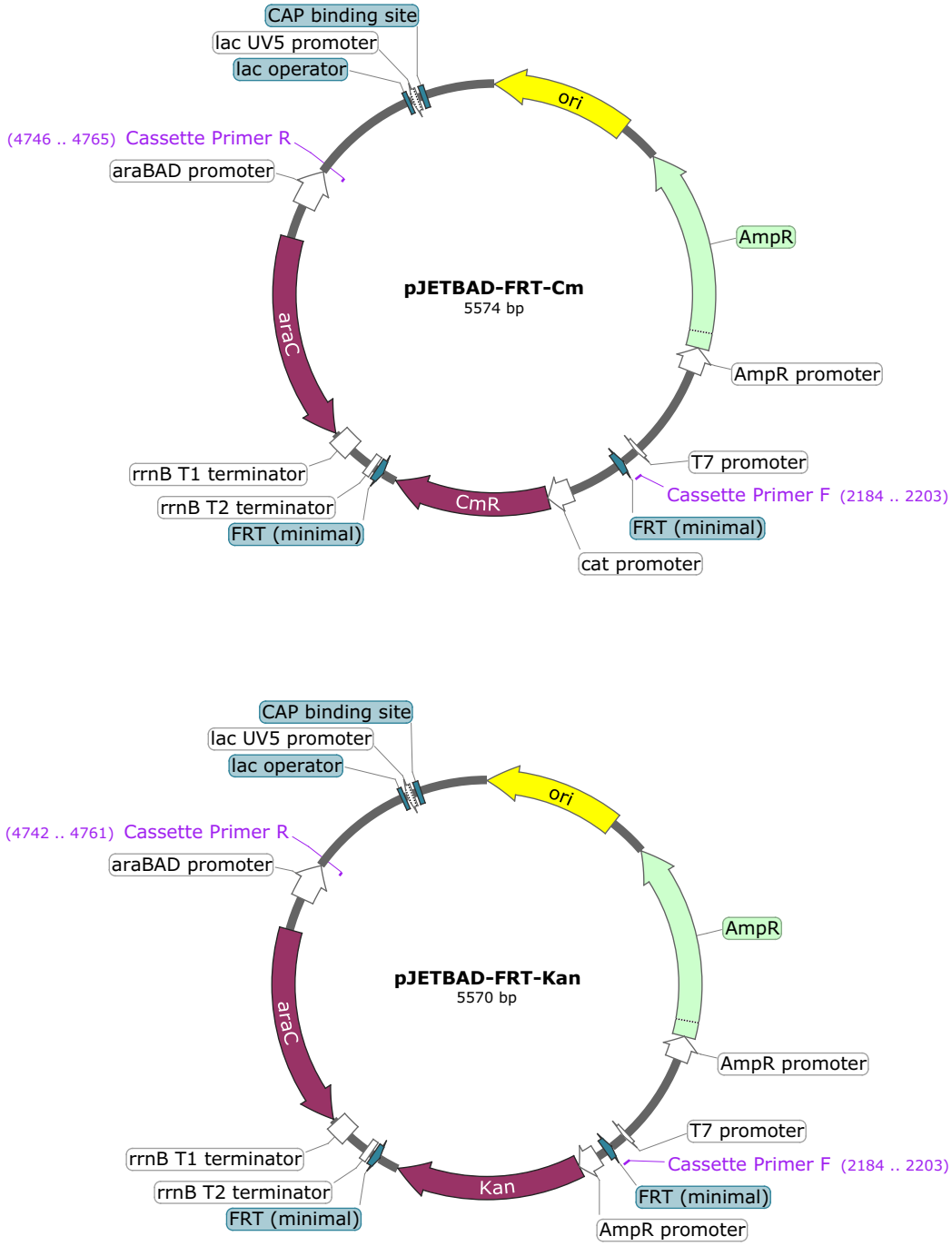


Figure 6.5 | Helper plasmids bearing the promoter/marker system for recombineering. Vector maps for the plasmids derivatised from the pJET-FRT plasmids which carry selectable markers flanked by FRT sites, such that the marker can be removed after a deletion has been achieved. The plasmids were derivatised in this study by incorporation of the araBAD promoter system.

therefore only one template cassette is therefore compatible. The linear oligos would need to be prepared on a cosmid-by-cosmid (or fosmid) basis, so as to incorporate the homology needed. This was done in such a way as to remove a chunk of the extraneous genomic sequence at the 5' end at the same time.

Unfortunately it became apparent in development of the process that recombineering multicopy replicons introduces an additional level of complexity than was even anticipated. It was not possible to unambiguously identify correctly engineered constructs from the resultant mix of modified, parental, and chimeric sequences even when compatible resistances are chosen. The issues surrounding this kind of engineering are evaluated further in the Discussion for this chapter and have been highlighted in the literature in the past, but ultimately after the significant effort invested in optimising the protocols and generating the required materials and constructs necessary, it was decided that the approach should be abandoned in favour of a more reliable and less time consuming alternative, one of which is explored in the next section.

6.2.1.2 Gibson assembly

To attempt to clone the PVCs in a manner that would not depend on restriction sites, a process of construct creation using the fairly new "Gibson" assembly technique was tested and optimised. Gibson assembly has been popularised ever since the advent of the first synthetic genome produced by Gibson *et al.* (2010, 2009) (hence the name). The method promises *in vitro* assembly in a largely sequence-agnostic manner, crucially without the requirement for restriction sites, for fragments up to the hundreds of kilobase range.

The technique works by mixing overlapping linear DNA, which have complementary ends (in a manner similar to recombineering), and using three isothermal enzymes to create 'sticky ends' which anneal, are 'gap-filled' and ligated, all in a single reaction pot. Firstly, T5 exonuclease creates the 3' overhangs, exposing the complementary overlaps as stretches of single stranded DNA. These complementary regions base pair, and the T5 nuclease is displaced by the Phusion polymerase (or *Taq*), which 'back fills' the gaps created by the exonuclease beyond the overlap sites. This leaves a double stranded fusion product of the adjoining fragments, but with disconnected backbones. As the last step of the isothermal assembly reaction, the *Taq* ligase enzyme seals the backbones, yielding two fully connected fragments with a high degree of sequence fidelity at the joins.

Several attempts were made, initially following the manufacturers guidelines, but later, additional optimisations were included which improved assembly size and efficiency. The most optimal of these is the protocol given in Section 2.2.6 on page 75. Briefly, substantially longer primers than the suggested overlaps were used - this was for two reasons. Firstly, the longer (70 bp) primers (35 bp of each neighbouring fragment), were considerably easier to PCR with the NEB Q5 enzyme, especially at longer lengths, where the shortest PCR fragment was ≈4.5 kb. Secondly, the increased overlaps provided the first indications of any assembly whatsoever, suggesting better efficiency, especially for such long fragments. The suggested overlaps of ≈12-18 bp were attempted initially, using standard *in silico* construction tools, but yielded no assembled colonies at all.

Two key optimisations which increased colony recovery and fragment incorporation were a step-wise assembly process (rather than the 'one-pot' process recommended - see Section 2.2.6 on page 75), as well as DpnI treatment of vector backbone templates for PCR,

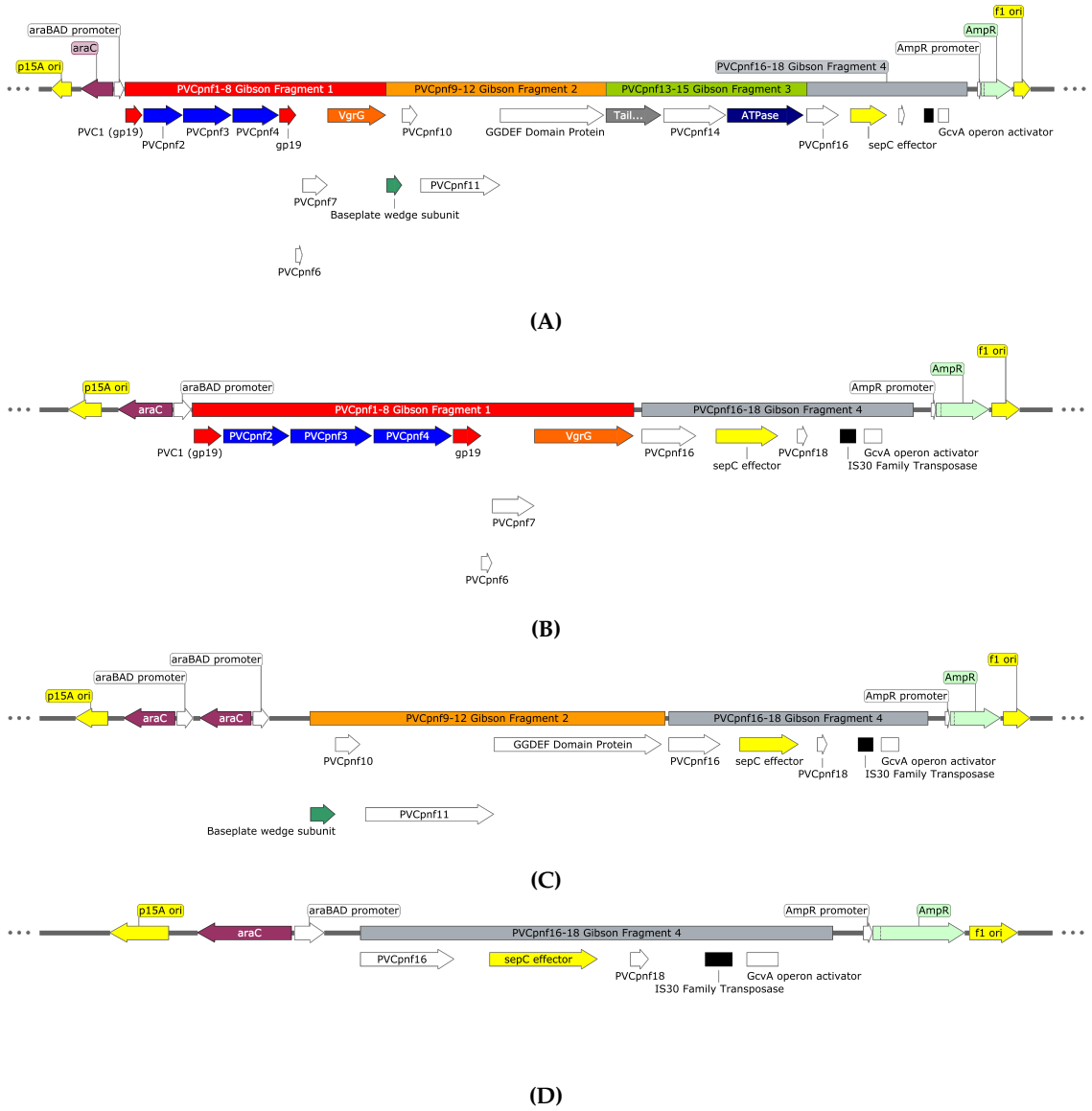


Figure 6.6 | Vector maps of obtained Gibson constructs

Vector maps for a selection of obtained Gibson constructs, assembled from short reads. Assemblies in various states of completeness were obtained, demonstrating that the overlaps, fragment lengths, and protocol used for Gibson assembly appears viable. **(A)** The construct scheme devised for the PVCpnf operon from ATCC43949. The operon was broken in to 4 fragments, ranging from ≈4-8 kb, so as to avoid interrupting CDSs, and these are colour coded, with the genes below. **(B)** A Gibson construct which successfully incorporated fragment 1 and fragment 4, but no middle fragments. It successfully incorporated the largest fragment. Fragment 4 appears easy to clone as it is recovered in all the constructs (and more not shown). **(C)** A Gibson construct which successfully incorporated fragment 2 which has several unknown genes, and fragment 4 again. No constructs have been found with Fragments 1 and 2 surprisingly. **(D)** A smaller Gibson construct which has captured only fragment 4. This was the most commonly recovered construct, which is a little surprising as it carries a toxin and a number of putative regulators/transposaes etc. Conspicuously absent from any recovered constructs is fragment 3.

which drastically reduced the empty background.

Despite these optimisations however, it was not possible to obtain entirely complete assemblies, though several partial assemblies came close. Figure 6.6 on page 249 shows a selection of the 'most successful' constructs that were obtained. Interestingly though, these assemblies offer some clues as to why the PVCs appear recalcitrant to cloning.

Specifically, the cloning approach was able to recover an example of all of the PVC fragments that were cloned (see Figure 6.6A on page 249), with the exception of fragment 3 (green), and possible reasons for this are explored in the discussion.

6.2.2 Population heterogeneity in PVC activity

A pre-existing hypothesis for the manner in which PVCs are deployed was that the populations may demonstrate extreme variability/heterogeneity, as is often the case with other aspects of *Photorhabdus* physiology. In order to investigate this further, a selection of PVC operons from both *P. luminescens* and *P. asymbiotica* were chosen to have their promoters cloned in to fusions with GFP. Briefly, the promoterless GFP bearing plasmid pGAG had been previously derivatised by the lab to remove the ribosome binding site and start codon of the GFP so that full promoter region fusions could be made. Primers were then designed to incorporate ≈500 bp upstream of 4 PVC operons, and the first 5 amino acids of each PVC1 locus from both *P. luminescens* and *P. asymbiotica*, thus reconstituting the now fused GFP coding sequence and ensuring a reliable fusion. Since all the constructs follow the same basic format, Figure 6.7 on the following page shows the map of the multiple cloning site for a single example. Vector descriptions and primers can be found in Chapter 2 on page 61.

Images were captured on a Leica DMi-8 inverted epifluorescent microscope with a Hamamatsu Flash4 4K camera, under phase contrast and GFP fluorescent channels - and the overlays are shown in the upcoming figures. Due to heterogeneity in the sample prep on agar pad-based slides, some images were darker than others. Consequently, images have been postprocessed to normalise their intensities to a control reference image which had an acceptable grey intensity for reproduction clarity. Images have also had their colour depth reduced for clarity, as well as being downsampled to 35 pixels per inch, from 72 (to reduce image file size).

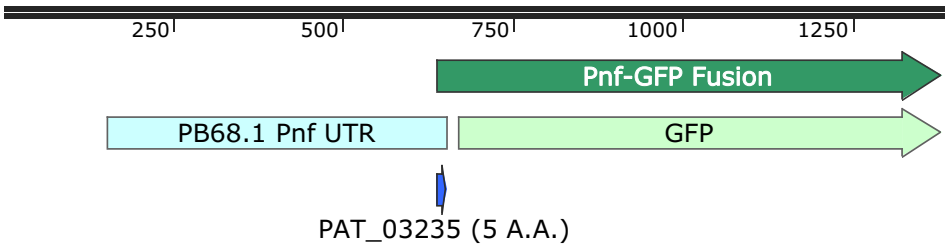


Figure 6.7 | The promoter region fused to GFP for *P. asymbiotica* PB68.1 Pnf in pAGAG
The construct scheme for promoter fusions of PVC operons. pAGAG has the RBS and start codon for GFP removed, allowing it to be replaced with the full promoter region of various PVCs. In this example, ≈500 bases upstream of the "pnf" PVC from the *P. asymbiotica* strain PB68.1 is inserted in to the BamHI and KpnI sites of the vector, creating a fusion that also has a small linker of 5 amino acids, following on from the first 5 amino acids of the PVC first locus.

Four images per slide/time point are shown to give a representative sampling of the slide as best as possible. There is a key to the semi-quantitative fluorescence levels below each column. In a departure from the sequence discussed in the chapters so far, the *P. asymbiotica* strain used here is not the 'model' strain ATCC43949, but instead a Thai isolate denoted PB68.1 sequenced in house, with PVCs identified. This was used as it is considerably easier to transform than the ATCC43949 strain, allowing for an *P. asymbiotica* and a *P. luminescens* strain comparison.

The degree of fluorescence is scored based on the intensity seen in the stills captured, but also indirectly incorporates a dimension of time, as the semi-quantitative values that are given are also taking in to consideration the time taken to scan the slide and look for regions of interest. To be more explicit, if a slide was found with one or two extremely bright cells, but it took considerably longer to find those two cells amongst the population, that sample will be given a lower brightness score, reflecting the heterogeneity and scarcity of the fluorescent cells. A selection of 'empty vector' controls, without promoters or GFP start codons are also shown, to ensure no background activity or auto-fluorescence.

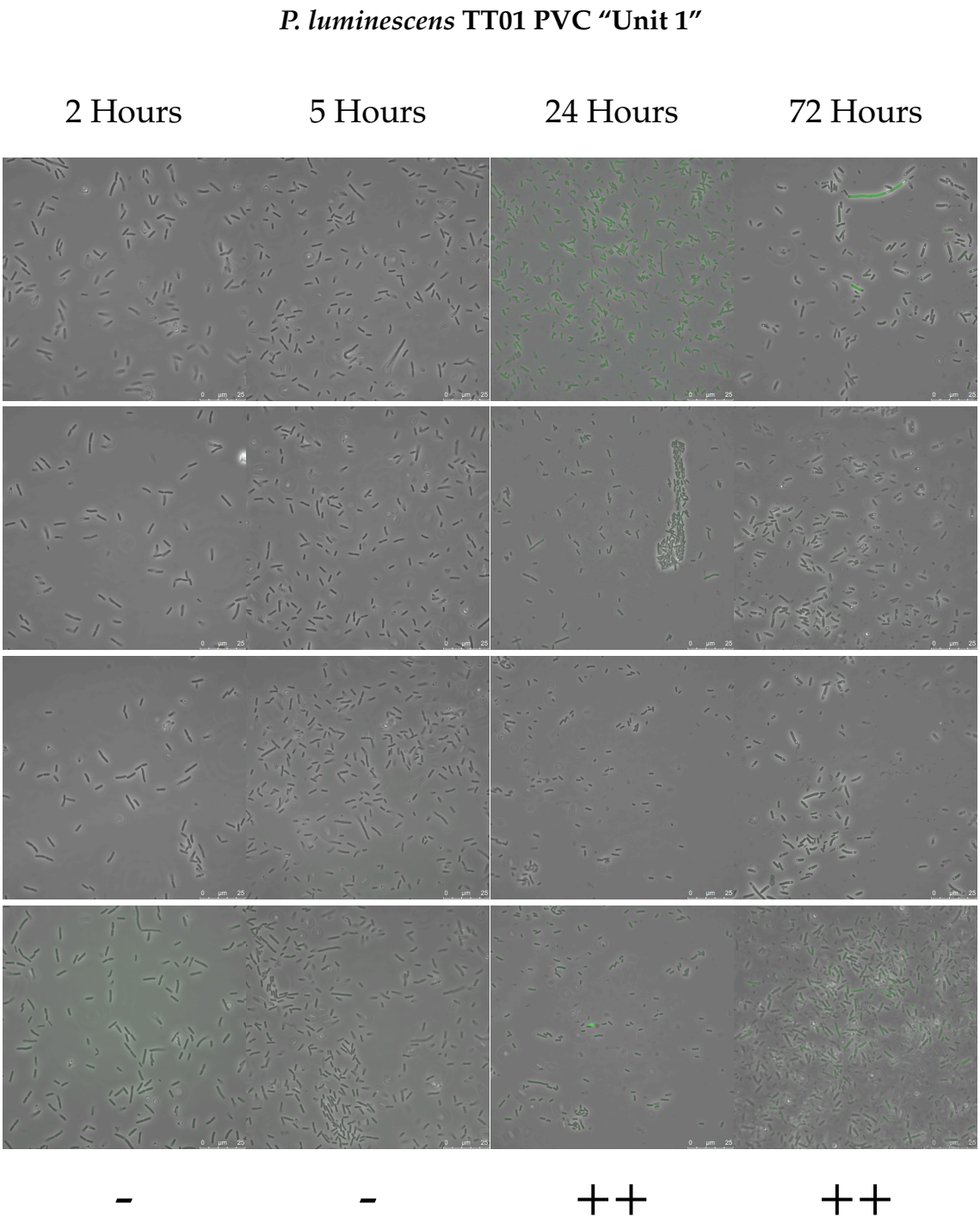


Figure 6.8 | Reporter microscopy for the *P. luminescens* TT01 "Unit 1" promoter.

A representative selection of images for 4 time points, for the PVC "Unit 1" promoter fusion. Quadruplicate images are displayed vertically as representative of the whole slide sample. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

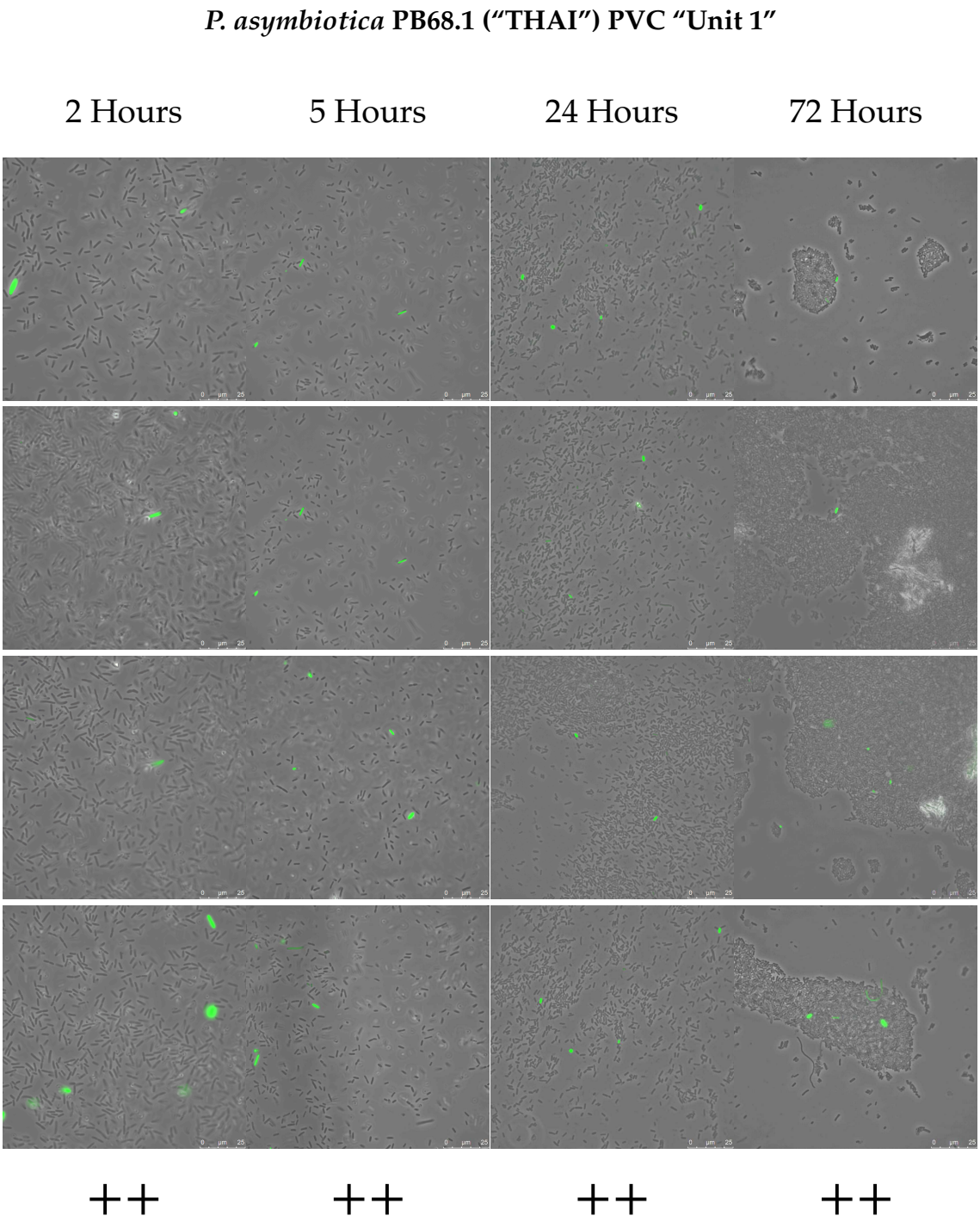


Figure 6.9 | Reporter microscopy for the *P. asymbiotica* PB68.1 "Unit 1" promoter.
A representative selection of images for 4 time points, for the PVC "Unit 1" promoter fusion. Quadruplicate images are displayed vertically as representative of the whole slide sample. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

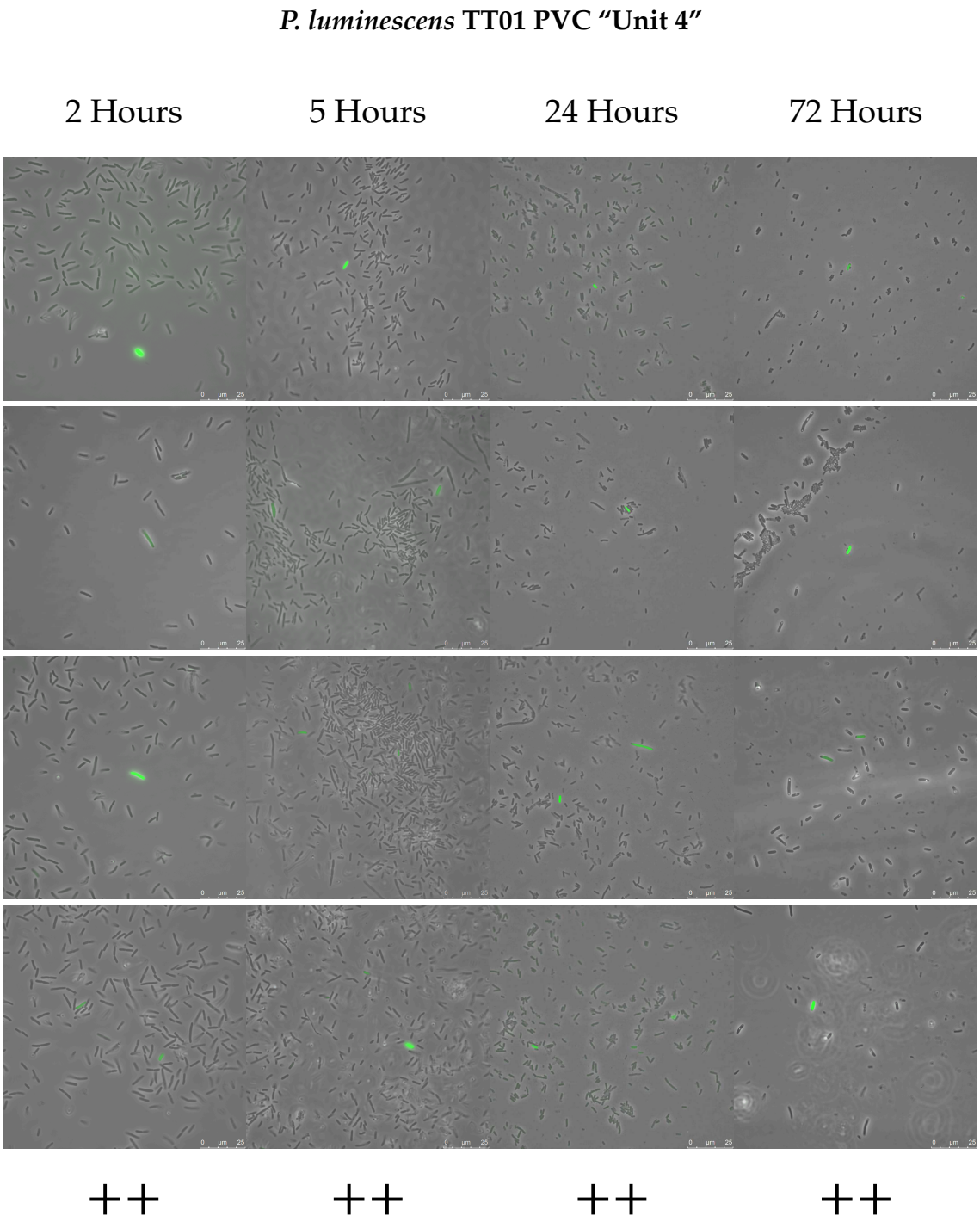


Figure 6.10 | Reporter microscopy for the *P. luminescens* TT01 "Unit 4" promoter.
A representative selection of images for 4 time points, for the PVC "Unit 4" promoter fusion. Quadruplicate images are displayed vertically as representative of the whole slide sample. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

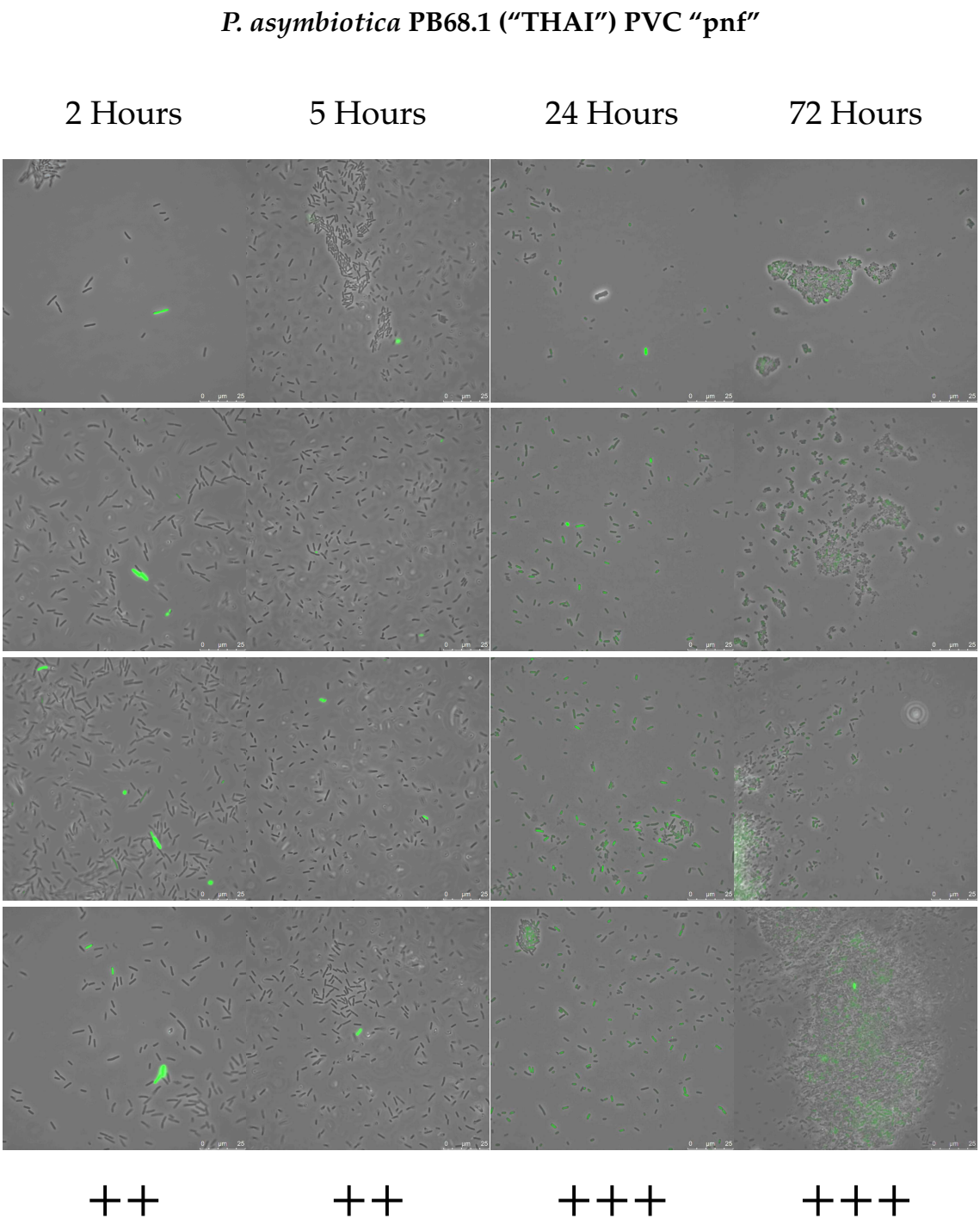


Figure 6.11 | Reporter microscopy for the *P. asymbiotica* PB68.1 "pnf" promoter.

A representative selection of images for 4 time points, for the PVC "pnf" promoter fusion. Quadruplicate images are displayed vertically as representative of the whole slide sample. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

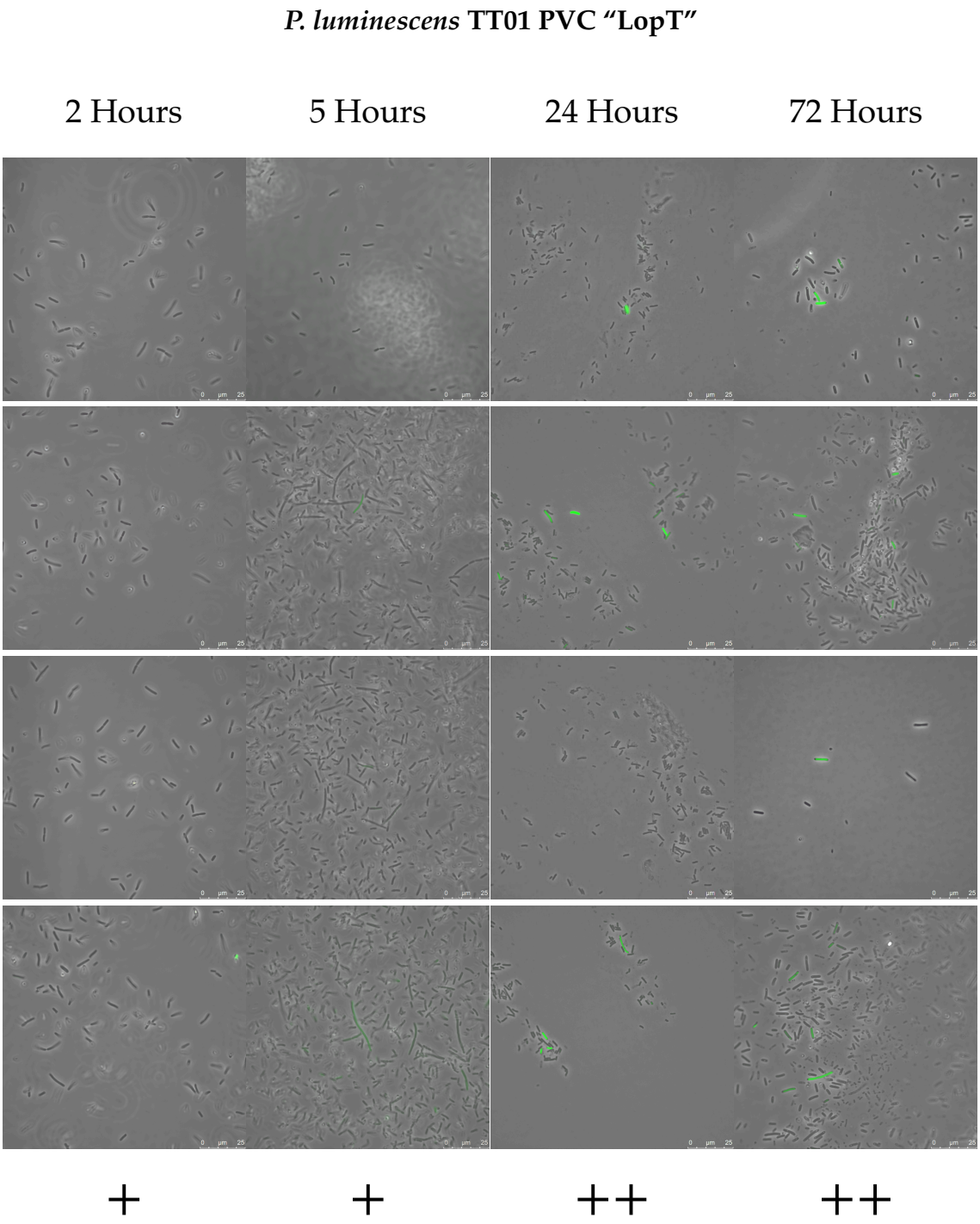


Figure 6.12 | Reporter microscopy for the *P. luminescens* TT01 "LopT" promoter.
A representative selection of images for 4 time points, for the PVC "LopT" promoter fusion. Quadruplicate images are displayed vertically as representative of the whole slide sample. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

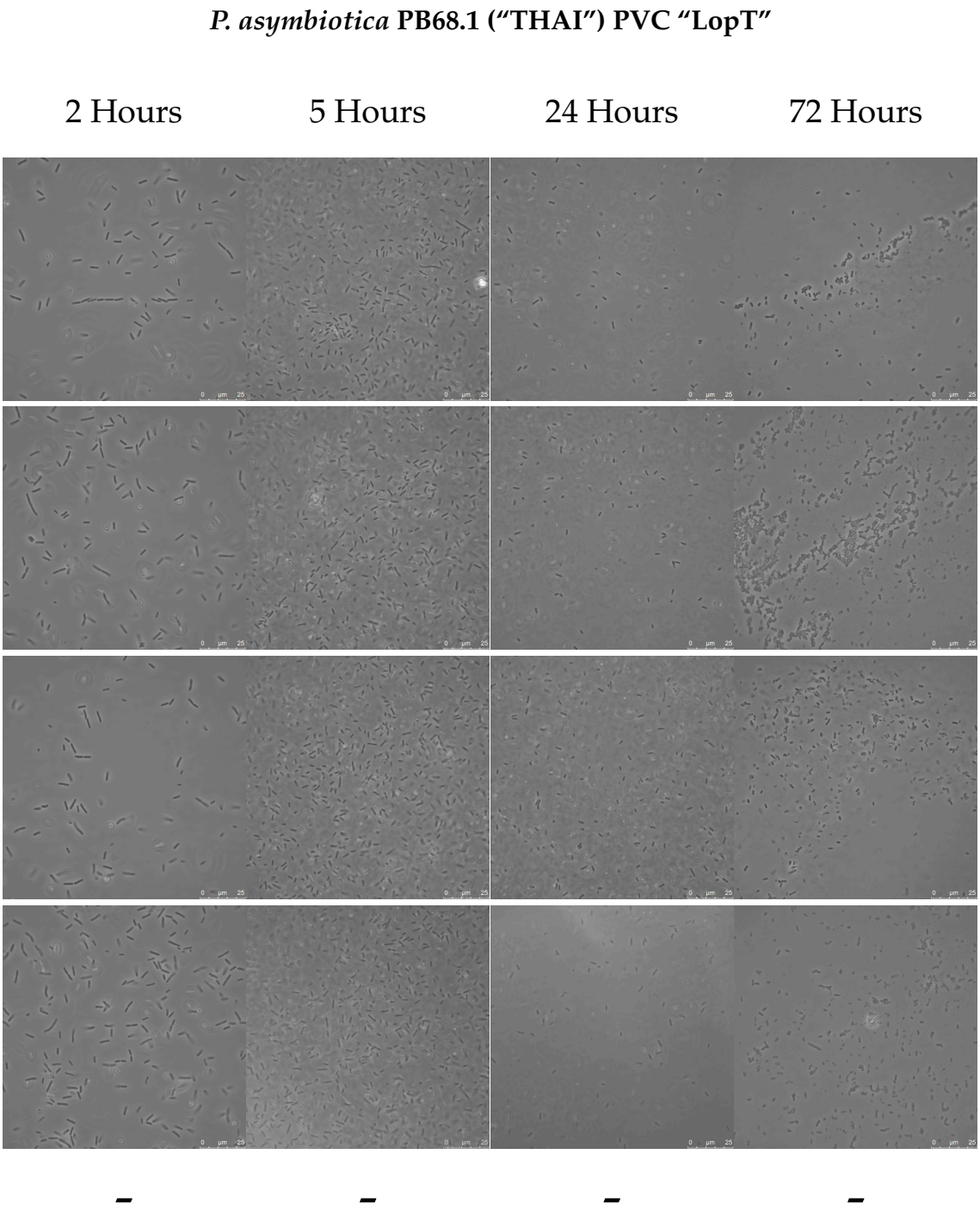


Figure 6.13 | Reporter microscopy for the *P. asymbiotica* PB68.1 "LopT" promoter.

A representative selection of images for 4 time points, for the PVC "LopT" promoter fusion. Quadruplicate images are displayed vertically as representative of the whole slide sample. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

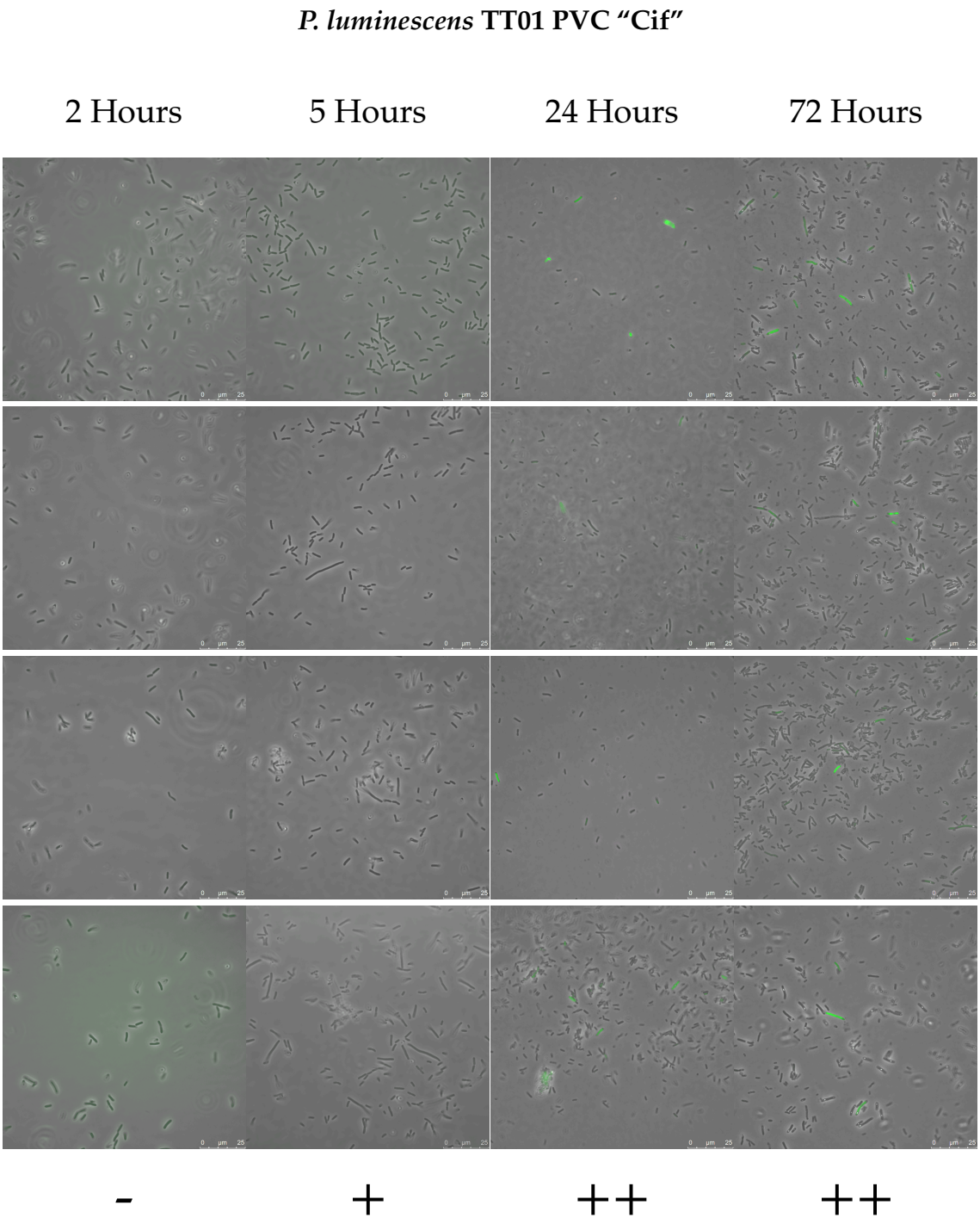


Figure 6.14 | Reporter microscopy for the *P. luminescens* TT01 "Cif" promoter.
A representative selection of images for 4 time points, for the PVC "Cif" promoter fusion. Quadruplicate images are displayed vertically as representative of the whole slide sample. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

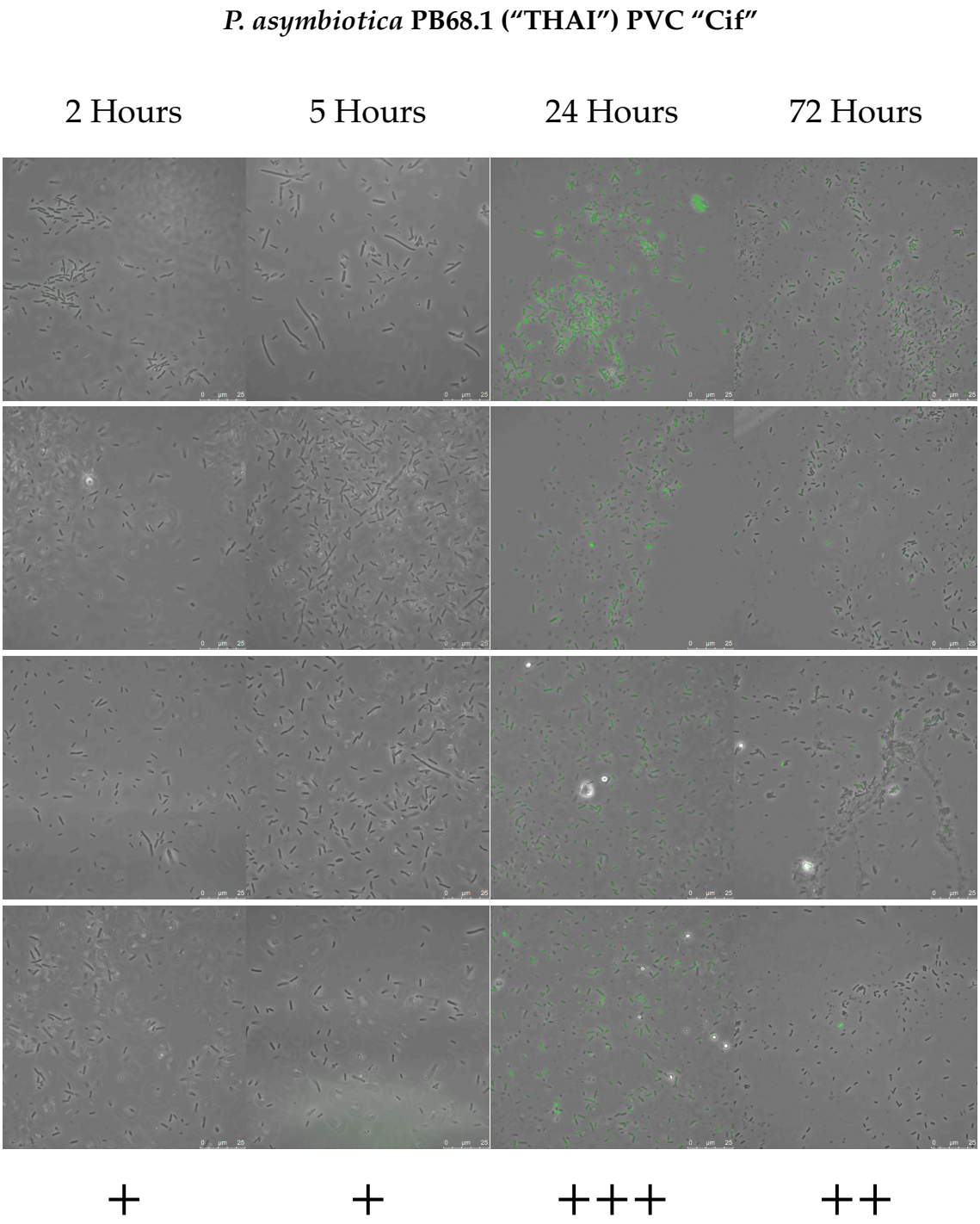


Figure 6.15 | Reporter microscopy for the *P. asymbiotica* PB68.1 "Cif" promoter.

A representative selection of images for 4 time points, for the PVC "Cif" promoter fusion. Quadruplicate images are displayed vertically as representative of the whole slide sample. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

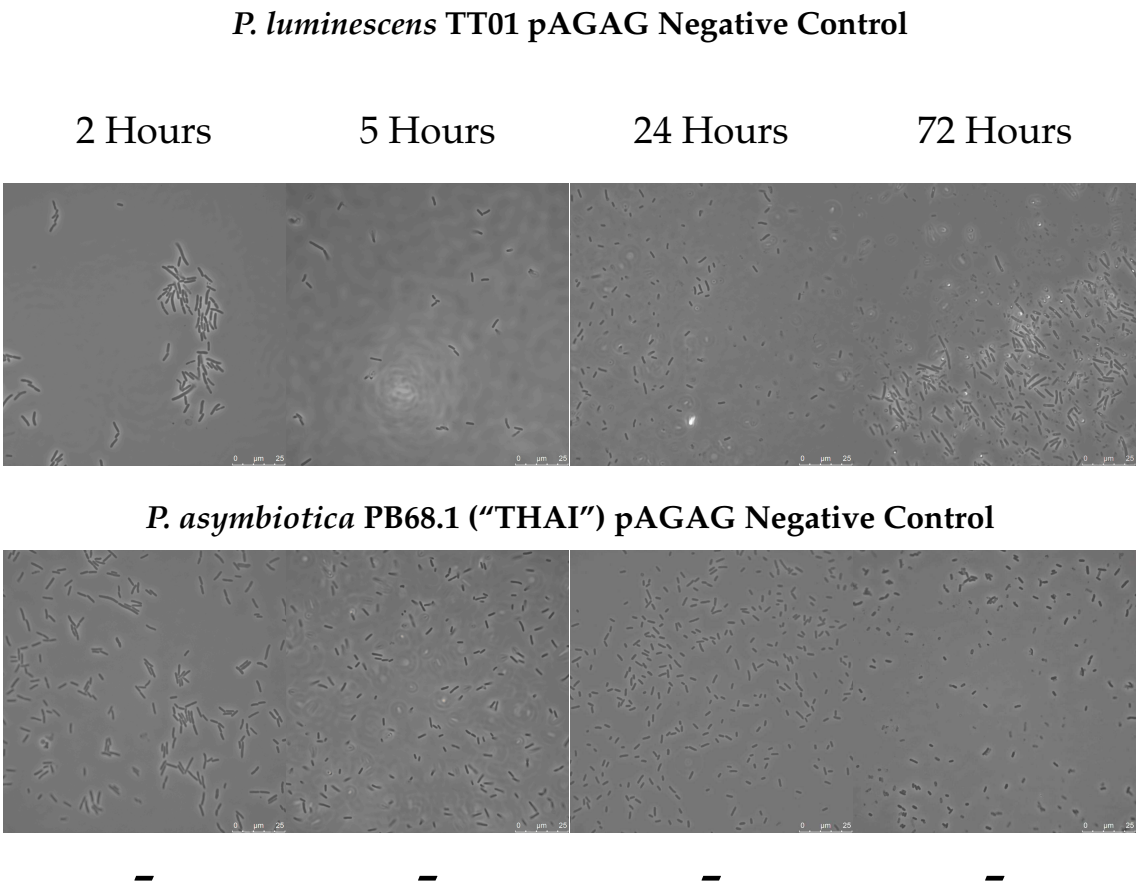


Figure 6.16 | Reporter microscopy for empty vector control plasmids.

A representative selection of images for 4 time points, for *P. luminenscens* TT01 and *P. asymbiotica* PB68.1 bearing 'empty' vectors, lacking promotors to ensure no background fluorescence or leaky expression. Key to qualitative fluorescence indication: "-" - no fluorescence, "+" - low level fluorescence in isolated cells. "++" - low level fluorescence in many cells or few brighter cells, "+++" - intermediate to high fluorescence in almost all cells, or very bright isolated cells.

The microscopy shows clearly that the manner in which different PVC operons are regulated/deployed can vary enormously, and not just between operons, but also between individual cells translating the same sequences. There is a general trend across almost all of the reporters demonstrating increased expression levels in individuals and the population at the later time points (24-72 hours), as the cultures enter stationary phase, in some cases diminishing slightly past 24 hours (Section 6.2.2 on page 251, the PB68.1 "Cif" reporter, is a good example of this).

Generally, there seem to be 2 predominant patterns of fluorescence (though they are also not mutually exclusive). On the one hand, in many of the panels, particularly in the later time points, low level fluorescence appears in most or all of the cells, for example in *P. luminescens* TT01 "Unit 1" and *P. asymbiotica* PB68.1 "Cif". On the other hand, even amongst these reporters, there are still individual cells which fluoresce noticeably brighter, revealing more of the variability in transcription/translation of the PVCs.

No expression is seen at all from the "LopT" construct of PB68.1, and very little can be seen from the equivalent *P. luminescens* reporter. This appears to be more of an exception though since it seems that there is a general trend of increased expression of most PVC operons in the *P. asymbiotica* strains, compared to *P. luminescens*. Aside from "LopT", the only PVC shared across both species is the "Cif" operon, and while it is expressed in *P. luminescens*, *P. asymbiotica* seems to be expressing it to a much greater degree, especially evident in the 24 hour time point.

In the majority of the panels it is possible to see isolated, extremely bright cells. This is most pronounced in the *P. asymbiotica* PB68.1 "Pnf" and "Unit 1" panels, where intense expression in certain cells is seen across the growth phases, including, unusually, as early as 2 hours in to measurement. This correlates nicely with the fact that in *P. asymbiotica* ATCC43949 "pnf" was the most bioactive PVC when assayed in the legacy cosmid screening, and may also go some way to explaining the lack of stability. Interestingly, the intensity of expression in most cells appears to reduce late on in the growth phases, but many more cells begin to express the operon at comparatively low level. Co-locality seems not to be a strong indicator of expression, with the degree of fluorescence varying even among very densely packed regions of cells - the bottom right panel of the PB68.1

"Pnf" reporter is an excellent example of this. The increased numbers of fluorescent cells in later time points could suggest that there may be a density dependence as the cultures become more turbid, but it is difficult to disentangle this observation since some cells appear to be behave differently to the bulk of the culture even if other cells begin to exhibit low level fluorescence.

6.2.2.1 Cellular morphology in reporter assays

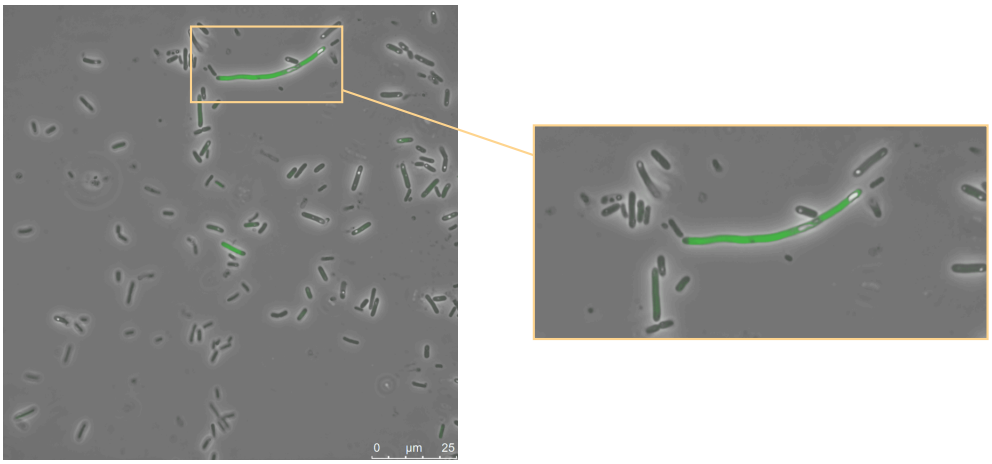
The reporter microscopy reveals a couple of additional phenotypic patterns aside from the apparent heterogeneity discussed in the previous section.

6.2.2.1.1 Cellular elongation

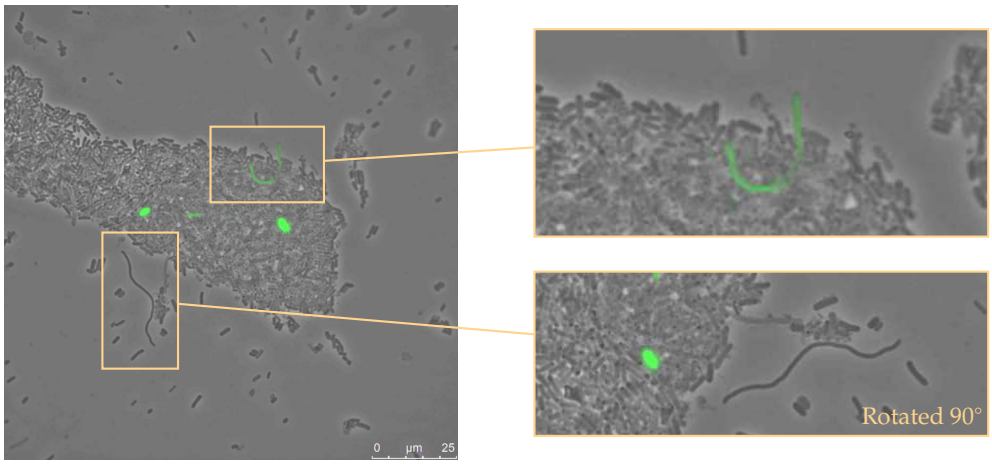
An unexpected observation from the microscopy is the appearance of a number of elongated cells in several of the panels. It is not entirely clear whether this is coincidental, however, in the control cultures, and in the case of the "LopT" construct for *P. asymbiotica* where no PVC expression is seen whatsoever, there are few if any elongated cells to speak of.

Subjectively, it seemed that the extended cell morphology correlated to some degree with GFP fluorescence, though examples of non-fluorescent elongated cells can be seen too. Figure 6.17 on the next page shows some inset magnifications of a few examples of this phenotype.

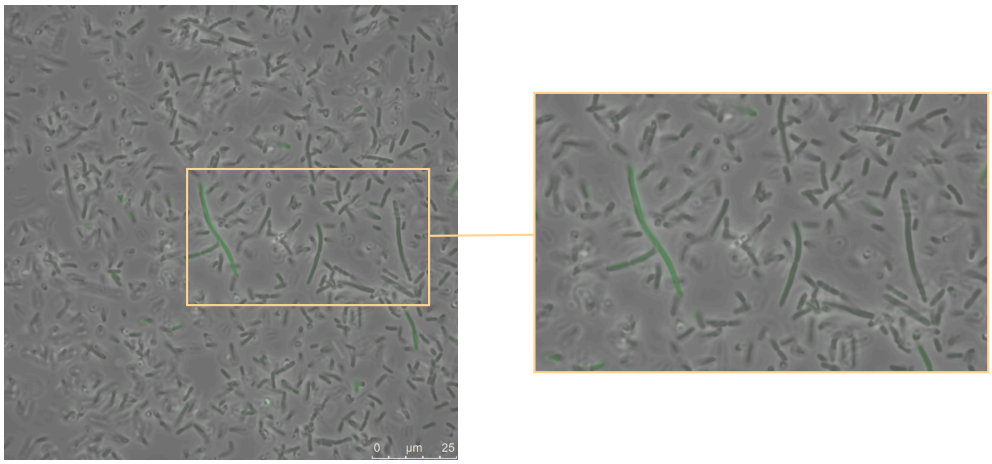
The elongated phenotype does not seem to be restricted to any one time point, however it does seem somewhat greater in the 5 hour time points. Initially it was thought that the elongation may be a general effect of stress response, however its emergence at early time points would run counter to this hypothesis. Instead, a speculative explanation for the observation may be some sort of failure to divide, which is then propagated through the life of the culture.



***P. luminescens* TT01 "Unit 1" Panel 1, 72 Hours**



***P. asymbiotica* PB68.1 "Unit 1" Panel 4, 72 Hours**



***P. luminescens* TT01 "LopT" Panel 4, 24 Hours**

Figure 6.17 | Elongated cell morphologies observed in reporter assays.
During the reporter microscopy assays, elongated cells could often be seen, and these elongated cells seemed to exhibit fluorescence in many cases. It is by no means a perfectly correlative relationship, as many elongated but non-fluorescent cells are visible in certain panels, but it did occur frequently enough so as to stand out. Those that did exhibit fluorescence and elongation tended not to be among the brightest visible cells however.

6.2.2.1.2 Culture integrity

Part of the hypothesis for the heterogeneity is attributed to the presence of a sacrificial *Photorhabdus* sub-population. It was not known, however, whether this results in lysis of these cells. Exactly how PVCs are released in to the extracellular milieu remains a mystery, but the most parsimonious hypothesis had been that the cells burst, releasing the PVCs, much like a phage would lyse an infected host bacterium.

Certainly, this was thought to be the case based on heterologous expression in *E. coli* as well, as the viability of these cultures diminished over time (discussed in Section 6.2.1 on page 240). However, from these images, no evidence of lysed cells was found. Moreover, those that express very intensely typically showed profoundly normal cellular morphology. No such reduction in viability of *Photorhabdus* has been observed, even when heterologously over-expressing certain PVC operons, though unusual pigmentation and other phenotypes indicative of stress have been noted in separate work.

This suggests that through an as-yet-unknown mechanism, *Photorhabdus* (but not *E. coli*), are somehow capable of releasing complexes the size of the PVCs in to the environment without lysing (Addison and Hapeshi, unpublished data). A rationale for the speculation in the previous section regarding cellular elongation may point to a loss of control of cell wall/membrane morphology, if they are being remodelled in order to allow the escape of the PVCs. For this to be definitive however, one would expect that the fluorescence indicating PVC expression to correlate with the unusual morphologies. Instead, the most intensely fluorescent cells, and therefore those that should theoretically be maximally producing PVCs, tended to have decidedly 'normal' cellular morphologies.

It is possible that, as the reporters are fused to just the first gene in the operon, that expression of the 5' region of the gene cluster is occurring, but not resulting in expression of the whole operon and therefore mature PVCs. There is no compelling reason to doubt that the reporters do not reflect the activity of the whole operon however, as the expression patterns (for the *P. asymbiotica* "Pnf" reporter for example) reflect the known expression activity of the operons from other data such as RNAseq and bioassays. Similarly, in separate studies, it has been demonstrated that *P. luminescens* TT01 "Unit 4" does indeed elaborate mature PVCs when expression is forced, with heigh yield and no loss in viability.

6.2.3 A putative role for transcript elongation in PVC regulation

As explored in the Introduction to this chapter, the closely related Afp tailocin discovered by Hurst *et al.* (2004) and the orthologue AfpX (Hurst *et al.*, 2018) both rely on the AnfA1 gene for expression (though it appears AfpX may harbour additional regulatory elements, as only it can be induced with mitomycin C), which is reminiscent of an RfaH or NusG-like protein, and located nearby on the pADAP plasmid (Hurst *et al.*, 2007). Given this close relationship, it was hypothesised that the PVCs may also be controlled in a similar fashion, though they are not carried on a plasmid and there is no obvious AnfA1 equivalent near any of the PVC operons identified to date.

Searching for probable orthologues within *Photorhabdus* via BLAST, with even extremely loose identity cut-offs reveals only matches to the chromosomally located RfaH protein, with no 'specialised' orthologues like that of AnfA1 identified. Moreover, without restricting BLAST results to the *Photorhabdus* genus specifically, a significant number of various NusG and RfaH orthologues match with better identity, such that the even the chromosomal RfaH identified doesn't even appear in the top 100 best matches. The next non-*Serratia* sequence to match is NusG (WP_087638535) from *Klebsiella penumoniae* at an E-value of 6.38×10^{-14} versus 4×10^{-12} for the first *Photorhabdus* sequences to match (BLASTp). The *Serratia* AnfA1 protein therefore represents a specialised NusG/RfaH-like homolog, and subject to its own selection pressures, due to its carriage on a mobile element (large plasmid).

Unusually, the PB68.1 strain has an additional *rfaH* paralogue further downstream on the same strand, associated with a number of pilus encoding genes. On further investigation, this additional gene copy appears to only be present in *P. asymbiotica* strains with an Eastern/Asian origin, being present in isolates from Nepal ("Nepal") and Australia ("Beaudesert" and "Kingscliff"), but not in the USA associated strains (ATCC43949/ATCC43951) or European isolates ("HIT" or "JUN"). This presumably points to a relatively recent acquisition of these pilus genes in the Asian strains which has also brought its own regulator along with it.

As the alignment in Figure 6.18 on page 267 shows, across 16 PVC operons, 4 positions are identical, with a further 2 differing only in 2 or 3 particular operons, and the consensus

sequence has a Hamming distance of 6 to the prototypical *E. coli* sequence.

If these few static sites are sufficient for RfaH (or a similar protein) to bind and act, it may be the case that the single identified copy of the protein is indeed responsible for regulating the PVCs, as it would for any other long operon.

However, a diversified operator sequence would suggest a diversified cognate binding protein, so as to ensure the binding partnership is maintained. The putative *ops* sequence for the PVCs does not adhere to what is canonically considered the conserved sequence with degeneracy in some conserved sites, whilst the highly similar *Serratia* Afp site does (with a sequence of `5'-GGCGGTAGCATG-3'`, only varying in the 2 degenerate N sites shared by *E. coli* (underlined)). Thus, one plausible explanation is that *Photorhabdus* is using an even more diverse orthologue of a NusG/RfaH-like protein, and the efforts to identify the *Photorhabdus* equivalent of AnfA1 simply is not sensitive enough. Figure 6.18 on the following page shows a multiple sequence alignment of multiple permutations of the RfaH *ops* site (with an N at every position) aligned against ≈500 bp of sequence upstream of the first PVC CDS, from 16 operons. Thus, the sequence is sufficiently conserved that it captures the *ops* site with various redundancy, not resulting in alignment to elsewhere in the sequence. Despite the variability in the *Photorhabdus* equivalent site, the canonical *ops* site can be aligned, highlighting its position in 16 PVC operons.

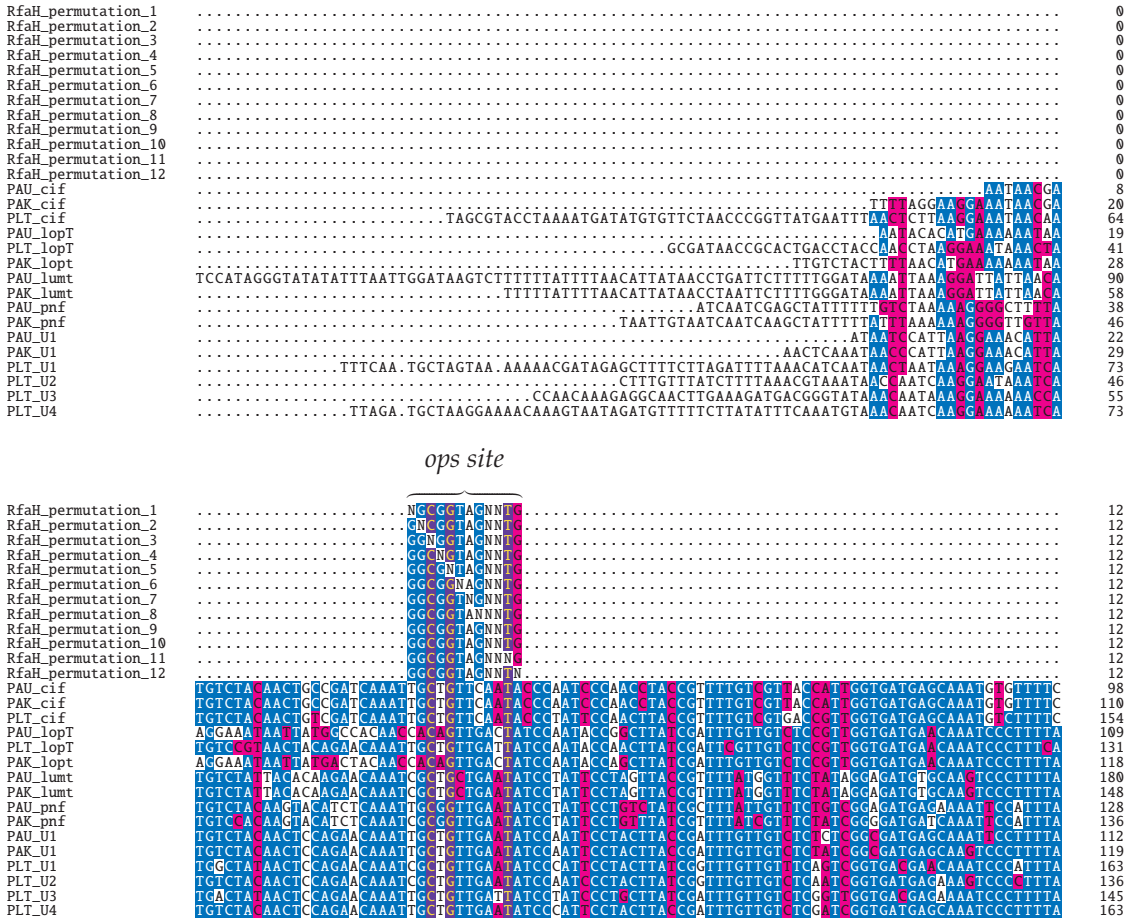


Figure 6.18 | Identifying RfaH *ops* sites via profile MSA.

A sequence alignment of ≈500 bases upstream of the first locus of the PVC operons (only ≈200 bases reproduced here), with redundant permutations of the RfaH *ops* binding site profile aligned. 5′ regions of the PVCs were first aligned together, then permutations of the canonical *E. coli* RfaH binding site were generated and aligned together against the existing MSA to identify the binding site and its conservation (or lack thereof). Visualised here via the `TexShade` LaTeXpackage, identical residues are yellow-on-purple, residues conserved at over 80 % are white-on-blue, residues conserved at over 30 % are black-on-pink, and uncoloured residues are less conserved than this cutoff.

As a final observation, contiguous alignment of the prototypical JUMPStart motif discussed in the Introduction to this chapter was not observed within the region around the *ops* site, suggesting that the PVCs do not utilise this extended regulatory structure. Nor, when selecting a 20 base pair span immediately upstream of this *ops* site, could any compelling RNA hairpins be detected. Rich secondary structure is predicted downstream of the *ops* site (not shown), suggesting terminators are present to prevent transcription of the operons in the absence of RfaH-like proteins. Interestingly, for the Afp however, an almost perfect JUMPStart motif was found, which does not appear to have been previously reported. Figure 6.20 on the next page shows the results of Vienna-RNA's RNAfold package, along with their folding energies. RNAs are shown for 16 PVC operons, as well as the Afp. The 3′ *ops* sequence is highlighted in blue, and an example of the canonical JUMPStart motif is reproduced from Wikimedia in Figure 6.19. In all cases for the regions upstream of the putative *ops* sites of PVCs, they have less favourable folding energies than does the Afp site. Perhaps the most striking indication of all, is that of the PVClumt operons (panels H and I) form absolutely no secondary structure whatsoever. Of the PVC sequences, most of them sequester part of the *ops* site within a hair pin structure, which is not the case for the canonical JUMPStart sequence, nor for the Afp. The Afp does produce a small stem in the ops site at the site of 2 G-C pairs interacting, but in subsequent analysis, these pairs were found to have high structural entropy, and are therefore less favourable, meaning that it would be simple to restore the proper JUMPStart architecture. Moreover, the Afp sequence not only follows the structural characteristics of the JUMPStart sequence, it largely obeys the redundancy rules laid out in Figure 6.19, reinforcing its role as a JUMPStart sequence, not just an *ops*.

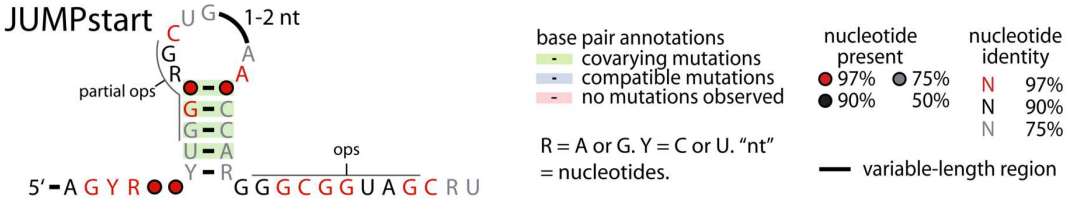


Figure 6.19 | The canonical sequence and structure of a JUMPStart motif

A diagram of the canonical JUMPStart motif, adapted/reproduced from Wikimedia (`https://en.wikipedia.org/wiki/JUMPStart_RNA_motif`). The *ops* site is clearly visible at the 3′ end of the region. The hairpin structure of the motif is generally well conserved, though can differ in length by around 6 nucleotides, and has a pseudo-degenerate sequence. Certain positions are free to vary, whilst others are not. In a few positions, there is some degeneracy, permitting just 2 of the 4 bases for example.

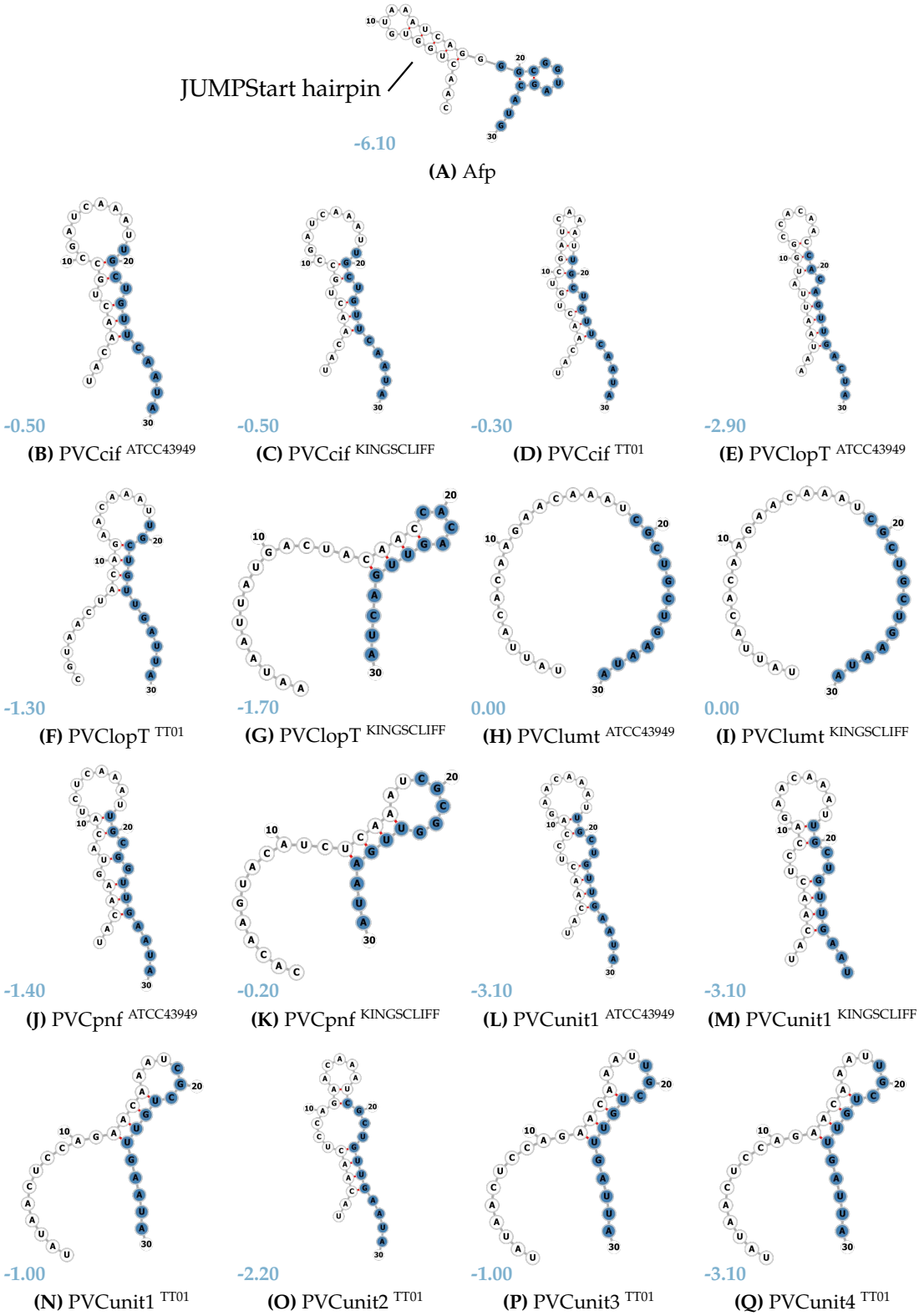


Figure 6.20 | Absence of JUMPStart motifs in PVC promoters.

A comparison of sites corresponding to JUMPStart motifs in PVCs and the Afp. The Afp has a compelling JUMPStart sequence, not just the *ops* sequence which has been found previously. The PVCs however, do not appear to contain such a structure. The *ops* site is depicted in blue. Structures are labelled with their folding free energies in kcal mol^{-1}.

6.2.3.1 Controlling RfaH activity

With only one candidate for an AnfA1 orthologue detected within *Photorhabdus* it was decided to try and create over-expressable constructs of the protein, to study whether a greater number of the population could be encouraged to express the operon through direct induction. This was pursued in a couple of different ways.

Turning once again to the Keio collection, it was noted that a Δ*rfaH* strain was available (BW25113 strain JW3818[1])(Baba *et al.*, 2006). This Keio strain was fortunately available from a neighbouring lab, and was secured for assaying. The reporter constructs produced for use in Section 6.2.2 on page 250 were transformed in to the deletion JW3818 strain.

Upon assaying the deletion mutant of *E. coli* however, it was observed microscopically that the BW25113 strain (both wild-type and mutant), despite being a K-12 lineage *E. coli*, had unusual cellular morphologies. Rather than the typical length rod shape that is expected, a morphology where a central 'waist' appeared constricted was observed (making the cells appear to have a semi-coccoid, figure-of-eight shape), and the cells appeared shorter than expected. It was decided that this unusual morphology may interfere with any further study, possibly due to BW25113 being a 'wilder' strain, and this experimental thread was abandoned. There is also the additional consideration that there may be further players involved in the regulation from elsewhere in the *Photorhabdus* genome involved, and *E. coli* data might not be entirely meaningful.

Having abandoned the *E. coli* approach, an alternative strategy with *Photorhabdus* was devised. A recombineering approach and the use of a suicide plasmid were attempted in order to introduce an inducible promotor upstream of the *rfaH* locus in the TT01 and PB68.1 strains of *P. luminescens* and *P. asymbiotica*, respectively. Due to *P. asymbiotica* PB68.1's paralogy in *rfaH*, only the gene located in the equivalent site as the TT01 orthologue was targeted. It is unknown what effect a secondary gene copy might have at this stage.

Photorhabdus can be recalcitrant when cloning at the best of times, but in other efforts the PB68.1 strain has proven the most tractable *P. asymbiotica* strain. The suicide plasmid approach which has been effective in other work, was designed to incorporate an inducible

[1] `https://cgsc2.biology.yale.edu/Strain.php?ID=109043`

promoter in front of the gene. Ideally, a *P. asymbiotica* strain such as ATCC43949, which is the more prototypical *P. asymbiotica* strain used in the lab (and has only 1 copy of *rfaH*), could have been used to avoid the problem of paralogy, however there have never been successful attempts to engineer it. The suicide plasmid was conjugated in to *Photorhabdus* via the *E. coli* propagation strain, S17 *λpir*, as per standard protocols.

Since it was unknown whether the region of the genome would be tractable or not, recombineering was also attempted as an alternative. Recent papers have shown that recombineering is feasible in *Photorhabdus*, and endogenous orthologus of the exo, Beta, and Gam proteins have also been found (Yin *et al.*, 2015).

Despite all positive indications from previous literature attesting to the potential non-essentiality of RfaH, and former successes in engineering *Photorhabdus* however, it was not possible to recover any successfully recombined colonies and thus it was not possible to further this line of inquiry. As was observed with the initial attempts at recombineering the *att* site, discussed in Section 6.2.1.1 on page 242, it appears that some regions within the chromosome are simply too recalcitrant. In *Photorhabdus* it may be the case that this RfaH locus is too tightly controlled due to its involvement with other genetic pathways, and that any attempt to interfere with it resulted in non-viable cells. The locus is also quite densely populated with genes on both strands, as can be seen in Figure 6.21, which shows the *P. luminescens* TT01 chromosomal region for the attempted engineering. The 5′ region of RfaH is only ≈600 bp (highlighted in cyan in the figure), and likely contains promoters for the *ubiD* gene on the opposite strand. Thus, there is potentially a high degree of transcriptional activity and secondary structure in the region which could prevent efficient recombination.

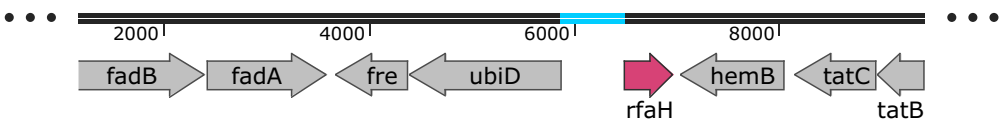


Figure 6.21 | The chromosomal locale of *rfaH* in *P. luminescens* TT01

The chromosomal locale of the *rfaH* in *P. luminescens* TT01, demonstrating the narrow upstream region targeted, unsuccessfully, for engineering (cyan). *P. asymbiotica* PB68.1 has an equivalent locus, and an additional paralogue elsewhere in the genome, closely associated with a pilus, only found in certain *Photorhabdus* strains.

6.3 Discussion

This chapter has attempted to get a handle on the manner in which PVCs are deployed naturally, and devise new ways of heterologously cloning these operons, such that they could be synthetically expressed at will in future. Being able to express a large, multipartite, biological structure such as the PVCs will be crucial to their continued study, but there are practical considerations to be made that are not usually a concern for more 'standard' protein expression.

6.3.1 Heterologous expression and control of PVC operons

Heterologous expression of PVCs presents some slightly unique challenges. Unlike phage, which can be easily recovered from plaque assays and infected cultures (though do have their own set of considerations such as identifying a infectable host), the synthesis of PVCs is being directed by the cell, rather than 'its own' DNA. Both the R-type pyocin and Afp have the advantage that purification is made much more straight-forward by virtue of there only being single examples of the structures in any one genome. Thus, the host organism can be used for production and any genetic manipulations, with confidence in the 'final product'. In the case of *Photorhabdus*, one cannot be entirely certain that, when isolating PVCs, the product of a single operon is obtained, and not a mix from the 4-5 that are present elsewhere in the genome, even if only expressed at low levels. The paralogy this results in also means that, without making up to 6 simultaneous modifications/mutations, it is also impossible to rule out conserved elements of different PVCs naturally 'transcomplementing' one another. By way of example, for just the inner sheath proteins, there may be as many as 12 paralogues within a single genome, contributed from other PVCs - and this neglects to include any prophages and similar proteins which might be able to complement the structure. In the case of the Afp operons, they were also readily available on natural plasmids, making them considerably more genetically tractable, as they could just be subcloned as a single restriction fragment after miniprep.

Consequently, moving out of the host organism, and in to an expression strain such as a laboratory *E. coli* is desirable. Broadly speaking there are two approaches to cloning long operons such as these: 'bottom up' - cloning small pieces, and stitching them together,

or 'top down' - capturing intact long segments of DNA and then engineering on to these constructs as necessary.

In this study, both of these approaches were implemented, utilising techniques that have not previously been brought to bear on the PVCs - namely, recombineering and Gibson assembly. Unfortunately, both approaches proved difficult to implement to completion, though partial successes were achieved in certain cases.

Firstly, the 'top-down' recombineering approach initially showed promise, as the technique was able to be successfully optimised in the lab, and a number of chromosomal knockouts in *E. coli* DH10β were achieved. However, it later became apparent that recombineering non-chromosomal targets introduces another layer of complexity. There is literature precedent for recombineering multicopy replicons and the additional complexities associated have been well reported. λRed recombineering originally found most utility for engineering of Bacterial Artificial Chromosomes (BACs) which are also single copy. In numerous cases, though correct engineering of the plasmids has been achieved, the resulting pool of transformants contains a mixture of parental, non-engineered plasmids, correct forms, as well as multimers of aberrantly joined plasmids (Thomason *et al.*, 2007; Lee *et al.*, 2001; Yosef *et al.*, 2004; Vetcher *et al.*, 2005).

In the cases where higher efficiency techniques are reported, there are often additional factors at play, which do not apply to the engineering of existing cosmids. For example in Yosef *et al.* (2004), the introduced cassettes relied on the re-introduction of a disrupted promoter, to restore a resistance marker on the targeted plasmid - this is obviously predicated on the target replicon having the prerequisite genetics to enable these efficiency-enhancing methods.

By the time that the recombineering protocol had been applied to the cosmids, a significant amount of time had been invested. The only means by which correctly engineered cosmids could be retrieved from the mix of multimeric forms seemed to be either to fortuitously identify single colonies which had only retained a correctly engineered cosmid, but with potentially low efficiency this could require screening of large numbers of transformants. Or alternatively, to liberate multimerised cosmids by restriction with single-cutting enzymes, followed by electrophoresis and gel extraction as reported in the

papers previously cited. This presented further challenges however, as the cosmids are significantly larger than the plasmids which had been engineered in the other studies. This means finding reliable single cutting enzymes is made difficult, not least because there are also no complete sequences of the cosmid inserts and their backbones (end sequencing of the inserts only). Furthermore, electrophoresis and extraction of large fragments can be problematic, resulting in low DNA yields, which would subsequently require re-ligation and transformation. All in all, it became apparent the process was going to become decreasingly robust and it was not worth investing more time.

As a random genome capture technique, the cosmids also often harboured additional upstream and/or downstream sequences, including whole genes and their associated promoter regions which could well present issues for further downstream cloning and activity assays. On top of this, the vector backbones for the cosmids already contained at least one, and sometimes up to 3 resistance markers, left over from the initial selection of the library (ampicillin, kanamycin and neomycin, or chloramphenicol). Since any subsequent edits to the cosmids would also require the introduction of another selectable marker in order to be able to screen for successful recombinants, this limited the scope of sequences and resistance markers that could be used, as well as significantly reducing the options for secondary plasmids which could be used in concert with the cosmids (e.g. in future trans-complementation studies).

This left the 'bottom up' approach. As discussed, avoiding restriction based cloning seemed wise, such that what could and could not be cloned was not reliant on the sequence of the PVCs specifically. Constructs were obtained in various states of completeness. It was common to obtain constructs with just a single fragment from the four that were used in the construction, belying the reduced efficiency of incorporation of increasing numbers of fragments, especially at this length. A useful observation was that the use of long primers (70 bp) required for introducing sufficient overlaps had the advantage of PCR-ing with minimal optimisation when used in concert with the NEB Q5 enzyme, allowing fragments of 7-8 kb (and in one case 12 kb) to be PCR'd simply.

Based on this partial success, the 'bottom up' approach seems like the best to pursue in future. With further optimisation and experimentation with other vectors it may be

possible to finally clone the entire region in a sequence independent manner. Some further validation of this approach can be inferred from mixed success that a collaborator has had in producing a PVC operon in the pBAD30 vector, using sequential classical cloning. Though initial attempts had failed due to PCR and assembly errors, construction of the PVCpnf operon has been achieved in at least two cases since.

The optimisations made here to the Gibson process look to be promising. The length of overlaps allowed for fragment joining (35 bp), and the size of the fragments are not prohibitive, as evidenced by obtaining multiple clones where fragment 1 (the longest of all four) was successfully incorporated (just one example is shown in Figure 6.6 on page 249. Additionally, fragments 2 and 4 contain a number of features which are either not well understood, or looked likely to cause cloning issues - for instance, a putative GGDEF-domain protein which binds cyclic di-GMP and is a ubiquitous bacterial regulator (Paul *et al.*, 2004), a transposase, "Gcv operon activator", and the SepC-like toxin. Consequently, it was a pleasant surprise to see that these regions appeared to clone with relative ease, and no detectable adverse effects.

The most successful constructs obtained in this manner incorporated some ≈12 kb of PVC sequence in two of the four total fragments. Interestingly, the consistently missing fragment 3 contains the ClpV-like ATPase, which has been shown for the Afp and AfpX to be essential to assembly/activity (Rybakova *et al.*, 2015; Hurst *et al.*, 2018). This observation also matches with similar patterns that were seen in the original cosmid libraries. It was common to find large 3' regions of the operon deleted, which encompassed the ATPase, suggesting that it at least partially contributed to the toxicity that was apparent in the clones. An attractive hypothesis, therefore, is that cloning the 3' end of the operon with the ATPase is the 'straw that broke the camel's back', and inclusion of this final piece of the puzzle leads to production of the PVCs; effectively restoring the toxic effect that is seen in the fully intact cosmids which prompted all these efforts in the first place. The use of an *araBAD* promoter, despite repression with glucose, may be insufficient. Firstly, the promoter is a significant distance from the ATPase site, and it is unknown whether operon interior architecture might contribute to expression of 3' genes; and secondly, the *araBAD* promoter system, despite being a drastic improvement over the

lac for example, is still capable of leaky expression, as arabinose can be imported by the cells can creates a binary expression response. The population of expressing cells is concentration dependent, but the level of expression within any individual cell is not (Siegele and Hu, 1997; Khlebnikov *et al.*, 2000). It is possible that this contributes to the toxicity, and in future it may be advantageous to use a fully titratable promoter system such as pRHA (rhamnose inducible (Giacalone *et al.*, 2006)), such that a greater number of cells produce low levels of PVCs, which might reduce the toxicity to an individual cell.

The consistently missing fragment 3 (*pvc*13-15) was designed such that it encompassed the putative tail fibres, ClpV-like AAA+ ATPase, and *pvc*14 which is proposed to be the tape measure protein. The results from Chapter 5 on page 187 demonstrate that the tail fibres can be cloned (including the tail fibre from the Pnf operon specifically), and over-expressed without detriment to the culture. While there is no current experimental evidence for PVC14, it seems likely that a putatively structural protein is unlikely to cause issues in cloning - leaving the ATPase as a likely culprit. Rybakova *et al.* (2015) and Hurst *et al.* (2018) demonstrated that the ATPase is absolutely essential for assembly and activity of the Afp/AfpX in *Serratia* and *E. coli*.

An alternative rationale for the ATPase toxicity mentioned above comes from the fact that the orthology that the gene attracts is to ATPases characteristically known for their diverse roles within a cell. It is probable, therefore, that expression of this gene could result in aberrant behaviour of any number of other cellular processes. Moreover, in separate cloning efforts, *pvc*1-5, *pvc*11 and *sepC* have all been shown to be cloneable without detriment to the cells. As a further attempted optimisation, fragments 2 and 3 were PCR'd together as a single fragment of ≈12 kb, in the hope that a reduced number of fragments might have increased efficiency. No clones could be obtained which harboured this fused fragment, suggesting that the inclusion of the ATPase within this fragment reduced the efficiency of cloning fragment 2 based genes significantly.

6.3.2 Understanding the natural regulation and deployment of PVCs

Miscroscopy unambiguously confirms population heterogeneity in the cultures, but presents conflicting theories about the manner in which the PVCs are deployed with little categorical agreement between expression activity and assorted cellular morpholo-

gies.

Taking an ensemble view of all the reporters, it seems to be the case that late in the growth cycle an increased number of cells begin to fluoresce, particularly at a lower level. It is tempting therefore to connect this to the stringent response, as nutrients become limiting and there is perhaps a pressure to become more virulent in an effort to kill additional prey organisms or fend off saprophytes which may make sense when the natural environment of *Photorhabdus* is considered: it may be attempting to kill and bioconvert organisms invading the carcass of recently killed insects, to then use as a new source of nutrients. As the PVCs are virulence factors, their expression is likely intrinsically linked to the stress response of the bacteria (Dalebroux *et al.*, 2010; Chatnaparat *et al.*, 2015). For example, the Type III secretion system has been shown to be controlled by the alarmone ppGpp (Ancona *et al.*, 2015). While a hallmark of the stringent response is often a redirection of resources toward biosynthetic gene clusters in an attempt to mitigate starvation, it seems that *Photorhabdus* perhaps opts to increasing its virulence factor production. Regardless, it seems that this cannot account for all of the expression seen in the reporters as a number of them exhibit high levels of expression in just a handful of cells right from the very beginning of the growth cycle.

A rationale for the population heterogeneity observed in *Photorhabdus* cultures may also be linked to this unusual life cycle. Since only a small number of bacteria ever re-associate with the nematode in a 'wild' infection, a significant proportion of the population are sacrifical. This may take the form of as food for the nematodes, or simply as factories of small molecules and enzymes associated with the protection of the cadaver. In the case of the PVCs, given their size, there is likely to be a lysis mechanism (and some operons encode lysins etc.), which results in the death of the cell in order to release such a cocktail of virulence factors etc.

The observation was made that there are many elongated cells observed in the cultures, and though they are often fluorescent, it is by no means a perfect correlation, with examples of elongated non-fluorescent cells also readily apparent. In fact, to complicate matters further, many of the brightest cells appear to have decidedly normal cellular morphologies. This is unexpected, as, at least naively, one might expect that a cell which

is expressing a large number of a sizeable macromolecular complex like a PVC might be 'fit to burst' and that this might be reflected in the cellular morphology. In phage lytic lifecycles, strategies for escape include the expression of lytic enzymes within the host, and they are known to have an elongating or remodelling effect (Young *et al.*, 2000). In the case of the PVCs, certain operons are found to have lysozyme and bacteriophage lysis proteins proximal. In the case of the troublesome pnf operon from *P. asymbiotica* ATCC43949, a lysozyme 'rrrD' and 'bacteriophage lysis protein' (according to genome annotations) are found just beyond the payload region. This may well account for their impact on culture viability in the original *E. coli* cosmid library. Though, as is so often the case with the PVCs and *Photorhabdus* biology, this too, is not a hard-and-fast rule, with many examples of operons that do not carry *cis*-encoded lysins. It is possible that the lytic enzymes may not need to be located *cis* to the operon in order to provide the required effect however. In short, this implies that somehow *Photorhabdus* is capable of releasing PVCs in to the environment, and that they are effectively 'secreted', and not lysed, as was initially suspected. Whether or not the cells which are producing the bulk of the PVCs in culture remain viable, even if they are not lysed, is unknown. It may be the case that these cells still represent a sacrificial population, but not in a lytic manner.

Turning to the putative role of RfaH-like in PVC regulation, *Photorhabdus*' unusual biology means that identifying a region which is even remotely consistent with a canonical *ops* site, is very much a 'smoking gun'. With the *Serratia* Afp as the 'nearest cousin' to the PVCs, and the role for the RfaH/NusG orthologue, AnfA1, established, it seems probably that a similar antitermination mechanism will be in place to enhance the expression of the long operons. Despite being its nearest cousin however, the differences between the PVCs and Afps are not insignificant. For one, the presence of only a single Afp locus on a mobile element would perhaps indicate that the Afp hasn't been evolving with the *Serratia* genome for as long as the PVCs may have been with the *Photorhabdus* one, and consequently it also needs to bring its regulatory mechanisms with it on the mobile element. With the PVCs, they are purely chromosomally located (as far is known to date), and highly paralogous which would imply that substantially difference selection pressures and mechanisms could be at play. The results from Chapter 4 on page 152

suggest that the PVCs are more ancestral and less mobile than originally expected, and by being chromosomal, it may be the case that the PVCs regulation is more tightly 'interwoven' in to the chromosome also. This may account for why a dedicated RfaH-like orthologue, similar to AnfA1 has not been found - the chromosomal RfaH may be capable of filling this role. What's more, the sequence analysis conducted in this chapter shows that, in fact, the Afp actually may utilise the lesser known JUMPStart sequence, whereas the PVCs do not. To add to this, the AfpX has been shown to respond to mitomycin C induction, whereas the original Afp does not, and AfpX also has a subtly different target range (Hurst *et al.*, 2018).This shows that even amongst two of the most closely related examples of caudate structures, there are exquisite differences in the control and effect of the complexes, and ultimately how they have been evolving, it is perhaps no surprise therefore, that the PVCs exhibit differences in how they're regulated that are also very subtle, but potentially significant.

6.3.3 Summary and future work

Much of the work presented in this chapter is preliminary, and sets the stage for quite a number of future experiments. Though much further study will be needed to conclude the roles of, for instance, RfaH and the *ops* sequence, the subtleties of PVC promoter differences, and the potential activity and toxicity of the ATPase, this chapter identified these factors and issues associated with their study. Without a good understanding of the regulation and a reliable way of heterologously expressing the PVCs, work on them will proceed considerably slower.

The evident population heterogeneity in the reporter constructs is possibly the most significant result presented here. It had been hypothesised for some time that the PVCs might be released by lysis from a small subset of the culture - a sacrificial population. While no evidence of lysis was found, it does appear to be the case that a subpopulation of cells do the lion's share of the PVC production.

6.3.3.1 PVC cloning

Overall, the attempts to robustly clone the PVCs in a restriction-free manner were ultimately unsuccessful, though not without some promise. It seems that constructs of

this size are somewhat amenable to assembly in this manner. Toxicity of the constructs may have played a part in preventing obtaining the full structure rather than it being a particular limitation of the assembly process/efficiency. If the project was to be re-done with the benefit of this hindsight, a promising alternative mechanism might be to do much of the assembly in yeast. This is the process that was followed in the early stages of the synthetic genome project, whereby short fragments were continually recombined in to progressively larger components before being transplated to *E. coli*, and it is generally considered a 'go-to' for homologous recombination-based assemblages (Kuijpers *et al.*, 2013). It is also possible that the gene products which are thought to be potentially toxic to the *E. coli* may not exert such an effect in a yeast, since the cellular structure and biochemical pathways are somewhat 'orthogonal' - though whether this would lead to additional, different, toxicities due to PVC's typical anti-eukaryotic activity is unknown.

The widespread literature regarding Gibson assembly attests to its utility for cloning large operons in a single reaction using overlapping component oligonucleotides. With further optimisation of the overlapping sequence lengths, molar fragment ratios, and enzyme mix, Gibson assembly still seems to be a robust approach to cloning, notwithstanding any toxic effects of the subcloned regions. Other labs have had success in assembling multiple fragments (albeit of considerably smaller size) by the use of custom made enzyme mixes. In this study, only ready mixed assembly kits from NEB were tested, so there may be scope to improve the assembly process by addressing the reaction mix itself. Additionally, it might be potentially enlightening to attempt to reclone the PVCs without the ATPase, to see if this is the single limiting factor (though whether they would be functional or not is a different matter). An extension of this that it would also be intriguing to test is to define a 'minimal' operon; the natural PVC variants show deletions in one of the sheath proteins, tail fibre proteins and some more enigmatic differences in the operon 'core' toward the 3' end. Streamlining the operon to the minimal components required would make for simpler cloning in future, as well as possibly shedding some light on the assembly requirements/choreography.

On a related note, in future it would be interesting to clone the ATPase separately on its own, to see if the toxicity it appears to be responsible for is indeed attributable

to this protein. Furthermore, it would be useful to have this protein cloned separately from the operon since its role is not well understood. Similar proteins in the Type VI Secretion System, for example, are responsible for the structures disassembly and recycling. Since the PVCs are essentially 'disposable' and single use, there does not appear to be a compelling reason to maintain an equivalent protein. The alternative hypotheses include, firstly, that the ATPase is some form of 'loading pump' for the toxin payloads, perhaps functioning to unfold them before passage in to the tube lumen. If this is the case, it might be expected that assembly could occur without the ATPase, unless the PVCs are assembled around the payloads concomitantly, as is the case for the T6SS (Basler, 2015). A natural first step would be to delete the ATPase from the PVC operon and then transcomplement its expression, to see if intact PVCs could be generated, in similar fashion to the previous Afp studies (or even if lack/dysregulation of the ATPase is detrimental to PVC function, as this is still unknown definitively). This would also help to answer whether there is some sort of internal regulation of these proteins toward the 3' end (such as in-operon promoters), and whether or not their expression needs to be tightly controlled to ensure correct construction versus simply being present in sufficient amount.

In future, an optimised long range PCR approach might be an option. During this project some attempts to PCR up to 18 kb were successful, though subsequent extraction of the DNA proved problematic, let alone cloning. As with many of the trickier molecular techniques, long range PCR is something of a 'dark art', requiring considerable optimisation.

Certain PVC operons have now been cloned in their entirety, either on cosmids or inducible plasmids, attesting to the fact that they are eminently cloneable. Achieving robust and reliable cloning and expression of the operons will be vital to their continued study, and is clearly possible. It seems at this point, it is just a matter of identifying precisely the right vectors to carry the sequences, identifying the best promoters, and what components of the operon can perhaps be regulated separately and directly. It may be the case, for example, that PVCs will be more stable on extremely tightly repressible vectors which are very low copy number. If indeed the ATPase is the 'problem child' of

the operon, this chapter may have provided valuable insight in to finally recreating the operons in a robust manner, and the 'missing ingredient' could be as simple as controlling the ATPase separately or identifying the ideal promoter/vector backbone.

6.3.3.2 PVC natural regulation and population dynamics

The reporter constructs produced for this study really are just the beginning of picking apart how the PVCs are deployed in the wild. In this chapter, the expression dynamics have only been explored with the variables of time and strain, but only under standard lab rich media culture conditions. There will no doubt be an effectively endless list of culture conditions that could be tested to examine the PVC expression response. A body of legacy RNAseq is available which demonstrates the PVC operon expression for *P. asymbiotica* and *P. luminescens*, with different blood/sera (e.g. human and insect), and at different temperatures. A logical first step would be to recapitulate these conditions and see if the transcriptional signal matches.

Beyond this, any number of conditions would be interesting to study, for example, the expression patterns when an infection cycle is in progress in complex with a nematode and inside a prey insect. Previously, GFP labelled bacteria where the Pnf effector toxin had been fused, had been detected in the spiracles (breathing tubes) of insects, potentially belying a preferential site of PVC deployment, however, in the RNAseq it can be seen that the effectors are capable of being expressed without the whole PVC operon, so this does not unambiguously confirm the role of the whole operon. Thus, it would be informative to examine the potential spatial/environmental differences in PVC activity using the PVC reporters instead, to see if there appears to be preferential deployment of different operons in different niches within an organism.

As a slight aside, the "Unit 4" operon of *P. luminescens* has been suspected of actually having some role in facilitating the symbiosis of the bacteria with the nematodes. They appear to have an effect on nematodes in the lab, but no activity against human cells, and this particular operon carries a number of effectors, 2 of which are reminiscent of "halovibrins", which are used by *Allivibrio fisheri* for signalling within the light organ of the bobtailed squid (Stabb *et al.*, 2001; Reich and Schoolnik, 1996). While this is still extremely speculative (and it is unclear whether the halovibrin-like homology is spurious

or not, as well as the operon containing other effectors which may be responsible), it would be interesting to explore the activity of this PVC within the nematode specifically. Given the effect that Shikuma *et al.* (2014) saw with the similar MAC complex and a helminth, it does not seem too unreasonable to put some weight behind this hypothesis.

More generally, the microscopy here has unambiguously revealed significant heterogeneity in PVC promoter activity within a culture, but is semi-quantitative at best. Future efforts will include flow cytometric quantification of the cultures, to understand exactly what proportion of the cells are producing the PVCs, to what level, and in as many culture conditions as possible.

Lastly, the efforts to understand the potential role of RfaH (or RfaH-like proteins) in the regulation of PVCs were frustrated experimentally by a number of issues. The feasibility of disrupting the protein in the *Photorhabdus* genome is questionable, as it is difficult to genetically engineer at the best of times. The only candidate RfaH-like protein detected so far in the *Photorhabdus* genome, is the chromosomal copy of RfaH itself. Given the broad role this protein is likely to play in all manner of operons it is almost certainly going to be problematic to attempt to make knockouts or controlled expression constructs. Non-ribosomal peptide and polyketide synthases (Challinor and Bode, 2015), are typically long operons which have previously had RfaH/NusG implicated in their regulation (Goodson *et al.*, 2017); given *Photorhabdus*' proclivity for secondary metabolite synthesis, RfaH may well be an integral part *Photorhabdus* biochemistry and physiology. One study that might be particularly interesting to understand the transcriptional landscape of PVCs (and other operons such as the aforementioned NRPS/PKSs) would be to use long-read technology direct RNA sequencing to look for evidence of long transcripts in the operons.

Because of this, study in *E. coli* is likely to be a better approach, but the existing Keio knockout strain exhibited a phenotype that was a cause for concern. If this were to be revisited, it would probably be worthwhile to recreate this mutant from scratch, using a different lab strain. There is always the possibility that the behaviour of the PVCs in any *E. coli* strain may not be truly representative however, since it has already been observed with the cosmids that they to not appear to be well regulated when expressed exogenously.

Perhaps the optimal study would therefore be to attempt to mirror the Afp study as closely as possible, by transcomplementing PVC expression constructs with candidate proteins which bind to *ops* sites. This would require identifying any NusG/RfaH homologues in *Photorhabdus* that could be implicated however, and so far this search has been fruitless.

Part IV

Discussion & Future Directions

Chapter 7

Discussion

As exemplified by Sarris *et al.* (2014) and Hurst *et al.* (2004), protein-translocating, phage-like macromolecular complexes are increasingly widespread, with many examples identified to date, including the Metamorphosis Associated Complex of *Pseudoalteromonas luteoviolacea*, the Antifeeding prophage(s) of *Serratia entomophila*, the Type 6 (and other secretion systems) found in many genera, and the PVCs themselves. The Afps and PVCs are somewhat different in that they are released from the host cells as individual 'needles', and capable of exerting their effects 'at range'. *Photorhabdus* and the PVCs remain more unique still for being the only examples identified to date where multiple variants are encoded within a single genome. What's more, while so far the Afps have only been observed on plasmids, PVCs have only even been found chromosomally integrated.

As a highly effective insect pathogen, *Photorhabdus* deploys the PVCs as just a small, but potentially highly versatile, part of its arsenal, facilitating pathogenicity and possibly other aspects of its life cycle. Given the bacteria's unusual life cycle as both a pathogen and a mutualist, it has to be capable of virulence against assorted insect prey, 'self-defence' to ensure the protection of the killed prey in the soil, all the while retaining symbiosis capability, to ensure continued re-association with the nematode. Consequently, *Photorhabdus* requires such a large repertoire of molecules, so it can use to influence the biology of a great many other organisms. Preliminary data in the lab has also begun to suggest that the PVCs may have roles beyond purely virulence, and that one of the PVCs may be responsible in part for the association between the bacteria and the nematode host.

This thesis has attempted to take a slightly different tack than many of the publications to date. Specifically, works like the studies of Sarris *et al.* (2014), and even the original PVC discovery paper by Yang *et al.* (2006), have tried to contextualise the PVCs in the pantheon of known caudate structures. Certainly, to begin to understand the basics of the structures this is a necessary step, and elements of the thesis continue in this vein. But, by leaning on this existing knowledge, a more novel approach of comparing and contrasting the variability between different PVC operons from various genomes, offers a completely different way of analysing these remarkable biological entities.

In summary, this work has examined the nuanced differences between PVC operons, both experimentally and computationally identifying various 'shades' of PVCs, not only in terms of their structural and payload differences, but also potentially in their regulation and deployment. Below some of the key insights from this work are summarised below, from each chapter discussion, and the relevant future study.

7.1 Chapter 3: New insights on PVC assembly and structure

7.1.1 A new role for an inner sheath paralogue

Subtle differences in the obtained structural models suggest that one of the inner sheath components, likely PVC5, may not actually form part of the inner sheath *per se*, instead forming an interfacing collar, similar to that of gp48/54 in the T4 phage. To date, it has been assumed that PVC1-5 simply contribute to the tube proper, and stoichiometric necessity was the given hypothesis. It seems that this may not be the case, when the electrostatics of PVC5 are examined, and the observation is made that it is translationally coupled to the rest of the spike/baseplate cluster of genes. If this is the case, this would also cast doubts on the contributions of the exterior sheath proteins, potentially implicating one or more of them in other roles such as other baseplate interfacing 'adapter' proteins. Evidence in the case of the outer sheath proteins is sparser however, as little difference could be identified in their simulated structures.

7.1.2 PVCs have profoundly negatively charged surfaces

Electrostatic observations of the exterior aspect of the putative outer sheath proteins reveals an overwhelming negative charge. This is borne out by subsequent experimental observation of their ability to bind to quaternary amine-based ion exchange columns. An extremely attractive hypothesis for this is that it may represent an 'evolved' surface, minimising the effects of trypsinisation and endocytosis, therefore potentially increasing the longevity of circulating released PVCs in an infection scenario (Del Tordello *et al.*, 2016; Kaur *et al.*, 2012).

7.1.3 PVCs resemble 'hybrid' caudate structures

While not strictly a new observation, since it has been known for some time that PVCs resemble elements of both phage and T6SSs, Chapter 3 on page 95 has demonstrated conclusively that the PVCs are reminiscent of a patchwork of different structures. Their outer sheaths, for instance, are highly similar to those of the R-type pyocin, yet their inner sheaths are not (or at least less so, having more in common with the T4 gp19 ortholog). The R-type pyocins, in turn, are thought to be descended from the P2 phage, rather than T4. Similarly, the models obtained for the VgrG-like spike complex appear 'stripped down' and simpler, much more like that of the T6SS than the T4 phage. Now, this is not to suggest that *Photorhabdus* has co-opted different components from different sources, rather it probably represents a honing and convergent evolution to handle the subtly different functions (translocating protein versus DNA for instance). Some questions remain however, for instance, if the spike complex is more similar to T6SS, but the sheath is more similar to non-T6SS structures, is there really the need for collar adapter proteins (Renault *et al.*, 2018). In short, no single orthologous structure dominates the 'homology signal' detected in querying PVC proteins.

7.1.4 Structural evidence for PVC14 as a 'tape measure protein'

While the models are likely far from accurate, and resolving the structures of tape measure proteins *in vitro* is an obstinate challenge, there is useful information in the models obtained. The work of Rybakova *et al.* (2015) on the Afp equivalent locus provides some

compelling evidence for their role as tape measure proteins, controlling the polymerisation of the Afp/PVC tubes. Now this *in silico* work exemplifies the striking helical secondary structure, and characteristic hydropathy of the locus, adding weight to the hypothesis.

7.1.5 Identification of possible new roles for certain loci

Structural modelling has revealed the striking structural similarity of PVC9 to a tube initiator protein, forming part of the baseplate complex, and revealing some spurious annotations attributed to certain protein structures within the PDB. Some preliminary evidence from HMM searches and backed up with preliminary modelling work has also identified PVC10 as a remote orthologue, on the very limits of sequence-based detection, to PAAR-spike tip proteins. Despite not containing the eponymous Pro-Ala-Ala repeats, potentially conserved metal binding residues were observed, along with characteristic secondary structure profiles. This has helped to further unpick some of the remaining 'dark matter' of the PVCs.

7.2 Chapter 4: Understanding PVC variability & mobility

7.2.1 PVCs as highly variable operons

Performing gene-by-gene phylogenetics within the PVC operons necessitated that the genes be clustered by synteny and orthology as a first step. This process alone reveals that the PVC operons are home to significant differences. There are, for example, numerous gene deletions in several of the operons, and curating the "Lumt" operon revealed that its 'operon core' (the last, approximately, six genes), are substantially different (and additional). Once the analysis proper had been completed, it showed that the putative tail fibre proteins, PVC13, are by far the most variable and incongruent genes in the whole operon. It was also revealed that both the "Lumt" and "Pnf" operons are frequent outliers, being somewhat sequentially different from the rest of the PVCs (and from each other). This has some potential consequences for the use of the "Pnf" operon as the lab model, however its function and structure remain among the better characterised, so this is unlikely to change. A good candidate for an alternative model operon would be the

"Cif" operons. Firstly, they would make good/better models as they are present in the *P. luminescens* genomes as well as the *P. asymbiotica* ones (unlike "Pnf"), and frequently cluster together, with the orthologue from *P. luminescens* as the relative outgrouping which is to be expected. The topologies for many of the other PVCs change within the trees, and this is especially true for "Pnf" and "Lumt".

7.2.2 PVCs are likely older than first thought

This workflow revealed a surprising conclusion. Inspecting the PVC sequences unequivocally reveals the presence of many tandem repeats, insertion elements, transposons, and other paraphernalia associated with horizontal gene transfer. However, the fact that the PVC genes are by-and-large congruent with the consensus tree and the known species topology, suggests that the PVC operons have been 'co-speciating' with their host genomes for some time. The hallmarks of horizontal gene transfer seen in the genomes probably speak to the mechanism of initial acquisition of the sequences. Certainly, the "Unit1" to "Unit4" operons in *P. luminescens* which are tandem to one another are proposed to have been carried in by the integration of a large plasmid (Yang *et al.*, 2006).

7.3 Chapter 5: PVCs have hyper-variable, chimeric tail fibres

7.3.1 PVC tail fibre proteins share hallmarks of real tail fibre proteins

In Chapter 5, the first experimental confirmation of these proteins is presented. Formation of homotrimers, and a high degree of thermal and chemical stability reveals their true nature as *bona fide* tail fibres, both of which are hall marks of previously resolved tail fibre proteins. Furthermore, their secondary structures, as determined by circular dichroism, follow similar patterns of minimal α-helix, instead being dominated by β-sheets and turn motifs.

7.3.2 Improved annotations lend confidence to a chimeric fibre structure

PVCs harbouring tail fibres has been proposed since their discovery, however annotations were typically mixed, often with low confidence scores, leaving much to be desired in terms of a conclusive explanation. A couple of promising models were able to be obtained

in Chapter 3, but the variability and putative chimerism meant that the simulation quality varied enormously, leaving questionable confidence in even the better models. Quality of hits for both identified domains of the tail fibres have improved in recent years, with regions of T4-like and Adenovirus-like homology now able to be demarcated quite clearly. What seemed like a spurious annotation some years ago (Adenovirus homology within a bacterial operon), now appears increasingly plausible. Of course, to say the PVC fibres resemble Adenoviral domains is probably incorrect, and it is simply that the Protein DataBank is saturated with resolved structures from phage and clinically relevant human pathogenic viruses. However, its clear that PVC tail fibres are not purely phage like. Hopefully their structures will be able to be resolved fully in future work, as promising results were achieved in crystallography attempts. If so, these proteins would represent the first natural chimeras of a prokaryotic and eukaryotic viral motif known.

7.3.3 PVC tail fibres potentially interact with cell surface proteins/sugars

The chapter showed some promising preliminary results for establishing potential binding partners for these fibres, and hopefully this will identify the cell and tissue specificities that the PVCs have when exerting their effects. Preliminary data (albeit for a single PVC fibre) suggested a specificity for galactose- and sialic acid-bearing sugars, as well as components of the desmosome, a cell surface and inter-cell junction complex. The identification of sugar binding activity in particular is intriguing, as phage fibres are known to associate with lipopolysaccharides, and a number of studies have performed similar array based experiments to good effect. In a separate study not discussed here, the tail fibres were also shown not to interact with lipids, meaning that the sugar and protein targets identified are all the more likely to be worth pursuing.

7.4 Chapter 6: PVC regulation is complex and heterogenous

7.4.1 PVCs are difficult to clone!

One of the early goals of this thesis was to develop and test new methods for heterologous expression of the PVCs in a robust manner. Unfortunately, these methods are either not appropriate or require considerably more optimisation. In the mean time, success has

been achieved by cloning using classical piece-wise restriction based methods. In future, a mixture of PCR/Gibson based assembly, combined with the higher efficiency of restriction based cloning likely remains the optimal way of producing these constructs.

7.4.2 Antitermination and operon polarity suppression is implicated in PVC production

Identification of a surprisingly canonical *ops* site (surprising, because *Photorhabdus* rarely likes to do much of anything canonically), suggests a role for anti-termination in the expression of PVC operons. Several long operons, particularly those which yield products destined for the extracellular environment, are known to be regulated by anti-termination mechanisms, often through the RfaH transcription factor, and the *ops* site. Antitermination enhances the transcription of long operons, by enhancing RNA polymerase processivity for approximately an additional 20 kb downstream. Further structural investigation showed that the Afp actually harbours an extended *ops* motif known as the JUMPStart sequence, forming a very particular secondary structure, but this is not found in PVCs. This may explain why Hurst *et al.* were able to tightly control the expression of Afps on the native pADAP plasmid, by manipulating AnfA1, but no equivalent protein (other than a chromosomal copy of RfaH) can be found in *Photorhabdus* - there may actually be subtle differences to their regulation. Unusually, subsequent cloning in the lab has shown that functional PVCs can be produced without the presence of the natural PVC 5′ UTR altogether, including the *ops* sequence; indicating that RfaH-like proteins/*ops* sites, are not *required* for PVC production, but may serve to fettle the expression in an as-yet-undetermined manner.

7.4.3 PVCs production is subject to significant population heterogeneity

Reporter microscopy has finally answered a long-pondered question. Namely, whether PVCs were produced by (potentially sacrificial) subpopulations of a culture. It seems that not only is this true, but the degree to which the operons are produced (and assuming this gives rise to expressed proteins/functional PVCs), differs greatly amongst the cells. Since at least part of an active *Photorhabdus* infection is sacrificial, as food for the nematode, it was proposed that these cells may have to lyse to release active PVCs - given their size.

No evidence of lysis has been detected in these assays or any others in the lab to date, suggesting that, miraculously, *Photorhabdus* has also found a mechanism of releasing these complexes, whereas heterologously expressing *E. coli* cultures suffer massive viability reduction. However, heterogeneity in expression remains. Not only this, but there are stark differences in the pattern of expression of different PVC operons. Combined with legacy RNAseq data from different inducing conditions, this implies that PVCs are indeed deployed under different and specific circumstances, by subsets of the population.

Chapter 8

Outlook

This thesis has covered many different aspects of PVC biology, and also represents a 'revival' of a project that had lain idle for a number of years. Consequently, there are many avenues which have generated preliminary data which could be built on significantly in future.

As far as Chapter 3 is concerned, the process could be iterated essentially *ad infinitum*, each time benefitting from better homologies through new depositions in the various databanks. In the 3-4 years of this work alone, new homologies to almost every protein have been found. Of course, a key advancement will be the final experimental elucidation of a full PVC structure, hopefully in the not too distant future, as the collaboration with the Max-Planck in Dortmund continues. Once this is done, it would be worthwhile repeating the process for many, if not all PVC proteins, to obtain more accurate homology models for future use. The models obtained in this study will prove useful many aspects of future lab work potentially; providing a reference for the effects of, for instance, mutagenesis, or for the design of anti-peptide antibodies against elements expected to be available on the surface of the complex.

For Chapter 4, it would be interesting to repeat the workflow with a greater number of PVC loci from a greater diversity of genomes. While not reported here, progress has been made on automatically detecting and 'isolating' PVC sequences from other genomes, based in part on the criteria defined in this chapter. It would also be advantageous to attempt to fully automate the process of syntenic orthologue clustering and matrix

creation for calculation of the Adjusted Wallace Coefficient - at present these tasks are more subjective than is ideal. Hand in hand with this, as ever, it would also be useful to sequence even more genomes for further study.

As the first experimental chapter, Chapter 5 has opened up a great many future avenues for study. Setting aside the practicalities for a moment, the most obvious tasks are to try and repeat the studies here for as many putative PVC tail fibres as possible. That includes binding partner characterisation (both proteomic and glycan array-based). Some progress along this path is underway, as the "Pnf" fibre that was cloned in this work, but slightly more recalcitrant to work with, has only recently been subjected to the same affinity experiments as the "Lumt" fibre. Since then, a further tail fibre from the "Unit 4" operon of *P. asymbiotica* TT01 has been cloned (as this is the same PVC which is currently under study in Dortmund). Just counting the PVCs which have been studied from the three genomes focussed on in this PhD, that leaves approximately another 13 tail fibre proteins which could be cloned and assayed. Much easier said than done, of course, would be further attempts to crystallise and finally resolve the structures of these proteins once and for all, instead of relying on indirect measurements like circular dichroism and SDS-PAGE. Exciting future work will be to recapitulate similar studies of engineering tail fibres, with the intent of altering the specificity of the PVCs, such that they may be able to convey their cargoes to cell/tissue types of our choosing. Current experiments ongoing in the lab are swapping the fibres between different PVC structures, to examine any effects on assembly or binding/function (it's possible for example, that they won't be able to transduce a contraction signal). It is not known for certain that the contraction signal is conveyed through the tail fibres, though this is the case for the T4 phage, so identifying if the fibres are implicated in this function will also be well worth testing in future. If they remain functional with 'transplanted' fibres however, this becomes a tool for exploring the role of different PVCs and their natural specificities for different cells.

Time was limited for much of the reporter microscopy work undertaken in the final chapter, which leaves much left to do. The natural progressions for this study are to, firstly, perform the microscopy assays under as wide a variety of potential inducing conditions as possible, to see there are differential responses in the deployment of the

PVCs - so far, only standard lab rich media conditions have been tested (albeit over a long growth cycle). Secondly, to obtain more quantitative data about the population heterogeneity, these reporters should be subjected to flow cytometry or flow-assisted cell sorting. The FC/FACS process can therefore also extend all the different conditions that would be tested. Sequencing of sub-populations isolated by FACS could potentially offer insights in to the genomic/transcriptomic state of the proposed 'sacrificial populations', which would not only be enlightening in terms of PVC biology, but might also have information to offer for understanding the link between a sacrificial population and association with the nematode, or other fundamental aspects of its life cycle. On the subject of transcriptomics, an interesting experiment would be direct RNA sequencing, now made possible by the likes of Oxford Nanopore. This would shed further light on the potential role of transcript elongation in the PVC operons. The current theory is that the PVCs are produced progressively extending transcripts (to produce large amounts of the 5' genes) or from one long transcript which is potentially re-used.

Finally, it will no doubt be necessary to continue grinding away at the heterologous expression of as many PVC operons as possible. It seems the most promising way to do this is still 'the old fashioned way', and some successes so far, both in terms of their construction, but also their functionality, provide confidence that this is the most viable way to proceed.

In summary, 'the PVC project' is nothing short of enormous. There is a great deal that can be done to extend this thesis, and a great deal more besides, that this thesis hasn't even touched on. The project will no doubt yield many more PhDs-worth of data and insight to come.

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., 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. (1994). Tail length determination in bacteriophage T4. *Virology*, 199:301–310.

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

Adhya, S., Gottesman, M., and De Crombrugghe, B. (1974). Release of polarity in Escherichia coli by gene N of phage lambda: termination and antitermination of transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 71(6):2534–8.

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

Akhurst, R. J. (1980). Morphological and functional dimorphism in Xenorhabdus spp., bacteria symbiotically associated with the insect pathogenic nematodes Neoaplectana and Heterorhabditis. *Journal of General Microbiology*, 121(May):303–309.

Aksyuk, A., Leiman, P. G., Kurochkina, L. P., Shneider, M. M., Kostyuchenko, V., 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.

Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences*, 104(29):11963–11968.

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.

Ancona, V., Lee, J. H., Chatnaparat, T., Oh, J., Hong, J. I., and Zhao, Y. (2015). The bacterial alarmone (p)ppGpp activates the type III secretion system in Erwinia amylovora. *Journal of Bacteriology*, 197(8):1433–1443.

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.

Artsimovitch, I. and Landick, R. (2002). The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand. *Cell*, 109(2):193–203.

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., 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.

Bailey, M. J., Hughes, C., and Koronakis, V. (1996). Increased distal gene transcription by the elongation factor RfaH, a specialized homologue of NusG. *Molecular Microbiology*, 22(4):729–737.

Bailey, M. J., Hughes, C., and Koronakis, V. (1997). RfaH and the ops element, components of a novel system controlling bacterial transcription elongation. *Molecular microbiology*, 26(5):845–851.

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.

Banerjee, S., Chalissery, J., Bandey, I., and Sen, R. (2006). Rho-dependent transcription termination: more questions than answers. *Journal of microbiology*, 44(1):11–22.

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. (2015). 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.

Bergmann, S., Schümann, J., Scherlach, K., Lange, C., Brakhage, A. A., and Hertweck, C. (2007). Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. *Nature Chemical Biology*, 3(4):213–217.

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. and Akhurst, R. J. (1988). Biochemical and Physiological Characterization of Colony Form Variants in Xenorhabdus spp. (Enterobacteriaceae). *Microbiology*, 134(May):751–761.

Boemare, N. E., Akhurst, R. J., and R., M. G. (1993). DNA Relatedness between Xenorhabdus spp. (Enterobacteriaceae), Symbiotic Bacteria of Entomopathogenic Nematodes, and a Proposal To Transfer Xenorhabdus luminescens to a New Genus, Photorhabdus gen. nov. N. *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. *Journal of Virology*, 78(14):7727–7736.

Cárcamo-Oyarce, G., Lumjiaktase, P., Kümmerli, R., and Eberl, L. (2015). Quorum sensing triggers the stochastic escape of individual cells from Pseudomonas putida biofilms. *Nature Communications*, 6(May 2014).

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.

Chaban, Y., Lurz, R., Brasilès, S., Cornilleau, C., Karreman, M., Zinn-Justin, S., Tavares, P., and Orlova, E. V. (2015). Structural rearrangements in the phage head-to-tail interface during assembly and infection. *Proceedings of the National Academy of Sciences*, 112(22):7009–7014.

Challinor, V. L. and Bode, H. B. (2015). Bioactive natural products from novel microbial sources. *Annals of the New York Academy of Sciences*, 1354(1):82–97.

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.

Charlop-Powers, Z., Milshteyn, A., and Brady, S. F. (2014). Metagenomic small molecule discovery methods. *Current Opinion in Microbiology*, 19(1):70–75.

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).

Chatnaparat, T., Li, Z., Korban, S. S., and Zhao, Y. (2015). The Stringent Response Mediated by (p)ppGpp Is Required for Virulence of Pseudomonas syringae pv. tomato and Its Survival on Tomato. *Molecular Plant-Microbe Interactions*, 28(7):776–789.

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. *British Biotechnology Journal*, 2(1):1–12.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823–6.

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).

Chung, T. H., Wu, S. H., Yao, M., Lu, C. W., Lin, Y. S., Hung, Y., Mou, C. Y., Chen, Y. C., and Huang, D. M. (2007). The effect of surface charge on the uptake and biological function of mesoporous silica nanoparticles in 3T3-L1 cells and human mesenchymal stem cells. *Biomaterials*, 28(19):2959–2966.

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.

Coca-Abia, M. M. and Romero-Samper, J. (2016). Establishment of the identity of Costelytra zealandica (White 1846) (Coleoptera: Scarabeidae: Melolonthinae) a species commonly known as the New Zealand grass grub. *New Zealand Entomologist*, 39(2):129–146.

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*, 2(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.

Dalal, S., Balasubramanian, S., and Regan, L. (1997). Transmuting α helices and β sheets. *Folding and Design*, 2(5):R71–R79.

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.

Dalebroux, Z. D., Svensson, S. L., Gaynor, E. C., and Swanson, M. S. (2010). ppGpp Conjures Bacterial Virulence. *Microbiology and Molecular Biology Reviews*, 74(2):171–199.

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.

Del Tordello, E., Danilchanka, O., McCluskey, A. J., and Mekalanos, J. J. (2016). Type VI secretion system sheaths as nanoparticles for antigen display. *Proceedings of the National Academy of Sciences*, 113(11):3042–3047.

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.

Dodd, S. J., Hurst, M. R. H., Glare, T. R., O'Callaghan, M., and Ronson, C. W. (2006). Occurrence of sep insecticidal toxin complex genes in Serratia spp. and Yersinia frederiksenii. *Applied and Environmental Microbiology*, 72(10):6584–6592.

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., Brunet, Y. R., Spinelli, S., Lensi, V., Legrand, P., Blangy, S., Kumar, A., Journet, L., Cascales, E., and Cambillau, C. (2016). Structure and specificity of the Type VI secretion system ClpV-TssC interaction in enteroaggregative Escherichia coli. *Scientific Reports*, 6(October):1–13.

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). Students May Have Answer for Faster-Healing Civil War Wounds that Glowed.

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, 22(10):1315–1316.

Eleftherianos, I., Ffrench-Constant, R. H., Clarke, D. J., Dowling, A. J., and Reynolds, S. E. (2010). Dissecting the immune response to the entomopathogen Photorhabdus. *Trends in microbiology*, 18(12):552–60.

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, B., Olivera, E. R., Sandoval, Á., Arias-Barrau, E., Arias, S., Naharro, G., and Luengo, J. M. (2004). Strategy for cloning large gene assemblages as illustrated using the phenylacetate and polyhydroxyalkanoate gene clusters. *Applied and Environmental Microbiology*, 70(8):5019–5025.

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. (2015). 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.

Giacalone, M. J., Gentile, A. M., Lovitt, B. T., Berkley, N. L., Gunderson, C. W., and Surber, M. W. (2006). Toxic protein expression in Escherichia coli using a rhamnose-based tightly regulated and tunable promoter system. *BioTechniques*, 40(3):355–364.

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.

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., Benders, G. A., Montague, M. G., Ma, L., Moodie, M. M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-pfannkoch, C., Denisova, E. a., Young, L., Qi, Z.-q., Segall-Shapiro, T. H., Calvey, C. H., Parmar, P. P., Hutchison, C. a., Smith, H. O., Venter, J. C., Iii, C. A. H., Smith, H. O., and Venter, J. C. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science (New York, N.Y.)*, 329(5987):52–56.

Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, 6(5):343–345.

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.

Goodson, J. R., Klupt, S., Zhang, C., Straight, P., and Winkler, W. C. (2017). LoaP is a broadly conserved antiterminator protein that regulates antibiotic gene clusters in Bacillus amyloliquefaciens. *Nature Microbiology*, 2(February):1–10.

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. (2014). 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., Belin, D., Carson, M. J., and Beckwith, J. (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.

Heinrich, A. K., Glaeser, A., Tobias, N. J., Heermann, R., and Bode, H. B. (2016). Heterogeneous regulation of bacterial natural product biosynthesis via a novel transcription factor. *Heliyon*, 2(11).

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.

Holm, L. and Sander, C. (1996). Mapping the protein universe. *Science*, 273(5275):595–602.

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 Artsimovitch, I. (2017). A screen for rfaH suppressors reveals a key role for a connector region of termination factor rho. *mBio*, 8(3):1–12.

Hu, K. and Webster, J. M. (2000). Antibiotic production in relation to bacterial growth and nematode development in Photorhabdus 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. H., Beard, S. S., Jackson, T. A., and Jones, S. M. (2007). Isolation and characterization of the Serratia entomophila antifeeding prophage. *FEMS Microbiology Letters*, 270:42–48.

Hurst, M. R. H., Beattie, A., Jones, S. A., Laugraud, A., van Koten, C., and Harper, L. (2018). Characterization of Serratia proteamaculans strain AGR96X encoding an anti-feeding prophage (tailocin) with activity against grass grub (Costelytra giveni) and manuka beetle (Pyronota spp.) larvae. *Applied and Environmental Microbiology*, 84(10):AEM.02739–17.

Hurst, M. R. H., 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., 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.

Illergård, K., Ardell, D. H., and Elofsson, A. (2009). Structure is three to ten times more conserved than sequence - A study of structural response in protein cores. *Proteins: Structure, Function and Bioinformatics*, 77(3):499–508.

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). Biosythese induite et mode d'action dune Pyocine, antibiotique de Pseudomonas pyocynia. In *Annales de l'Institute Pasteur*, volume 86, pages 149–160. Masson editeur 120 Blvd. St. Germain, 75280 Paris.

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. *Biopolymers*, 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. (1990). Mechanism of Length Determination in Bacteriophage Lambda Tails. *Adv. Biophys.*, 26:1–18.

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

Kaur, T., Nafissi, N., Wasfi, O., Sheldon, K., Wettig, S., and Slavcev, R. (2012). Immunocompatibility of bacteriophages as nanomedicines. *Journal of Nanotechnology*, 2012(i).

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.

Kerpedjiev, P., Hammer, S., and Hofacker, I. L. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31(20):3377–3379.

Khlebnikov, A., Risa, Ø., Skaug, T., Trent, A., Keasling, J. D., and Carrier, T. A. (2000). Regulatable Arabinose-Inducible Gene Expression System with Consistent Control in All Cells of a Culture. *Society*, 182(24):7029–7034.

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. (2015). 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.

Kuijpers, N. G. A., Solis-Escalante, D., Bosman, L., van den Broek, M., Pronk, J. T., Daran, J. M., and Daran-Lapujade, P. (2013). A versatile, efficient strategy for assembly of multi-fragment expression vectors in Saccharomyces cerevisiae using 60 bp synthetic recombination sequences. *Microbial Cell Factories*, 12(1):1–13.

Kumar, S. N., Nambisan, B., Kumar, B. S. D., Vasudevan, N. G., Mohandas, C., Cheriyan, V. T., and Anto, R. J. (2013). Antioxidant and anticancer activity of 3,5-dihydroxy-4-isopropylstilbene produced by Bacillus sp. N strain isolated from entomopathogenic nematode. *Archives of Pharmacal Research*, 1(3):1–11.

Kumar Sarkar, S., Takeda, Y., Kanamaru, S., and Arisaka, F. (2006). Association and dissociation of the cell puncturing complex of bacteriophage T4 is controlled by both pH and temperature. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1764(9):1487–1492.

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., Gibert, M., Popoff, M. R., Boquet, P., and Gauthier, M. (2003). Expression of cnf1 by Escherichia coli J96 involves a large upstream DNA region including the hlyCABD operon, and is regulated by the RfaH protein. *Molecular Microbiology*, 47(6):1653–1667.

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.

Langer, A., Moldovan, A., Harmath, C., Joyce, S. A., Clarke, D. J., and Heermann, R. (2017). HexA is a versatile regulator involved in the control of phenotypic heterogeneity of Photorhabdus luminescens. *PLoS ONE*, 12(4):1–23.

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).

Lautru, S., Deeth, R. J., Bailey, L. M., and Challis, G. L. (2005). Discovery of a new peptide natural product by streptomyces coelicolor genome mining. *Nature Chemical Biology*, 1(5):265–269.

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. *Infection and Immunity*, 67(2):717–725.

Leeds, J. A. and Welch, R. A. (1996). RfaH enhances elongation of Escherichia coli hlyCABD mRNA. *Journal of Bacteriology*, 178(7):1850–1857.

Leeds, J. A. and Welch, R. A. (1997). Enhancing transcription through the Escherichia coli hemolysin

operon, hlyCABD: RfaH and upstream JUMPStart DNA sequences function together via a postinitiation mechanism. *Journal of Bacteriology*, 179(11):3519–3527.

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 Defence 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, D. and Huang, L. (1992). Trypsin-induced lysis of lipid vesicles: Effect of surface charge and lipid composition. *Analytical Biochemistry*, 202(1):1–5.

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.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1).

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.

Machado, R. A. R., Wüthrich, D., Kuhnert, P., Arce, C. C. M., Thönen, L., Ruiz, C., and Zhang, X. (2018). Whole-genome-based revisit of Photorhabdus phylogeny : proposal for the elevation of most Photorhabdus subspecies to the species level and description of one novel species Photorhabdus bodei sp . nov ., and one novel subspecies Photorhabdus laumondii subs. *International Journal of Systematic and Evolutionary Microbiology*, pages 1–18.

Mahony, J., Alqarni, M., Stockdale, S., Spinelli, S., Feyereisen, M., Cambillau, C., and Van Sinderen, D. (2016). Functional and structural dissection of the tape measure protein of lactococcal phage TP901-1. *Scientific Reports*, 6(August):1–10.

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.

Marokházi, J., Lengyel, K., Pekár, S., Felföldi, G., Patthy, A., Gráf, L., Fodor, A., and Venekei, I. (2004). Comparison of proteolytic activities produced by entomopathogenic Photorhabdus bacteria: Strain- and phase-dependent heterogeneity in composition and activity of four enzymes. *Applied and Environmental Microbiology*, 70(12):7311–7320.

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.

Netta, M., Ikedab, H., and Moore, B. S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Natural Product Reports*, 26(11):1362–1384.

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.

Nieto, J. M., Bailey, M. J. A., Hughes, C., and Koronakis, V. (1996). Suppression of transcription polarity in the Escherichia coli haemolysin operon by a short upstream element shared by polysaccharide and DNA transfer determinants. *Molecular Microbiology*, 19(4):705–713.

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.

Nudler, E. and Gottesman, M. E. (2002). Transcription termination and anti-termination in E. coli. *Genes to Cells*, 7(8):755–768.

O'Farrell, P. Z. and Gold, L. (1973). Bacteriophage T4 Gene Expression. *Journal of Biological Chemistry*, 248(15):5502–5511.

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.

Reich, K. A. and Schoolnik, G. K. (1996). Halovibrin, secreted from the light organ symbiont Vibrio fischeri, is a member of a new class of ADP-ribosyltransferases. *Journal of Bacteriology*, 178(1):209–215.

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.

Renault, M. G., Beas, J. Z., Douzi, B., Chabalier, M., Zoued, A., Brunet, Y. R., Cambillau, C., Journet, L., and Cascales, E. (2018). The gp27-like Hub of VgrG Serves as Adaptor to Promote Hcp Tube Assembly. *Journal of Molecular Biology*, page #pagerange#.

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.

Rodriguez, R., Chinea, G., Lopez, N., Pons, T., and Vriend, G. (1998). Homology modeling, model and software evaluation: three related resources. *Bioinformatics (Oxford, England)*, 14(6):523–528.

Rodríguez-Guerra Pedregal, J. and Maréchal, J.-D. (2018). PyChimera: use UCSF Chimera modules in any Python 2.7 project. *Bioinformatics*, 1(January):1–2.

Rose, G. D. and Creamer, T. P. (1994). Protein folding: Predicting predicting.

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.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94.

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, University of Auckland.

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. (2015). Afp14 is involved in regulating the length of Anti-feeding prophage (Afp). *Molecular Microbiology*, pages n/a–n/a.

Ryjenkov, D. A., Tarutina, M., Moskvin, O. V., and Gomelsky, M. (2005). Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: Insights into biochemistry of the GGDEF protein domain. *Journal of Bacteriology*, 187(5):1792–1798.

Sandmeler, H. (1994). Acquisition and rearrangement of sequence motifs in the evolution of bacteriophage tail fibres. *Molecular Microbiology*, 12(3):343–350.

Santangelo, T. J., ČuboÅová, L., Matsumi, R., Atomi, H., Imanaka, T., and Reeve, J. N. (2008). Polarity in archaeal operon transcription in Thermococcus kodakaraensis. *Journal of Bacteriology*, 190(6):2244–2248.

Santangelo, T. J. and Roberts, J. W. (2002). RfaH, a bacterial transcription antiterminator. *Molecular Cell*, 9(4):698–700.

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., 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.

Shore, D. (2000). The Sir2 protein family: A novel deacetylase for gene silencing and more. *Proceedings of the National Academy of Sciences*, 97(26):14030–14032.

Siegele, D. A. and Hu, J. C. (1997). Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proceedings of the National Academy of Sciences*, 94(15):8168–8172.

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.

Skibinski, D. a. G., Golby, P., Chang, Y.-s., Sargent, F., Hoffman, R., Harper, R., Guest, J. R., Attwood, M. M., Berks, B. C., and Andrews, S. C. (2002). Regulation of the Hydrogenase-4 Operon of Escherichia coli by the sigma-54 -Dependent Transcriptional Activators FhlA and HyfR. *Society*, 184(23):6642–6653.

Smigielski, A. J., Akhurst, R. J., and Boemaret, N. E. (1994). Phase Variation in Xenorhabdus nematophilus and Photorhabdus luminescens: Differences in Respiratory Activity and Membrane Energization. *Applied and Environmental Microbiology*, 60(1):120–125.

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). Why Some Civil War Soldiers Glowed in the Dark.

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. *International Journal*

of Systematic Bacteriology, 49:1433–1438.

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.

Stabb, E. V., Reich, K. A., and Ruby, E. G. (2001). Vibrio fischeri genes hvnA and hvnB encode secreted NAD+-glycohydrolases. *Journal of Bacteriology*, 183(1):309–317.

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.

Szumanski, M. B. and Boyle, S. M. (1990). Analysis and sequence of the speB gene encoding agmatine ureohydrolase, a putrescine biosynthetic enzyme in Escherichia coli. *Journal of Bacteriology*, 172(2):538–547.

Taylor, N. M., Prokhorov, N. S., Guerrero-Ferreira, R. C., Shneider, M. M., Browning, C., Goldie, K. N., Stahlberg, H., and Leiman, P. G. (2016). 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*.

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.

Thomason, L. C., Costantino, N., Shaw, D. V., and Court, D. L. (2007). Multicopy plasmid modification with phage Lambda-Red recombineering. *Plasmid*, 58:148–158.

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.

Turlin, E., Pascal, G., Rousselle, J. C., Lenormand, P., Ngo, S., Danchin, A., and Derzelle, S. (2006). Proteome analysis of the phenotypic variation process in Photorhabdus luminescens. *Proteomics*, 6(9):2705–2725.

Uratani, Y. and Hoshino, T. (1984). Pyocin Ri Inhibits Active Transport in Pseudomonas aeruginosa and Depolarizes Membrane Potential. *Journal of Bacteriology*, 157(2):1–6.

Van Lanen, S. G. and Shen, B. (2006). Microbial genomics for the improvement of natural product discovery. *Current Opinion in Microbiology*, 9(3):252–260.

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.

Vetcher, L., Tian, Z.-q., Mcdaniel, R., Rascher, A., Revill, W. P., Hutchinson, C. R., and Hu, Z. (2005). Rapid Engineering of the Geldanamycin Biosynthesis Pathway by Red / ET Recombination and Gene Complementation Rapid Engineering of the Geldanamycin Biosynthesis Pathway by Red / ET Recombination and

Gene Complementation. *Society*, 71(4):1829–1835.

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.

Vogelaar, N. J., Jing, X., Robinson, H. H., and Schubot, F. D. (2010). Analysis of the crystal structure of the ExsCExsE complex reveals distinctive binding interactions of the pseudomonas aeruginosa type III secretion chaperone ExsC with ExsE and ExsD. *Biochemistry*, 49(28):5870–5879.

Wallace, D. L. (1983). A Method for Comparing Two Hierarchical Clusterings: comment. *Journal of the American Statistical Association*, 78(383):553–569.

Wang, H., Li, Z., Jia, R., Hou, Y., Yin, J., Bian, X., Li, A., Müller, R., Stewart, A. F., Fu, J., and Zhang, Y. (2016). RecET direct cloning and Red$\alpha\beta$ recombineering of biosynthetic gene clusters, large operons or single genes for heterologous expression. *Nature Protocols*, 11(7):1175–1190.

Wang, J., Brackmann, M., Castaño-Díez, D., Kudryashev, M., Goldie, K. N., Maier, T., Stahlberg, H., and Basler, M. (2017). Cryo-EM structure of the extended type VI secretion system sheath-tube complex. *Nature Microbiology*, 2(11):1507–1512.

Wang, L., Jensen, S., Hallman, R., and Reeves, P. R. (1998). Expression of the O antigen gene cluster is regulated by RfaH through the JUMPstart sequence. *FEMS Microbiology Letters*, 165(1):201–206.

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.

Wenzel, S. C. and Müller, R. (2009). The impact of genomics on the exploitation of the myxobacterial secondary metabolome. *Natural Product Reports*, 26(11):1385.

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., 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.

Wilkinson, R. G., Gemski, P., and Stocker, B. a. (1972). Non-smooth mutants of Salmonella typhimurium: differentiation by phage sensitivity and genetic mapping. *Journal of general microbiology*, 70(3):527–54.

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 Bacteriology*, 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. (2014). Structure and function of bacteriophage T4. *Future Microbiology*, 9(12):1319–1327.

Yin, J., Zhu, H., Xia, L., Ding, X., Hoffmann, T., Hoffmann, M., Bian, X., Müller, R., Fu, J., Stewart, A. F., and Zhang, Y. (2015). A new recombineering system for Photorhabdus and Xenorhabdus. *Nucleic Acids Research*, 43(6):e36.

Yosef, I., Bloushtain, N., Shapira, M., and Qimron, U. (2004). Restoration of gene function by homologous recombination: From PCR to gene expression in one step. *Applied and Environmental Microbiology*,

70(12):7156–7160.

Young, R., Wang, I. N., and Roof, W. D. (2000). Phages will out: Strategies of host cell lysis. *Trends in Microbiology*, 8(3):120–128.

Yu, D., Ellis, H. M., Lee, E. C., Jenkins, N. a., Copeland, N. G., and Court, D. L. (2000). An efficient recombination system for chromosome engineering in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 97(11):5978–83.

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.

Zhang, Y. and Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309.

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.

Ziemert, N., Alanjary, M., and Weber, T. (2016). The evolution of genome mining in microbes a review. *Nat. Prod. Rep.*, 33(8):988–1005.

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

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.

As a brief abstract for the reader; the paper concerns attempts to create mimics of biological 'antifreeze' proteins, for application in areas such as cryopreservation. Ice crystal growth is a significant problem which leads to reduction in cell viability, and compromises the quality of vital resources such as stored donor blood, among others. Current methods employ toxic compounds such as glycerol and dimethylsulfoxide which can cause undesirable downstream consequences for the patient, and also require extensive dialysis which high in cost, both in terms of time and money, and is low in efficiency.

The paper demonstrates that polyproline helices exhibit ice recrystallisation inhibition activity, and it is postulated that this is due to alternating patches of hydrophobicity and -philicity (amphipathy). When used in concert with one of the standard cryopresevation additives, 10% DMSO, polyproline additives increased cellular recovery from thawing by up to 30%.

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.*

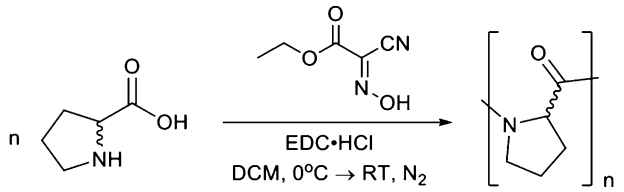
Tissue engineering, gene therapy, therapeutic protein production, and transplantation rely on the successful storage and transport of donor cells.[1] For example, in the production of therapeutic proteins, a specific cell line must be developed for each protein.[2] Given that any in vitro culture will 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

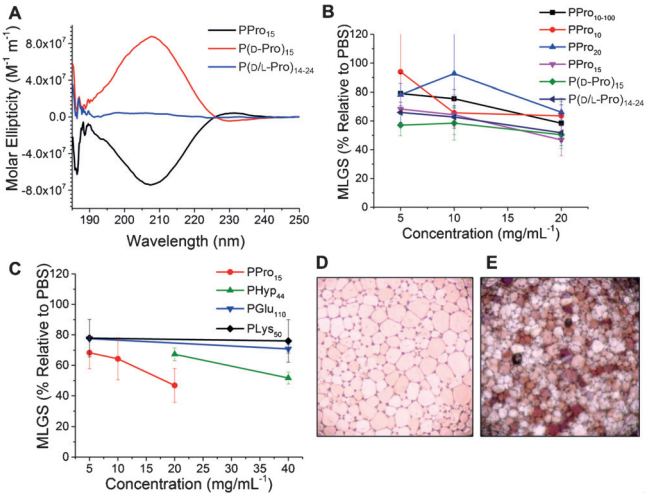


Figure 1. A) Circular dichroism spectra. B) IRI activity of the polyproline series. C) IRI activity compared to other homopolypeptides. D) Cryomicrograph of a PBS negative control. E) Cryomicrograph of 20 mg mL^{-1} polyproline. Photographs taken after 30 min at −8 °C. Error bars represent ± standard 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-

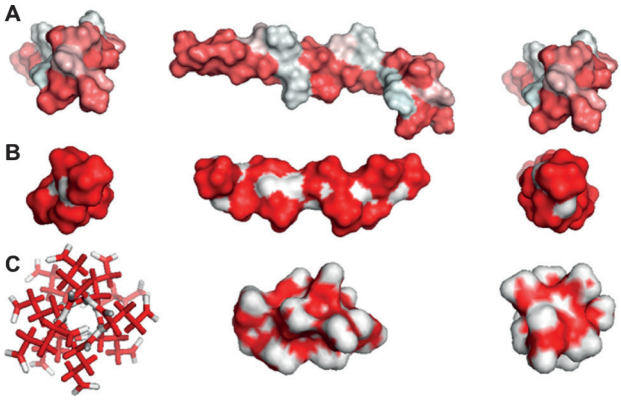


Figure 2. Hydrophobic surface mapping of A) recombinant type I sculpin AFP, B) $PPro_{10}$, and C) $PGlu_{10}$, showing charged hydrophilic surfaces. Hydrophobic regions (red), hydrophilic regions (white).

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

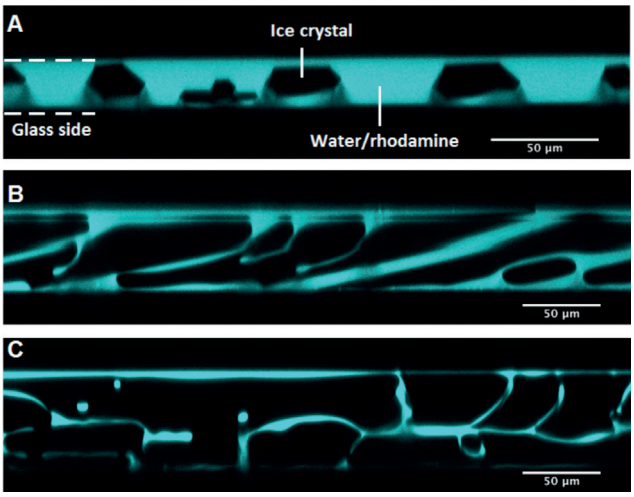


Figure 3. Cross-section of ice crystals perpendicular to the temperature gradient: A) ZrAc (positive control), B) $PPro_{19}$, C) PBS (negative control). The ice crystals expel the dye while growing, appearing in black, while the remaining liquid fluoresces.

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.

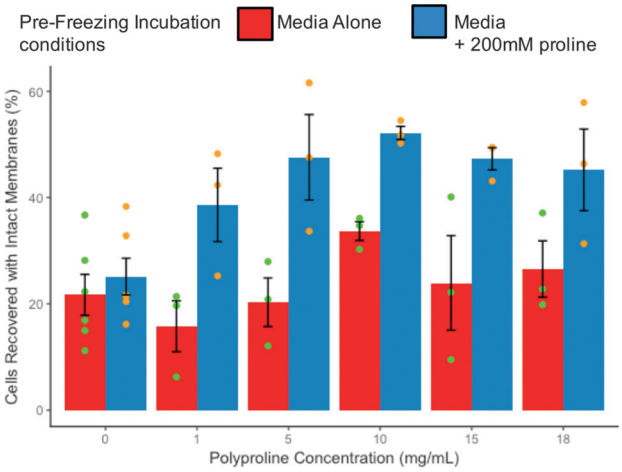


Figure 4. A549 cryopreservation. Cell recovery determined by trypan blue assays. Cells were first incubated either in the medium alone or with 200 mM proline for 24 h. They were subsequently cryopreserved by addition of 10% DMSO with the indicated $PPro_{11}$ concentration. Error bars ± S.E.M. from $n = 3$ 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

Appendix B

Chapter 4 Appendices

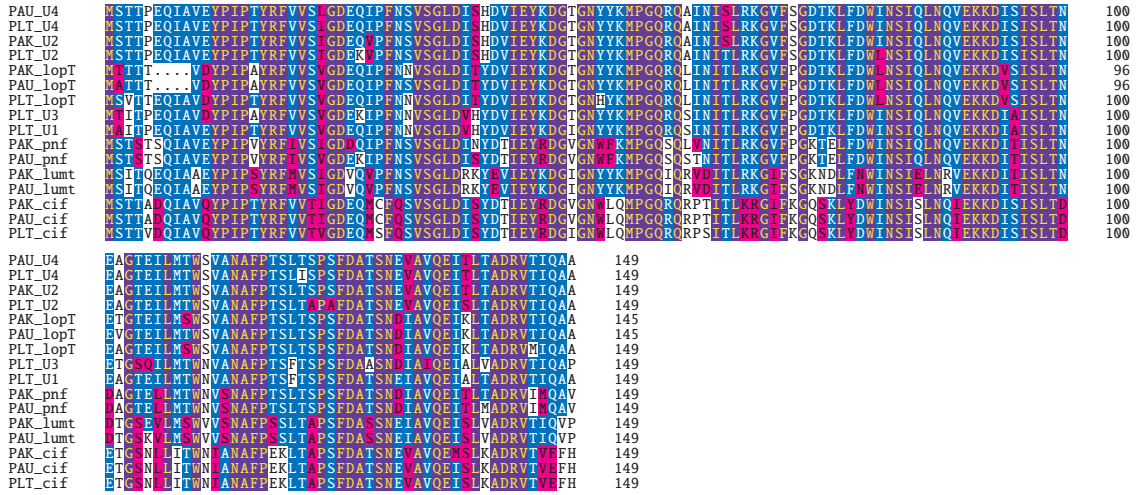
B.1 Clusterings of PVC genes for phylogenetic studies

Table B.1 | The final clustering of CDS features for phylogenetic analysis. Where a cell is blank, a gene deletion was observed.

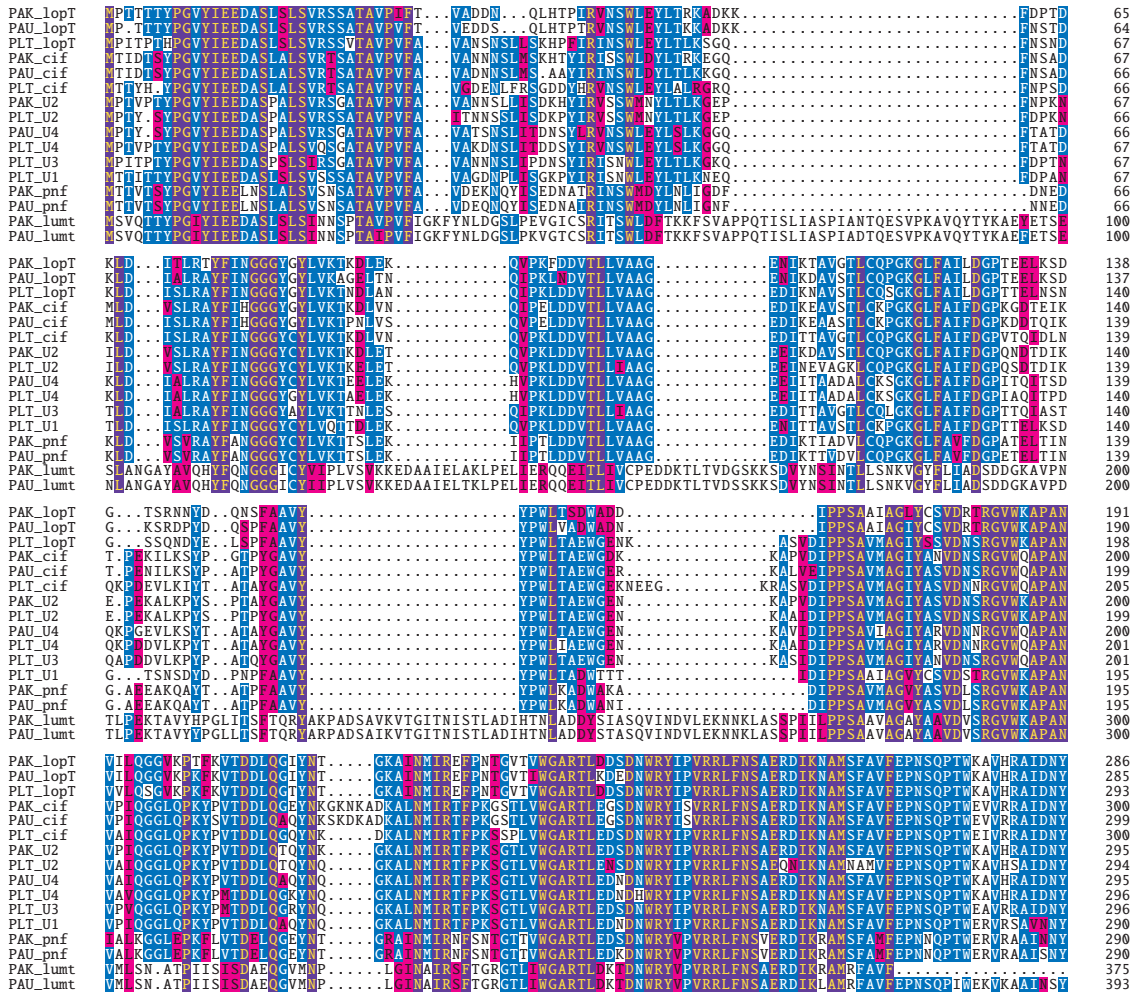
	"PNF"		"CIF"			"LOPT"			"UNIT4"		"UNIT2"		"UNIT1"	"UNIT3"	"LUMT"	
ORF	43949	Kingscliff	43949	Kingscliff	TT01	43949	Kingscliff	TT01	43949	TT01	Kingscliff	TT01	TT01	TT01	43949	Kingscliff
PVC1	PAU03392	PAK03203	PAU01961	PAK01787	PLT02568	PAU02074	PAK01896	PLT02424	PAU02775	PLT01696	PAK02606	PLT01736	PLT01758	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

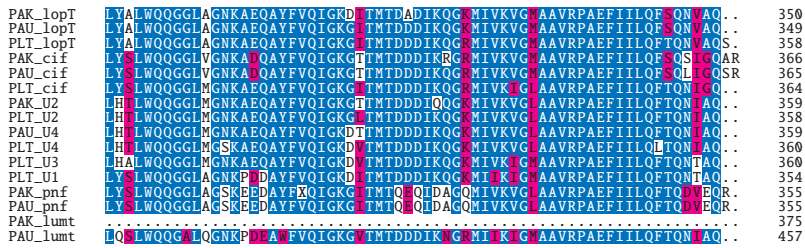
B.2 Multiple Sequence Alignments for PVC proteins

PVC Locus 1

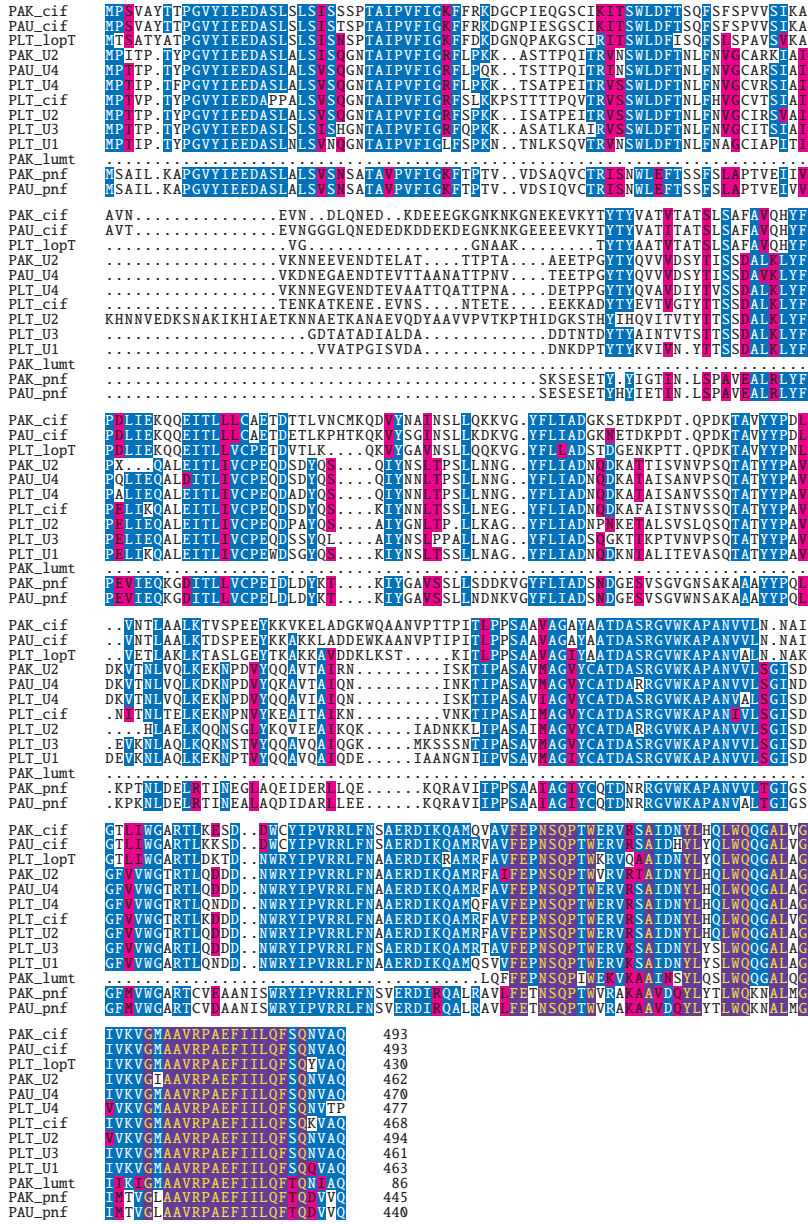


PVC Locus 2

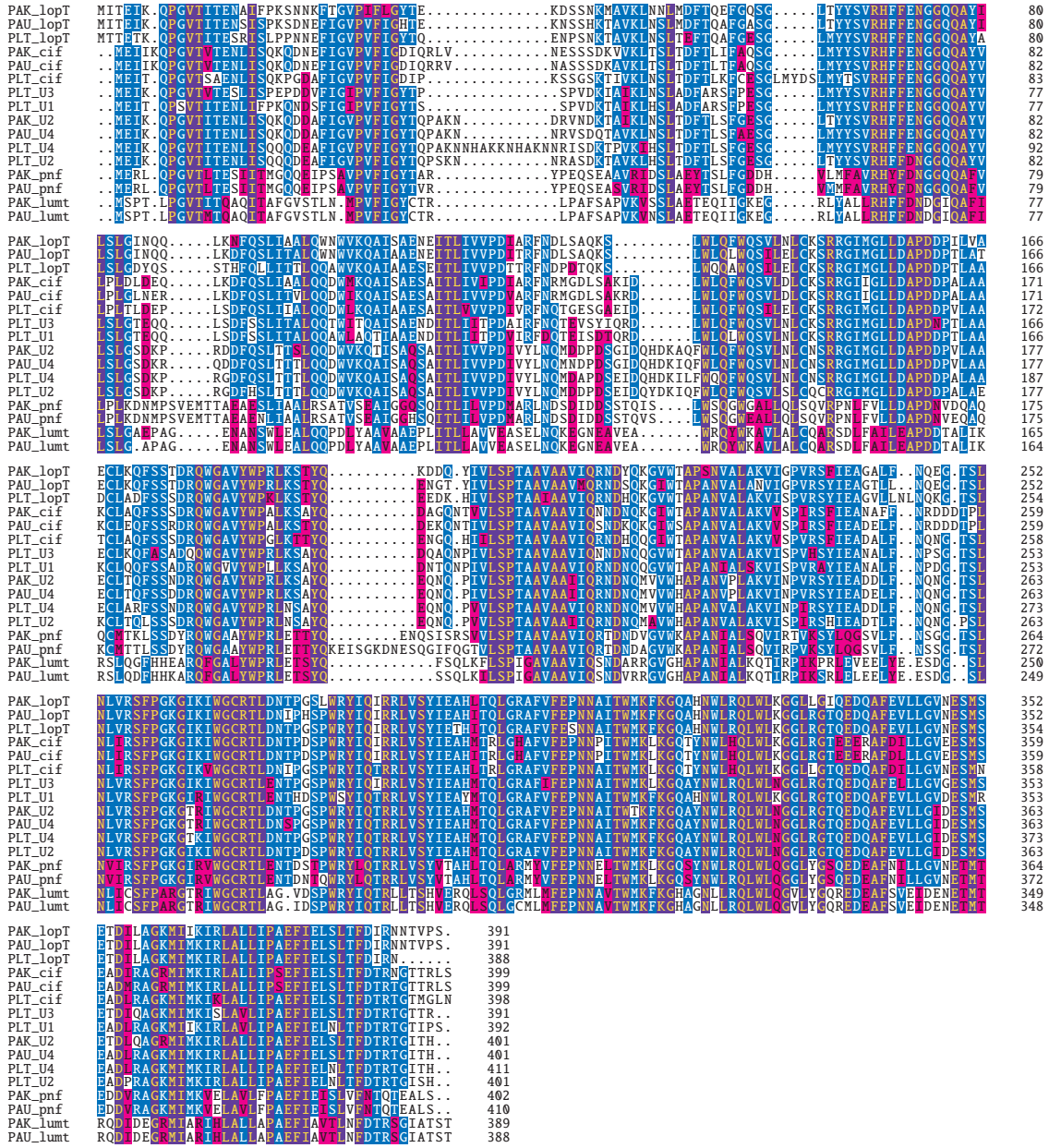



PAK_lopT
PAU_lopT
PLT_lopT
PAK_cif
PAU_cif
PLT_cif
PAK_U2
PLT_U2
PAU_U4
PLT_U4
PLT_U3
PLT_U1
PAK_pnf
PAU_pnf
PAK_lumt
PAU_lumt

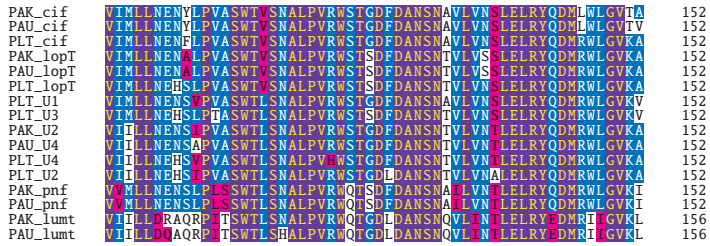
PVC Locus 3


PAK_cif
PAU_cif
PLT_lopT
PAK_U2
PAU_U4
PLT_U4
PLT_cif
PLT_U2
PLT_U3
PLT_U1
PAK_lumt
PAK_pnf
PAU_pnf

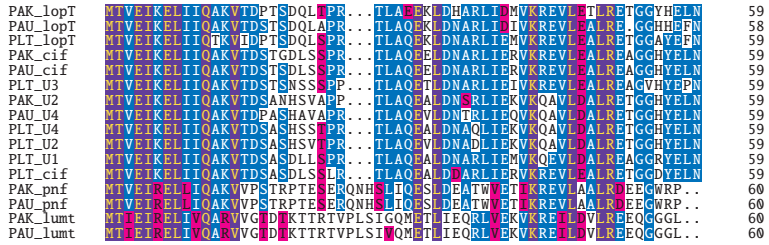
PVC Locus 4



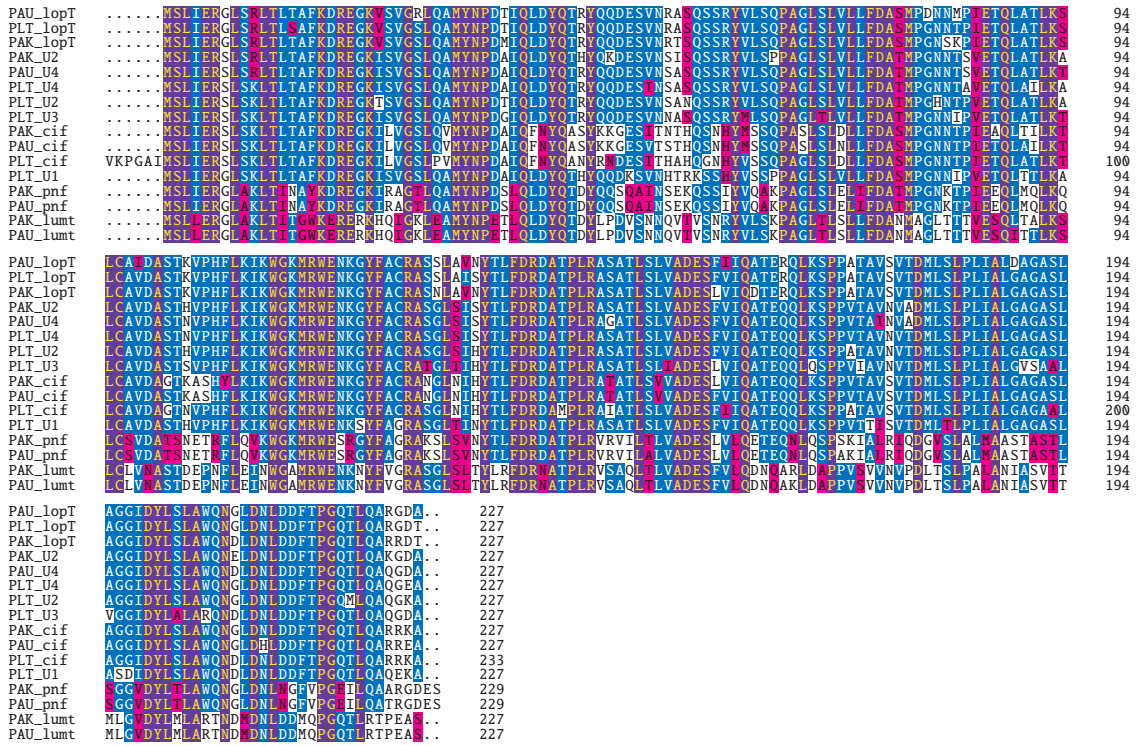
PVC Locus 5



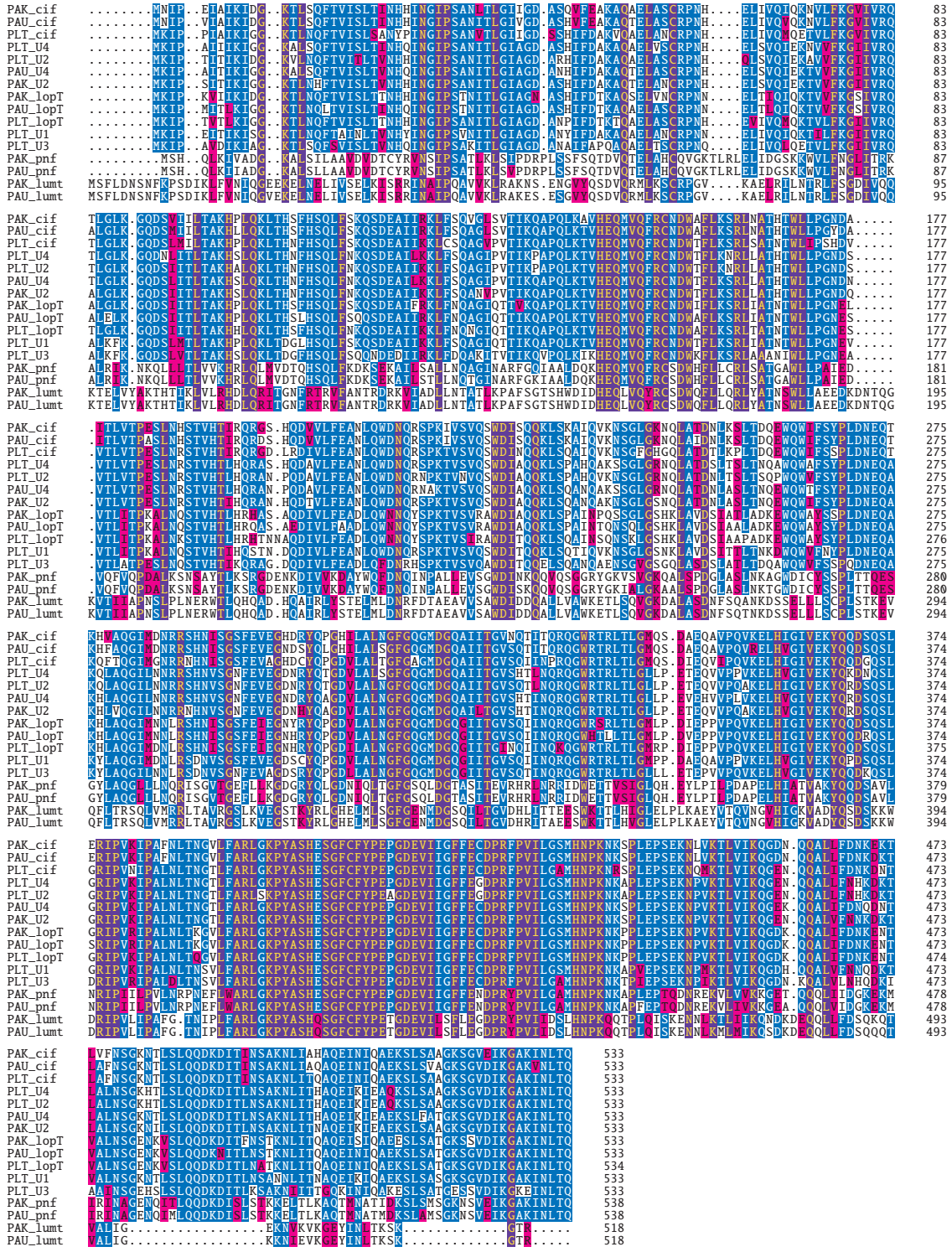
PVC Locus 6



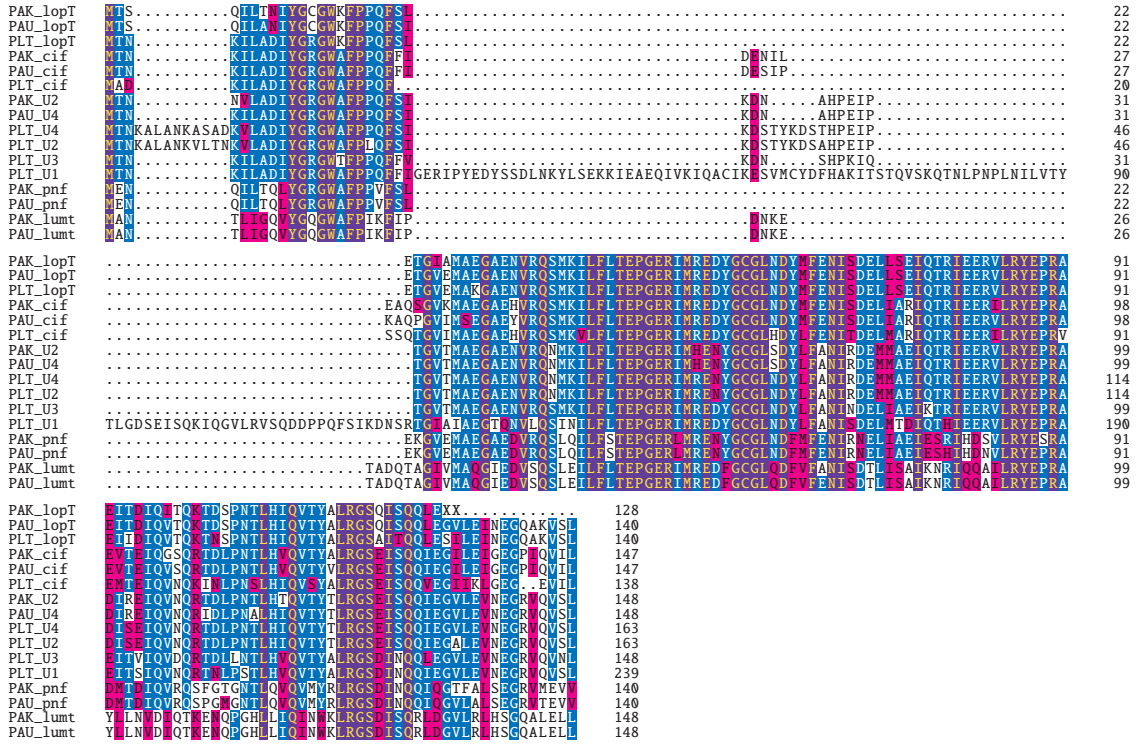
PVC Locus 7



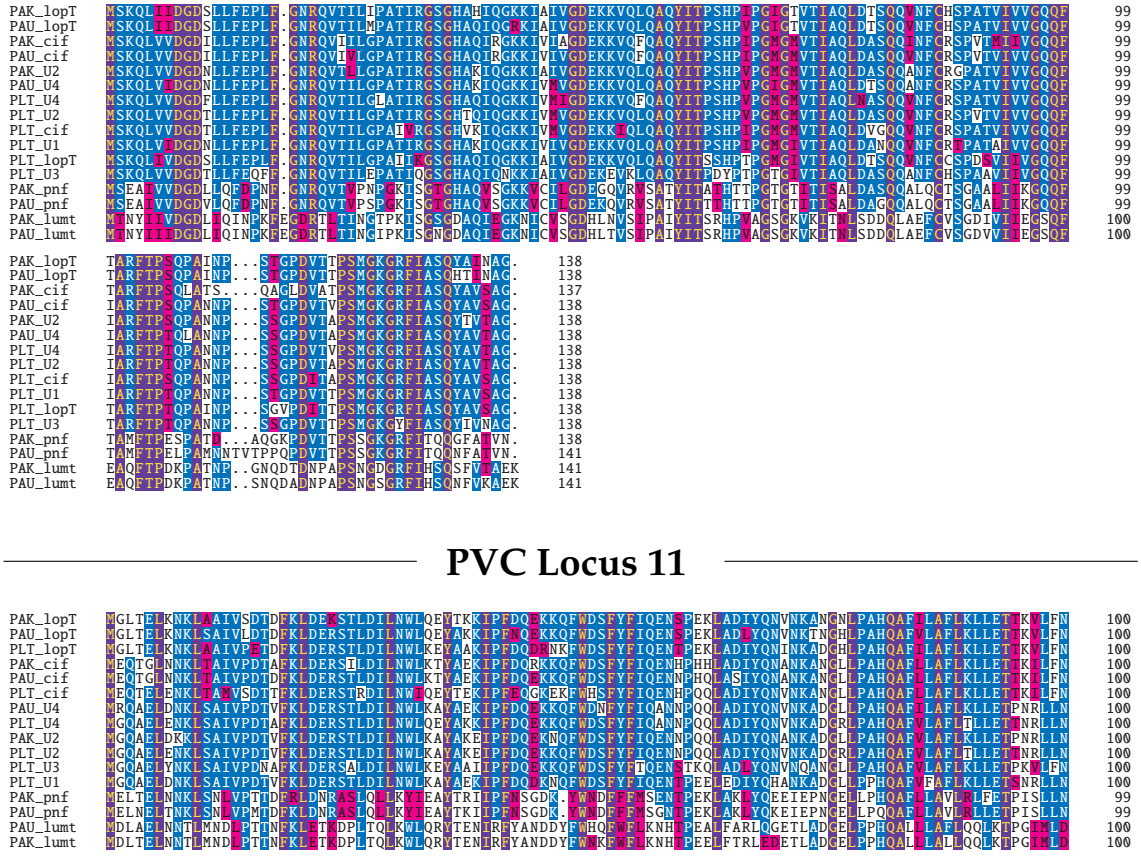
PVC Locus 8



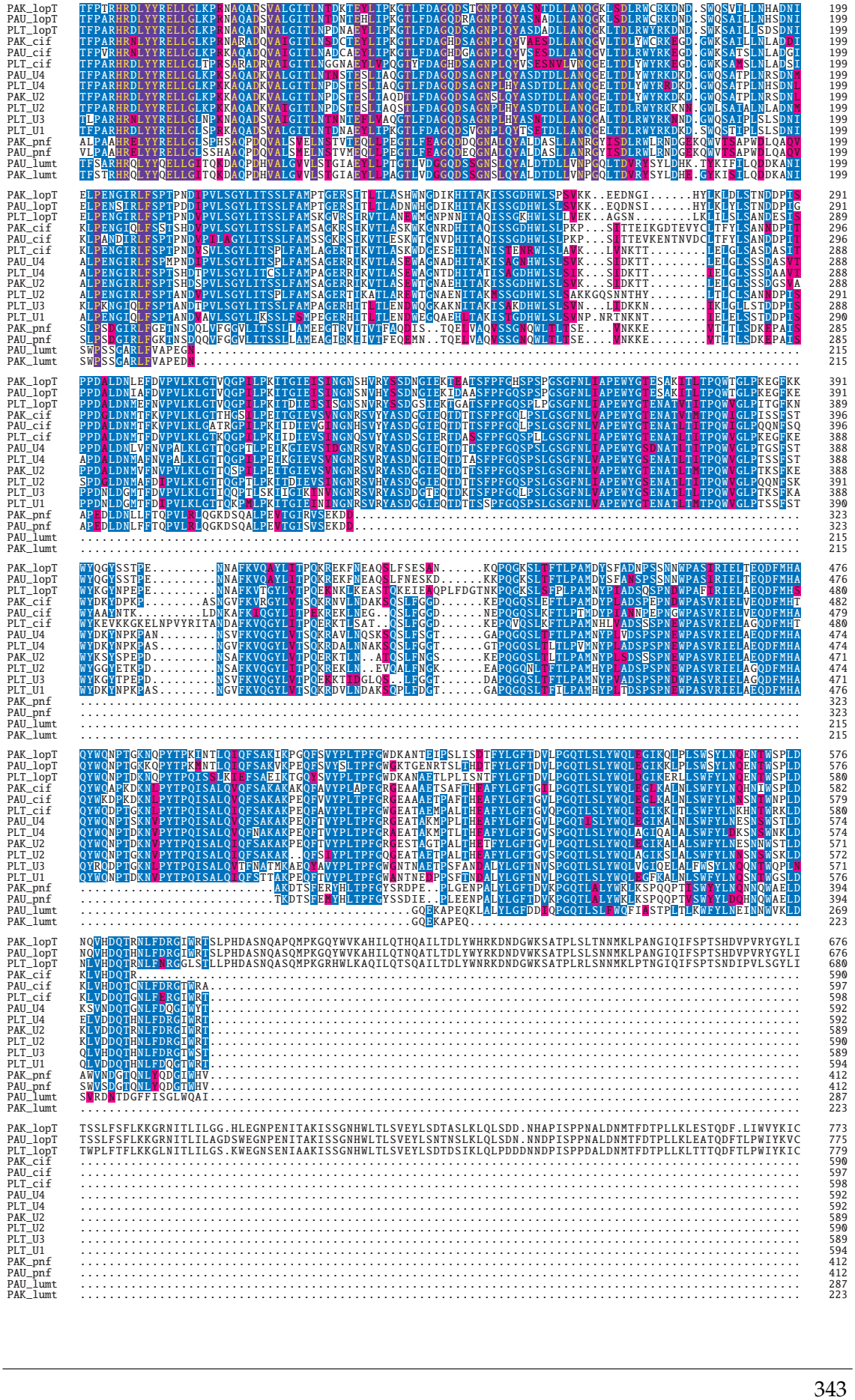
PVC Locus 9

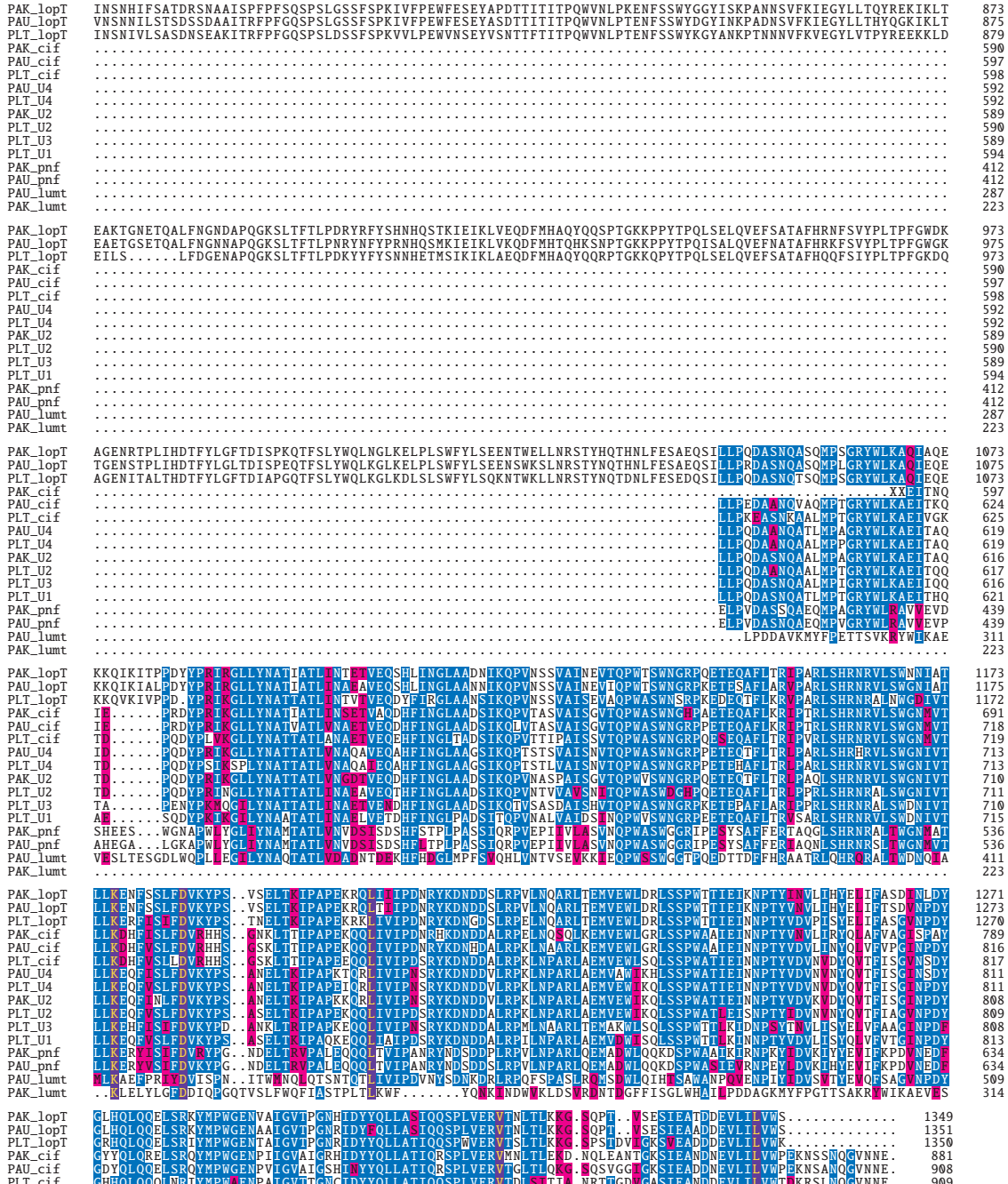


PVC Locus 10

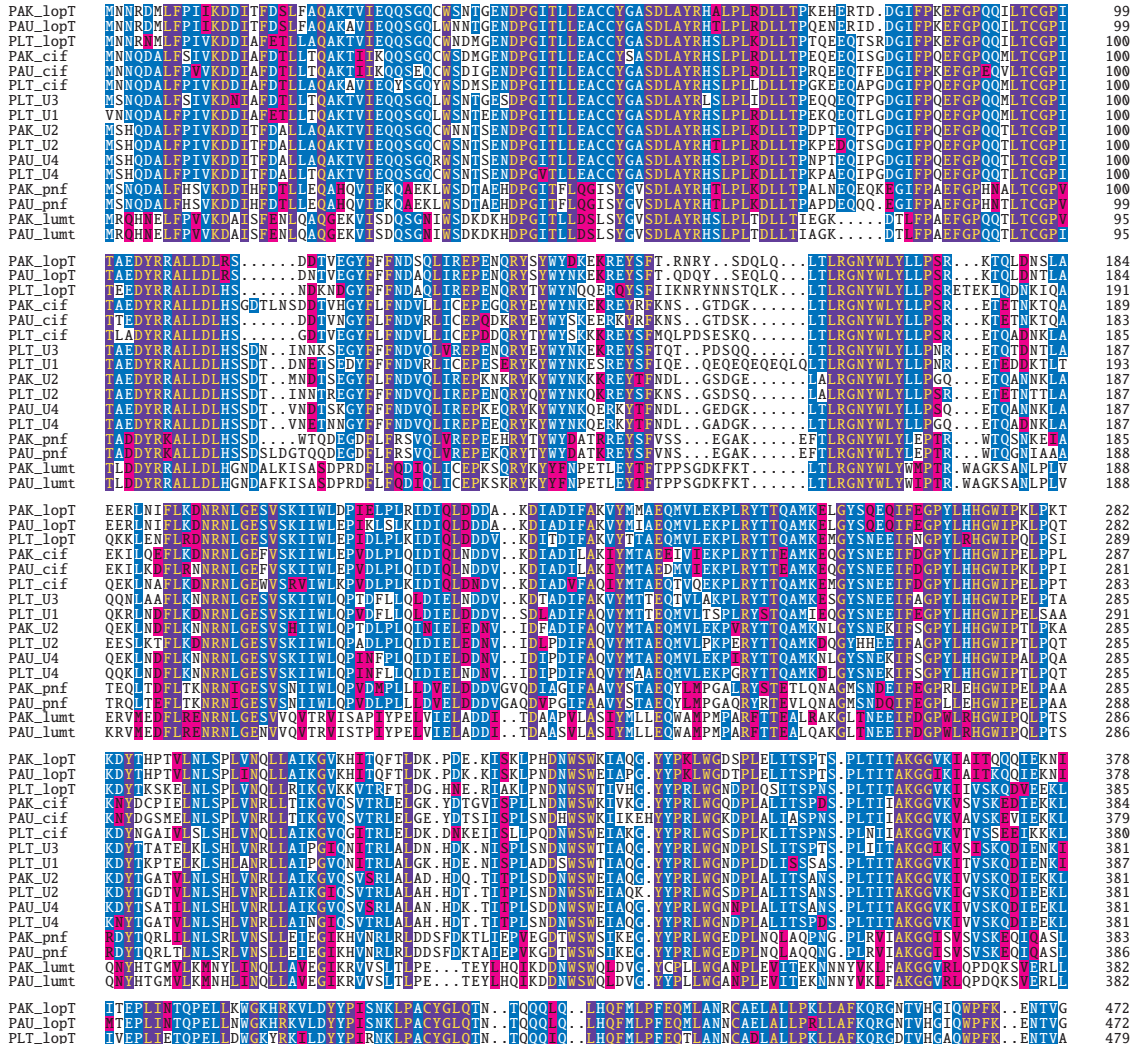


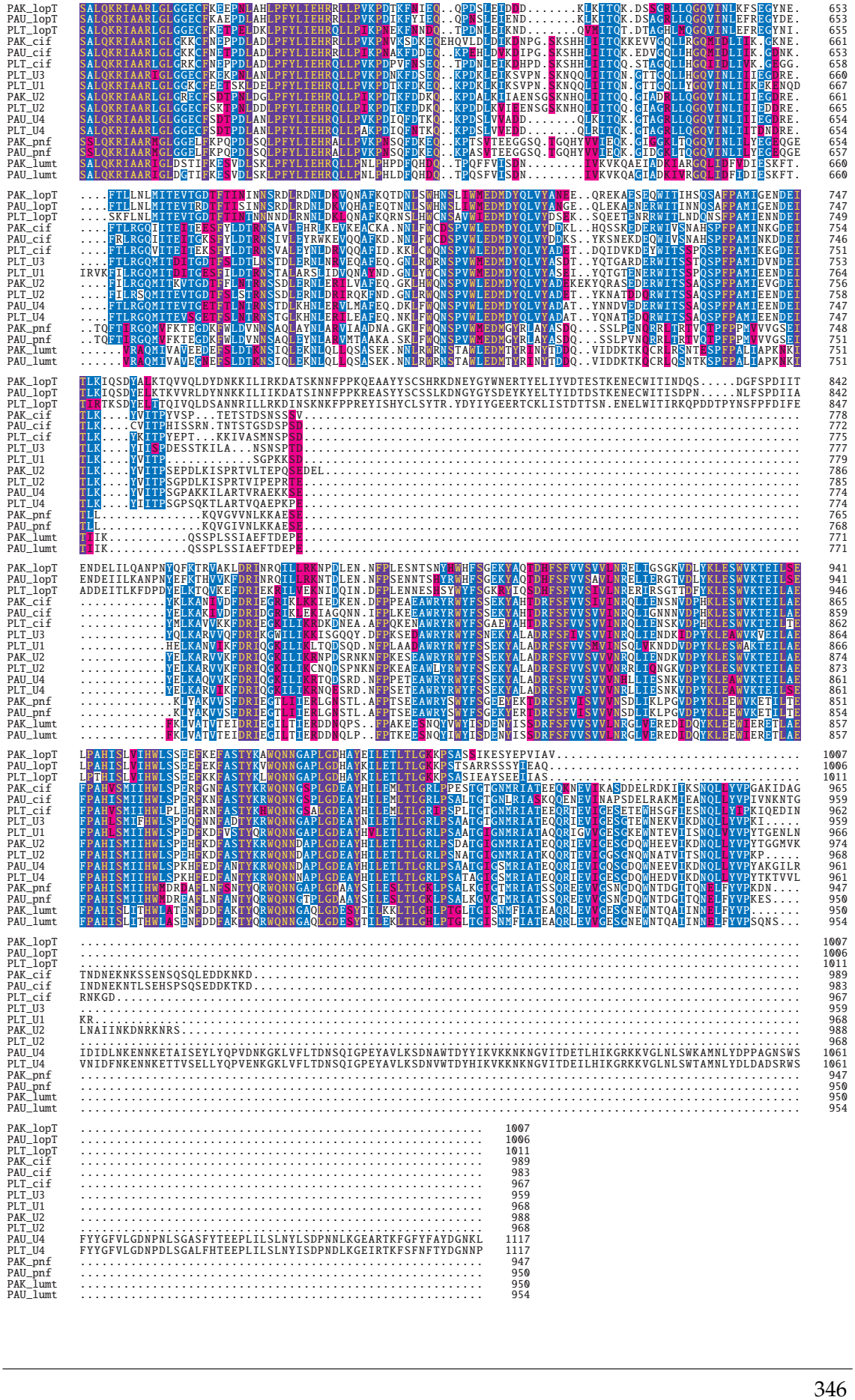
PVC Locus 11



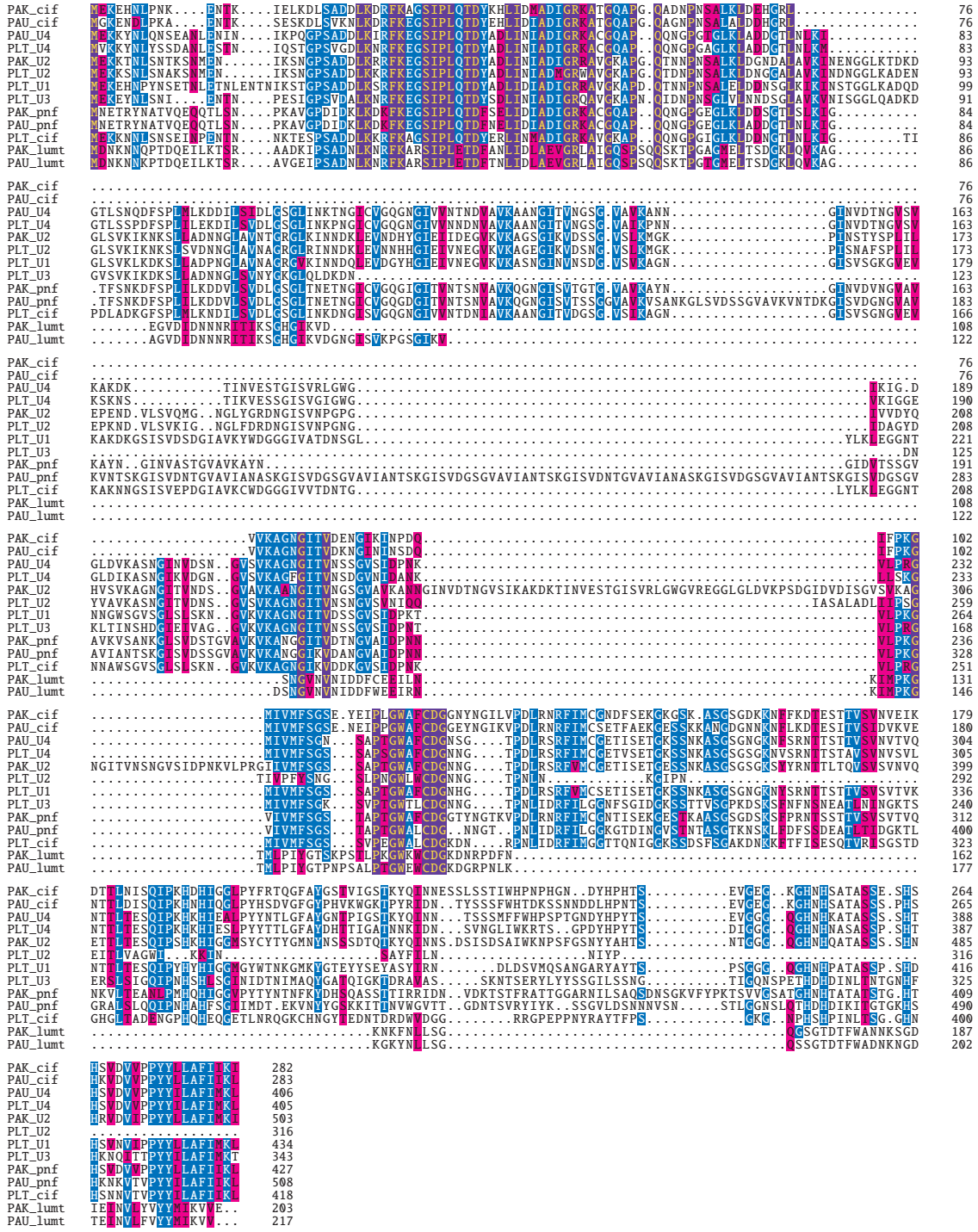


PVC Locus 12

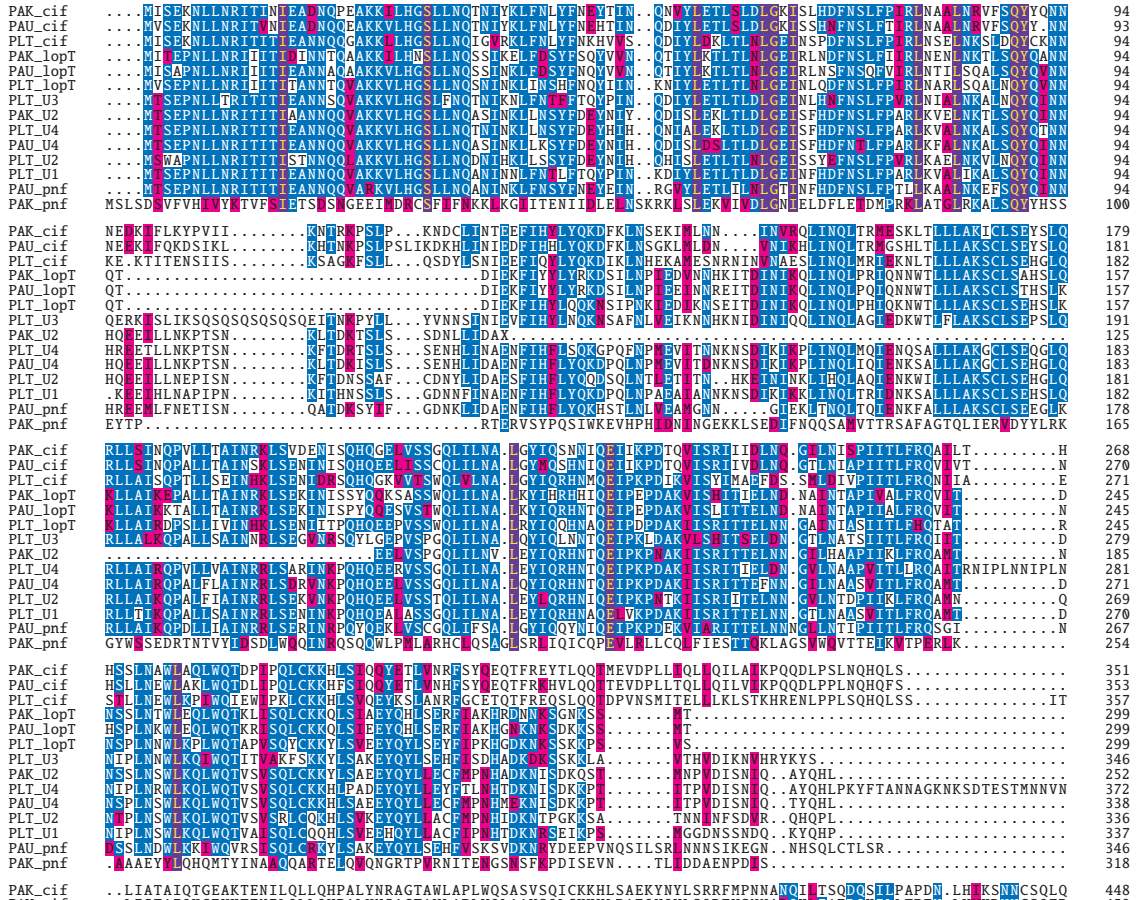




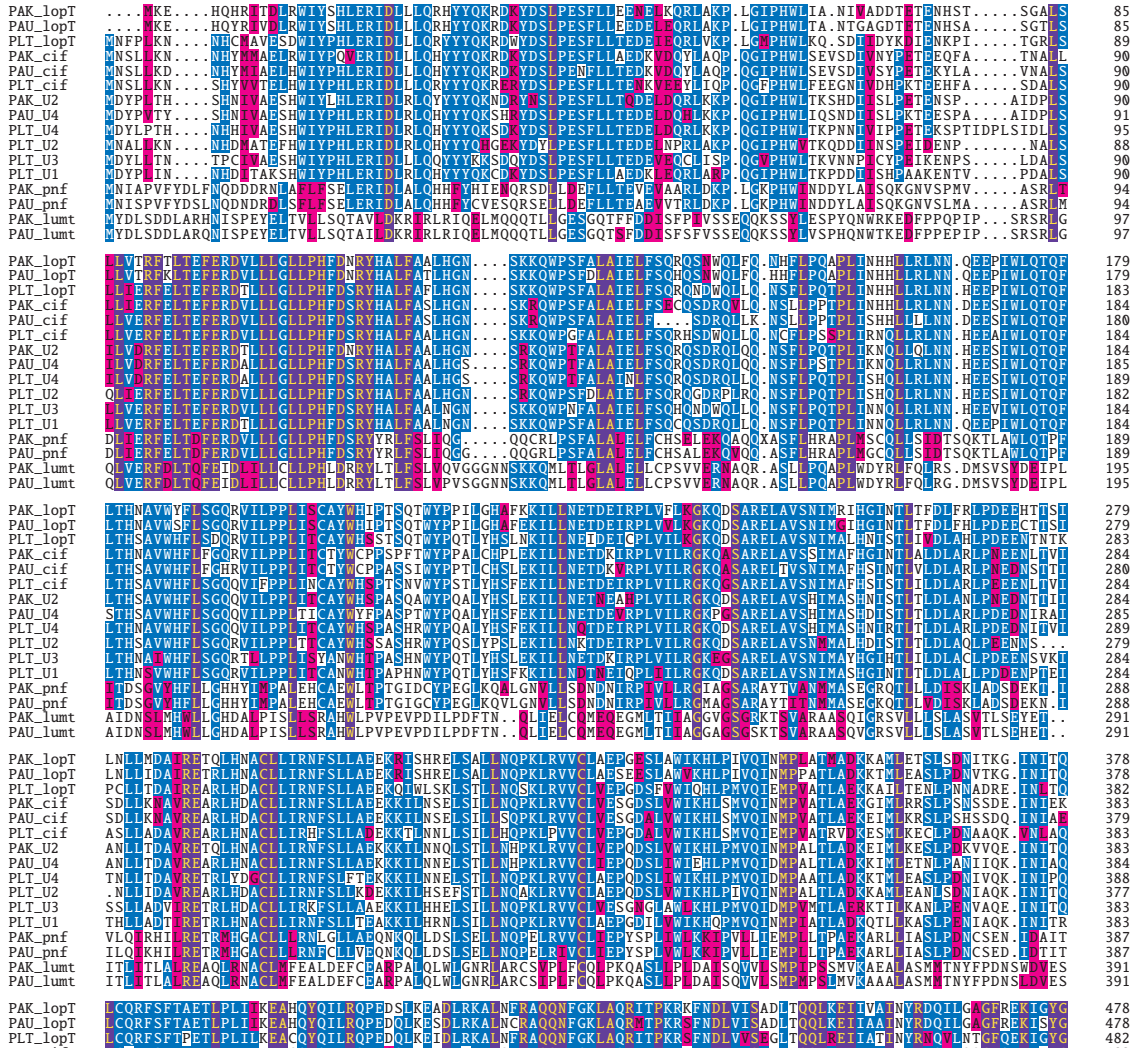
PVC Locus 13

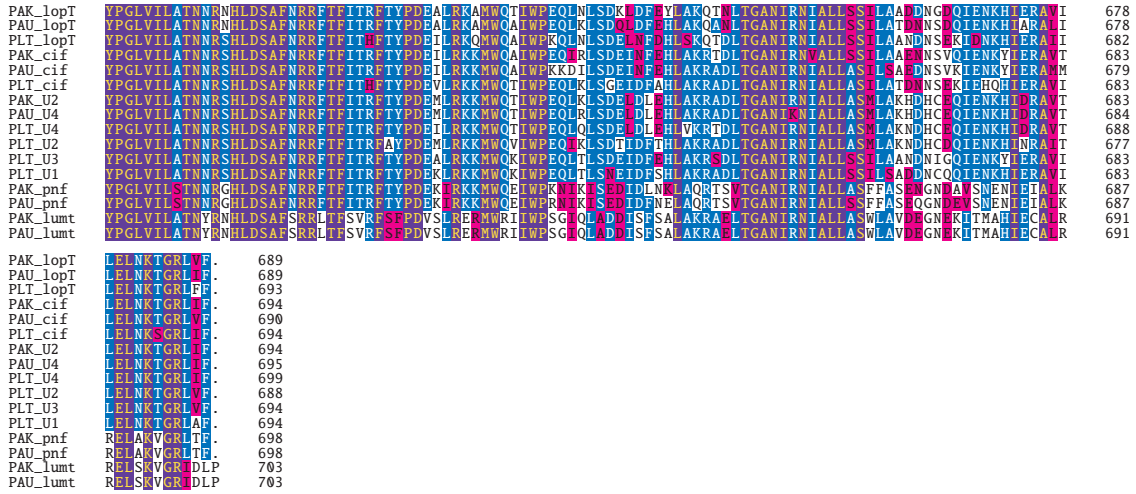

PAK_cif
PAU_cif
PAU_U4
PLT_U4
PAK_U2
PLT_U2
PLT_U1
PLT_U3
PAK_pnf
PAU_pnf
PLT_cif
PAK_lumt
PAU_lumt

PVC Locus 14



PVC Locus 15





PVC Locus 16