Bioinformatic insights into the diversity and evolution of bacterial toxins

by

Michael James Mansfield

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Biology

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Dr. David Guttman
Professor, Department of Ecology & Evolutionary Biology
University of Toronto

Supervisor(s): Dr. Andrew Doxey
Professor, Department of Biology
University of Waterloo

Internal Member: Dr. Trevor Charles
Professor, Department of Biology
University of Waterloo

Internal Member: Dr. Brendan McConkey
Professor, Department of Biology
University of Waterloo

Internal-External Member: Dr. Elizabeth Meiering
Professor, Department of Chemistry
University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see the Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Michael Mansfield was the sole author of Chapters 1 and 5, which were written under the supervision of Dr. Andrew Doxey and were not written for publication.

This thesis consists in part of seven manuscripts written for publication and exceptions to sole authorship of material are as follows.

Chapter 1

This chapter has been written for the thesis solely by Michael Mansfield, with input from Dr. Andrew Doxey. Some of the historical notes and citations were part of a published review by Doxey et al. [1], and Figure 1 from this published review has been adapted for the thesis as Figure 1.1. The contributions of all co-authors is as follows:

1. Doxey, A. C., Mansfield, M. J., & Montecucco, C. (2018). Discovery of novel bacterial toxins by genomics and computational biology. Toxicon, 147, 2-12. [1].

   https://doi.org/10.1016/j.toxicon.2018.02.002

   (a) Michael Mansfield created all figures, performed bioinformatic analyses, and assisted with the writing of the manuscript with input from all co-authors. Dr. Andrew Doxey and Dr. Cesare Montecucco were co-corresponding authors and contributed to the writing of the manuscript and figure design.

Chapter 2

This chapter concerns botulinum neurotoxins. It is structured around four subsections. The first, second, and third subsections are published manuscripts, with the fourth currently in preparation. Citations to the published manuscripts are indicated below with a description of all co-authors' contributions. The contributions of all co-authors is as follows:

1. Zhang, S., Lebreton, F., Mansfield, M. J., Miyashita, S. I., Zhang, J., Schwartzman, J. A., Tao L., Masuyer G., Martínez-Carranza M., Stenmark P., Gilmore M.S., Doxey A.C., & Dong M.D. (2018). Identification of a botulinum neurotoxin-like toxin in a commensal strain of *Enterococcus faecium.* Cell host & microbe, 23(2), 169-176. [2].

   https://doi.org/10.1016/j.chom.2017.12.018

(a) Michael Mansfield is co-first author with Dr. Sicai Zhang and Dr. Francois Lebreton. Dr. Andrew Doxey and Michael Mansfield discovered the BoNT/En sequence and Dr. Pål Stenmark identified it independently. Michael Mansfield performed the bioinformatic analyses and created the figures used in the thesis. The thesis includes a minimally modified excerpt of the published article's text sufficient to interpret the bioinformatic portion. This chosen excerpt excludes the extensive work performed by the study's co-authors, the details of which are available in the published paper [2]. The contributions of the authors in the published paper are as follows:

i. Dr. Andrew Doxey and Dr. Pål Stenmark identified BoNT/En independently. Dr. François Lebreton, Dr. Julia Schwartzman, and Dr. Michael Gilmore established the collection, sequenced IDI0629, and carried out comparative genome analysis. Dr. Andrew Doxey, Dr. Pål Stenmark, Michael Mansfield, and Dr. François Lebreton carried out bioinformatic analysis. Dr. Sicai Zhang carried out all other experiments. Dr. Jie Zheng assisted with DAS assays. Dr. Liang Tao and Dr. Shin-Ichiro Miyashita assisted with protein purification. Dr. Geoffrey Masuyer, Markel Martínez-Carranza, and Dr. Pål Stenmark generated the BoNT/X antibody. Dr. Min Dong, Dr. François Lebreton, and Dr. Andrew Doxey wrote the manuscript with input from all co-authors.

2. Mansfield, M. J., Wentz, T. G., Zhang, S., Lee, E. J., Dong, M., Sharma, S. K., & Doxey, A. C. (2019). Bioinformatic discovery of a toxin family in *Chryseobacterium piperi* with sequence similarity to botulinum neurotoxins. Scientific reports, 9. [3].

https://doi.org/10.1038/s41598-018-37647-8

(a) Michael Mansfield is co-first author with Travis Wentz. Travis Wentz and Dr. Shashi Sharma performed the genomic sequencing and assembly of the *Chryseobacterium* genome (which contributed to Figure 2.8a and Table A.2). Michael Mansfield curated the neurotoxin-related protein data set and performed all bioinformatic analysis except for the homology models in Figure 2.5a, which were created by Elliot Lee. Biochemical assays on Cp1 were performed by Dr. Sicai Zhang. All authors contributed to the writing of the manuscript. Dr. Andrew Doxey, Dr. Shashi Sharma, and Dr. Min Dong conceived of and coordinated the project.

3. Mansfield, M. J., & Doxey, A. C. (2018). Genomic insights into the evolution and ecology of botulinum neurotoxins. Pathogens and disease, 76(4), fty040. [4].

https://doi.org/10.1093/femspd/fty040

   (a) Michael Mansfield compiled and analyzed the data, created Figure 2.12, and wrote the manuscript with Dr. Andrew Doxey.

4. Mansfield, M. J., Lee, E. J., & Doxey, A. C. (2019). Comparative genomics and evolution of the BoNT-associated P47/OrfX gene cluster. Manuscript in preparation.

   (a) Elliot Lee discovered the sequence in *Bacillus* sp. 2SH and performed preliminary bioinformatic analyses, which have not been included in the thesis. Michael Mansfield curated the data set, performed all bioinformatic analyses presented here, created the figures, and wrote the manuscript with input from Dr. Andrew Doxey.

Chapter 3

This chapter concerns the expansion of the diphtheria toxin gene family. The first subsection of the chapter is consists of a published manuscript, and the second section consists of a portion of a manuscript currently in preparation. The contributions of all co-authors is as follows:

1. Mansfield, M. J., Sugiman-Marangos, S. N., Melnyk, R. A., & Doxey, A. C. (2018). Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus. FEBS letters, 592(16), 2693-2705. [5].

   https://doi.org/10.1002/1873-3468.13208

   (a) Michael Mansfield generated and analyzed the data set, created all figures, and wrote the manuscript with input from all co-authors.

The second subsection of Chapter 4 presents an analysis of the crystal structures of two diphtheria toxin-related sequences. These structures are the subject of ongoing work in Dr. Roman Melnyk's lab. The proteins were expressed, purified, crystallized, and their structures were solved by Dr. Seiji Sugiman-Marangos and Shivneet Gill. Dr. Sugiman-Marangos and Shivneet Gill generously provided their solved structures. Michael Mansfield analyzed the data, created all figures, and wrote the subsection here with input from Dr. Andrew Doxey.

Chapter 4

This chapter concerns large clostridial toxins, and particularly their translocase domain. This chapter consists of a manuscript accepted for publication. The contributions of all co-authors is as follows:

1. Orrell, K. E., Mansfield, M. J., Doxey, A. C., & Melnyk, R. A. (2019). The *C. difficile* toxin translocase is an evolutionarily conserved apparatus for bacterial protein delivery into host cells. Nature Communications (accepted).

   (a) Michael Mansfield created and analyzed the sequence data set, and created the bioinformatics-focused figures with input from all co-authors. Kathleen Orrell performed all biochemical assays and created the corresponding figures under the supervision of Dr. Roman Melnyk. All authors contributed to writing the manuscript.

Abstract

Bacterial toxins are a broad category of molecules ranging from small organic compounds and peptides to large multi-domain or multi-meric enzymes. Several important diseases are caused primarily by bacterial toxins including botulism and diphtheria. Paradoxically, the same toxins have proven useful for the treatment of muscular disorders and cancer, respectively. Given their importance in medicine and their utility as drugs, it is desirable to attain a greater functional and mechanistic understanding of toxin families. However, a full description of any sequence's functionality must incorporate an understanding of the evolutionary processes that produced them, and currently little is known about these forces. Using a bioinformatic approach, this thesis presents analyses of three bacterial toxin families: clostridial neurotoxins, which cause botulism and tetanus; diphtheria toxins, which cause diphtheria; and large clostridial toxins, which contribute to the infections produced by various clostridia, including *Clostridioides difficile*. The detection of toxin-related sequences from bacterial genomes allows the discovery of toxin variants that may have gone undetected by other methods of toxin identification. Based on the available genomic data, toxin families that cause disease in humans appear to be broader than previously imagined. Toxin-related sequences are capable of performing unique functions compared to the toxin variants more traditionally associated with human disease. By examining human toxins in evolutionary terms, it is possible to identify the functional innovations that have occurred to result in human specificity, as well as delve more deeply into the relationships between toxin sequences and their functions. Thus, the studies presented here provide examples of how the field of toxin biology, like many other disciplines, has much to gain from the genomic revolution.

Acknowledgements

This thesis document and the research it contains have been made possible through the collective effort of many people. I would first like to express my deepest and most sincere gratitude to my supervisor Dr. Andrew Doxey. His endless enthusiasm and optimism have been a great source of encouragement throughout my graduate studies, during which he exhibited considerable patience and understanding. I would also like to thank my committee members Dr. Trevor Charles and Dr. Brendan McConkey for their feedback and support over the past five years. I thank Dr. Min Dong for hosting me at his lab at Boston Children's Hospital and his contributions to the chapter on botulinum neurotoxins. I also thank the Dong lab members for making my stay in Boston productive and enjoyable.

I am grateful to have received funding to support my research and conference activities, including from the University of Waterloo (and more generally, the people and governments of Canada), the International Neurotoxin Association, and the Botulinum Research Center.

I am grateful for the many collaborators that contributed to the research in this thesis. I would particularly like to thank Dr. Sicai Zhang, Travis Wentz, and Dr. Shashi Sharma for their contributions to the chapter on neurotoxins, Dr. Sugiman-Marangos and Dr. Roman Melnyk for their contributions to the chapter on diphtheria toxin, and Kathleen Orrell and Dr. Melnyk for their contributions to the chapter on large clostridial toxins.

I thank past and present Doxey lab members for their collaborations, and for making the research process more amusing.

I would like to thank all of the friends and family that supported me, encouraged me, and helped me to maintain an appropriate work-life balance over the years.

Finally, I thank my wife Katherine Mansfield. Without her love and support, my research and this thesis would not have been possible. I am enormously thankful for and humbled by her continued support of my endeavours in science and in life.

Dedication

I dedicate my thesis to my family, to those who are able to read these words and to those that are unable. Their example taught me that excellence and competency are to be admired, but never taken too seriously; that beautiful and gentle things are to be cherished; that fair judgment requires humility. To them above all I owe my life and my life's work.

Table of Contents

List of Tables

List of Figures

Glossary

BoNT Botulinum neurotoxin. Causes flaccid paralysis by cleaving proteins involved in neuronal exocytosis at neuromuscular junctions. 15, 20, 41, 58, 103, 124

CNT Clostridial neurotoxin. Refers to the homologous protein family including botulinum neurotoxins and tetanus neurotoxin. 20, 41

DT Diphtheria toxin. Causes cell death by transferring an ADP moiety to eukaryotic elongation factor 2, thereby preventing translation. 33, 73, 94, 103, 126

eEF-2 Eukaryotic elongation factor 2. Required for translation in eukaryotes. Contains a posttranslationally modified residue termed diphthamide that is the target of ADP-ribosylation by diphtheria toxin. 74, 100, 127

HA Hemagglutinin. Characteristic of one type of botulinum neurotoxin gene clusters. Forms a progenitor toxin complex with the neurotoxin and its associated non-toxic non-hemagglutinin protein. 15, 42, 58

HB-EGF Heparin-binding epidermal growth factor. Transmembrane form functions as receptor for diphtheria toxin. 74, 100, 127

LCT Large Clostridial Toxin. A family of cytotoxins that includes *Clostridioides difficile* TcdA and TcdB, *Clostridium perfringens* TpeL, *Paraclostridium sordellii* TcsH and TcsL, and *Clostridium novyi* TcnA, which intoxicate cells by glucosylation of Rho-GTPases. 102, 127

NTNH Non-toxic non-hemagglutinin. A paralog of botulinum neurotoxin, and component of progenitor toxin complexes. 15, 20, 58, 126

OrfX OrfX proteins are characteristic of one type of botulinum neurotoxin gene cluster. Their functions are mostly unknown but potentially relate to membrane binding. 15, 42, 58, 125

SNAP25 Synaptosomal nerve-associated protein 25. Component of SNARE complexes related to exocytosis. Proteolytic target of botulinum neurotoxin serotypes A, C, E, and weakly BoNT/En. 15, 20, 44, 125

SNARE Soluble N-ethylmaleimide-sensitive factor Attachment Protein Receptor. Contribute to protein complex that mediates exocytosis by facilitating endosomal membrane fusion. 15, 41, 58, 123, 136

TeNT Tetanus neurotoxin. Causes spastic paralysis by cleavage of proteins related to neuronal exocytosis in the spinal column, causing spastic paralysis. 20, 41, 58

VAMP Vesicle-associated membrane protein, also known as synaptobrevin. Proteolytic target of botulinum neurotoxin serotypes B, F, G, D, X, as well as tetanus neurotoxin. 15, 20, 44, 125

Chapter 1

Introduction

Material in this chapter has been published as part of Doxey et al. [1]. The published manuscript is available from here:

1. Doxey, A. C., Mansfield, M. J., & Montecucco, C. (2018). Discovery of novel bacterial toxins by genomics and computational biology. Toxicon, 147, 2-12. [1].

   https://doi.org/10.1016/j.toxicon.2018.02.002

1.1 Overview

In this chapter, I introduce key terms that are used throughout the thesis. I frame these discussions around the terms' historical definitions, and consider how their definitions have changed over time. I begin by discussing the history of pathogenicity and virulence factors. I then focus upon toxic proteins, both in terms of the history of toxin identification as well as their mechanisms, structures, and functions. I discuss the impact of genomics on the field of bacterial toxins, followed by a brief overview of the theoretical and methodological background for the bioinformatic techniques necessary to interpret the data and results of subsequent chapters. In the last section, I provide an outline of the thesis document.

1.2 Pathogens, virulence factors, and toxic proteins

In its simplest terms, pathogenicity can be defined as the capacity for a microbe to cause damage to a host [6]. The modern usage of the term arose shortly after the development of the germ theory of disease. At this time, experiments on toxigenic bacteria contributed significantly to the burgeoning field of microbiology, and strongly informed ideas about pathogenesis. Klebs and Löffler's identification of the microbe that causes diphtheria led to the discovery of its toxin, a substance produced by the microbe in infected tissues but capable of causing damage far from the site of infection [7, 8]. Around the same time, Koch's isolation and growth of the tuberculosis bacterium [9] systematized the process for establishing pathogen-disease causality, and developed into what is now known as Koch's postulates. Stated plainly, the fulfilment of Koch's postulates requires: 1) isolation of a microbe from diseased tissues, common to all cases of the disease; 2) the organism must be grown in pure culture; 3) inoculation of the microbe into a host must produce the same disease; 4) the organism must be isolated from the infected animal and grown in pure culture once more [10]. Together, the observations that specific microbes cause diseases through toxins or through infection supported the idea that pathogenicity was an intrinsic property of specific microbes.

However, it now seems clear that the idea that specific microbial species are pathogens and others are not is an oversimplification [6, 11, 12]. Although some bacterial lineages are strictly pathogenic - for example, obligate intracellular parasites like chlamydia species [13] - pathogenicity is a property found broadly across the kingdom Bacteria, and is not limited to any particular taxon. Conversely, closely related bacteria can vary greatly in pathogenicity, as is the case for strains of *Escherichia coli* [14]. Some pathogenic microbes produce asymptomatic infections, and bacteria that would normally not be considered pathogens can cause serious infections in immunocompromised individuals [15]. Therefore, it is perhaps more appropriate to consider pathogenicity the result of an interaction between a particular microbe with a particular host. On the pathogen side, variables like strain-specific differences must be considered, and on the host side, immunity/susceptibility and the protective effects of the host's microbiome must be considered [16].

In a broad sense, all properties of a pathogen that contribute to pathogenicity may be considered virulence factors [17, 18]. However, this all-encompassing definition may lack specificity. As pointed out by Weiss and Hewlett [19], although core metabolic properties common to many bacteria like purine metabolism are not typically considered virulence factors, they are required for virulence in *Salmonella* [20]. Similarly, even though lipopolysaccharides are a component of all gram negative membranes, *Salmonella* specifically modifies its O antigens to evade immune detection [21], and its O antigens are therefore considered virulence factors. The apparent difficulty in defining what constitutes a virulence factor has led some to call for the development of a virulence factor classification scheme akin to enzyme commission numbers [18], and some resources that attempt to categorize virulence factors like the Virulence Factor Database are widely used [22]. Nevertheless, in its most common usage, the virulence factors of a pathogen include general physiological and life cycle properties like flagella [23], sporulation [24], dormancy [25], and membrane structure [26], as well as microbial products that are specifically produced to target and modify a host's physiology. Toxins are one example of such a targeted virulence factor.

The word "toxin" is derived from the ancient Greek τοξικός (*toxikós*) relating to the bow or use of the bow, or more specifically, poison for smearing on arrows. The word was introduced into the modern scientific lexicon by Brieger, who primarily worked on toxic metabolites derived from putrefied tissues ("ptomains") [27, 28]. Brieger introduced the term toxin to describe tetanus toxin, which he had unsuccessfully attempted to isolate; evidently, Brieger thought it was worth distinguishing from other poisonous substances. Indeed, biological toxins can be broadly classified into several categories: small organic molecules, modified amino acids, and other low molecular weight metabolites; short polypeptide chains, including bacteriocins; and large protein toxins. It is the category of large protein toxins that is of interest for this thesis, and more specifically large protein toxins produced by bacteria. The word "toxin" is used throughout the thesis to refer to this case.

1.3 History of bacterial toxin identification

In the early 1880s, Klebs and Löffler had identified the microbe responsible for diphtheria, and had demonstrated that the disease was caused by microbial toxin production, but they were unable to isolate the toxin itself. Roux and Yersin achieved the isolation of diphtheria toxin (DT) in 1888 [29]. It took nearly one hundred years to develop an explanation for Roux and Yersin's success. After DT production was found to be related to culture iron concentration [30], the molecular mechanism for this phenomenon was proposed by Murphy et al. in 1976: DT expression could be controlled by an iron-inducible repressor [31, 32]. Thus, DT is only expressed in conditions of iron depletion. Apparently, Klebs and Löffler had not allowed their cultures to reach saturation.

The discovery of diphtheria toxin provided evidence that bacterial pathogens might produce substances primarily (or solely) responsible for infectious diseases. A similar approach was quickly and successfully applied to other diseases and toxin families, including tetanus toxin [27, 33, 34, 35] and botulinum toxin [36]. A generalization of the approach, which can be thought of as an application of Koch's postulates, is visualized in Figure 1.1 (adapted from Figure 1 of Doxey et al. [1]). After observing a clinical phenotype, the suspected microbial agent must be isolated and cultivated. Then, its toxin must be purified and shown to elicit the associated disease phenotype. Once the toxin is identified, specific

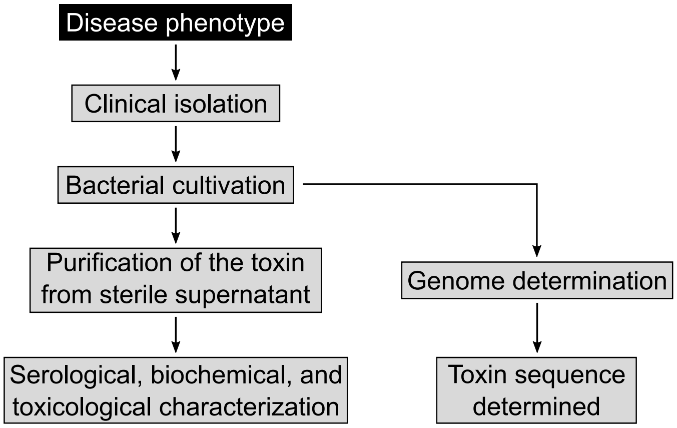

Disease phenotype
Clinical isolation
Bacterial cultivation
Purification of the toxin from sterile supernatant
Serological, biochemical, and toxicological characterization
Genome determination
Toxin sequence determined

Figure 1.1: A generalized bacterial toxin identification method. Following an outbreak of a disease, the causative microbe must be isolated and cultivated, then production of a toxin must be determined, in accordance with Koch's postulates. Once this toxin is determined to be the cause of disease, additional biochemical (and recently, genomic) characterization is possible. Adapted from Doxey et al. [1].

mechanistic details can be determined, which is discussed in greater detail below. The initial success of this toxin identification strategy held much promise for understanding toxin-related diseases, but proved difficult for cases like anthrax, cholera, and pertussis. The major toxins that contribute to those illnesses were only described much later (in 1954, 1959, and 1982, respectively) [37, 38, 39, 40, 41].

Bacterial toxins are now understood to be the protein products of transcribed and translated genes, encoded within bacterial genomes. As such, toxin characterization methods developed with the emergence of genetic, microbiological, and molecular biology techniques. Techniques like molecular cloning, protein expression, purification and crystallography, molecular imaging, and sequencing allow different aspects of toxin function to be interrogated in much greater resolution. Since their discovery in the 1880s, the mechanisms used by many toxin families have been elucidated. In general, toxins can be divided into two types: those that act on the extracellular plasma membrane, and those that act on intracellular substrates within host cells [42, 43]. Toxins that act on intracellular substrates must first gain access the host cytosol. Various bacterial secretion systems are used to inject effectors directly into host cells, some of which are toxic. Secretion systems used to transport effectors into host cells are mostly associated with gram negative pathogens, although similar secretion systems are found in gram positive pathogens [44, 45]. For example, type III secretion systems (T3SS) are employed by pathogens like *Salmonella* [46], *Pseudomonas syringae* [47], and *Escherichia coli* [48], while the type IV secretion system is used in *Legionella* species. Toxins that are not transported directly into host cells by secretion systems must translocate into cells some other way, mediated by a specialized toxin domain or complexed partner protein(s). Toxins with their own translocation mechanism can be classified as “AB” toxins, containing functionally distinct active (“A”) and binding/translocating (“B”) components. The AB toxin architecture is common among bacterial toxin families, but the specific mechanisms for toxin binding, translocation, and enzymatic activity are variable.

1.4 AB toxins: structure and function

Many bacterial toxin families possess AB toxin architecture. The A component is an enzyme that chemically modifies a host substrate. The B component is responsible for binding one or more receptors on the host cell surface, as well as translocating the A component into the cytosol. The A and B components can be encoded in a single gene or encoded by separate genes. If the A and B components are encoded in separate genes, either A or B component may be a hetero- or homomeric complex comprised of one or more A/B components. As an example, the cholera toxin of *Vibrio cholerae* is a complex of a single A fragment and a homopentameric B fragment (AB_5) [49], while the pertussis toxin of *Bordetella pertussis* is comprised of a single chain A subunit with a heteropentameric B complex containing four different proteins (technically $A_{S1}B_{S2}B_{S3}B_{2xS4}B_{S5}$, although this notation is never used) [50].

Toxin A components modify host substrates through a wide range of enzymatic activities. Known toxin enzyme types include proteases, ADP ribosyltransferases, adenylate cyclases, glucosyltransferases, and DNases [1]. Enzymatic modification of the host substrate elicits a change in the host’s physiology, which can have subtle or dramatic phenotypic consequences depending on the toxin’s target. For example, pertussis toxin disrupts immune processes yielding increased and more persistent infectivity [54], whereas a single molecule of diphtheria toxin is sufficient to halt protein synthesis and kill human cells [58]. In general, the targets of toxins are involved in conserved, constitutively expressed cellular processes like cell cycle regulation or protein synthesis [42], contributing to the severity of intoxication. Toxin A components with the same enzyme type can have similar sequences

[1] For reference: diphtheria, pertussis, and cholera toxins are ADP-ribosyltransferases [51], botulinum toxin and anthrax lethal factor are metalloproteases [52, 53], *Bordetella pertussis* CyaA and anthrax edema factor are adenylate cyclases [54, 55], large clostridial toxins are glucosyltransferases [56], and cytolethal distending factors are DNases [57].

and function, distantly related sequences but similar functions, or else have limited sequence and functional similarity [2]. The enzymatic specificity of the toxin A component plays a role in determining host resistance and susceptibility, as the orthologous targets of toxins may vary between different hosts. This explains the resistance of rats and chickens to tetanus [59]. However, the enzymatic specificity of the A component by itself is inadequate to explain host specificity or pathological severity. Although some botulinum toxins and tetanus toxin cleave the same substrate at the same site, they yield nearly opposite phenotypes (flaccid versus spastic paralysis). In this case, the key difference is where cleavage occurs: for botulinum toxin, the toxin acts directly at neuromuscular junctions and paralyzes muscles, but tetanus is instead transported to the spinal cord, causing unbalanced muscular contraction [60, 61].

The B component of AB toxins is responsible for localizing the toxin to host receptors, and then facilitating the A component's entry into host cytoplasm. The binding receptors for toxins include a large variety of proteins and lipids, which can be glycosylated or non-glycosylated. Toxin binding may involve or require multiple simultaneous receptors (as in botulinum toxins and large clostridial toxins) [62, 63, 64], or toxin multi-merization (as in anthrax toxin, pertussis toxin, and cholera toxin) [65, 66, 67]. As well, closely-related toxins within the same family may utilize different receptors [68, 69, 70]. After a toxin binds its receptor(s), it is internalized by receptor-mediated endocytosis [71, 72]. Further trafficking through the host varies between toxin families. Within a single cell, some toxins translocate through early or late endosomes while others are transported to the endoplasmic reticulum prior to translocation. Toxins may additionally be trafficked between different host cells [73]. For many toxins, it is the acidification of late endosomes that facilitates translocation, as the decrease in pH induces conformational changes in the toxin B component. These changes allow the B component to form a pore through which the A component is translocated [74, 75], which may involve partial or complete unfolding of the A component [76, 77]. Toxin A and B components are generally linked by one or more disulfide bonds. If the A and B components of the toxin are produced as a single polypeptide chain, it must first be cleaved by bacterial or host proteases, and linked to one another by a disulfide bond. The details of proteolytic activation are known for some toxins

[2] An example for each of the three cases. 1) Botulinum and tetanus toxin A components are homologous protease toxins that share around 30% amino acid identity and have significant functional similarities; 2) The A components of diphtheria toxin and *Pseudomonas aeruginosa* ExoA components are distantly related ADP-ribosyltransferases that share around 18% amino acid identity, but both target the diphthamide residue of eukaryotic elongation factor 2; 3) diphtheria toxin and cholera toxin A components are both ADP-ribosyltransferases, but have little else in common, even using different catalytic mechanisms for ADP-ribosyltransferase activity. See Simon et al. [51] for an in-depth review of ADP-ribosyltransferase toxins.

(diphtheria toxin is activated by the host's furin protease [78], and large clostridial toxins contain an autoproteolytic cysteine protease domain [79]) and unknown for others (the activating protease for botulinum toxins is currently unknown). At any rate, the disulfide bond(s) linking the A and B components becomes reduced during or after translocation, thereby liberating the A component within the cytosol. Once inside the cytosol, the A component is able to perform its toxic enzymatic function.

1.5 The genomic era of toxin biology

All of microbiology has been profoundly influenced by the development of whole-genome sequencing. In the last 15 years, the cost to determine an organism's whole genome has greatly decreased, and new genomes are being sequenced at an incredible rate (Figure 1.2). Third generation sequencing technologies have the potential to further improve on some of the issues associated with previous platforms (namely, the limitations caused by small read lengths [80, 81]). The field of bacterial toxin biology has benefited greatly from genomic sequencing. Indeed, toxigenic human pathogens were among the first bacterial genomes sequenced, starting with *Helicobacter pylori* in 1997 [82], and *Pseudomonas aeruginosa* and *Vibrio cholerae* in 2000 [83, 84] (see Figure 1.2 for other milestones). Whole genome sequencing has become a routine technique for monitoring and surveying toxigenic microorganisms, including isolates from clinical and non-clinical environments [85]. For some toxin families, whole genome sequencing has revealed toxin families like botulinum toxins can vary by 70% or more amino acid identity [86, 87, 88].

Nevertheless, a sequence-centric understanding of virulence and toxin biology has become widely adopted, aided in part by whole genome sequencing. The sequence-centric paradigm necessitates modifications to ideas like Koch's postulates [89, 90] and pathogenicity [17, 6, 11, 12]. New sequences are now produced much more quickly than can be functionally characterized. As a result, methods for rapidly identifying sequences of interest from genomes or other sources are a necessary component of toxin classification. Bioinformatic techniques have been adopted to address this problem. In addition to quickly identifying and classifying toxin sequences, bioinformatic sequence-based toxin identification contributes a parallel to the classical toxin identification model (Figure 1.2): toxins can be discovered based on their similarity to other toxin sequences, in the absence of an observed diseased phenotype. Including consideration for toxins that do not produce obvious pathology is important because of the possibility that the primary function of some "toxins" may be to prevent and avoid host responses. The bioinformatically-predicted typhoid toxin of *Salmonella* Typhimurium [91] may be an example of this phenomenon, as

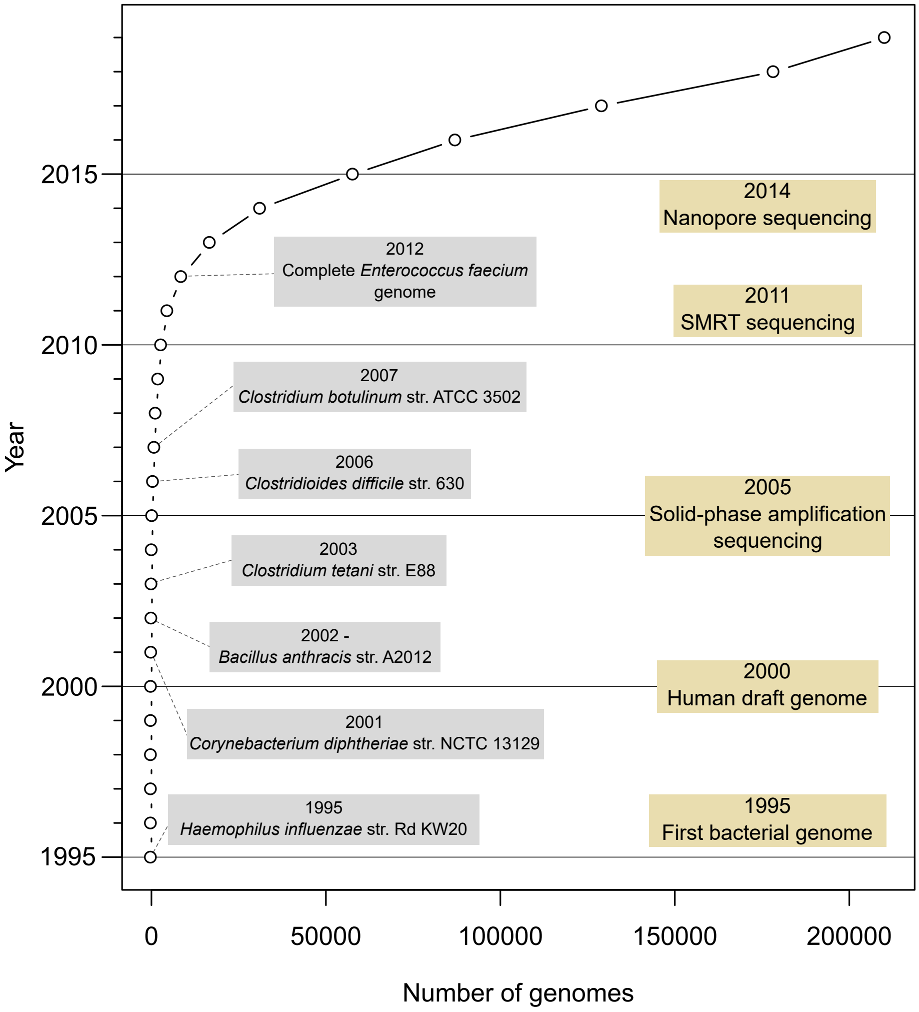


Figure 1.2: The graph above shows the exponential increase in the NCBI microbial genome database over time (as of September 22, 2019). The release dates for the first complete genome sequences of several important toxigenic bacteria are indicated. Approximate market release dates for several sequencing technologies are also noted, including solid-phase amplification sequencing (Illumina), single-molecule real-time sequencing (SMRT; Pacific Biosciences), and nanopore sequencing (ONT; Oxford Nanopore Technologies).

it is possible that the role of typhoid "toxin" may be to promote asymptomatic infection [92].

The bioinformatic approach to toxin identification has already proven successful for identifying novel toxins related to several toxins families. In 2008, a novel ADP-ribosyltransferase related to *Pseudomonas aeruginosa* ExoA was discovered in *Vibrio cholerae* through a combination of suppression subtractive hybridization and bioinformatic comparison to other toxin sequences [93]. This sequence, termed cholix toxin, was found to function as an ADP ribosyltransferase targeting eukaryotic elongation factor 2 [94], similar to diphtheria toxin and ExoA. Cholix toxin has been suggested to play a significant role in the *V. cholerae* life cycle, and belongs to a large and diverse family of toxin-related sequences [95]. In another example, the genome of a virulent strain of *Bacillus cereus* was sequenced and *in silico* searches revealed a locus containing genes similar to anthrax protective antigen and anthrax lethal factor [96]. Anthrax lethal factor contains an inactive vegetative insecticidal protein-related ADP ribosyltransferase and an active C-terminal metalloprotease domain, which cleaves MAPK kinases [53]. The lethal factor-related sequence in *B. cereus*, termed certhrax, was found to enter cells using a homolog of anthrax protective antigen, and possesses an active ADP ribosyltransferase but lacks the metalloprotease domain [96]. The target of certhrax was determined to be vinculin [97], implying a role in immune evasion, although recent studies suggest its role may be to decrease virulence rather than enhance it [98]. Finally, a study by our group in 2015 reported the first botulinum toxin-like sequences outside of *Clostridium*, found in the genome of a strain of *Weissella oryzae* [99]. In spite of its low sequence identity to BoNTs (around 18% amino acid sequence identity), the protein was found to be capable of cleaving VAMP2, a known target of botulinum toxins [100], at a unique cleavage site [101]. Structural characterization of the metalloprotease domain by Košenina et al. [102] revealed a uniquely open and charged substrate binding pocket, which the authors suggest to potentially confer additional unique functionality. In each of the examples listed above, the sequence-centric approach to toxin identification has revealed toxin-related sequences with unique properties. All of these discoveries were enabled by genomic sequencing and bioinformatic analysis.

1.6 Background for bioinformatic methodology

The enormous influx of sequencing data necessitates the development of new techniques for organizing and interpreting biological information. As described in the previous section, toxin biology has already gained from the adoption of bioinformatic techniques. In this section, I describe the background for bioinformatic approaches to sequence comparison and analysis, which are used throughout the thesis.

Sequence similarity and homology

Two sequences that share an ancestor are homologous, and are termed homologs. Generally, homologous sequences are more similar to one another than they are to non-homologous or unrelated sequences. As such, the most straightforward way to detect homology is to detect sequence similarity, normally through the process of sequence alignment. The alignment's quality is then assessed according to some scoring criteria. Ideally, the scoring criteria award high scores to high quality alignments, and are capable of distinguishing them from low-score, poor quality alignments. Sequence alignment is fundamental to bioinformatics and the core of the widely-used BLAST algorithm [103]. An alignment between two sequences is assigned a score, and the statistical significance of that score is normally evaluated by calculating its expectation value (E-value). For BLAST, the E-value for an alignment with score S between two sequences with lengths m and n is given by the following formula:

$$E = Kmne^{-\lambda S}$$

Here, K and λ represent scaling factors for the search space size and scoring scheme. Alignments with scores higher than what is expected by chance given the length of the alignment will produce low E-values. In addition, the BLAST algorithm scales the E-values by database length, filters matches in repetitive regions, and employs heuristic filters to avoid exhaustive computation [103]. Besides BLAST, there are many alternative sequence search algorithms with their own advantages and disadvantages [104, 105, 106, 107, 108].

The detection of statistically significant sequence similarity between two sequences, as measured by E-values or otherwise, provides evidence that sequences are homologous [103]. Sequence similarity implies functional similarity, although the relationship between sequence similarity and function is not altogether straightforward [109]. Since sequence similarity implies functional similarity, sequences displaying sufficient evidence for homology are often assumed to perform similar functions. The process of transferring known functions to similar sequences of unknown function is termed functional annotation. Irrespective of the particular search algorithm or sequence similarity metric, many bioinformatic analyses involve collecting, comparing, and annotating sets of homologous sequences (including all of the chapters presented in this thesis).

Remote homology detection

Inferring homology by comparing individual query sequences to particular target sequences is limited in its ability to capture all members of a sequence family, as any pairwise comparisons may produce alignments with identities below statistical significance [110, 111]. Algorithms better suited to detect more distant relationships usually incorporate information from multiple target sequences, and can perform iterative searches to discover more distantly related sequences. Model-based searches like PSI-BLAST's position-specific scoring matrices [112] or HMMER's hidden Markov models [113] use this strategy to search more broadly than particular query sequences. Profile-to-profile searches, such as those found in the HHsearch suite [114], may be capable of detecting even more distant relationships [115]. In some cases, matching a query sequence to one or more sequences with solved structures can help to identify underlying structural similarities between distantly related sequences (for example, the structural threading approach used by PHYRE [116]).

Functional annotation

After one has identified a set of homologous sequences, it is desirable to estimate what their biological function(s) might be. Most functional annotation methods use homology for function prediction, which relies on databases of annotated sequences. Annotated databases include Pfam [117], CATH [118], SCOP [119], COG [120], and PANTHER [121]. More specialized resources can be used for other types of information, including KEGG for metabolic pathways [122], the PDB for protein structures [123], and MEROPS for proteases [124]. Meta-search utilities like InterProScan combine information from multiple databases to provide more comprehensive annotations [125, 126]. At low sequence identities, it becomes more difficult to infer homology and infer function. In these cases, the addition of information extraneous to the sequence itself can be helpful to infer the function of the sequence. This might include the conservation of key functional motifs, partial or complete protein domains, or association with particular genes or genomic elements [127].

Evaluating relationships between homologous sequences

Homologous sequences can be classified through their mechanism of divergence [128]. Orthologs are homologous sequences that have diverged as a result of speciation between organisms, paralogs have diverged after a gene duplication event, and xenologs have diverged following a lateral gene transfer event [3]. Ascertaining the particular nature of a homologous relationship can be difficult, since the processes of gene duplication, speciation, and lateral gene transfer occur continuously and simultaneously. Further complicating matters are cases of partial gene deletion, recombination, and fusion, which create lineages with partial similarities to multiple gene families. Some of these difficulties can be addressed, or at least detected, through careful analysis and inspection of multiple sequence alignments and phylogenetic trees. Tools for generating multiple sequence alignments include MUSCLE [129], ClustalO [130], and MAFFT [131] programs.

Phylogenetic techniques, which generally rely on multiple sequence alignments, are commonly used for analyzing sets of homologous sequences. These methods aim to construct a model (a tree) that describes the relationship between a set of sequences, given a model of how those sequences are evolving. Parsimony methods generate a tree with the assumption that the smaller numbers of changes are more likely than larger numbers of changes, and neighbour-joining methods attempt to generate a tree based on distances between sequences [132]. Model-based methods like maximum likelihood and Bayesian generate trees based on explicit statistical assumptions about the types and rates of substitutions [133, 134, 135, 132]. An accurate phylogeny should reproduce the ancestry the sequences in the tree, indicating their order of divergence and making it possible to evaluate the relationships between particular substitutions and functions. There are many useful applications for phylogenetics. One is the generation of species trees based on shared taxonomic markers, which attempt to trace back speciation events within or between taxonomic groups, and potentially a tree of all life [136, 137]. Phylogenies can be used to detect sites or branches that evolve more quickly than expected and thus potentially represent adaptations (positive selection) [133, 134], or identifying evolutionary events that produce functional differences between different family members (functional shifts) [138].

1.7 Thesis outline

The broad theme of the work presented in this thesis is the discovery and bioinformatic analysis of bacterial toxin homologs. The thesis is structured as five chapters. This first introductory chapter provides historical and methodological background information to aid interpretation of the following data chapters. The three data chapters are organized by toxin family (described in Table 1.1). The first data chapter (Chapter 2) presents analyses of botulinum neurotoxins and contains four subsections, the first two concerning the

[3] The 2005 review by Koonin [128] features a nuanced and detailed discussion of the molecular evolutionary dynamics surrounding the concept of homology.

discovery of neurotoxin-related sequences in *Enterococcus faecium* and *Chryseobacterium piperi*, the third concerning overview on the family's molecular evolution, and the fourth an analysis of the evolution of the neurotoxin gene cluster. The second data chapter (Chapter 3) focuses on diphtheria toxins and contains two subsections, the first of which is related to the discovery of diphtheria-like sequences, and the second related to an analysis of the structures of two diphtheria-like sequences. The third data chapter (Chapter 4) is related to the translocation apparatus of large clostridial toxins and its distribution in a number of pathogenic species. The final chapter (Chapter 5) features a discussion of general and specific information gained from preceding chapters, including technical and methodological considerations that affect results and their interpretation, and concludes with a short prospective for the field.

Table 1.1: Summary of thesis chapters.

Chapter	**Toxin family**	**Section**	**Description**
Chapter 1		1	Introduction
Chapter 2	BoNT	2.1	Identification of a botulinum neurotoxin-like toxin in *Enterococcus faecium*
		2.2	Bioinformatic discovery of a toxin family in *Chryseobacterium piperi* with sequence similarity to botulinum neurotoxins
		2.3	Genomic insights into the evolution and ecology of botulinum neurotoxins
		2.4	Comparative genomics and evolution of the BoNT-associated P47/OrfX gene cluster
Chapter 3	DT	3.1	Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus
		3.2	Structural characterization of diphtheria toxin homologs
Chapter 4	LCT	4.1	The *C. difficile* toxin translocase is a conserved apparatus for bacterial protein delivery into host cells
Chapter 5		5	Conclusions

Chapter 2

Evolution and diversity of botulinum neurotoxins

Material in this chapter has been published or is currently in preparation for publication. The published materials are available from the following sources:

1. Zhang, S., Lebreton, F., Mansfield, M. J., Miyashita, S. I., Zhang, J., Schwartzman, J. A., Tao L., Masuyer G., Martínez-Carranza M., Stenmark P., Gilmore M.S., Doxey A.C., & Dong M.D. (2018). Identification of a botulinum neurotoxin-like toxin in a commensal strain of *Enterococcus faecium*. Cell host & microbe, 23(2), 169-176. [2].

   https://doi.org/10.1016/j.chom.2017.12.018

2. Mansfield, M. J., Wentz, T. G., Zhang, S., Lee, E. J., Dong, M., Sharma, S. K., & Doxey, A. C. (2019). Bioinformatic discovery of a toxin family in *Chryseobacterium piperi* with sequence similarity to botulinum neurotoxins. Scientific reports, 9. [3].

   https://doi.org/10.1038/s41598-018-37647-8

3. Mansfield, M. J., & Doxey, A. C. (2018). Genomic insights into the evolution and ecology of botulinum neurotoxins. Pathogens and disease, 76(4), fty040. [4].

   https://doi.org/10.1093/femspd/fty040

4. Mansfield, M. J., Lee, E. J., & Doxey, A. C. (2019). Comparative genomics and evolution of the BoNT-associated P47/OrfX gene cluster. Manuscript in preparation.

2.1 Identification of a botulinum neurotoxin-like toxin in *Enterococcus faecium*

2.1.1 Introduction

BoNTs are one of the most dangerous potential bioterrorism agents (category A and tier 1 select agents) [139]. They have also been utilized to treat many medical conditions as well as for cosmetic applications [140]. There are seven well-established serotypes of BoNTs (BoNT/A-G). They are composed of a light chain (LC) and a heavy chain (HC) [141, 142, 143], connected via an inter-chain disulfide bond. The LC is a zinc-dependent metalloprotease. The HC contains the translocation domain (H_N) and the receptor-binding domain (H_C). BoNTs target neurons and block neurotransmission by cleaving host proteins VAMP1/2/3 (BoNT/B, D, F, and G), SNAP25 (BoNT/A, C, E), or syntaxin 1 (Syx 1, BoNT/C). These three proteins mediate fusion of synaptic vesicles to plasma membranes and are the prototype of the SNARE family proteins (soluble NSF attachment protein receptor) [144, 145].

Genes encoding BoNT proteins reside within two types of gene clusters [146]. Both include a gene encoding NTNH (non-toxic non-hemagglutinin protein), which forms a complex with BoNTs and protects them in the gastrointestinal (GI) tract [147]. One type of gene cluster expresses additional proteins HA17, HA33, and HA70, which facilitate the absorption of toxins across the epithelial barrier [148, 149]. The other type encodes proteins with unknown functions; these are designated OrfX1, OrfX2, OrfX3, and P47 [146]. Multiple mechanisms contribute to horizontal gene transfer and the recombination of BoNT clusters, including being located on plasmids or phages and the presence of transposases. Recent genomic studies revealed a growing number of subtypes and mosaic toxins [150, 151, 152, 153, 154]. A new serotype, BoNT/X, was also recently identified in a *Clostridium botulinum* strain [155].

The evolutionary origin of BoNTs remains a mystery. Recent studies reported a homolog of BoNT in a gram-positive bacterium *Weissella oryzae*, designated BoNT/Wo [99, 101]. However, BoNT/Wo is quite distinct from BoNTs. First, the sequence identity between BoNT/Wo versus other BoNTs is around 14%-16%, below the normal range for the members of the BoNT family (28%-65%). Second, the two cysteines that form the essential inter-chain disulfide bond in BoNTs are not conserved in BoNT/Wo, suggesting a distinct mode of action. Third, the BoNT/Wo gene is not in a typical BoNT gene cluster.

Enterococcus faecium is a core commensal member in the human gut and widespread in most terrestrial animals [156, 157, 158, 159]. Since the 1970s, *E. faecium* has become a leading cause of hospital-acquired multi-drug-resistant (MDR) infection of the bloodstream, urinary tract, and surgical wounds [160, 161]. Compounding the problem, the enterococci serve as collection and distribution points for mobile elements, exemplified by acquiring and transmitting a variety of antibiotic resistance to gram-positive and gram-negative species [162].

As a part of an ongoing diversity study, we have collected and sequenced a growing number of enterococcal strains. One strain, IDI0629, was recently isolated from cow feces in South Carolina in the US. Genomic sequencing revealed that it contains a BoNT-like toxin gene (GenBank: OTO22244.1), tentatively designated BoNT/En. BoNT/En represents the first neurotoxin-related sequence to be found in *Enterococcus* species, as well as the closest relative to the BoNT family outside of the genus *Clostridium*. The gene encoding BoNT/En is located within an OrfX-type gene cluster flanked by transposases, suggesting that the unit may have been laterally transferred into *Enterococcus*, and may be capable of additional transfer events.

2.1.2 Methods

Bioinformatic identification and analysis of BoNT/En

BoNT/En was discovered using BLASTp with BoNT/X as a query sequence against the nr database with default parameters (BLOSUM62, gap existence = 11, gap extension = 1, with conditional compositional score matrix adjustment) [112]. As of April 2017, this search space covered a total of 231,827,651,552 bases and 200,877,884 sequences. Domains were annotated using the hmmsearch command of the HMMER package against the Pfam database (v31.0, [117]). Genomic architecture visualized using genoPlotR (v0.8.6, [163]) in R. BoNT sequences representing all major lineages (A-G, F5A, and X) were aligned in a multiple alignment using ClustalO (v1.2.1, [130]), then pairwise identity between BoNT/En and the others was calculated in a 50 amino acid sliding window across the length of the multiple alignment with a step of 1. Regression splines were calculated using the splines base package in R.

2.1.3 Results

BoNT/En shows 29%-38.7% identity with the other BoNTs and is most closely related to BoNT/X (Figure 2.1A). All key BoNT motifs are conserved in BoNT/En (Figure 2.1B), including the zinc-dependent protease motif HExxH (residues H225-H229) in the LC [52], two cysteines that may form an inter-chain disulfide bond (C424 and C438), and a ganglioside-binding motif SxWY in the HC (residues S1250 to Y1253) [164]).

The gene encoding BoNT/En is located within a typical OrfX gene cluster, preceded by a gene encoding NTNHA and containing putative *orfX2*, *orfX3*, and *p47* genes (Figure 2.1C). A gene located 5′ to *orfX2* showed a relatively low degree of sequence similarity to *orfX1* and was therefore designated as an *orfX1*-like gene. The BoNT/En gene cluster is flanked by a 1,719-bp direct repeat sequence (90.1% nucleotide identity), with two truncated non-functional copies of a repB gene on each side. This region also contains a putative phage endolysin, an insertion element (IS204), three putative site-specific recombinases, and additional hypothetical genes. There is also a putative Phd-Doc cassette within this region, an addiction module (a type of toxin-antitoxin system) usually utilized to maintain mobile elements. The occurrence of direct repeats flanking the pBoNT/En cluster suggests that this region was acquired by a repUS15 plasmid precursor through homologous recombination within conserved repB sequences, potentially mediated by associated putative recombinases. This also suggests that the BoNT/En cluster may be mobile by additional mechanisms beyond conjugation.

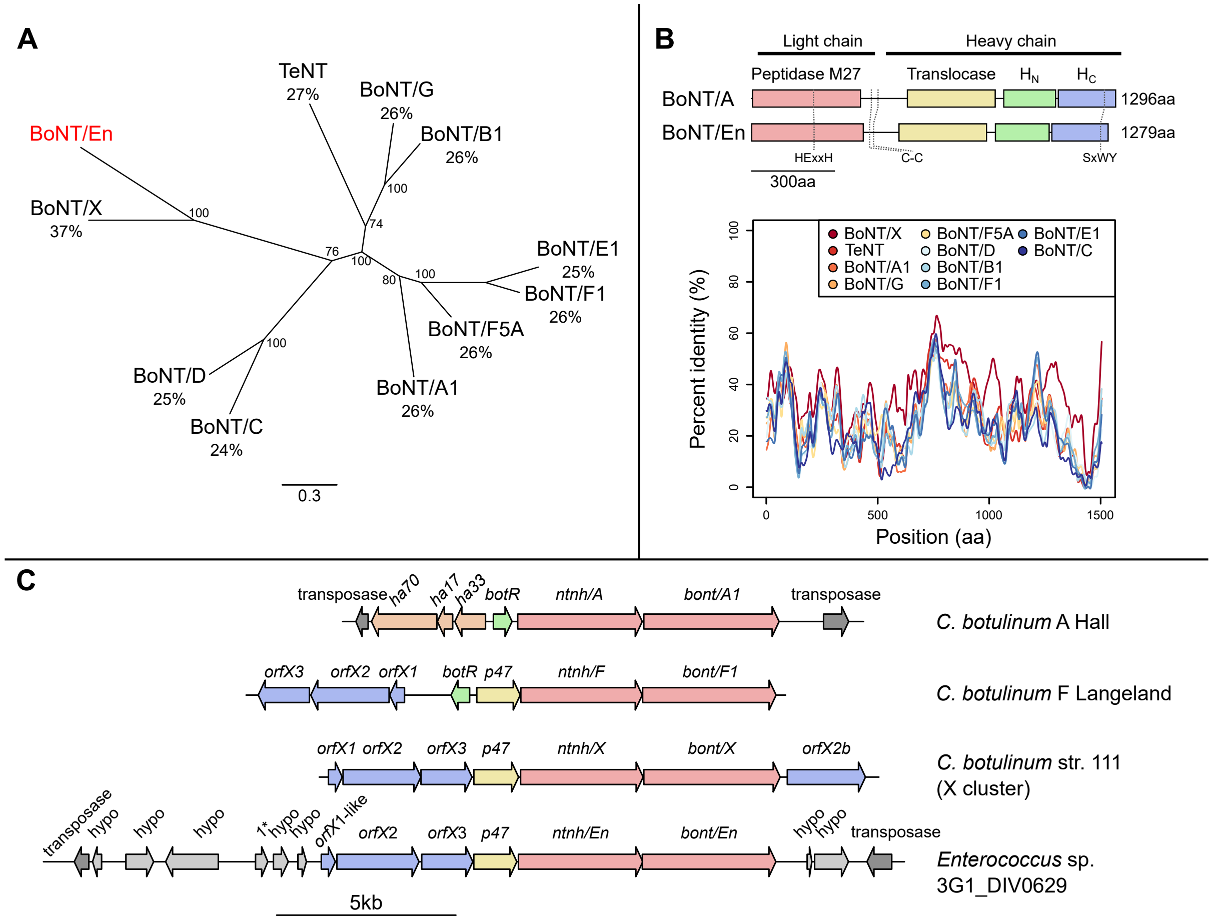


Figure 2.1: BoNT/En is a unique BoNT serotype. (A) The maximum likelihood phylogeny of full-length BoNT amino acid sequences demonstrates that BoNT/En forms a distinct lineage, grouping most closely with BoNT/X. The percentages of protein sequence identity for each BoNT serotype with BoNT/En are noted. Rapid bootstrap support values are indicated at the base of each group, with the scale bar representing the number of estimated substitutions per site. (B) A schematic drawing of the domains of BoNT/En in comparison with BoNT/A. Shared, conserved motifs necessary for BoNT function are highlighted. The sliding window analysis, which compares segments of 50 amino acids in BoNT/En to all other BoNT serotypes, demonstrates that BoNT/En is not a mosaic of known toxin serotypes. (C) BoNT/En is encoded within an OrfX-type gene cluster (blue, with *p47* in yellow). As observed in other BoNT clusters, the cluster is flanked by insertion sequences (dark grey).

2.1.4 Discussion and Conclusions

BoNT-like gene clusters have not previously been identified in any bacterial species outside of *Clostridium* and no toxins of *E. faecium* have been reported before now. It is disconcerting to find a member of potent neurotoxins in this widely distributed gut microbe, which is a leading cause of hospital-acquired infections [157, 161]. The rarity of BoNT/En-producing *E. faecium* in strains sequenced so far may reflect its recent acquisition, or may be due to the relatively limited sampling of clade B strains from wild ecologies. To know the scope of the natural diversity of genes harbored by enterococci and to monitor the emergence of new strains, it will be critical to survey the enterococci beyond lineages that commonly cause infection now. Many important questions remain unknown including the evolutionary origin of BoNT/En and the host species/cell types targeted by BoNT/En. Nevertheless, the capability of *E. faecium* to acquire a BoNT gene cluster could create emerging strains with severe consequences. Furthermore, the possibility of introducing a BoNT cluster into MDR *E. faecium* strains could pose a significant biosecurity threat.

2.2 Bioinformatic discovery of a toxin family in *Chryseobacterium piperi* with sequence similarity to botulinum neurotoxins

2.2.1 Introduction

Clostridial neurotoxins (CNTs), including botulinum neurotoxins (BoNTs) and tetanus neurotoxin (TeNT), respectively, are the causative agents of botulism and tetanus and are the deadliest known biological toxin family, with LD_{50} values ranging from 0.1 to 1.0 ng kg^{-1} [165]. Owing to their extreme toxicity, BoNTs are potential bioterrorism agents, and yet also have enormous utility as protein therapeutics [166, 100]. BoNTs are produced by *Clostridium botulinum*, a polyphyletic taxon classified solely by the presence of the neurotoxin, and several other species of *Clostridium*. Neurotoxin genes reside in distinct gene clusters encoded on the chromosome, plasmids or phages. All BoNTs are neighboured by genes encoding non-toxic non-hemagglutinin proteins (NTNHs), a homolog of BoNT that lacks the HExxH motif and forms part of the progenitor toxin complex. There are currently seven universally accepted, antigenically distinct BoNT serotypes, designated BoNT/A-G, as well as several recombinant mosaics (C/D, D/C, and F5A). A new BoNT serotype (BoNT/X) has been recently identified in the genome sequence of *C. botulinum* strain 111 [155]. A subtype numeral (e.g. BoNT/A1) is also designated to label a growing number of divergent sequences within serotypes [167].

The extreme toxicity of BoNT is a consequence of its unique structure and function (Figure 2.2). BoNTs are initially produced as a single polypeptide chain, which is then cleaved by bacterial or host proteases to result in a light-chain (LC) and heavy-chain (HC) which remain linked by a disulfide bond. The HC contains two functional domains: the N-terminal translocation domain (H_N) and the C-terminal receptor-binding domain (H_C). The receptor-binding H_C domain can be further divided into two subdomains, consisting of an N-terminal laminin-like jelly roll fold (H_{CN}) and a C-terminal ricin-type beta-trefoil fold (H_{CC}). BoNTs recognize motor nerve terminals by targeting neuronal receptors, including SV2 for BoNT serotypes A/D/E/F, and synaptotagmin I/II for BoNT serotypes B/G/DC, with polysialogangliosides as co-receptors [168, 169, 68, 170, 69, 70, 171, 172, 173, 174, 175, 176]. After neuronal binding, BoNTs are internalized within endocytic vesicles. At low pH, the H_N, which forms an all alpha-helical bundle structure, transports the partially unfolded LC into the cytosol. The LC, composed of a ∼400 residue N-terminal zinc metalloprotease domain, then cleaves intracellular SNARE proteins including VAMPs, SNAP25, and syntaxin 1 to prevent exocytosis of synaptic vesicles [177, 178, 179, 180], resulting in flaccid

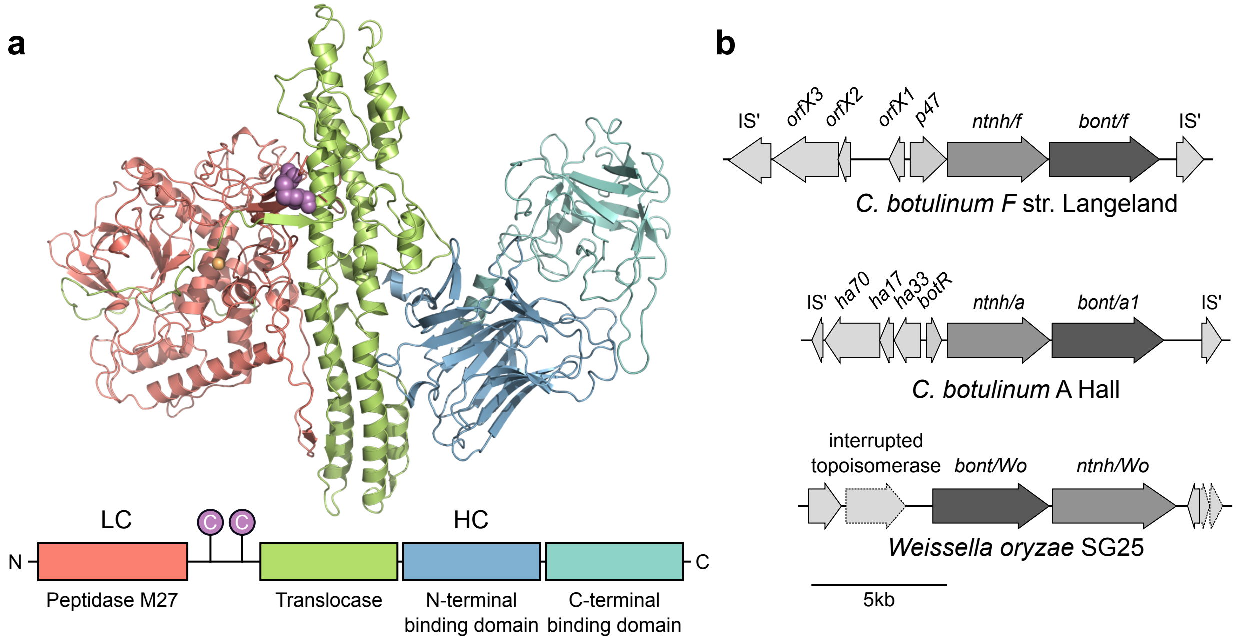


Figure 2.2: Botulinum neurotoxin (BoNT) protein and gene structure. (a) BoNTs are composed of four distinct structural domains (PDB identifier 3BTA). A single BoNT protein is cleaved into a smaller enzymatic component (the light chain, LC, which encodes a zinc endopeptidase) and larger binding and translocating component (the heavy chain, HC, which encodes the translocase domain H_{CN} and binding domains H_{CC}). The receptor-binding domain is further divided into two subdomains, H_{CN} and H_{CC}, which adopt a laminin G-like beta-sandwich fold and a beta-trefoil fold, respectively. The light chain and heavy chain are linked by a disulfide bond. (b) The genes encoding BoNT proteins are generally found in one of two main gene architectures. The *bont* genes are always located next to a paralogous non-toxic non-hemagglutinin (*ntnh*) gene, but the two types are distinguished by their surrounding components, which consist of hemagglutinin (*ha*) or *orfX* genes. Currently, the only known example of altered synteny is in the unique *Weissella oryzae* BoNT homolog BoNT/Wo, where the *bont*/*ntnh* gene order has been reversed.

paralysis [142].

Recent work by our research group [99] reported a divergent BoNT homolog in the genome of *Weissella oryzae*, which suggested that BoNT-related proteins are not limited to the genus *Clostridium* [101]. This hypothesis has been further supported by the recent discovery of BoNT/En, a novel BoNT in *Enterococcus faecium* strain IDI062 [2, 181], which was demonstrated to cleave both SNAP25 and VAMP2. The presence of BoNT homologs in *Weissella* and *Enterococcus* raises the intriguing possibility that a larger family of BoNT-related toxins exists in a broader range of bacterial taxa [1]. These homologs may include not only toxins with globally conserved domain architectures, but potentially distant homologs of BoNTs with more divergent domain architectures, sequences and functions [154, 4].

Here we present a large-scale bioinformatic screen for putative toxin genes in all currently available genomes. Unlike previous studies, we did not limit our searches to the detection of complete homologs, but also considered detectable similarities involving individual BoNT domains to increase the chance of detecting distant homologs. Our analysis identified hundreds of putative toxins, and revealed a novel toxin family from *Chryseobacterium piperi* [182] that exhibits distant homology to BoNTs and has a similar domain architecture. We re-sequenced the genome of *C. piperi* to confirm and further analyze the genomic context of these toxins, and also examined their potential toxicity by transfection assays into human cells. These toxins target a yet unknown class of substrates, potentially reflecting divergence in substrate specificity between the metalloprotease domains of these toxins and the related metalloprotease domain of clostridial neurotoxins.

2.2.2 Methods

Ethics statement

Experiments were performed in accordance with the procedures approved by the Institutional Animal Care and Use Committee (IACUC) at Boston Children's Hospital (protocol #3030). All experiments were performed in BSL-2 laboratory settings.

Detection, comparison, and analysis of *bont*-like genes

Sequences were retrieved using PSI-BLAST with default parameters (BLOSUM62 scoring matrix; expect threshold 10; gap open 11; extension 1) from the *nr* database on (March 26, 2017) [112]. Initial homologs were discovered by searching with BoNT/A1 (NCBI accession number ABS38337.1) with up to two rounds of PSI-BLAST. Then, in order to retrieve all possible sequences from each sequence family, different queries were used to search for specific BoNT homolog subfamilies (*Chryseobacterium*: WP_034681281.1, Actinobacteria: *Streptomyces* sp. NBRC 110027 GAO13068.1, fungal: *Metarhizium anisopliae* KFG81441.1) and reiterating to convergence. BoNT homologs identified this way were added to a set of known BoNT and NTNH proteins representing all known serotypes, including the recently discovered BoNT/F5A (KGO15617.1), BoNT/X (KGO12225.1), and BoNT/En (WP_086311652.1). Sets of M91 peptidases and diphtheria toxins were also retrieved via PSI-BLAST, with diphtheria toxin (PDB accession number 4AEO.1) and *E. coli* NleD (WP_069191536.1) as the original queries. These sets of M91 peptidases and diphtheria toxins were pruned to remove identical sequences using Jalview [183].

All-by-all sequence pairwise alignments were generated with needle (of the EMBOSS package, v6.6.0.0; [106]) with default parameters (gap open = 11, gap extend = 1, EBLOSUM62 scoring matrix). In Figure 2.3, percent similarity was used over percent identity in order to allow divergent homologs to cluster more accurately. Principal coordinate analysis was performed in R on a distance matrix of pairwise similarity values using the default dist() and cmdscale() functions.

Domains were annotated with hmmscan (v3.1b2, available from `http://hmmer.org/`) against the Pfam database v31.0 [117] with an E-value cutoff of 1×10^{-6}. Annotations were subsequently confirmed by comparison to the Conserved Domain Database with relaxed cutoffs (v3.16, [184]), and alignment to BoNTs. For Figure 2.3, the BoNT homologs with the most BoNT-like annotations were depicted to facilitate comparison between categories.

Comparison of proteases from BoNTs, BoNT-like proteins, and M91 peptidases

All BoNT homologs possessing a putative peptidase domain (i.e., possessing an HExxH motif) were aligned with BoNT and M91 peptidases using Clustal-Omega with defaults (v1.2.1, [130]), manipulated and colored in Jalview [183]. Only regions corresponding to the peptidase domain boundaries were used, the positions of which were estimated based on alignment with domain boundaries of BoNT/A1 (PDB structure 3BTA). The same alignment procedure was used to identify the putative translocation region of BoNT homologs (Figure 2.9). After identifying putative domains in BoNT homologs, the segments were combined and realigned.

A maximum likelihood phylogeny for BoNT, BoNT homolog, and M91 peptidases was generated using RAxML (v8.2.4, [135]) with automatic model selection and 4 gamma-distributed rate categories (see simplified cladogram in Figure 2.5, for the full tree see Figure 2.6). Bootstrap support was calculated using 1000 rapid bootstraps. The same alignment was used to infer a Bayesian phylogeny using MrBayes [185] (with the ML-selected substitution model VT, 4 gamma-distributed rate categories, and 1,000,000 MCMC samplings; the consensus tree with 25% burn-in is depicted in figure 2.6).

Pairwise global alignments were generated using BoNT/A1 (ABS38337.1) against the 220,362 metallopeptidase sequences available in the MEROPS database (retrieved Feb. 7, 2018) using needle from the EMBOSS package (v6.6.0.0, [183]). The alignment parameters were as follows: a gap open penalty of 11 and gap extension 1, with the BLOSUM62 substitution matrix. Raw alignment scores were averaged across peptidase families according to their MEROPS group, and visualized in R.

Structural modelling of BoNT-like proteins and M91 peptidases

Structural templates were identified for Cp1 (*C. piperi*, accession WP_034687872.1) using the LOMETS meta-server [186] on July 18, 2017. Templates (PDB identifiers: 3BTA:A, 1XTG:A, 5BQN:A) were selected based on highest significant threading alignments (normalized Z-scores: 5.12-1.21, identity: 17-21%). Structural modelling and refinement was done through I-TASSER [187], and the model with the lowest C-score was selected. For *E. coli* NleD, structural templates were identified through GeneSilico Metaservers (PDB identifiers: 1Z7H:A, 1EPW:A, 3BWI:A, 3DEB:A, 3BON:A, 2QN0:A, 2A97:A, 3DDA:A, 1XTG:A, 1ZB7:A, 1F0L:A, 3FFZ:A, 1YVG:A, 2FPQ:A, 2G7K, 5BQN:A, 2NYY:A, 1T3C:A, 3V0A:A, 3FIE:A, 1RM8:A, 1E1H:A, 3VUO:A, 2A8A:A, 3D3X:A, 3DSE:A) were selected from COMA (score $\leq 5.4\times10^{-7}$, identity: 19%), HHblits (score: 100, identity: 13-20%), and HHsearch (score: 96.3, identity: 13-19%) on July 15, 2017. Structural modelling was carried out through PRIMO's pipeline [188]. Identified template sequences were aligned to M91 with T-Coffee Expresso [189], which uses 3D-Coffee to incorporate structural information during alignment. A total of 20 homology models were created with slow refinement based on the resulting alignment using MODELLER [190]. The model with the lowest DOPE Z-score was selected. Structural quality was assessed with Ramachandran plot analysis using PROCHECK [191]. Models were visualized using Chimera [192].

Re-sequencing and annotation of the *Chryseobacterium piperi* genome

Methods, materials, and platforms used in the sequencing and assembly of *C. piperi* are described in Wentz et al. [182]. The closed genome is accessible at the DDBJ/ENA/Genbank under the accession number CP023049. MiSeq and RS-II reads utilized in assembly are available at NCBI SRA under accessions SRX3229522, SRX3231351, SRX3231352. Figure 2.9 was generated using the program Circos [193].

HEK 293T cell transfection and cell number counting

HEK 293T cells were dispensed on 24-well plate at the density of approximately 0.2 × 10^6 cells/well. After 24 h, cells were transfected with 0.5 µg vehicle vector (pcDNA3.1(+)), Cp1-LC WT(1-398 aa), Cp1-LC H209A and Cp1-LC E210Q plasmids with PolyJet reagent. Pictures were taken 48 hours after transfection. Cell numbers were counted and combined from three different pictures.

HEK 293T cell death assay

HEK 293T cells were dispensed on 60 mm dish at the density of 2.5 × 10^6 cells/dish. Cells were transfected with 2.5 µg vehicle vector (pcDNA3.1(+)), Cp1-LC WT, Cp1-LC H209A and Cp1-LC E210Q plasmids by using 5 µL PolyJet. Cells were harvested 24 hrs after transfection and washed with cold phosphate-buffered saline (PBS). Cell density was adjusted to 1 × 10^6 cells/mL. Hoechst 33342, YO-PRO-1 and propidium iodide stock solution (1 µL Invitrogen) were added into 1 mL cell suspension. Cells were incubated on ice for 20-30 min. Stained cells were analyzed by flow cytometry (BD/Cytek FACSCalibur DxP 11). UV excitation was used for detection of 460 nm emission of Hoechst 33342 dye, 488 nm excitation was used for detection of the 530 nm emission of YO-PRO-1 dye, and 575 nm emission of propidium iodide. Cell populations separated into three groups: live cells showed low levels of blue fluorescence, apoptotic cells showed bright green and blue fluorescence, and necrotic cells showed bright red fluorescence.

Cleavage of SNARE proteins by Cp1-LCs

HEK293T cells were dispensed on 24-well plate at the density of 0.3 × 10^6 cells/well. 24 h later, cells in a single well were transfected with 0.5 µg vehicle vector (pcDNA3.1(+)), Cp1-LC WT (1-398 aa), Cp1-LC H209A, Cp1-LC E210Q, together with syntaxin 1, SNAP25, VAMP2 in pEGFP-C1 as indicated in Figure 2.10. Cells were harvested 48 hours after transfection and lysed in RIPA buffer (50 mM Tris, 1% NP40, 150 mM NaCl, 0.5% sodium deoxycholate, 0.1% SDS, 400 mL per 10 cm dish) plus protease inhibitors. Cleavage assay was conducted by mixing cell lysates of vehicle vector, Cp1-LC WT, Cp1-LC H209A, Cp1-LC E210Q and GFP fused syntaxin 1, SNAP25, VAMP2 respectively and incubating the mixtures at 37 °C for 30 minutes. Samples were analyzed by immunoblot.

2.2.3 Results

Genomic data mining uncovers proteins with BoNT-like domains

We screened the NCBI GenBank database (March 26, 2017) comprised of 94,396 prokaryotic, 4,123 eukaryotic, and 7,178 viral genomes, for potential homologs of BoNTs over one or more domains. Using PSI-BLAST with selected BoNT sequence queries (see Methods), we detected a total of 311 protein sequences displaying at least partial homology to BoNTs with an E-value below 0.001 (see table in Appendix A A.2). The data set includes all known BoNT serotypes, and an additional 161 predicted toxin sequences, all of which are experimentally uncharacterized to date. We performed all-by-all pairwise alignments and clustered the toxins using principal coordinates analysis (PCoA). The toxins clustered largely into three main groups, which differ in domain composition and detectable similarities to BoNTs (Figure 2.3). Group I includes a large family of ADP-ribosyltransferase toxins, including diphtheria toxin-like sequences [5] and putative ADP-ribosyltransferase toxins (ADPRTs) from entomopathogenic fungi (Table S1). These sequences possess partial similarity only to the BoNT translocation domain (17.3% maximum sequence identity, PSI-BLAST E-value = 7×10^{-40}). Group II is formed by M91 family peptidases such as the *Escherichia coli* type III effector toxin NleD, which cleaves host JNK and p38 [194]. These sequences possess remote detectable homology to the BoNT-LC with 14.9% maximum sequence identity and a PSI-BLAST E-value of 4×10^{-5} (see Methods).

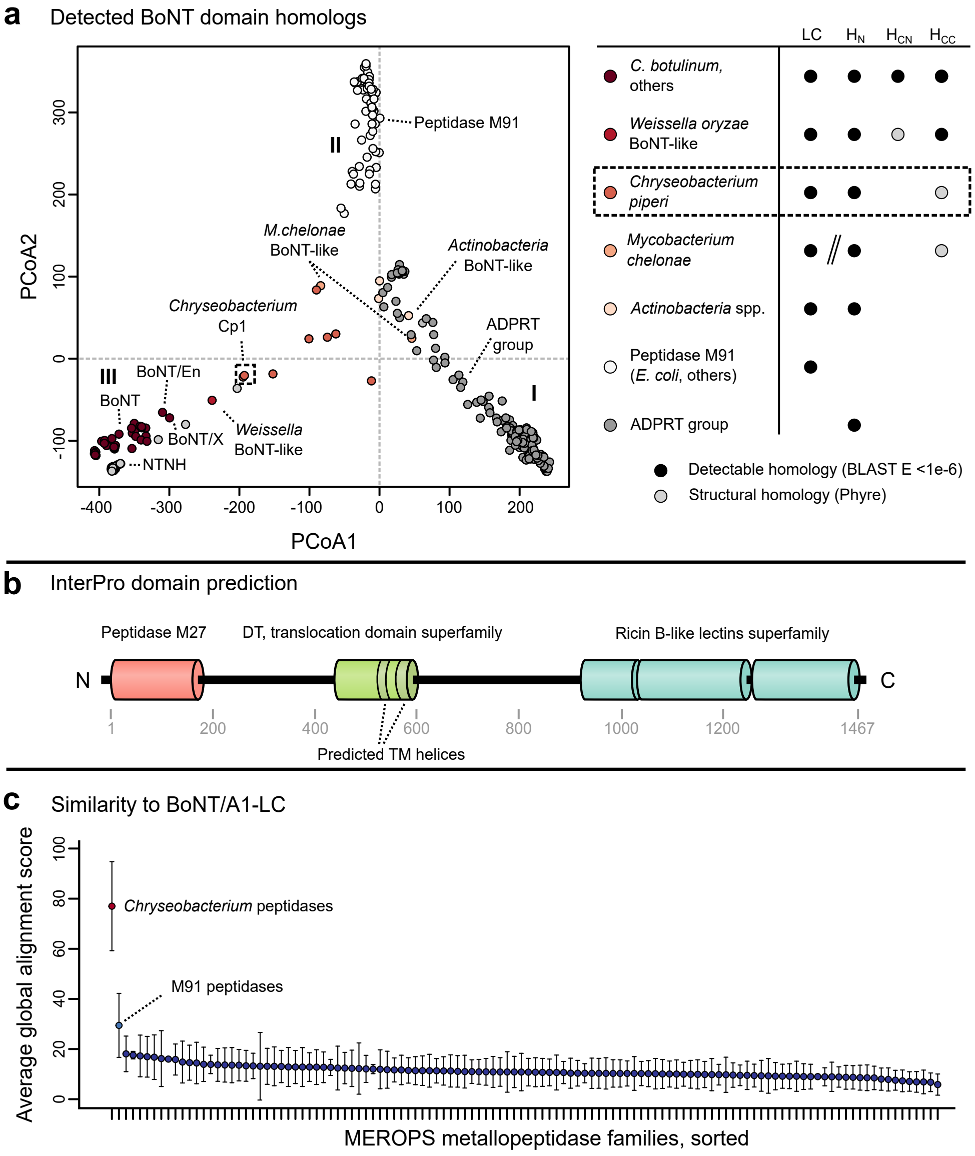

a
Detected BoNT domain homologs
PCoA2
PCoA1
300
200
100
0
-100
-400
-300
-200
-100
0
100
200
II
I
III
Peptidase M91
M.chelonae
BoNT-like
Actinobacteria
BoNT-like
Chryseobacterium
Cp1
ADPRT
group
BoNT/En
BoNT
BoNT/X
NTNH
Weissella
BoNT-like
LC
H_N
H_CN
H_CC
C. botulinum, others
Weissella oryzae BoNT-like
Chryseobacterium piperi
Mycobacterium chelonae
Actinobacteria spp.
Peptidase M91 (E. coli, others)
ADPRT group
Detectable homology (BLAST E <1e-6)
Structural homology (Phyre)
b
InterPro domain prediction
Peptidase M27
DT, translocation domain superfamily
Ricin B-like lectins superfamily
N
C
1
200
400
600
800
1000
1200
1467
Predicted TM helices
c
Similarity to BoNT/A1-LC
Average global alignment score
0
20
40
60
80
100
Chryseobacterium peptidases
M91 peptidases
MEROPS metallopeptidase families, sorted

Figure 2.3 ***(previous page)*****:** Bioinformatic detection of *bont*-related genes in microbial genomes. (a) PCoA ordination of pairwise percent similarities reveals groups of sequence subfamilies with partial similarities to BoNTs. A large family of ADP-ribosyltransferase toxins (grey), which includes diphtheria toxin-like proteins and a large family of predicted toxins from entomopathogenic fungi, possess similarity only over the translocase domain (cluster I). M91 peptidases have similarity to the BoNT LC and group separately (white; cluster II). BoNT and NTNH form distinct groups with more divergent relatives (red to orange; cluster III). Within cluster III, BoNT-related sequences possess the following characteristics, in order of decreasing similarity to BoNT: the *Weissella* BoNT-like protein displays evidence for all BoNT domains; the *Chryseobacterium* sequences have strong similarity in the LC, H_N, and H_{CC} domains; two genes in *Mycobacterium chelonae* encode separate LC-like peptidase and HC-like H_N/H_{CC} domains; *Actinobacteria* spp. sequences are similar only within the LC and H_N domains. (b) InterPro domain predictions for *Chryseobacterium piperi* BoNT-like protein Cp1 (WP_034687872.1), reveals a similar predicted architecture to BoNT. Its translocase-like domain is annotated as diphtheria-like, and contains two predicted transmembrane helices. (c) Comparison of BoNT/A1 peptidase domain to *C. piperi* putative peptidases and proteases in the MEROPS database. Except for peptidase M27 (not pictured), the peptidases from *Chryseobacterium* produce the highest-scoring global alignments, followed by peptidase M91. Both *C. piperi* peptidases and M91 peptidases score higher than all other known peptidase families.

Group III contains BoNTs, NTNHs, the *Weissella* toxin and several uncharacterized proteins (Figure 2.3) that share multiple domains in common with BoNTs (Figure 2.4) and are therefore of considerable interest [1, 147]. Among the uncharacterized proteins are nine partial and full-length homologs from the genome of *Chryseobacterium piperi*, two from *Mycobacterium chelonae*, and five from other Actinobacteria. Several of these organisms are associated with disease; some *Chryseobacterium* species are known opportunistic pathogens [195, 196], *Acaricomes phytoseiuli* is a pathogen of mites [197], and *Mycobacterium chelonae* is a human pathogen associated with skin, soft tissue, and bone infections [198]. We termed these proteins "divergent BoNT homologs" given their distant but significant detectable evolutionary relationship to BoNTs (Fig. 1a 2.3) and similarity of domain architecture (Fig. 1b 2.3). As shown for a representative protein from this group (putative *Chryseobacterium* toxin, "Cp1", NCBI accession number WP_034687872.1), these proteins are predicted by InterProScan [125] to contain a BoNT-like three domain architecture composed of a metalloprotease domain, central translocation domain and a C-terminal ricin-type beta-trefoil domain (Figure 2.3), each of which are analyzed in greater detail below. Cp1 for example possesses detectable homology to BoNTs spanning multiple domains (Figure 2.4), but has low sequence identity (17% identity between Cp1 and BoNT/A compared to >=28% identity between BoNT family members) indicative of a distant evolutionary relationship.

To further confirm this detected relationship, we compared BoNT/A1-LC to the *C. piperi* toxin peptidase domains, as well as all known metallopeptidase families from the

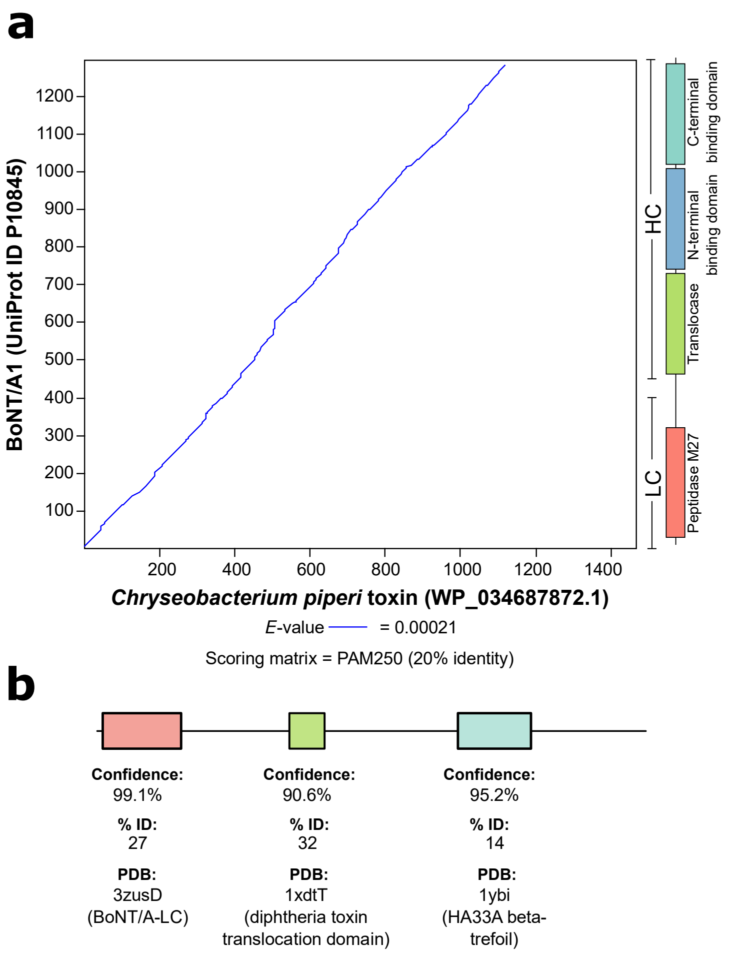


Figure 2.4: Pairwise local alignment and associated *E*-value between BoNT/A1 and *Chryseobacterium* toxin computed using lalign/plalign from the FASTA package. The alignment shown in (a) was generated using a PAM250 scoring matrix, which was selected since it models remote relationships with sequence identities of ∼20%. The pairwise alignment spans multiple domains of BoNTs including the BoNT LC peptidase (red), translocation domain (green) and binding domain (with subdomains blue and cyan). (b) Domain structure predicted by Phyre version 2.0.

MEROPS database [199], consisting of 220,362 sequences from 102 families (Figure 2.3). Based on alignment scores, the *C. piperi* toxins displayed stronger similarities to BoNT-LC than to all other known protease families, and the M91 protease family ranked second.

Chryseobacterium **toxins are a novel toxin family distinct from but related to BoNTs**

Next, the alignment of protease domains from BoNTs and the divergent BoNT homologs was analyzed further perform sequence, structural, and phylogenetic analysis (Figure 2.5). Phylogenetically, BoNT-LCs grouped into a distinct clade, with BoNT/X and BoNT/En forming divergent early branching lineages (Figure 2.6). The BoNT clade is outgrouped by lineages consisting of the predicted toxins from *Weissella*, *Chryseobacterium*, and *Mycobacterium*, although the precise branch order is difficult to resolve with the available data. Nonetheless, the BoNTs together with the *Weissella*, *Mycobacterium*, and *Chryseobacterium* toxins form a distinct clade from the peptidase M91 group with high statistical confidence (83% maximum likelihood bootstrap support and 100% Bayesian posterior probability) (Figure 2.5 and Figure 2.6). Protease domains from the actinobacterial toxins group more distantly, and the clade of distantly related M91 family proteases forms a lineage distinct from BoNTs and the divergent BoNT homologs (Figure 2.5). Despite some variable segments and low sequence identity (BoNTA1/Cp1: 17.6%), the protease domains from *C. piperi* and other divergent BoNT homologs possess detectable homology to the BoNT-LC (bl2seq *E*-value = 2×10^{-6} between Cp1 and BoNT/A1) and conserve key functional residues found in BoNTs (Figure 2.5). These residues include: the critical HExxH zinc-coordinating active site motif; the third zinc ligand Glu-261; the functionally important Glu-350 which shapes active site fine structure, the active site stabilizing motif R363-x-x-Tyr366 [200], and the cysteine residues that form the disulfide bridge between BoNT LC and HC [201] (Figure 2.5).

Consistent with phylogenetic analysis, the predicted structure of the protease domain of *C. piperi* toxin is most similar to BoNT-LC (7.0 Å RMSD versus 12.0 Å for *E. coli* NleD; both models based on BoNT template structures) (Figure 2.5). Although experimental structure determination is required to confirm these models, several obvious structural differences can be inferred based on the models and sequence alignments. One insertion is common to BoNT-LCs and the divergent BoNT homologs, and absent in NleD and other M91 peptidases (Figure 2.5), which makes extensive contacts with SNAP25 (51 inter-residue contacts <2 Å) and VAMP2 (91 inter-residue contacts <2 Å) in BoNT co-crystal complexes (Figure 2.7). This insertion may therefore have contributed to an ancestral shift in substrate specificity between M27 and M91 peptidase families. A second C-terminal

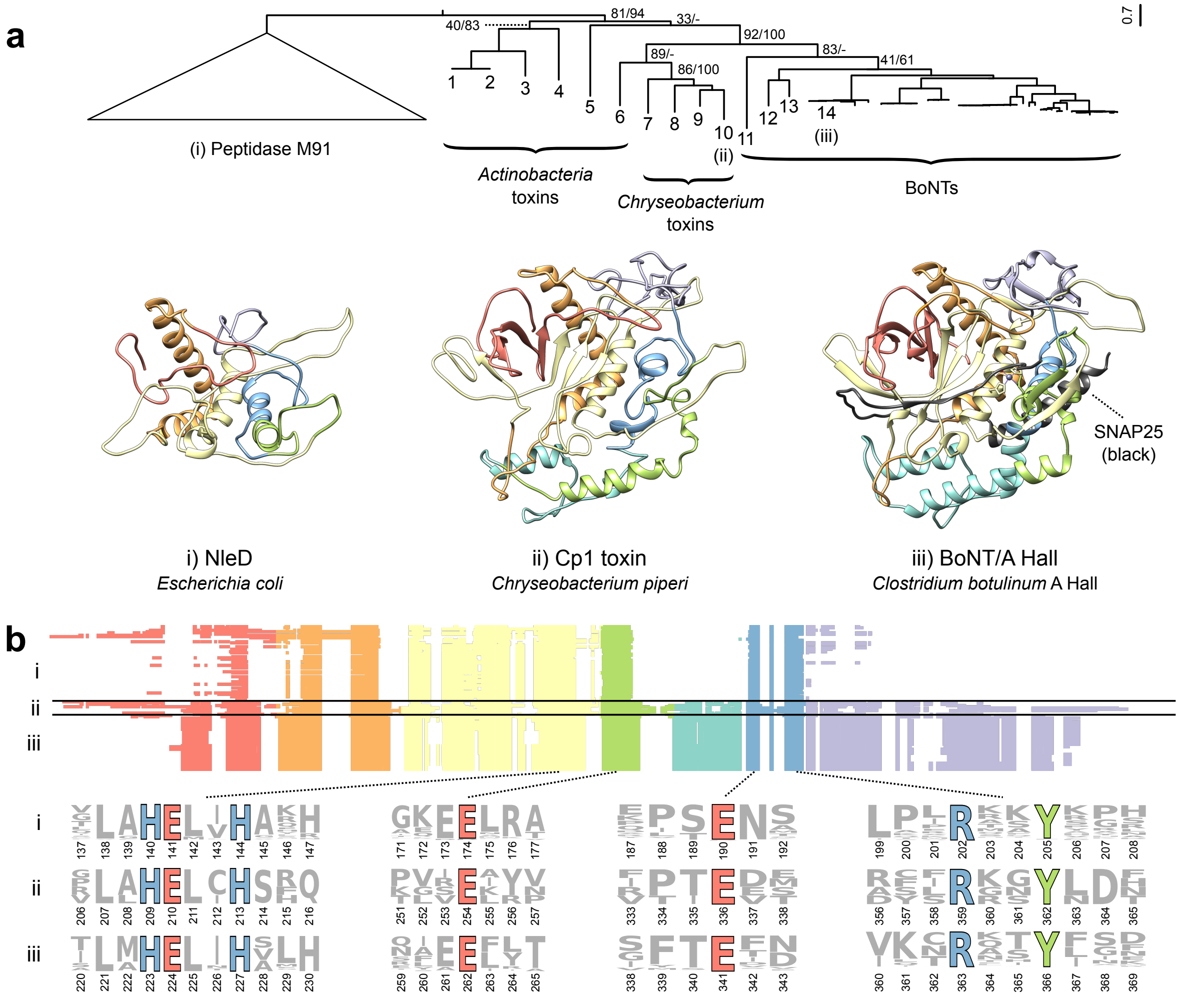


Figure 2.5: Comparison of the BoNT-LC with homologous domains from BoNT-like toxins and M91 family proteases. (a) Phylogenetically, the peptidases of BoNT and BoNT homologs group separately from distantly related peptidase M91 sequences. Statistical support for the tree is indicated as maximum likelihood bootstrap value and Bayesian posterior probability (percentage). Structural comparison of BoNT/A (iii; PDB 1XTG) with homology models of *Chryseobacterium* Cp1 (ii) and *E. coli* NleD (i) reveals two regions that are unique to BoNTs and BoNT-like proteins: the lower alpha-helical region, which interacts directly with SNARE substrates, and the C-terminal region that plays a role in catalytic product removal. (b) M91 peptidases (i), divergent BoNT homologs (ii), and BoNT-LCs (iii) have conserved features. These include the HExxH zinc-coordinating and catalytic residues, the zinc ligand E261, the active site-refining E350, and RxxY motif. The cartoon of the multiple alignment above reveals two regions unique to BoNTs and divergent BoNT homologs, shown in teal and purple respectively. The identities of proteins labeled 1-14 are: 1 - WP_037712107.1, *Streptomyces* sp. AA4; 2 - EFL04418.1, *Streptomyces* sp. AA4; 3 - WP_083906476.1, *Acaricomes phytoseiuli*; 4 - GAO13068.1, *Streptomyces* sp. NBRC_110027, 5 - WP_055473237.1, *Streptomyces pathocidini*; 6 - WP_070931163.1, *Mycobacterium chelonae*; 7 - WP_034681281.1, *Chryseobacterium piperi*; 8 - WP_034687877.1, *Chyrseobacterium piperi*; 9 - WP_034687193.1, *Chryseobacterium piperi*; 10 - WP_034687872.1, *Chryseobacterium piperi*; 11 - WP_027699549.1, *Weissella oryzae*; 12 - WP_086311652.1, *Enterococcus* BoNT/En; 13 - BAQ12790.1, BoNT/X; 14 - ABS38337.1, BoNT/A1.

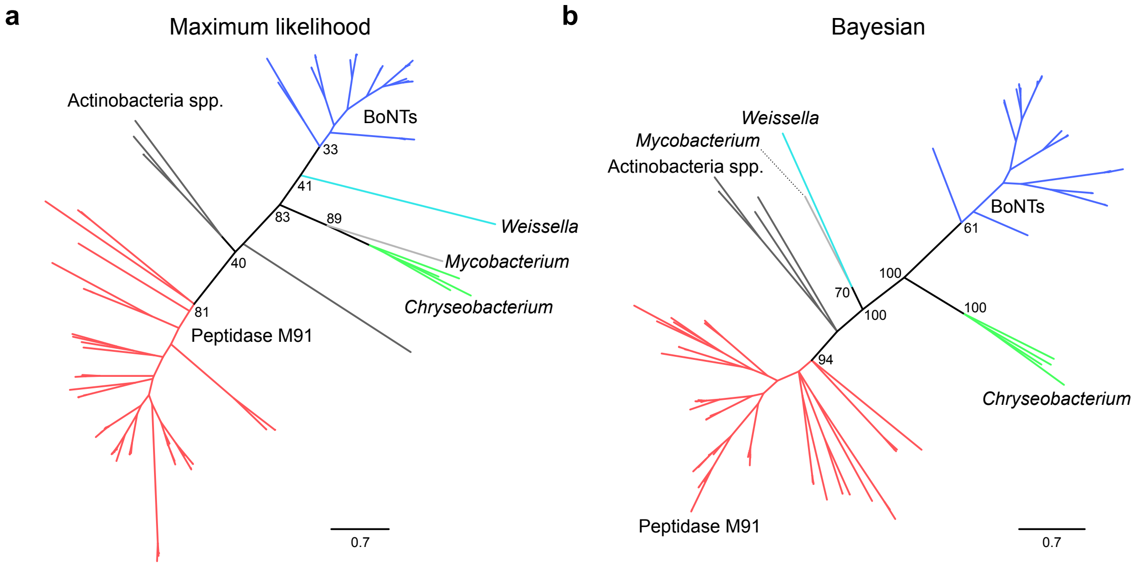


Figure 2.6: Maximum likelihood and Bayesian consensus phylogenies of peptidase amino acid sequences from BoNT, distant BoNT homologs, and M91 peptidases. In both the maximum likelihood (a) and Bayesian (b) phylogenies, the peptidase M91 sequences form a monophyletic group (red) distinct from the BoNTs and BoNT homologs (various colours). Although the branch order differs between the two methods, the grouping of BoNT with BoNT homologs from *Weissella* (cyan), *Chryseobacterium* (green), and *Mycobacterium* (light grey) is well-supported. The differences in branch order may be related to long branch attraction, since the sampling of some lineages is low, and so are BoNT to BoNT homolog percent identities. However, in both trees, the distant BoNT homologs from phylum Actinobacteria are basal to the group of BoNT and other BoNT-like sequences. The ML phylogeny was generated using RAxML (v8.2.4) with 1000 rapid bootstraps, and the Bayesian tree with 1,000,000 MCMC generations using MrBayes (v3.2.4). Support values values are indicated at each node (bootstrap values for maximum likelihood and posterior probability percentage for Bayesian), and branch lengths represent estimated numbers of substitutions per site.

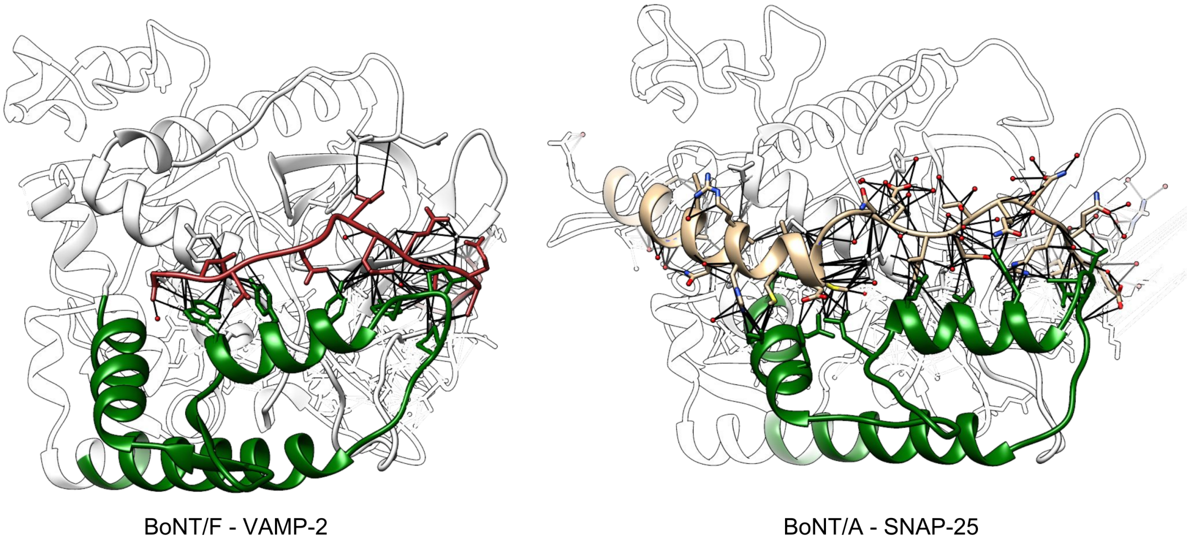


Figure 2.7: Structural visualization of BoNT-LC specific insertions. The two insertion regions identified by comparing the BoNT LC to peptidases from distant BoNT homologs and peptidase M91 sequences (dark green) share extensive contacts (black lines) with VAMP-2 (red) and SNAP25 (tan) in co-crystal complexes with BoNT/F (PDB identifier 3FIE) and BoNT/A (PDB identifier 1XTG).

insertion common to BoNT-LC and the divergent BoNT homologs (Figure 2.5) forms part of the hydrophobic SNAP25 binding pocket [202], and has been shown to mediate catalytic activity and product removal [203]. Lastly, a region corresponding to the "belt" region present in BoNTs was identified in *C. piperi* toxins based on multiple sequence alignment, but this region did not display significant sequence similarity to BoNT, suggesting that it is highly divergent or unrelated.

In addition to conservation in the peptidase domain, the divergent BoNT homologs from *Chryseobacterium* and other species possess significant similarity to the BoNT translocation domain (22% identity, bl2seq E-value $<1\times10^{-5}$), particularly across the region 593-686 in BoNT/A1 (Figure 2.8). This region has been suggested to form a channel-forming amphipathic alpha helical motif that assists in translocation [204, 205, 206]. Unexpectedly, BLAST searches of this region also detected a segment of the diphtheria toxin (DT) translocation domain (residues 286-325, helices TH5-TH6/TH7, PDB identifier 4AE0) [207], which was consistent with structural predictions for this region made by PHYRE (Figure 2.4). Sequence alignment revealed a common region of sequence similarity flanking a motif ([K/R]x(8)PxxG) within the translocation domains of the divergent BoNT homologs, BoNT, and DT (Figure 2.8). Although the functional significance of this shared motif is unclear, the detectable similarity to both BoNT and DT translocation domains strongly suggests a translocation-related function for this region in Cp1 and other divergent BoNT homologs.

Lastly, following the translocation domain, the divergent BoNT homologs possess a receptor-binding domain that is predicted to adopt the same fold as the BoNT H_{CC} subdomain (Figure 2.3 and Figure 2.4). A ricin-type beta-trefoil fold was predicted for the C-terminal region of the putative *C. piperi* toxins by three separate methods (HHpred, Pfam, and Phyre with E-values <0.001). The beta-trefoil domains from *C. piperi* toxins exhibited at most 17% sequence similarity to the H_{CC} in BoNT; therefore, the extent to which they are homologous is unclear at this point. Interestingly, a ricin-type beta-trefoil domain from the *C. botulinum* hemagglutinin, HA33, was identified as the top template by PHYRE (2.4), indicating that if this domain is not related to the H_{CC} in BoNT, it may be homologous to other ricin-type beta-trefoils that are encoded within BoNT gene clusters.

Genome resequencing of *C. piperi* confirms presence of toxin gene clusters

The putative *C. piperi* toxins are located on three separate contigs (2, 44, and 59) in the original draft *C. piperi* genome (NCBI accession JPRJ01, 89 total contigs). To verify a *C. piperi* origin for these contigs, determine extrachromosomal content, and enable further genome-wide analysis, *C. piperi* was acquired from ATCC and sequenced on Illumina MiSeq and Pacific Biosciences RS-II sequencing platforms. A closed genome 4.5 Mbp in length, 35.3% GC content, and 250X coverage was produced and analysed (Figure 2.9). No plasmids were observed. The assembly revealed a toxin gene cluster (GC1) located at 1399-1432 kbp which contains the Cp1 gene as well as 6 other genes with detectable similarity to BoNTs and an alternating pattern of presence/absence of the HExxH motif (Figure 2.9 and Table A.2). A second toxin gene cluster (GC2) is located at 3,287-3,312 kbp containing two additional genes with detectable similarity to BoNTs (Figure 2.9). The genes located in this region have similarity to BoNT over different regions, and may contain or lack the catalytic HExxH motif. Similar to *bont* gene clusters, HExxH-positive homologs in *C. piperi* are flanked by genes that also possess detectable partial homology to BoNTs, but lack the active site motif. The paralogous nature and genomic arrangement of these gene pairs resemble that of *ntnh* and *bont*, which raises the possibility that the HExxH-negative genes may play a similar role to *ntnh* in *bont* gene clusters. Further, these HExxH-negative proteins uniquely contain an IBC1 ("Isoprenoid_Biosyn_C1 super family") domain at the N-terminus similar to class I terpene-synthases, whose role is unclear. No additional neurotoxin-associated genes from the *ha/p47/orfX* families are present in these clusters.

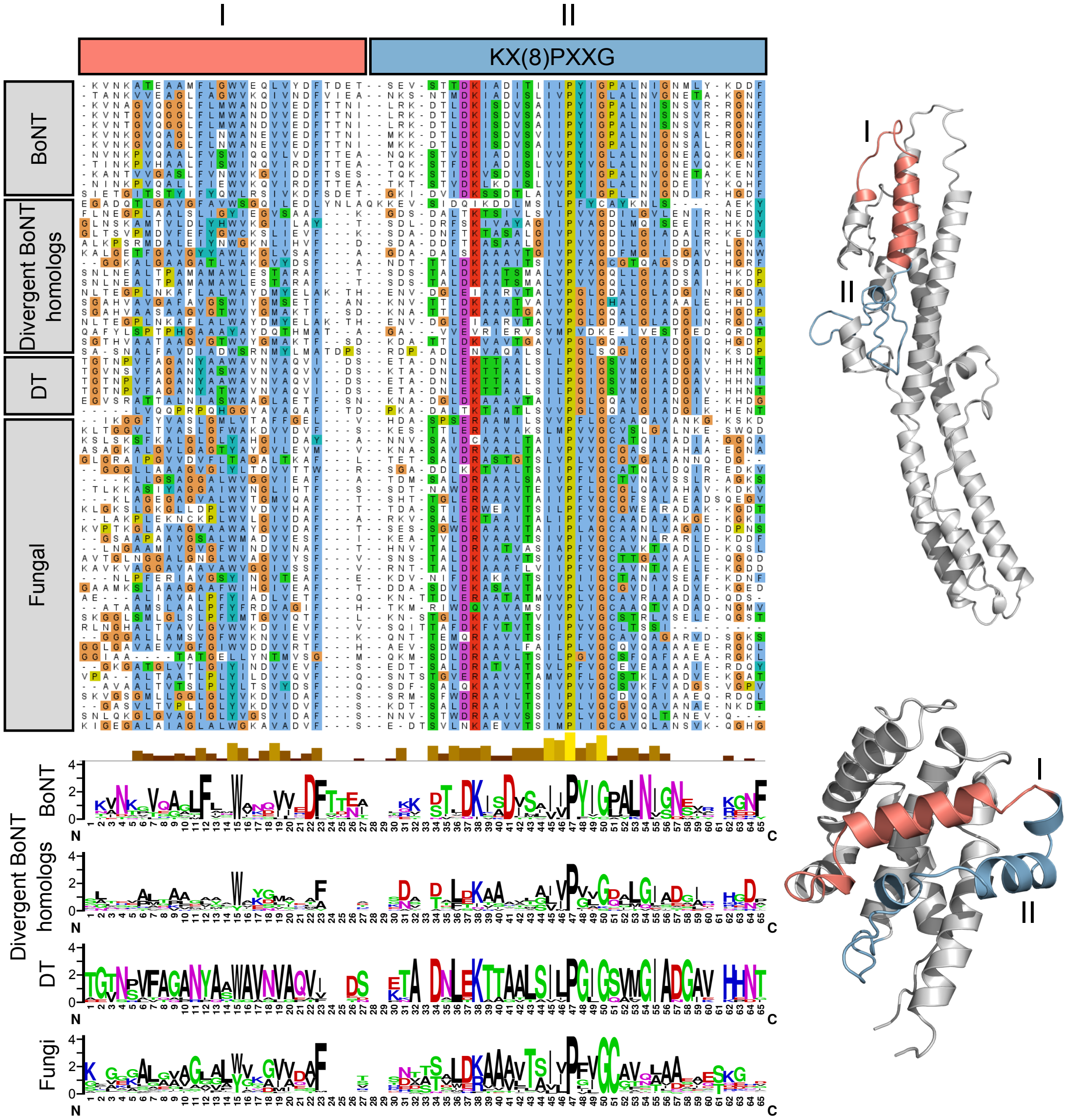


Figure 2.8: Translocase similarity between BoNTs, BoNT-like toxins and diphtheria toxins. A subdomain of BoNT translocases contains significant similarity to diphtheria T domains (DT in the above) as well as the translocase region of BoNT-like toxins. One segment of similarity (I) corresponds to the transmembrane TH5 helix in diphtheria toxins and a loop in BoNTs. A second region (II), including a K(x)8PxxG motif, is also strongly conserved. For ease of visualization, only a portion of the translocase alignment has been pictured here.

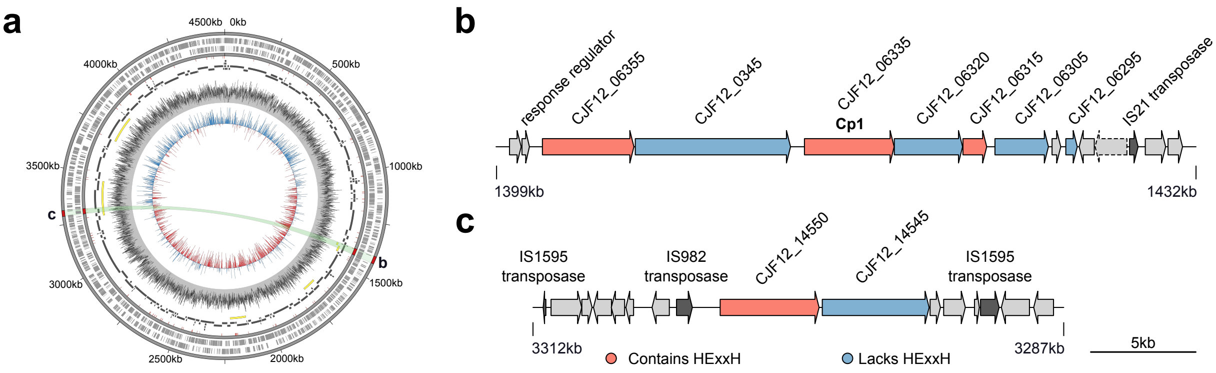


Figure 2.9: BoNT-like toxins reside within two toxin gene clusters in the genome of *Chryseobacterium piperi*. (a) Re-sequencing with a combination of Illumina and PacBio resulted in a closed genome with a single circular chromosome. Yellow bands reflect the local alignment of the toxin containing contigs (shown in b and c) from the initial JPRJ01 assembly against an intermediate assembly (black), and the closed genome (CP023049). Annotated insertion sequences (red ticks), SRX3229522 read mapping to CP023049 (gray histogram), and GC skew (blue/red histogram) are also indicated. The closed genome contains two separate toxin gene clusters (shown in b and c). The chromosomal loci that contain the toxin gene clusters have several notable features. First, there is evidence of gene duplication and pseudogenization among toxin genes, including an apparent split of a BoNT-like gene homolog into two genes, consisting of an HExxH-positive peptidase domain (CJF12_06315) and a "heavy chain" coding sequence containing a diphtheria-like translocase and a C-terminal ricin-type beta trefoil domains (CJF12_06305). Genes neighbouring the toxin gene cluster include putative response regulator genes (e.g., CJF12_06360, CJF12_06365, CJF12_06370, CJF12_06375), and several different types of transposases (e.g., CJF12_06285, CJF12_06290, CJF12_14525, CJF12_14555, CJF12_14620). These genes may potentially have a role in the expression or lateral transfer of these toxin gene clusters. Full gene annotations for the clusters are available in Table A.2.

Several genomic features surrounding the *C. piperi* toxin gene clusters indicate an origin via mobile element insertion. First, homologous regions to GC1 and GC2 were not detected in any other available *Chryseobacterium* genomes, suggesting non-*Chryseobacterium* origins. Second, numerous transposases are present including two IS110 family transposases, a IS200/IS605 transposase 18-40 kbp upstream (CJF12_06430, CJF12_06460, CJF12_06500), and an IS1182 family transposase (CJF12_05985) 30 kbp downstream of GC1. IS110 transposases have been previously shown to flank other *bont* gene clusters [151]. A disrupted IS982 family transposase pseudogene (CJF12_14555) is located immediately upstream of CJF12_14550 and flanking GC2 are complete and partial IS1595 family, ISChpi, insertion sequences (CJF12_14525 and CJF12_14620). Third, genes neighbouring *Chryseobacterium* toxin gene clusters were found to possess homology to genes in *M. chelonae* (e.g., the closest homolog of CJF12_14560 was *M. chelonae* WP_064393402.1, with 71% amino acid identity), consistent with the detected similarity between the *C. piperi* toxins and *M. chelonae* genes (Figure 2.5).

C. piperi **toxin is a novel metalloprotease toxin that induces necrotic cell death**

Given the substantial sequence variation between predicted *C. piperi* toxins, we selected WP_034687872 (Cp1) for experimental characterization based on it having the greatest sequence similarity to BoNTs among the *C. piperi* toxins over catalytic and functional sites (around 35% amino acid similarity). Initial protease assays of recombinant Cp1-LC against known BoNT substrates, including syntaxin 1, VAMP2, and SNAP25, yielded negative results (Figure 2.10). Although Cp1-LC did not display activity against canonical BoNT targets, the conservation of the active site residues and similarity to M27 and M91 metallopeptidases suggested the possibility of other targets. We elected to test for broad, metallopeptidase-induced toxicity via transfection and subsequent expression of the Cp1 LC cDNA in human embryonic kidney HEK293T cell line. Two Cp1-LC mutants containing point mutations at the HExxH motif (H209A and E210Q), were utilized as negative controls.

As shown in 2.11, the expression of wild-type Cp1-LC resulted in a cell death phenotype in HEK293T cells. These cells stopped proliferating and were visibly shrunken, eventually dying and detaching from culture plates. Cell counts after 48 hours revealed an almost 4-fold reduction in the number of cells (Figure 2.11). No cell death phenotype or significant reduction in cell number was observed in the H209A and E210Q mutants, confirming that the observed toxicity is likely metalloprotease-dependant. To further confirm the effect of Cp1-LC, we performed cell apoptosis assays by flow cytometry using Hoechst 33342, YO-PRO-1, and propidium iodide (Figure 2.11). In this assay, live cells, apoptotic cells

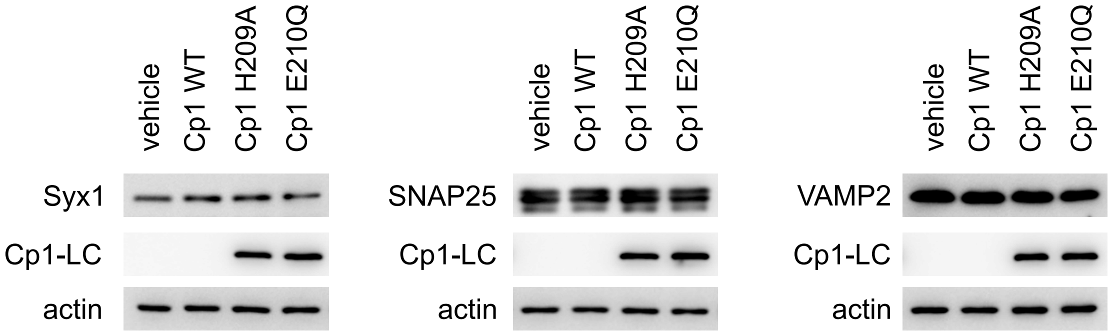


Figure 2.10: Cp1 does does not cleave canonical BoNT substrates. HEK293T cells were transfected with vehicle vector, Cp1-WT, and two Cp1 mutants that abolish metalloprotease activity (H209A and E210Q), and separately with plasmids containing syntaxin, SNAP25, and VAMP2. Cells were then lysed in RIPA buffer, and cleavage of SNARE proteins was evaluated by mixing lysates from SNARE-producing cells with vehicle, Cp1-WT and Cp1 mutant strains, followed by incubation at 37 °C for 30 minutes. The samples were then analyzed by immunoblot with anti-syntaxin 1, SNAP-25 and VAMP2 antibodies; Cp1 proteins were blotted with anti-FLAG antibody. Actin served as control for equal sample loading.

(green), and dead cells (bright red and some green) are visualized by fluorescence. The percentage of necrotic death was much higher in cells transfected with Cp1-LC than in cells transfected with control plasmid, H209A, or E210Q mutants (greater than 10-fold increase). In contrast, the percentage of cells labeled as apoptotic death did not change appreciably. These results suggest that expression of Cp1-LC leads to necrotic death of cells, and that the cell death depends on the protease activity of Cp1-LC, although the specific target(s) of Cp1-LC remains unknown at this point.

2.2.4 Discussion and Conclusions

Our survey of existing bacterial genomes reveals a diverse set of BoNT-related proteins outside of *Clostridium* such as those present in *Chryseobacterium* spp., a genus that includes pathogens of non-human hosts. Compared to the recently discovered BoNT in *Enterococcus faecium* (BoNT/En, ∼29% overall sequence identity to BoNTs across full length sequence) and the BoNT homolog in *Weissella oryzae* (∼20% overall identity to BoNTs), the *Chryseobacterium* BoNT-like toxins are more distantly related to BoNTs (∼15% overall identity to BoNTs). This is supported by sequence and phylogenetic analysis of the light chain, which demonstrates that the homologs from *C. piperi* and other species cluster outside of the BoNT family (Figure 2.5). In addition, there are numerous other features

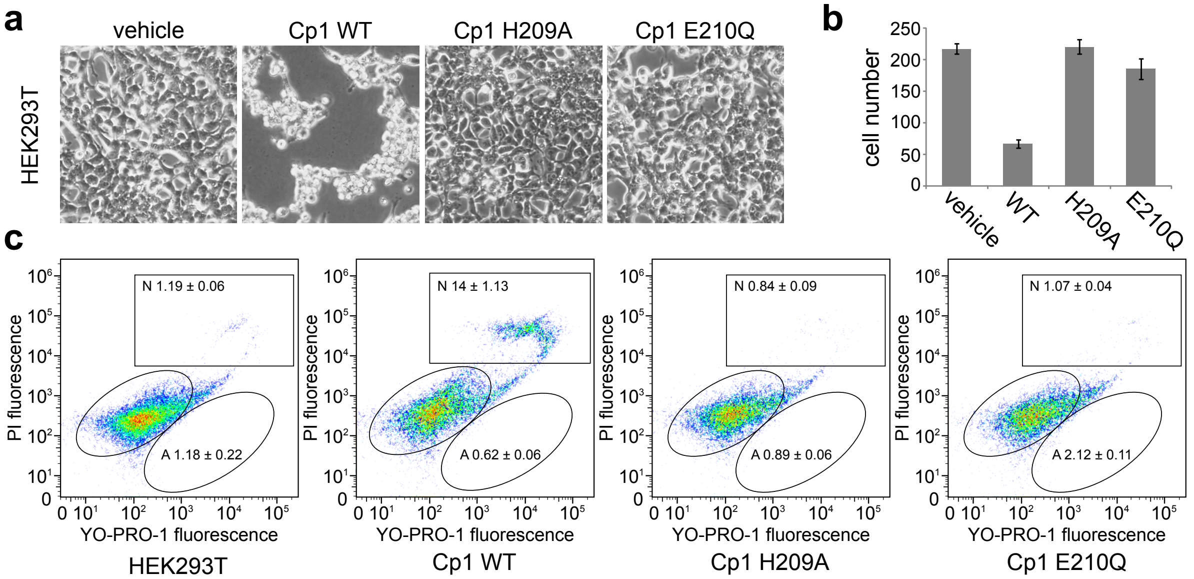


Figure 2.11: Expression of Cp1-LC in HEK 293T cell caused cell death, which is metalloprotease activity dependent. (a) HEK 293T cells were transfected with vehicle vector (pcDNA3.1(+)), Cp1-LC WT, and two mutants (H209A, E210Q) which abolished metalloprotease activity. Cells were observed and images were taken under inverted microscopy after 48 hrs. Cells transfected with Cp1-LC WT showed a dramatic cell death phenotype, becoming shrunken and detaching from the plate. Cells transfected with vehicle vector and protease activity-abolishing point mutants were not affected. (b) Cell numbers were counted in defined field of images taken in (a). The number of cells transfected with Cp1-LC WT were dramatically reduced compared to other treated cells. (c) Cell apoptosis assay was carried out with Chromatin Condensation/Membrane Permeability/ Dead cell Apoptosis kit. Transfected cells with vehicle vector, Cp1-LC WT and mutant plasmids were analyzed by flow cytometry. Transfection with the Cp1-LC WT plasmid increased the necrotic population of HEK 293 T cells, which is not observed in cells transfected with the vehicle vector or Cp1 point mutants.

present in the BoNT family that are lacking in *C. piperi*; specifically, the *C. piperi* toxins have weaker alignments to the translocation domain, lack the belt region, lack a detectable LamG-like H_{CN} subdomain present in the BoNT family, and lack the toxin accessory genes neighbouring the proteolytic toxin gene (including NTNH, and HA or P47/OrfX proteins). These trait differences suggest that, if BoNTs and *C. piperi* toxins do indeed share common ancestry, one of two scenarios have taken place: either the divergent BoNT homologs have lost some of these key BoNT features, or alternatively, these features emerged along with the BoNT lineage and may have differentiated BoNTs from their evolutionary relatives. Although one cannot distinguish between these two scenarios conclusively, we postulate that the latter scenario is more likely given the increased taxonomic and sequence diversity observed in the divergent BoNT-like toxin lineages.

The cytotoxic activity of *C. piperi* toxin in HEK293 cells, combined with the lack of protease activity against common BoNT substrates, suggests that the *C. piperi* toxin may have different targets in human cells. Given the degree of sequence and structural conservation observed between the protease domains of *C. piperi* toxins and BoNT-LC, it is tempting to speculate that *C. piperi* toxins may target different proteins with characteristics similar to SNAREs such as coiled-coil motifs. The future identification of the substrates targeted by *Chryseobacterium* toxin, *Weissella* toxin and others, combined with determination of their structure, will be important for not only illuminating the function and mechanism of these new toxins, but understanding the evolutionary novelties that occurred within the BoNT LC responsible for its gain of activity against neuronal SNAREs. Further, it will be important to explore the functionality of full-length Cp toxins and to determine whether they are expressed in their native host organism. Finally, if the protease domains of BoNT-related toxins have altered specificity to BoNT-LC, this may have significant biomedical and biotechnology applications through the engineering of BoNT-derived therapeutics that target different cell types.

2.3 Genomic insights into the evolution and ecology of botulinum neurotoxins

2.3.1 Introduction

Clostridial neurotoxins (CNTs) include both botulinum neurotoxins (BoNTs) and tetanus neurotoxin (TeNT). Various *bont* gene types exist, and can be classified into seven distinct serotypes (plus TeNT) based on their serological properties, designated by the letters A-G. Each serotype contains subtypes designated with roman numerals (e.g. BoNT/A1) [167]. Several clearly identifiable recombination events have resulted in mosaic toxins comprised of more than one serotype, as in the C and D clades (termed C/D and D/C, based on which serotype represents a greater proportion of the sequence), as well as an F5/A hybrid [151, 208]. CNTs are found in various *Clostridium* spp. and *Clostridium tetani*. *Clostridium* species are Gram-positive spore-forming anaerobes and are ubiquitous in the environment (e.g. soil, water and sediments), but can also be found in the gastrointestinal microbiota of humans, animals and other species. CNTs and clostridia have undergone a separate evolution, with clostridia forming an ancient class of bacteria with a complex phylogenetic history, and CNTs emerging more recently in a subset of clostridial lineages. Genes encoding BoNT proteins are found in the species *C. botulinum*, *C. argentinense*, *C. sporogenes*, *C. baratii* and *C. butyricum*, although the species *C. botulinum* is polyphyletic, and not all strains of these species are neurotoxigenic [209, 210, 88]. Consistent with the mobility evidenced by their scattered phyletic distribution in *Clostridium*, *bont* genes can be encoded on the chromosome, plasmids or phages. They are found within characteristic gene clusters, typically flanked by genes that enhance mobility, such as transposases and insertion sequence elements [211, 212, 213, 214, 88]. TeNT is encoded on a virulence plasmid of *C. tetani* of variable composition, but all contain the tetanus regulator *tetR* and a collagenase gene [215].

BoNTs are the most potent known toxins, and have evolved a unique and complex mechanism of toxicity. BoNT is composed of an N-terminal zinc metalloprotease domain (the light chain) which is disulfide linked to a heavy chain composed of a translocation domain and two binding domains. The mechanism of action involves entry into a host (e.g. mammal, bird or fish) through a wound or absorption through the gut, circulation of the toxin through the bloodstream and lymphatic system, binding to receptors on the presynaptic membrane of neuromuscular junctions, delivery into neurons by receptor-mediated endocytosis, pH-mediated translocation into the cytoplasm and cleavage of intracellular SNARE proteins, thereby inhibiting neurotransmitter release and ultimately leading to flaccid paralysis.

Aiding BoNTs are additional neurotoxin-associated proteins (NAPs) which are encoded within *bont* gene clusters. These include an immediate downstream non-toxic paralog (non-toxic non-hemagglutinin, or NTNH), as well as either hemagglutinin (*ha*) genes (encoding HA proteins) or *orfX* genes (encoding OrfX proteins), dividing the gene clusters into two main types. Strains may possess one or more toxin gene clusters, which may be the same or different toxin cluster types. NTNH and the three HA proteins form complexes that associate with BoNT, and these complexes protect the toxin in the gut environment and mediate attachment to the gut epithelium via E-cadherin [148]. There is some evidence for additional roles for the "nontoxic" NAP complex [216].

Although much work has been done to elucidate these properties and other mechanistic details of the CNTs, the evolutionary history of the CNT protein family remains mostly unknown. Here, we synthesize the information about the evolution of BoNTs based on the available literature in combination with an analysis of recently described BoNT-related sequences derived from microbial genomes. Together, these data provide insights into how the CNT family may have originated and diversified, and at the same time highlights some of the outstanding questions in the evolution of CNTs. Finally, we propose potential future directions and hypotheses for additional explorations of CNT evolution.

2.3.2 Methods

Data sources

A data set representing known BoNT subtype amino acid sequences was collected, a list of which is available from Peck et al. [167], with the addition of BoNT/X, BoNT/En, and the BoNT-like sequences from *Weissella* (NCBI accession number WP_027699549.1) and *Chryseobacterium* (WP_034687872.1). A list of accession numbers used is available in Table A.3.

Phylogenetic analysis

Full-length BoNT sequences were aligned using Clustal Omega version 1.2.1 [130] with default parameters. A maximum-likelihood BoNT phylogeny was generated using RAxML version 8.2.4 [135] with automatic substitution model selection and four gamma-distributed rate heterogeneity categories, with 1000 rapid bootstraps. For Figure 2.12, the tree was midpoint rooted and visualized using FigTree (available from `http://tree.bio.ed.ac.uk/software/figtree/`). The features mapped on the tree were determined from a number of sources, available in Table A.3.

2.3.3 Results

Where did BoNTs come from?

Although there is considerable knowledge on the BoNT mode of action (see Rossetto, Pirazzini and Montecucco 2014 [142] and Pirazzini et al. 2017 [100] for recent reviews), the physiological benefit of *bont* genes to their bacterial hosts is less clear. The BoNT protein does not appear to provide an obvious fitness benefit to its producer, nor is it essential for survival [217, 154]. With this in mind, how and why did such a toxin evolve? This fundamental question about the origins of botulism and tetanus has been a mystery since these toxins were first discovered over a century ago. The origins of BoNTs have been elusive for several reasons, but the most important is that, prior to a few years ago, there were no known evolutionary relationships between BoNTs and other protein families that could shed light on their ancestry.

In a 2006 paper, B.R. DasGupta speculated on the evolutionary origins of these toxins, suggesting that their ancestors were not found within the genus *Clostridium*, but instead were ultimately derived from viral polyproteins. This hypothesis was based on several pieces of evidence, including the independent action of different domains within the toxin, the observation that viral metalloproteases also known to cleave cellular proteins, and the conceivable notion of viruses as quintessential vectors for the transfer of foreign DNA into a host genome. It was also suggested that there is no reason that such toxins should necessarily be limited to cleaving SNARE proteins from vertebrates, and that the toxin family may be far more ancient.

Roughly a decade later, our research group provided evidence for DasGupta's hypothesis by identifying a distant, but clearly related, homolog of BoNTs in the genome outside of the *Clostridium* genus [99]. This protein, encoded within the genome of an organism called *Weissella oryzae* SG25, possesses all the properties that are expected of a divergent homolog: remote sequence similarity but with detectable homology, conserved domain architecture, and conserved active sites and functional motifs. Follow-up work by Zornetta et al. [101] then confirmed the *W. oryzae* BoNT-like toxin to possess similar activity to BoNTs by cleaving VAMP2, but interestingly, at a unique W-W bond unlike any other BoNTs. Whether this protein is truly a neurotoxin, however, remains unknown, as does its preferred host or even cell type. Why such a toxin exists within *Weissella*, a lactic-acid fermenting species isolated from the grains of Japanese rice, is also unclear. Regardless of these uncertainties, the *Weissella* homolog provided an indication that there may exist a larger family of divergent BoNT-like toxins in non-clostridial species that could provide insights into their ancestry and evolution.

Following the discovery of the *W. oryzae bont*-like gene, numerous additional BoNT-like toxins have now been identified by genomic data mining methods. First, a novel BoNT-like protein in *C. botulinum* str. 111, nominally named BoNT/X, was shown to result in botulism-like flaccid paralysis when injected into mice [155]. This activity was caused by catalytic cleavage of VAMP, a canonical BoNT target, which was cleaved at a novel site. Furthermore, Zhang et al. recently identified and characterized a BoNT-like toxin in *Enterococcus faecium* str. DIV0629 [2]. This *Enterococcus* BoNT-like gene, “BoNT/En”, possesses all the hallmarks of a BoNT: the presence of all BoNT domains and key functional motifs, and is located within an *orfX*-like gene cluster. Injection of the full-length BoNT/En protein into mice did not result in paralysis, but injection of the BoNT/En light chain ligated to the BoNT/A heavy chain resulted in flaccid paralysis. Thus, BoNT/En is unable to specifically bind and enter mouse neurons, but is capable of cleaving neuron substrates. Indeed, expression of the toxin in cultured neurons reveals surprisingly broad protease specificity, cleaving VAMP2 and SNAP25 at unique sites, as well as several additional SNARE proteins (i.e. SNAP23) with lower efficiency. Like *Weissella* toxin, BoNT/En appears to have been laterally transferred into this strain of *Enterococcus* from an unknown source, and so its functional role in *Enterococcus* is unclear. Independently, the BoNT/En gene cluster has been identified bioinformatically by two other groups: Williamson et al. [218] and Brunt et al. [181] who proposed the name eBoNT/J for this toxin.

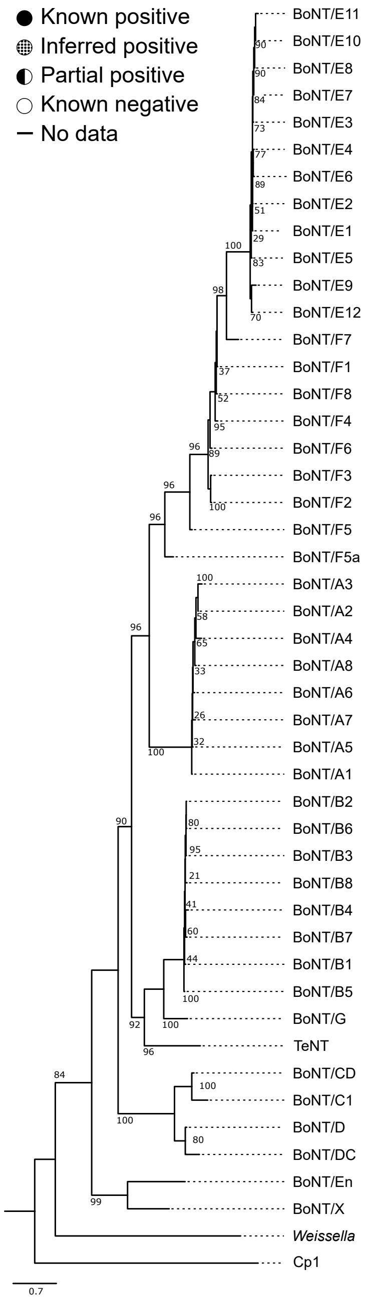
Mansfield et al. [219] and Wentz et al. [182] discovered a highly divergent group of BoNT-like toxins in the genome of *Chryseobacterium piperi* str. CTM, as well as additional BoNT-like genes in other Actinobacteria. One of the predicted *C. piperi* BoNT-like toxins induced necrotic cell death in human kidney cells, but was not found to cleave common SNARE substrates of BoNTs. Ultimately, the *C. piperi* toxins represent a distantly related group of toxins that share partial similarities to BoNTs, and provide a model for understanding the unique molecular changes that have occurred leading up to the BoNT lineage, particularly with respect to the evolution of the BoNT-LC. Additionally, it is important to note that, to date, the aforementioned BoNT-like proteins have only been examined biochemically in isolation from their host organism and biological context. Considerable work is needed to characterize the environmental conditions in which BoNT-like proteins are produced by their host organisms, the degree to which they are active, and their potential host and substrate specificity.

Lastly, fragments of *bont*-like genes have been detected in environmental metagenomic samples, including the termite gut [1]. Because these are only fragments and have not been assembled into larger contigs, it is unclear whether they are derived from full-length homologs of BoNTs. Further sequencing and analysis will be necessary to confirm whether these represent true BoNT homologs, and elucidate their relationship to other BoNTs. In this work, we provide an updated phylogenetic analysis that includes bona fide BoNTs and recent, computationally identified *bont*-like genes from non-clostridial species. Using the tree as an evolutionary framework, we discuss key questions regarding the evolution of BoNTs such as: What features emerged specifically in the lineage leading to BoNTs? How has host and substrate specificity evolved? What was the function of the BoNT ancestor? How have serotypes diverged? And what was the function of the BoNT's evolutionary precursors? Finally, we speculate on the possible ecological roles of BoNTs for their natural hosts, which may also shed light on the functions of other BoNT-like genes that will undoubtedly continue to be identified in future genomes and metagenomes.

A phylogeny of BoNTs and BoNT-like genes

Figure 2.12 reveals an updated phylogenetic tree of BoNTs and BoNT-like proteins. Although construction of an accurate phylogeny is complicated by recombination and lateral gene transfer events, this is nonetheless a useful representation of relationships between the BoNTs. Mapped onto the tree are various conserved and variable features that provide insights into BoNT evolution. By including recently identified divergent BoNT homologs (e.g. *Weissella* toxin), it is possible to root the BoNT phylogeny and infer evolutionary directionality in the tree, as well as estimating a possible order in which various clades emerged. The tree is largely consistent with previous phylogenetic analyses [220, 181], dividing the family into serotypes. BoNT/E and F form one major group, which forms a larger clade with BoNT/A. This clade is sister to a clade containing B, G and TeNT. The clade containing C and D appears to have diverged earlier in the history of the family. Finally, BoNT/En and BoNT/X cluster together along an early diverging lineage that forms a sister group to the other BoNTs. Beyond BoNT/X and BoNT/En is the *Weissella* toxin which diverged before the lineage leading to all BoNTs, and may even predate the origin of the BoNT clade. Finally, the *Chryseobacterium* toxin represents a divergent sister lineage to BoNTs and the *Weissella* toxin, outgrouping the entire tree. While it shares detectable homology with BoNTs, it also lacks numerous BoNT features, and is thus better considered a distant homolog.

● Known positive
⊕ Inferred positive
◐ Partial positive
○ Known negative
– No data



	Location			Gene cluster			Protein domains				Motifs		Substrate		
	Chromosome	Plasmid	Phage	NTNH	HA	ORFX	Peptidase	Translocase	N-terminal	C-terminal	LC-HC C-C	SxWY motif	VAMP	SNAP25	Syntaxin
BoNT/E11	●	○	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E10	●	●	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E8	⊕	○	○	⊕	○	⊕	●	●	●	●	●	●	○	⊕	○
BoNT/E7	⊕	○	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E3	●	●	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E4	●	○	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E6	⊕	○	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E2	●	○	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E1	●	●	○	●	○	●	●	●	●	●	●	●	○	●	○
BoNT/E5	●	○	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E9	⊕	○	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/E12	●	○	○	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/F7	●	○	○	●	○	●	●	●	●	●	●	●	⊕	○	○
BoNT/F1	●	○	○	●	○	●	●	●	●	●	●	●	●	○	○
BoNT/F8	●	○	○	●	○	●	●	●	●	●	●	●	⊕	○	○
BoNT/F4	●	○	○	●	○	●	●	●	●	●	●	●	⊕	○	○
BoNT/F6	●	○	○	●	○	●	●	●	●	●	●	●	⊕	○	○
BoNT/F3	–	–	–	●	○	●	●	●	●	●	●	●	⊕	○	○
BoNT/F2	○	●	○	●	○	●	●	●	●	●	●	●	⊕	○	○
BoNT/F5	–	–	–	●	○	●	●	●	●	●	●	●	●	○	○
BoNT/F5a	–	–	–	●	○	●	●	●	●	●	●	●	●	○	○
BoNT/A3	○	●	○	●	○	●	●	●	●	●	●	●	○	●	○
BoNT/A2	●	●	○	●	○	●	●	●	●	●	●	●	○	●	○
BoNT/A4	○	●	○	●	○	●	●	●	●	●	●	●	○	●	○
BoNT/A8	●	○	○	●	○	●	●	●	●	●	●	●	○	●	○
BoNT/A6	–	–	–	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/A7	–	–	–	●	○	●	●	●	●	●	●	●	○	⊕	○
BoNT/A5	●	○	○	●	●	○	●	●	●	●	●	●	○	●	○
BoNT/A1	●	○	○	●	●	●	●	●	●	●	●	●	○	●	○
BoNT/B2	●	●	○	●	●	○	●	●	●	●	●	●	⊕	○	○
BoNT/B6	○	●	○	●	●	○	●	●	●	●	●	●	⊕	○	○
BoNT/B3	–	–	–	⊕	⊕	–	●	●	●	●	●	●	⊕	○	○
BoNT/B8	–	–	–	⊕	⊕	–	●	●	●	●	●	●	⊕	○	○
BoNT/B4	○	●	○	●	●	○	●	●	●	●	●	●	⊕	○	○
BoNT/B7	–	–	–	⊕	⊕	–	●	●	●	●	●	●	⊕	○	○
BoNT/B1	○	●	○	●	●	○	●	●	●	●	●	●	●	○	○
BoNT/B5	○	●	○	●	●	○	●	●	●	●	●	●	⊕	○	○
BoNT/G	●	●	○	●	●	○	●	●	●	●	●	●	●	○	○
TeNT	○	●	○	○	○	○	●	●	●	●	●	●	●	○	○
BoNT/CD	○	○	●	●	●	○	●	●	●	●	●	○	○	●	●
BoNT/C1	○	○	●	●	●	○	●	●	●	●	●	○	○	●	●
BoNT/D	○	○	●	●	●	○	●	●	●	●	●	○	●	○	○
BoNT/DC	○	○	●	●	●	○	●	●	●	●	●	○	●	○	○
BoNT/En	○	●	○	●	○	◐	●	●	●	●	⊕	●	●	●	○
BoNT/X	●	○	○	●	○	●	●	●	●	●	⊕	●	●	○	○
Weissella	●	○	○	◐	○	○	●	◐	◐	◐	○	○	●	○	○
Cp1	●	○	○	◐	○	○	◐	◐	○	◐	⊕	○	○	○	○

Figure 2.12 ***(previous page)*****:** Maximum likelihood phylogeny of full-length BoNT amino acid sequences with bioinformatically identified BoNT-like homologs. Genomic locations, gene cluster contents, protein domains, conserved motifs, and light chain target substrates are indicated where possible. Partial positives: NTNH-like proteins in *Weissella* and *Chryseobacterium* might be unique paralogs rather than NTNH orthologs; BoNT domains are more difficult to detect in *Weissella* and *Chryseobacterium*. Inferred positives: the presence of conserved cysteine residues in BoNT/En, BoNT/X and Cp1 imply a similar disulfide bond, but this has yet to be observed experimentally. Most SNARE substrates are indicated as inferred positive, since we were unable to find experimental evidence of substrate cleavage for all subtypes.

Origin of TeNT and its surprising lack of diversity

In most respects, TeNT is similar to BoNTs. All BoNT domains are present in TeNT, demonstrating sequence homology across the length of the molecule, sharing an average of 36% identity with BoNT serotypes. Both TeNT and BoNTs function by binding and entering neurons and cleaving SNAREs; further, TeNT cleaves VAMP2 at the same site as BoNT/B [221]. The most important difference is the location of this cleavage event: BoNTs cleave SNAREs within cholinergic nerve terminals, while TeNTs undergo retrograde axonal transport and target inhibitory neurons in the central nervous system [222]. This is the cause of pathological differences between botulism, characterized by flaccid paralysis, and tetanus, which is characterized by uncontrollable spasms. The TeNT gene is found only in the species *C. tetani*, where it is encoded on a plasmid.

According to the tree, TeNT is not a sister lineage to BoNTs, nor is it even the most divergent member of the BoNTs (Figure 2.12). Rather, TeNT is nested within the BoNT tree as an early diverging member of the B + G clade. TeNT has a number of unique properties, the most obvious of which is the lack of NAPs. TeNT is not found in BoNT-like gene clusters, and is not flanked by *ntnh* or *ha/orfX* genes. Based on the phylogeny, the most parsimonious explanation is that the TeNT lineage has lost these NAPs rather than multiple individual gain events. The secondary loss of NAPs in the TeNT lineage suggests either that these proteins did not play an important role for the TeNT ancestor, or that TeNT gained a novel function to obviate the need for NAPs. Perhaps this reflects the differences in physiological environment between TeNT, which is produced in wounds, and BoNT, where NAPs may confer greater specialization for the gut environment. Overall, TeNT is most likely an "evolutionary outlier" in the history of the BoNT family, having lost ancestral traits while gaining novel functionality, and illustrates the potential for BoNT functional divergence.

Another anomalous characteristic in the TeNT lineage is the relative lack of sequence diversity among TeNT genes compared to the various BoNT serotypes. Comparative genomic analysis of *C. tetani* strains has confirmed this limited variation, and shown that despite variation elsewhere in the *C. tetani* genome, the tetanus toxin exhibits very low variation (>98% identity) [215, 223]. Despite this, it is notable that there is an increased proportion of variable sites in the binding domains, particularly in the H_{CC} domain, which could hint at positive selection for adaptive shifts in binding [224, 225].

Does the lack of TeNT sequence variation simply reflect a lack of sampling, where known TeNTs are simply a subset of a larger repertoire of TeNT-like sequences in the environment, or are there other functional or evolutionary constraints that limit TeNT sequence variation? The answer to these questions is still unclear.

Evolutionary diversification within the BoNT family

Given recently sequenced genomes of Clostridia [88] and the identification of BoNT-like toxins by genomic data mining [1], there has been considerable diversification within the BoNT tree. How and why have the different serotype lineages emerged and diverged from one another? Related to this question, why are the branches between serotypes so deep - between-serotype amino acid percent identities may be as low as 30% - with comparatively minimal variation within serotypes?

The structure of the BoNT tree is suggestive of an ancient adaptive radiation followed by extensive anagenesis in each lineage. One possibility that could explain this phenomenon is that different serotypes have been produced by host-pathogen co-evolutionary diversification [226, 227]. Co-evolution between host receptors and virulence factors is well established and is known to drive accelerated evolution and diversification [228, 229, 230, 231, 232]. Several lines of evidence support the idea of BoNT-host coevolutionary diversification. Different serotypes appear to exhibit host-specific adaptations: human botulism is mostly caused by serotypes A, B, E and rarely F; avian botulism is mostly caused by types C/D and E [233, 234, 235, 236]. Recent work suggests that even subtle changes in BoNT may result in an alteration or refinement of host specificity; for instance, under laboratory conditions human neuron sensitivity to BoNT/B was increased significantly by the mutation of a single residue in the binding domains [237]. While the extant serotype lineages may be the outcome of long-term host-pathogen coevolution, their initial diversification may have arisen through these types of minimal host specificity-altering amino acid substitutions.

Co-evolution has potentially occurred not only between BoNTs and their receptors but also between BoNTs and their SNARE substrates [224]. It has been suggested that sequence variation across vertebrate VAMP1 reflects a selective pressure to evade cleavage by BoNT [238]. In humans, however, BoNT cleavage sites are highly conserved, suggesting little to no selective pressure from BoNT [59].

Repeated evolution of substrate specificity in the history of BoNT

BoNTs have undergone several shifts and/or refinements in substrate specificity during the course of their evolution (Figure 2.12). The vesicle-associated SNARE VAMP, or synaptobrevin, is cleaved by serotypes B, D, F, G, X, TeNT, En and the *Weissella* BoNT-like protein, and the outer membrane-associated SNAP25 is cleaved by A, C, C/D and En (see Pirazzini et al. [100] for a comprehensive review of cleavage sites). BoNT serotypes typically cleave one of these SNAREs at a specific site, with the exceptions of BoNT/C, which also cleaves syntaxin, and BoNT/En, which cleaves both VAMP and SNAP25. BoNT/F subtypes are capable of cleaving VAMP2, but F5 and F5A cleave at a unique site.

The substrate specificity of BoNTs is diverse and has a scattered distribution on the phylogenetic tree (Figure 2.12), which may reflect inherent evolutionary plasticity in substrate recognition. By evolutionary parsimony, there are at least three substrate shifts that must have occurred in the evolution of BoNTs, with the BoNT ancestor most likely cleaving VAMP2 (which is also consistent with the recently determined specificities of divergent BoNT-like toxins). In this model, A, C and E have independently shifted toward a SNAP25 specificity from an ancestor possessing a VAMP2 specificity. This apparent parallel evolution toward SNAP25 specificity suggests that each of these lineages has separately adapted through different substitutions and substrate recognition mechanisms. This is consistent with the observation that A, C and E employ unique modes of SNAP-25 substrate recognition and cleavage [239].

It is puzzling that BoNTs have potentially undergone multiple substrate specificity shifts while retaining the conserved function of SNARE cleavage. Why, for example, have BoNTs not diverged in substrate specificity to cleave one of the thousands of other possible targets in the cell? Determining the events that have contributed to substrate specificity shifts is complicated by the extended contacts between substrates and the light chain, as exosites distant from the active site significantly contribute to binding efficiency, and the importance of different exosites varies among serotypes [240, 241].

Evolution of the BoNT ancestor

In addition to questions exploring the diversification of BoNT serotypes and subtypes, a fundamental evolutionary question concerns the origin of BoNT itself. It can be reasoned that if the BoNT clade is defined by its neurotoxic function, then the ancestral lineage leading to BoNT is that separating the *Weissella* and *Chryseobacterium* toxins from BoNTs (Figure 2.12). Thus, although the first neurotoxic BoNT likely originated within an ancestral *Clostridium* species, as suggested previously [220], earlier BoNT-like ancestors may have existed elsewhere (the phylum Actinobacteria has been suggested as one possible source [219]).

Although the function of the divergent BoNT-like toxins is still unclear, it is possible that some of them are cytotoxins rather than neurotoxins, and may even target non-neuronal SNAREs as it has been suggested previously [217]. Interestingly, *C. piperi* toxin induces kidney cell necrosis and although its target is still unknown, it does not cleave common SNAREs and thus does not appear to be a neurotoxin. Similarly, despite cleaving VAMP2, *Weissella* BoNT-like toxin has yet to be confirmed as a *bona fide* neurotoxin and lacks some key features (e.g. disulfide bond separating the LC and HC) that are essential for BoNT-like neurotoxicity.

The features found uniquely in the BoNT clade which are absent in these early diverging lineages are nonetheless of considerable interest, as these may differentiate neurotoxic from cytotoxic activity. By evaluating BoNT features in a phylogenetic context (Figure 2.12), it is possible to infer which features may have been gained in the lineage leading to BoNTs, and may thus have contributed to the evolution of neurotoxicity. These include (but are not limited to) the following:

1. Emergence of the BoNT gene cluster
2. Evolutionary shifts in the "light chain" protease domain
3. Modification of a pre-existing translocation domain
4. Gain of the SxWY motif

Emergence of the BoNT gene cluster

An important innovation that coincides with the emergence of the BoNT lineage is the appearance of the *bont-ntnh* synteny and flanking hemagglutinin or *orfX* genes. The *bont-ntnh* synteny is a derived feature of all CNTs except TeNT. It is interesting that the more divergent relatives of BoNTs (*Weissella* and *Chryseobacterium*) lack most of the hallmarks of *bont* gene clusters. That is, although the *Weissella* BoNT-like toxin gene is located next to a paralog with some similarities to NTNH, there are no neighbouring *ha* or *orfX* genes. Similarly, numerous pseudogenes and duplicates are found near the BoNT-like toxin genes in *C. piperi*, and the immediately neighbouring genes are perhaps analogous to NTNH (they possess M27-like protease domains but lack HExxH motifs). However, the surrounding genes and the "NTNH-like" analogs found in *C. piperi* do not appear to be direct orthologs of those in BoNTs.

Based on this, it appears that the BoNT OrfX and HA gene clusters evolved in the ancestral BoNT lineage (marked in Figure 2.12). How did this occur and from where did these associated genes originate? It is quite clear that NTNH is a paralog of BoNT and evolved by a tandem gene duplication followed by divergence [242, 147]. The origin of the OrfX and HA proteins, however, is more unclear. Bioinformatic analyses have suggested that some of these proteins may be derived from an ancestral toxin gene cluster, since statistically significant sequence similarity was detected between HA proteins and other clostridial toxins (Doxey et al. 2008). Notably, homology was detected between the HA70 component of BoNT gene clusters and the major virulence factor, CPE, *C. perfringens* enterotoxin. This prediction has been further supported by the recently solved structure of HA70 (Amatsu et al. [243]; PDB identifier 3WIN), whose closest structural neighbour in the Protein Data Bank is also CPE (PDB identifier 3AM2, p-value = 3.57×10^{-9}). The HA33 component of BoNT gene clusters is homologous to ricin type beta-trefoil domains found outside of the BoNT gene cluster [244, 242], such as *Bacillus thuringiensis* toxins (e.g. WP_088070506.1) and *Lysinibacillus sphaericus* mosquitocidal toxin. The OrfX gene clusters also possess intriguing similarities to other toxin gene clusters outside of *Clostridium* spp., found in species of *Paenibacillus*, *B. thuringiensis* and others.

Based on these similarities, the HA and OrfX gene clusters appear to be derived from pre-existing toxin gene clusters that are present in pathogens that are related to *Clostridium*. It is therefore possible that an ancestral BoNT-like toxin fused together with a pre-existing HA or OrfX gene cluster by recombination and perhaps lateral transfer. Given the phylogenetic distribution of HA and OrfX homologs, it appears that this ancestral gene fusion may have occurred within an ancestral *Clostridium* species. Further bioinformatic analysis of the distribution of OrfX and HA genes in genomes and metagenomes may shed further light into the origin and evolutionary diversification of the *bont* gene cluster.

Origin and evolution of the BoNT light chain

The BoNT LC forms a distinct sequence family, and is classified as peptidase family M27 under the "MA" clan based on MEROPS [199]. Notably, in addition to the BoNT and TeNT light chains, this clan contains additional protease families, many of which are found in toxins, and include anthrax lethal factor (family M34), bacterial collagenase (M9) and IgA proteases (M26 and M64). Proteases are classified under the MA clan based on a shared catalytic mechanism and gluzincin motif (Xaa-Xbb-Xcc-His-Glu-Xbb-Xbb-His-Xbb-Xdd, where Xaa is hydrophobic or Thr, Xbb is uncharged, Xcc is any amino acid except Pro and Xdd is hydrophobic [245].

Based on these similarities of function and mechanism, it should be expected that the BoNT LC is ultimately derived from another family of MA proteases. Mansfield et al. [219] recently provided evidence for this evolutionary hypothesis by identifying similarities to peptidase family M91, a family of type III effector peptidases that includes *E. coli* NleD and also *Xanthomonas* HopH1. Interestingly, the divergent homologs of BoNTs such as *Chryseobacterium* toxins possess overlapping peptidase M91 and peptidase M27 annotations. The Pfam database lists M91 as the only domain family with detectable similarity to M27 (see Pfam identifier PF01742), further supporting a link between these families. As shown by Mansfield et al. [219], the M91 domain of type III effector proteases shares numerous catalytic features found in BoNT-L, including not only the HExxH motif but also the third zinc ligand E261, the catalytically important E350 and an RxxY motif (residue numbers relative to BoNT/A, PDB identifier 3BTA). There are additional functional similarities between the two families, as both peptidase M91 type III effectors and the BoNT light chain are bacterial protease toxins that cleave intracellular eukaryotic targets.

However, it is the differences between the two that are perhaps more interesting from an evolutionary standpoint. The M91 peptidase type III effector proteins, which consist of a single protease domain, are injected into host cells by the type III secretion system directly, while BoNTs enter cells by receptor-mediated endocytosis and subsequent translocation into the cytosol by the activity of the heavy chain. M91 proteases are not known to cleave SNAREs, instead cleaving JNK and p38 in host cells [246, 194], thereby disrupting the host immune system. Although these substrates are quite different, as are the effects of their cleavage, there is another possible functional similarity: p38 is involved in endosomal fusion and trafficking [247]. Thus, although the function of the hypothetical M91-LC ancestor is unknown, these commonalities suggest that it may have been involved in disrupting vesicle function.

One of the clear differences between BoNT light chain peptidases and the peptidase M91 family is the presence of two insertion regions, conserved among all BoNTs and present to some extent in BoNT-like proteins. The first insertion is found from residues Phe 282-Leu 345 in BoNT A1 (PDB identifier 3BTA), and includes residues that contribute to the SNARE substrate binding pocket. The second is found at the C terminus (Lys 371 to the end of the LC), and changes in these regions have significant effects on toxin activity [248, 203, 249]. Considering the different substrates of the two peptidase members, it is possible that these insertion regions play a role in the SNARE specificity unique to BoNT

LCs, and quite possibly the BoNT-like toxins as well.

The species in which a hypothetical M27-M91 ancestral protease first evolved is also an interesting question, as type III secretion systems are limited to Gram negative organisms, and CNTs are limited to Gram positives. Perhaps the idea put forth by DasGupta [217] of a viral origin of BoNT provides a partial explanation for the occurrence of BoNT-like genes in such distantly related taxa. At this point, however, the repertoire of BoNT-like genes outside of *Clostridium* is still too narrow to be able to infer a lineage in which BoNT-like toxins first evolved, and the directionality of the relationship between M27 and M91 proteases (i.e. which came first) is also unclear.

Origin and evolution of the translocation domain

The translocation domain is one of the most puzzling aspects of BoNT function and evolution. Iterative PSI-BLAST searches starting with BoNT translocation domains as queries detect homologous regions in BoNT/X and BoNT/En, as well as the divergent BoNT-like toxins from *Weissella* and *Chryseobacterium*. Thus, at least some portion of the translocase domain appears to have been present in the common ancestor prior to the emergence of true BoNTs.

The complexity of the translocation mechanism makes it unlikely that the translocase domain evolved uniquely in the BoNT ancestor. A more likely explanation is the repurposing of pre-existing, functionally similar protein domains, most likely from other toxins. Indeed, homologs of the translocation domain can be found in a large family of putative toxins from entomopathogenic fungi [219] (e.g. the region from residues 379 to 513 of NCBI accession XP_008602550.1). Many of these fungal sequences contain predicted N-terminal ADP-ribosyltransferase domains rather than BoNT-like proteases. Assuming these genes originated from a common source, the large phyletic distance between fungi and bacteria necessitates an ancestral lateral transfer event, although it is difficult to determine the directionality of this transfer. Regardless, these fungal sequences possess translocase-like domains that are more similar to BoNTs than any others in current databases, suggesting that the translocase domain exists outside of the BoNT family alone.

In addition to these entomopathogenic fungal toxins, another possible distant relationship to the translocase domain may be the T domain of diphtheria toxin [219]. Statistically significant sequence similarity between the two can be detected after two PSI-BLAST iterations, and common motifs can be identified which span most of the diphtheria T domain and an N-terminal portion of the BoNT translocase domain. This sequence similarity, and the presence of conserved motifs, is surprising considering the structural difference between the two translocase domains - the diphtheria T domain possesses a different fold than BoNT translocases. However, it is possible that these similarities provide evidence for the "molten globule" model of BoNT translocation in which translocation involves the unraveling of tertiary structures [250, 251, 252, 253, 142]. Thus, it is possible that the detected relationship between the translocase domains of BoNT and DT reflects structural similarities that only become apparent in the molten globule membrane insertion intermediate state. Importantly, however, if there is a relationship between the BoNT and DT translocases, it only involves the N-terminal portion of the BoNT translocation domain. The C-terminal portion of the BoNT translocation domain, which adopts an extended alpha helical bundle structure, is still of unknown origin.

Evolution of the binding domains

Both of the heavy chain binding domains (the N-terminal binding domain H_{CN} and C-terminal binding domain H_{CC}) are detectable in all BoNTs, including divergent variants in the *Weissella* BoNT-like toxin, and thus were likely present early in the BoNT ancestor. In the *Chryseobacterium* toxins, ricin-type beta trefoil domains can be detected in the C-terminal regions, which may be homologous to the BoNT H_{CC} domain, although the H_{CN} (LamG) domain is not currently detectable based on sequence similarity or structure prediction approaches (e.g. threading). Thus, it is possible the BoNT ancestor possessed a domain architecture similar to the *C. piperi* toxins, consisting of a BoNT-like LC and translocase domain, but only part of the receptor binding machinery. This ancestral H_{CC} domain may have targeted different cell surface receptors, which is consistent with the absence of the SxWY motif in the early-diverging BoNT-like lineages (Figure 2.12). Adaptive evolution of the H_{CC} and gain of the LamG (H_{CN}) domain may have been key evolutionary innovations that resulted in the emergence of neuron binding in the BoNT ancestor. An ancestral gain of the H_{CN} LamG domain is entirely possible given that the broader LamG domain family is widespread outside of BoNTs; LamG domains homologous to H_{CN} domains are present in non-BoNT proteins including *Vibrio* MSHA biogenesis protein MshQ (ALM69800), and ALP-like superfamily proteins from *Bacillus* spp. (WP_071711091.1). The gain of the H_{CN} LamG domain at the base of the BoNT lineage, potentially from other bacterial LamG-domain containing proteins such as these, may have brought with it an increased ability to bind phospholipids on neuronal cell surfaces [254, 255].

The numerous links to insects

One of the recurring themes that appears when analyzing homologs of BoNTs and the other members of the gene cluster is a link to insects. Homologs of OrfX and HA proteins can be found in insecticidal gene clusters; partial homologs of BoNTs can be detected in entomopathogenic fungi; BoNT-like gene fragments have been identified in insect gut metagenomes. Do these links reflect evolutionary similarities to invertebrate-specific BoNTs or BoNT-like toxins? Is it possible that some of the identified BoNTs are already adapted for invertebrate hosts? Or, do these similarities reflect other roles that insects play in the BoNT life cycle?

What is the ecological function of BoNT in environmental clostridia?

One last fundamental question about the origin of BoNT concerns not the evolution of its molecular function, but rather the adaptive value it provides for the bacterium in its natural environment [154]. It has been hypothesized that BoNT operates as a method for the spread of the pathogen through rapid killing of vertebrate hosts, as is commonly seen in cases of avian botulism. As described by Rossetto, Pirazzini and Montecucco [142], botulism in the wild is propagated through a life cycle involving vertebrate decomposition and invertebrate predation. BoNT-producing clostridia are ingested or enter wounds, kill the animal, the animal carrying *C. botulinum* spores is decomposed by other organisms such as necrophagous fly larvae, which pick up the spores, intoxicate additional animals (e.g. birds) when they are ingested, and the cycle continues. However, there is still one unanswered question regarding the role of BoNTs in this life cycle - what is the adaptive value of neuroparalysis? After all, there are many possible toxin modes of action that could kill an organism, so what is the ecological value of paralyzing the host with such extreme specificity?

According to forensic entomology, animal decomposition takes on a predictable succession of stages: fresh, bloat, active decay, advanced decay and dry decay, and each stage is associated with specific arthropod species that have adapted to that stage to efficiently use resources and proliferate [256, 257]. How might BoNT-induced paralysis influence a decomposition cycle such as this in the wild? By paralyzing the host, *C. botulinum* may effectively favor certain species of necrophagous invertebrates such as blowflies, effectively “freezing” the host before later stages of decomposition occur. A paralyzed host would provide fresh tissue before later stages of decomposition are initiated by the microbial necrobiome (anaerobic bacteria, fungi), and thus a major competitive advantage to these early-stage necrophagous insects, especially if they are also the vectors of *C. botulinum* spores. It is tempting to speculate that BoNTs may therefore have originally evolved due to competition between scavengers in decomposition. It is also conceivable that some BoNTs may have evolved broader host specificity to target not only the host vertebrate but also target competing necrophagous invertebrates.

Members of the Calliphoridae are known to harbor *C. botulinum* spores, lending credence to the idea of a link to necrophagous insects [258, 259]. The exact nature of the relationship of BoNTs to insects and decomposition, however, remains to be seen. It is also important to note that there is not one but potentially many ecological cycles that involve different BoNT-producing taxa and different hosts.

With this in mind, it is tempting to speculate that the divergent BoNT-like toxins from *Chryseobacterium*, *Weissella* and *Enterococcus* could be part of similar ecological cycles in the wild but with different host organisms than those typically studied in the context of bacterial pathogenesis. Alternatively, it is also possible that these BoNT-like toxins reflect remnants of earlier cytotoxin lineages that predate the evolution of BoNT and its neuroparalytic role in decomposition. Future studies that examine the ecological life cycle of *C. botulinum* in the wild, and evaluate the fitness impact of BoNT on C. *botulinum*, its vectors and the broader "necrobiome" community will likely shed light on these important questions.

2.3.4 Conclusions

Traditionally, BoNTs have been defined based on their neuroparalytic activity and their occurrence in *C. botulinum* and related species. In the post-genomic era that has generated over a hundred thousand bacterial genomes, it is still true that *bona fide* BoNTs exist predominantly within *Clostridium*, perhaps with the exception of the recently identified BoNT from *E. faecium* (BoNT/En). However, BoNT-like toxins with divergent activities exist outside of the *Clostridium* genus and may be remnants of a much older lineage of toxins. Several key features differentiate BoNTs from BoNT-like toxins and likely played a role in their evolution, including acquisition of NAPs, substrate specificity changes in an ancestral LC, extension/modification of the translocation domain, gain of the H_{CN} binding domain and adaptive changes in binding specificity within the H_{CC} domain. The precise order that these evolutionary events occurred, all of which may have affected the function of the ancestral toxin, is for the moment difficult to discern. All of the BoNT domain families appear to have existed in some form within the ancestral precursor to the BoNT lineage. Thus, it is conceivable that a three or four-domain BoNT-like cytotoxin existed in an ancient species (*Clostridium* or elsewhere) which targeted non-neuronal SNAREs, and subsequently adapted to become a neurotoxin.

Assuming that BoNTs have functionally differentiated from precursors that were already toxins, their true adaptive benefit likely relates to their neuroparalytic effect on the host. To understand this adaptive role, it is important to examine in detail the role of BoNTs in its ecological life cycle. We suggest that this role may be to delay the occurrence of decomposition to provide a competitive advantage to necrophagous invertebrates that are specialized for early stages, which in turn benefit toxigenic *C. botulinum* by facilitating its dispersal.

2.4 Comparative genomics and evolution of the BoNT-associated P47/OrfX gene cluster

2.4.1 Introduction

Botulinum neurotoxins (BoNTs) are produced by several species of *Clostridium* and cause flaccid paralysis in vertebrates by cleaving neuronal SNAREs, thereby preventing neuronal exocytosis [52]. The BoNT family includes at least 8 serologically distinct serotypes denoted A-H [167]. The identification of several BoNT-related sequences in various bacterial genomes has further expanded the family: BoNT/X in *C. botulinum* str. 111 [155], BoNT/En in *Enterococcus faecium*, and most recently PMP1 in *Paraclostridium bifermentans* [260]. More distant relatives include the tentatively named BoNT/Wo in *Weissella oryzae* [99] and Cp1 in *Chryseobacterium piperi* [182, 3].

The canonical BoNT serotypes are encoded adjacent to a sequence encoding a non-toxic paralog known as non-toxic non-hemagglutinin (NTNH), and the *bont-ntnh* genes are located within one of two characteristic toxin gene clusters. The first type is classified by the presence of hemagglutinins (HAs), which are thought to protect the toxin in the gut and contribute to binding the gut epithelium [148], but may have their own toxic effects [216], and notably have similarity to clostridial collagenases and *Clostridium perfringens* enterotoxin [242]. The second BoNT cluster type is characterized by OrfX proteins, which remain mostly uncharacterized. Structural and functional experiments have suggested that they play a role in binding lipids and increasing membrane permeability [261, 262], and more recent evidence suggests a role in increasing the toxicity of the BoNT-like insecticidal toxin PMP1 [260]. Toxin clusters containing *orfX* genes typically encode three OrfX proteins (OrfX1, OrfX2, and OrfX3) and a P47 protein. All four of these proteins are partially homologous to one another, with OrfX1 as the shortest protein. BoNT/X, BoNT/En, and PMP1 are encoded in OrfX-type clusters, and their toxin sequences form an early-branching lineage within the BoNT family [2]. The proteins in *Weissella* and *Chryseobacterium* are not associated with typical BoNT clusters. Tetanus neurotoxin (TeNT) is also anomalous in that the *tent* gene does not associate with either NTNH or OrfX/HA [263].

The OrfX-like proteins in *C. botulinum* str. 111, *E. faecium*, and *P. bifermentans* are dissimilar to those in canonical BoNT gene clusters (varying by ∼70% amino acid identity or more). However, the retention of the OrfX gene cluster across large species barriers suggests that they perform a key function in the proliferation of *bont* genes.

In order to better characterize the OrfX protein family, as well as gain insights into what their functions may be, we performed comparative genomic analysis of OrfX proteins and the gene clusters encoding them. OrfX gene clusters are found in a wide variety of organisms and are associated with diverse toxin-related sequences, suggesting that OrfX gene clusters are not unique to clostridial neurotoxins. Most notably, OrfX gene clusters are found in several taxonomically distinct clades of pathogenic bacteria that target insect hosts, which appear to undergo lateral gene transfer. Together, these findings imply that the BoNT OrfX gene cluster is one type of a large class of insect-related toxin clusters, which may also relate to the role of *orfX* genes in BoNT-type gene clusters.

2.4.2 Methods

Data set retrieval and curation

A representative set of OrfX proteins were retrieved by querying the NCBI non-redundant protein database by two iterations of PSI-BLAST [112], combined with two iterations of jackHMMer against the UniProt database [264]. The proteomes for each organism with a match to either model were downloaded and redundancy was removed, in order to limit the contributions of identical organisms and multiple versions of the same assemblies. Then, the *Clostridium*-specific sets of OrfX1 sequences and OrfX2, 3, and P47 sequences were separated from the search results, aligned using the L-INS-i algorithm of MAFFT (v. 7.407, [131]), and used to create two HMM profiles with HMMER (v. 3.2.1, available from `http://hmmer.org/`; [113]). These two models were used to search the non-redundant proteomes in order to improve the sensitivity of detecting all OrfX family members. For each OrfX locus, the region 10kb up- and downstream of the outermost *orfX* gene boundaries were extracted. Nucleotide fragments smaller than 10kb were therefore excluded. In total, the non-redundant set of 212 proteomes yielded 213 nucleotide contexts (*Brevibacillus laterosporus* str. 1951 contains two unique OrfX loci), which encode a total of 3684 protein sequences, and 645 OrfX-related proteins. A summary of the final data set is available in Supplementary Table A.4.

Genomic context analysis

For each nucleotide context, all-by-all translated BLASTX (TBLASTX; [103, 265]) searches were performed. Comparisons were made at the protein level in order to allow alignments between more distantly-related nucleotide segments, as well as allowing for genomic insertions, deletions, inversions through six-frame translation. In order to normalize the bit scores for each genomic context, which are variable in length, the bit scores for each pairwise comparison were summed together and divided by the maximum bit score (that is, the score of a self-match). This results in a proportional bit score for every pairwise comparison (or, a 213×213 matrix). This pairwise matrix was used to correlate the degree of similarity between genomic contexts using Pearson correlation, and was visualized as a heatmap using the R package pheatmap (v. 1.0.12, [266]) and as a directed network using igraph (v. 1.2.4, [267]). For the simplified network in Figure 2.13, weak edges (edge weight <0.01) and poorly connected vertices (degree <3) were dropped from the graph. The graph was laid out using the Fruchterman-Reingold algorithm as implemented in igraph.

Phylogenetic analysis

All OrfX1, OrfX2, OrfX3, and P47 sequences were collected and deduplicated to yield only unique sequences (reducing the total set of 645 OrfX sequences to 462 non-redundant sequences). These sequences were aligned using the L-INS-i algorithm of MAFFT (v7.407, [131]). A maximum likelihood phylogeny was inferred with the autoMRE bootstopping criterion, automatic substitution model selection in RAxML [135] and depicted in FigTree (available from `https://github.com/rambaut/figtree/`).

Genomic context and protein annotations

Genomic context diagrams were created using genoPlotR (v. 0.8.9, [163]). Gene annotations provided for gene clusters in Figure 2.14 were generated by the NCBI Prokaryotic Genome Annotation Pipeline [268]. Protein annotations used in Figure 2.8 were retrieved through the EMBL InterProScan web server [126, 269]. In the case of overlapping annotations, the domain start and end sites were enlarged to maximize the domain's annotated region.

2.4.3 Results

OrfX gene clusters are found in diverse taxa

OrfX proteins were detected through two iterations of PSI-BLAST [112] against the NCBI non-redundant protein database as well as the UniProt database [264]. These results were pooled, mapped back to proteomes (or segments comprising a full BoNT gene cluster), and identical strains were removed. In order to improve the coverage of the OrfX search, the OrfX homologs from *Clostridium* were aligned and used to generate two HMM models: a combined model for OrfX2, OrfX3, and P47, and a second model for OrfX1 sequences which frequently went undetected with a combined model (see Methods). The set of proteomes was then searched with the two models to capture all OrfX family members more sensitively. For each OrfX cluster in a proteome, the 10kb region upstream of the first *orfX* gene and 10kb downstream of the last *orfX* gene was extracted, yielding a final set of 213 nucleotide contexts containing 3684 protein sequences, 645 of which are OrfX-like.

Gene clusters encoding sets of OrfX-like proteins (hereafter referred to as OrfX gene clusters, or OrfX clusters) can be found in diverse taxa (Figure 2.13a). This includes 8 different bacterial phyla, the fungus *Fusarium*, and the red alga *Gracilariopsis*. The largest number of OrfX clusters is found in the bacterial phylum Firmicutes, as most gene clusters in the data set are derived from the genomes of *Clostridium* species (94 genomic contexts, 487 proteins). Proteobacteria are the second largest contributor (83 genomic contexts, 282 proteins), followed by the Actinobacteria (16 genomic contexts, 55 proteins). Outside of *Clostridium*, OrfX clusters are most commonly associated with *Pseudomonas*, *Erwinia*, *Paenibacillus*, *Streptomyces*, and *Brevibacillus* genomes. Although some of these genera are associated with human pathogenicity, the particular species and strains that appear to contain OrfX gene clusters are not generally considered significant causes of human disease. However, nearly all of the genomes containing OrfX gene clusters are associated with disease in other hosts. Many of these species are associated with phytopathogenicity: species of *Erwinia* are mostly regarded as plant pathogens [270, 271]; the most numerous OrfX-containing *Streptomyces* species, *S. scabiei*, is mostly known as a phytopathogen [272]; *Fusarium* species are fungal phytopathogens [273]. Interestingly, many other species containing OrfX gene clusters are insect pathogens. The *Pseudomonas* group mostly consists of environmental isolates and *P. putida*, which is a rare human pathogen but also has the ability to kill insects [274]; *Paenibacillus larvae* causes lethal disease in honeybees, while other members of the genus produce insecticidal toxins that promote plant growth [275]; some members of the genus *Brevibacillus* are insecticidal and nematocidal [276]; the genus *Arsenophonus* includes both insect symbionts and pathogens [277, 278]; and, OrfX-related proteins are found in at least one strain of *Bacillus thuringiensis*, which is well known for its insecticidal activity [279].

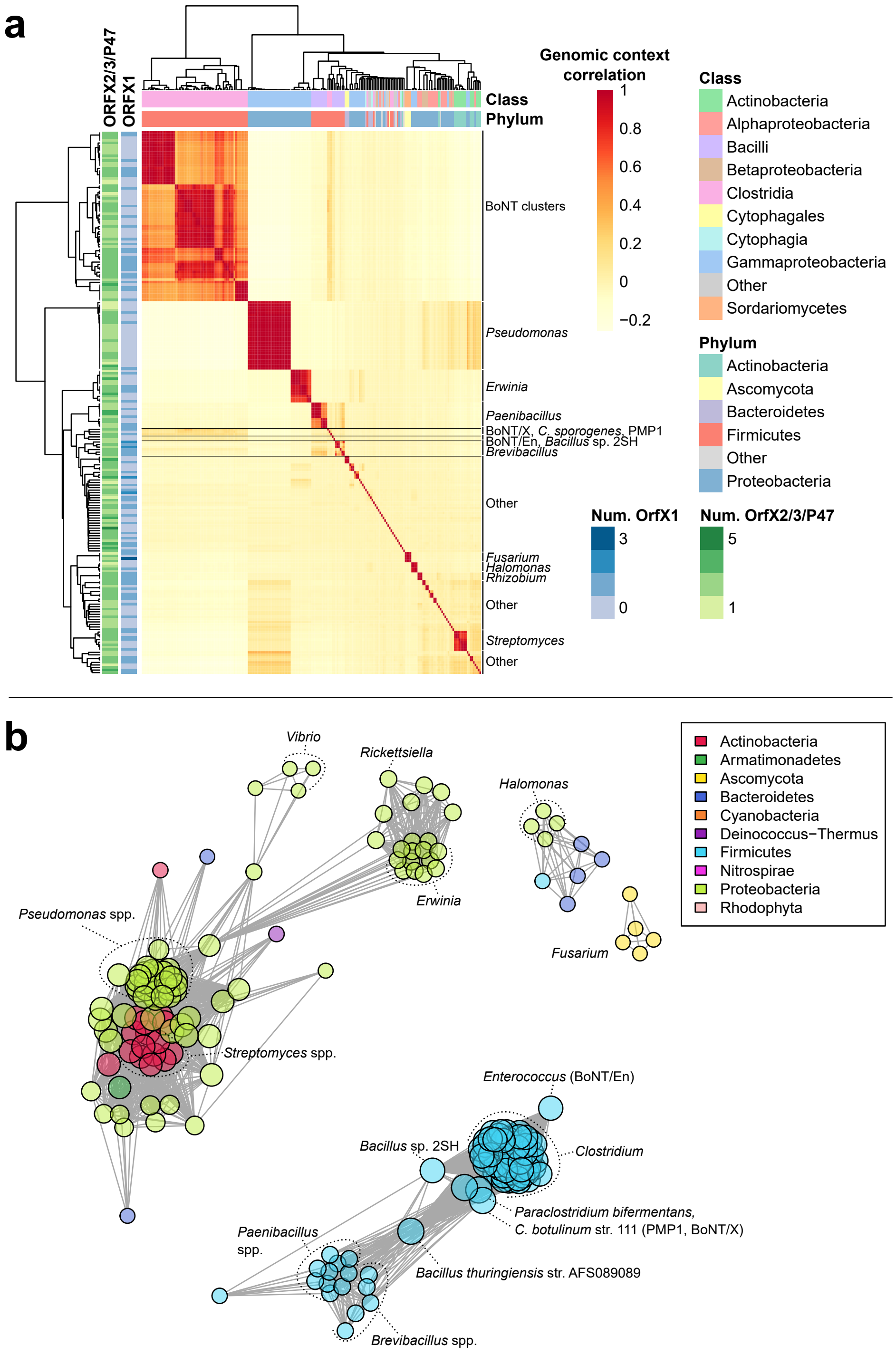

a
ORFX2/3/P47
ORFX1
Class
Phylum
BoNT clusters
Pseudomonas
Erwinia
Paenibacillus
BoNT/X, C. sporogenes, PMP1
BoNT/En, Bacillus sp. 2SH
Brevibacillus
Other
Fusarium
Halomonas
Rhizobium
Other
Streptomyces
Other
Genomic context correlation
1
0.8
0.6
0.4
0.2
0
−0.2
Class
Actinobacteria
Alphaproteobacteria
Bacilli
Betaproteobacteria
Clostridia
Cytophagales
Cytophagia
Gammaproteobacteria
Other
Sordariomycetes
Phylum
Actinobacteria
Ascomycota
Bacteroidetes
Firmicutes
Other
Proteobacteria
Num. OrfX1
3
0
Num. OrfX2/3/P47
5
1
b
Vibrio
Rickettsiella
Halomonas
Erwinia
Fusarium
Pseudomonas spp.
Streptomyces spp.
Enterococcus (BoNT/En)
Bacillus sp. 2SH
Clostridium
Paraclostridium bifermentans,
C. botulinum str. 111 (PMP1, BoNT/X)
Paenibacillus spp.
Bacillus thuringiensis str. AFS089089
Brevibacillus spp.
Actinobacteria
Armatimonadetes
Ascomycota
Bacteroidetes
Cyanobacteria
Deinococcus−Thermus
Firmicutes
Nitrospirae
Proteobacteria
Rhodophyta

Figure 2.13 *(previous page)*: OrfX gene clusters are found in diverse taxa. (a) A heatmap of correlations between scaled pairwise TBLASTX scores. Blocks of shared correlations between gene clusters are related to taxonomy, such as the large BoNT-containing OrfX clusters of *Clostridium* spp. The number of OrfX1-like or OrfX2/3/P47-like genes varies between OrfX gene clusters. (b) A simplified directed network (edge weights > 0.01, connectivity > 3, vertex size proportional to connectivity; see Methods) of correlations between pairwise TBLASTX scores from (a) highlights the association of BoNT-containing *Clostridium* clusters with other members of the phylum Firmicutes. The figure also highlights the relatively distant relationship between Firmicutes-type OrfX gene clusters and those found in other bacterial taxa.

The OrfX gene cluster content of the genus *Clostridium* is most closely related to the OrfX clusters of other members of the phylum Firmicutes in terms of sequence content (Figure 2.13). The closest relatives of the canonical BoNT OrfX gene clusters are found in *Enterococcus faecium*, *Paraclostridium bifermentans*, *C. botulinum* str. 111, which additionally encode BoNT/En, PMP1, and BoNT/X, respectively. The next nearest neighbours are OrfX clusters from *Bacillus* sp. 2SH and *B. thuringiensis* str. AF8089089, which additionally provide a link to the OrfX clusters in *Paenibacillus* and *Brevibacillus*. The weak correlations between any of the bacterial OrfX clusters with the OrfX loci in the fungus *Fusarium*, and the relatively strong correlations between *Streptomyces* spp. and *Pseudomonas* spp. together imply that lateral gene transfer events contribute to OrfX gene cluster diversity. Potential lateral transfer events have previously been noted for *Fusarium* species [280].

OrfX gene clusters associate with toxin-related genes

A common feature of OrfX clusters is the presence of toxin-related sequences. This includes BoNT-type OrfX clusters, as well as many diverse toxin-related sequences can be found in OrfX gene clusters (Figure 2.14). In *Brevibacillus laterosporus* (Figure 2.14) and *Paenibacillus larvae* several proteins have notable similarities to anthrax toxin components. Anthrax toxins are multimeric proteins comprised of protective antigen (PA) proteins, the adenylate cyclase anthrax edema factor (EF), and the metalloprotease anthrax lethal factor (LF) which additionally contains an inactive ADP-ribosyltransferase [55, 65]. Both *Paenibacillus* and *Brevibacillus* clusters encode proteins with similarities to anthrax PA, EF, and LF. RHS toxin-related sequences commonly associate with OrfX clusters, which contribute to virulence in *Vibrio* species [281], and comparative genomic analysis revealed that the family is highly diverse [282] and are also related to bacterial interspecific competition [283, 284]. Evidence of toxic functionality is also seen in more distantly related sequences, including ricin-type beta trefoil proteins in *Paenibacillus larvae* (a component of *Clostridium perfringens* enterotoxins [242] as well as ricin toxin itself [285]), type VI secretion system components in *Erwinia*, and a protein containing an adenylate cyclase domain in *Rhizobium* sp. Leaf386. As well, many OrfX clusters, including BoNT-type OrfX clusters, contain partial or complete transposases, which may contribute to their mobility (Figure 2.14).

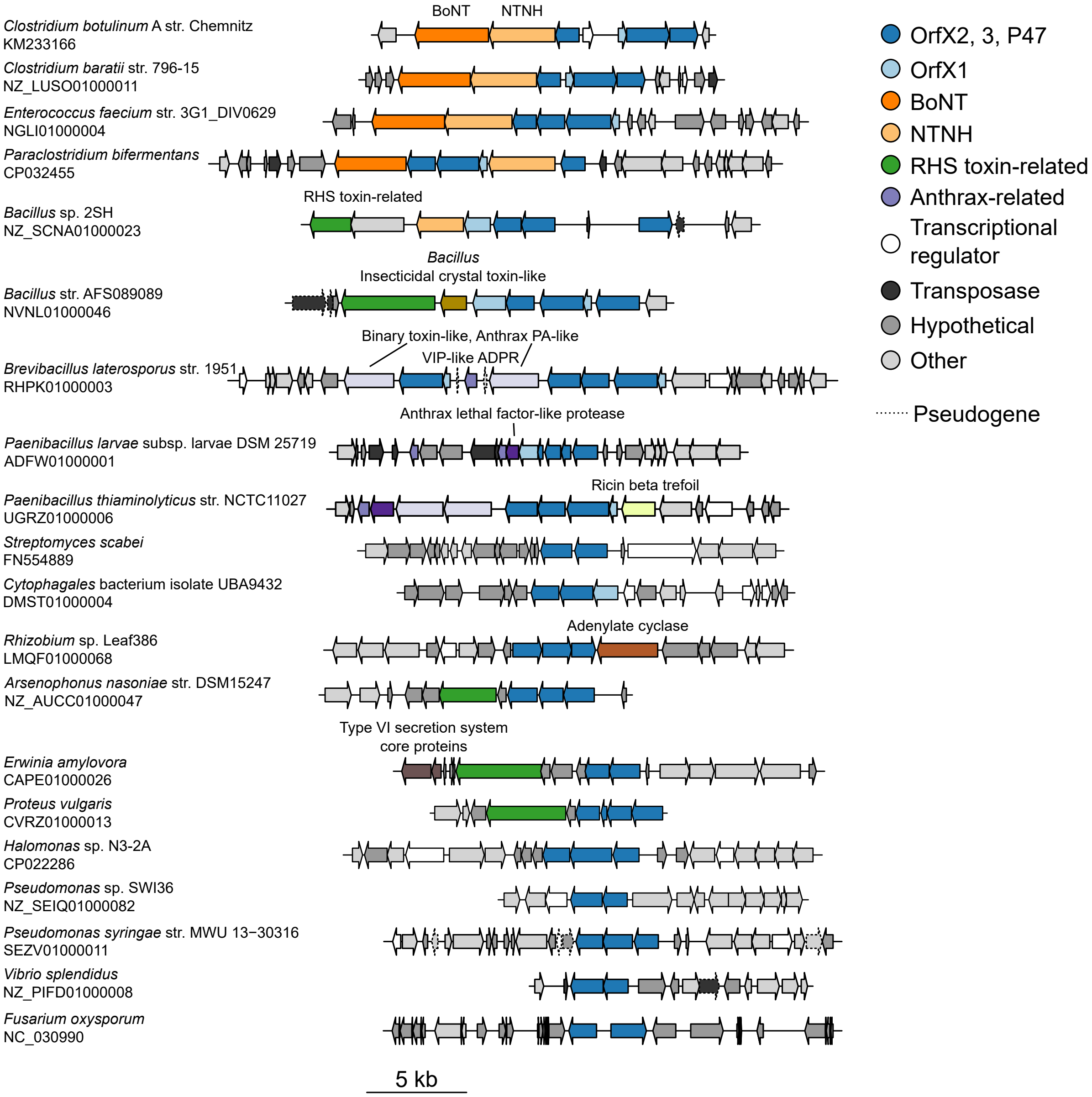

BoNT
NTNH
Clostridium botulinum A str. Chemnitz
KM233166
Clostridium baratii str. 796-15
NZ_LUSO01000011
Enterococcus faecium str. 3G1_DIV0629
NGLI01000004
Paraclostridium bifermentans
CP032455
RHS toxin-related
Bacillus sp. 2SH
NZ_SCNA01000023
Bacillus
Insecticidal crystal toxin-like
Bacillus str. AFS089089
NVNL01000046
Binary toxin-like, Anthrax PA-like
VIP-like ADPR
Brevibacillus laterosporus str. 1951
RHPK01000003
Anthrax lethal factor-like protease
Paenibacillus larvae subsp. larvae DSM 25719
ADFW01000001
Ricin beta trefoil
Paenibacillus thiaminolyticus str. NCTC11027
UGRZ01000006
Streptomyces scabei
FN554889
Cytophagales bacterium isolate UBA9432
DMST01000004
Adenylate cyclase
Rhizobium sp. Leaf386
LMQF01000068
Arsenophonus nasoniae str. DSM15247
NZ_AUCC01000047
Type VI secretion system
core proteins
Erwinia amylovora
CAPE01000026
Proteus vulgaris
CVRZ01000013
Halomonas sp. N3-2A
CP022286
Pseudomonas sp. SWI36
NZ_SEIQ01000082
Pseudomonas syringae str. MWU 13−30316
SEZV01000011
Vibrio splendidus
NZ_PIFD01000008
Fusarium oxysporum
NC_030990
5 kb
● OrfX2, 3, P47
● OrfX1
● BoNT
● NTNH
● RHS toxin-related
● Anthrax-related
○ Transcriptional regulator
● Transposase
● Hypothetical
● Other
········ Pseudogene

Figure 2.14 ***(previous page)*****:** Genomic context of a selected set of diverse gene clusters containing OrfX-related genes. As seen among BoNT-type OrfX gene clusters, OrfX gene clusters from other species typically contain more than one OrfX-related gene (blue). Althoug the gene clusters are heterogeneous, toxin-related genes are common among OrfX gene clusters (various colours). The BoNT-type OrfX clusters of *Clostridium* species typically encode BoNT and NTNH proteins (orange), while RHS toxin-like (green), *B. thuringiensis* Cry-like (brown), and anthrax toxin-like sequences are found in other species. Other proteins encoded in OrfX gene clusters may contain weaker evidence of toxin functionality such as domains associated with toxins, including ricin-type beta trefoils (yellow), adenylate cyclases (dark orange), and type VI secretion system proteins (red-grey.

orfX genes have undergone several independent lateral transfers and duplications

In order to disentangle distinct OrfX subfamilies, we generated a phylogeny based on a representative set of OrfX-related proteins (Figure 2.15). OrfX proteins associated with BoNT-type OrfX clusters group separately from those found in other taxa, and form well-supported and distinct clades consisting of OrfX1, OrfX2, OrfX3, and P47. The closest relatives of BoNT-type OrfX clades found among *Clostridium* spp. are typically the OrfX-like clades associated with BoNT-like toxin clusters, including those from *C. botulinum* str. AM1195, *C. botulinum* str. 111, *E. faecium*, and *P. bifermentans*. The relationship of OrfX-related proteins from BoNT and BoNT-like gene clusters is not monophyletic since OrfX-related proteins from *Bacillus* species, *Paenibacillus*, and *Brevibacillus* occasionally group more closely to one another. However, for each of OrfX1, OrfX2, OrfX3, and P47, the BoNT, BoNT-like, and *Paenibacillus-Bacillus-Brevibacillus* groups are well-supported and monophyletic. This reinforces that the BoNT-type OrfX gene clusters are closely related to the clusters of the bacilli of order Bacillales. Additionally, the OrfX1 clade contains only OrfX proteins derived from genomes of the Bacillales and Clostridiales, suggesting OrfX1 proteins may be an innovation specific to the gene clusters in these taxa.

The tree (Figure 2.15) provides strong support for clades of OrfX proteins found in specific taxa. Two groups of OrfX-related genes in *Erwinia* group are well-supported and appear to be most closely related to the clostridial P47 and OrfX2 clades. Two clades of OrfX-related sequences in *Pseudomonas* are also well-supported, with one clade appearing most similar to clostridial OrfX3 and the second forming a relatively independent lineage stemming off earlier than the OrfX2-OrfX3 clade. A large clade of *Streptomyces* sequences forms a group that diverges early in the OrfX3-related clade. The placement of these monophyletic groups within the tree implies patterns of lateral gene transfer. Further, *Pseudomonas*, *Streptomyces*, and *Erwinia* were more likely to have been the recipients of lateral transfer than to be the donors, as OrfX proteins are not common in these genera.

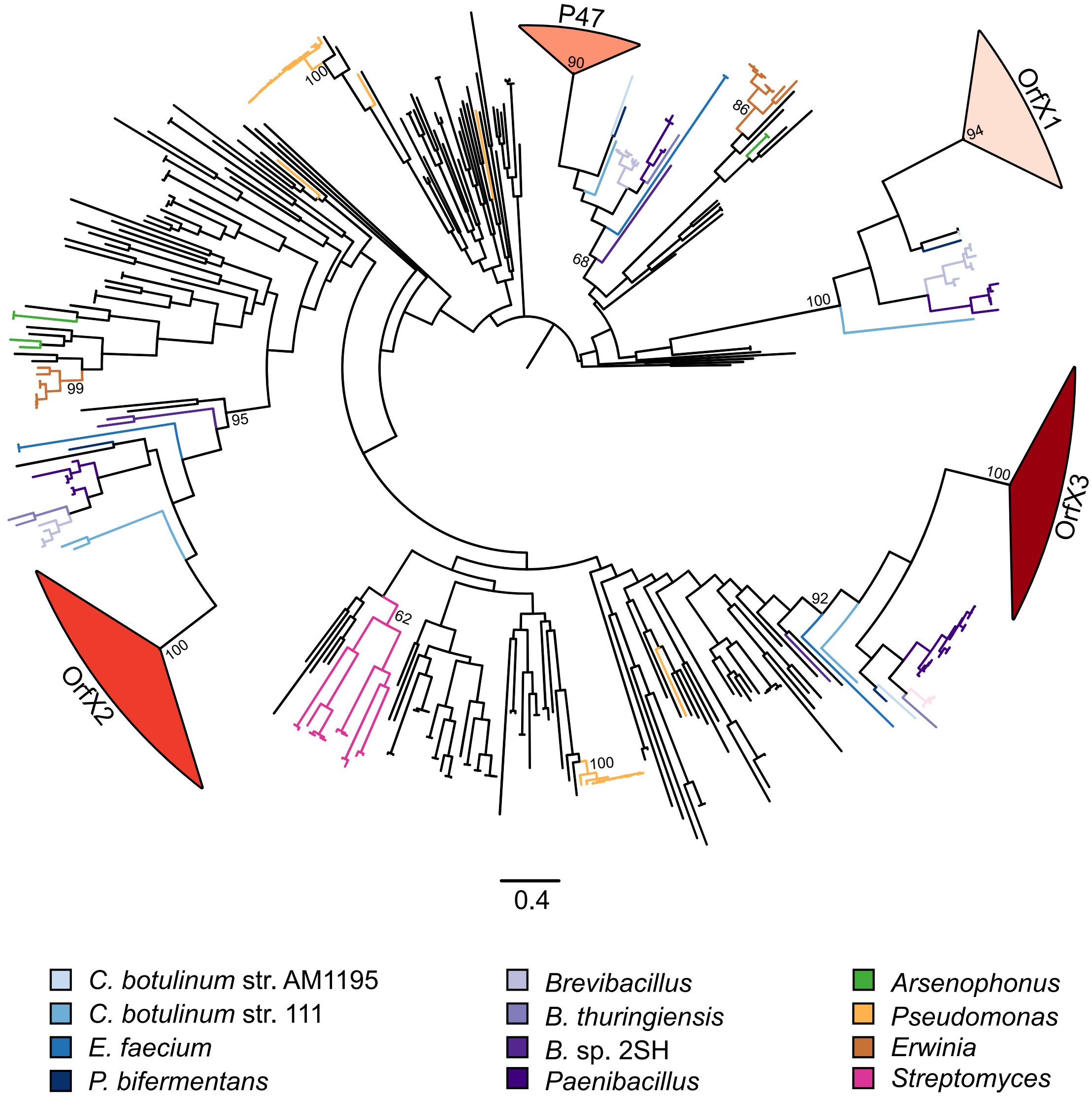


Figure 2.15: Maximum likelihood phylogeny of OrfX-related amino acid sequences. The clades of OrfX1, OrfX2, OrfX3, and P47 proteins from BoNT-type OrfX gene clusters are indicated. The closest relatives of BoNT-type OrfX proteins are generally the OrfX-related proteins of BoNT-like gene clusters (that is, OrfX-like proteins from *C. botulinum* str. 111, *E. faecium*, and *P. bifermentans*, encoding BoNT/X, BoNT/En, and PMP1, respectively). Other Firmicutes OrfX proteins are indicated in shades of purple, with other colours indicated for taxa containing numerous OrfX relatives outside of phylum Firmicutes.

As well, the tandem inheritance of multiple paralogs suggests that some of these events occurred after the duplication and divergence of OrfX family members. Based on their position in the tree, *Erwinia* most likely acquired a *p47-orfX2* cluster, while *Streptomyces* acquired only a copy of an *orfX3*-like gene. The pattern in *Pseudomonas* is more difficult to explain, as one clade appears to have diverged early from the OrfX2-OrfX3 clade, while the other falls within the OrfX3-like clade. One possible explanation for this pattern is the acquisition of an *orfX3*-like gene followed by *Pseudomonas*-specific duplication and divergence, although this is difficult to verify.

BoNT gene clusters as recipients and donors in lateral gene transfer

Based on patterns in OrfX gene cluster sequence content (Figure 2.13 and Figure 2.14) and phylogenetic placement (Figure 2.14), it seems clear that *orfX* genes undergo lateral gene transfer. To provide a clear example of this, two particular ORF gene clusters are highlighted: one from *C. botulinum* str. AM1195 genome and the other from *Bacillus* sp. 2SH (Figure 2.16). The strain of *C. botulinum* has been noted as containing a hemagglutinin-type BoNT B gene cluster [88], but also contains an additional OrfX gene cluster (Figure 2.16a). The OrfX proteins of this cluster are phylogenetically distinct from other BoNT-type OrfX proteins (Figure 2.15), and the locus lacks a neurotoxin-like gene. In its place, the OrfX gene cluster contains a hypothetical protein with significant similarities to insecticidal *Bacillus thuringiensis* δ-endotoxins. The acquisition of a *Bacillus* insecticidal toxin gene is certainly atypical for BoNT-type OrfX gene clusters, and the phylogenetic placement of its OrfX proteins suggests they are more closely related to the OrfX proteins found among the Bacillales.

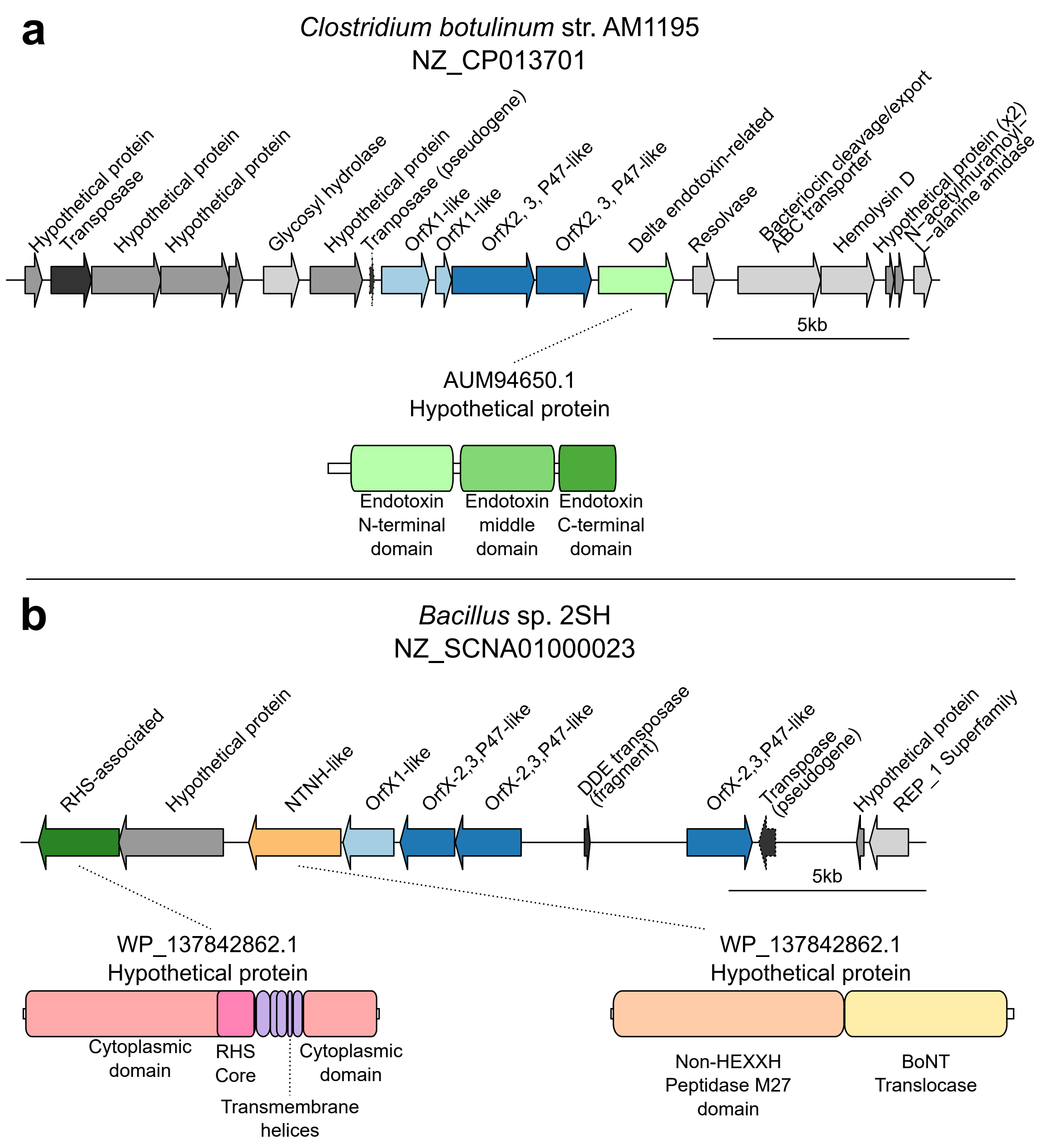

a
Clostridium botulinum str. AM1195
NZ_CP013701
Hypothetical protein
Transposase
Hypothetical protein
Hypothetical protein
Glycosyl hydrolase
Hypothetical protein
Tranposase (pseudogene)
OrfX1-like
OrfX1-like
OrfX2, 3, P47-like
OrfX2, 3, P47-like
Delta endotoxin-related
Resolvase
Bacteriocin cleavage/export
ABC transporter
Hemolysin D
Hypothetical protein (x2)
N-acetylmuramoyl-
L-alanine amidase
5kb
AUM94650.1
Hypothetical protein
Endotoxin
N-terminal
domain
Endotoxin
middle
domain
Endotoxin
C-terminal
domain
b
Bacillus sp. 2SH
NZ_SCNA01000023
RHS-associated
Hypothetical protein
NTNH-like
OrfX1-like
OrfX-2,3,P47-like
OrfX-2,3,P47-like
DDE transposase
(fragment)
OrfX-2,3,P47-like
Transpoase
(pseudogene)
Hypothetical protein
REP_1 Superfamily
5kb
WP_137842862.1
Hypothetical protein
Cytoplasmic
domain
RHS
Core
Cytoplasmic
domain
Transmembrane
helices
WP_137842862.1
Hypothetical protein
Non-HEXXH
Peptidase M27
domain
BoNT
Translocase

Figure 2.16 ***(previous page)*****:** OrfX clusters exhibit clear evidence of lateral gene transfer events. Here, the contents of two OrfX gene clusters are depicted: in (a), an OrfX gene cluster from *Clostridium botulinum* str. AM1195 lacking BoNT-related genes, and in (b) an OrfX gene cluster from *Bacillus* sp. 2SH containing an NTNH-like gene that is commonly associated with BoNT-type gene clusters. The OrfX gene cluster in *C. botulinum* str. AM1195 is unique from other clostridial OrfX gene clusters in that its OrfX proteins have low amino acid identity (∼30% identity), and they are phylogenetically distinct (see Figure 2.15). Further, the cluster contains a hypothetical protein with similarity to *B. thuringiensis* δ-endotoxin unlike other OrfX gene clusters associated with BoNTs. In *Bacillus* str. 2SH, the OrfX gene cluster is unique in containing an NTNH-like gene, which is only found in BoNT-type gene clusters. Together, these results imply that OrfX gene clusters can be laterally transferred with or without associated BoNT genes. The protein domain diagrams are annotated according to InterProScan annotations, and other gene colours assigned according to their NCBI annotations.

The OrfX cluster found within the genome of an environmental *Bacillus* isolate (strain 2SH, [286]) is perhaps an example of the opposite phenomenon (Figure 2.16b). The *Bacillus* sp. 2SH OrfX gene cluster contains three OrfX2, OrfX3, and P47-related sequences and a single OrfX1-related sequence, as well as a protein with an RHS core domain, which may be a potential toxin. More remarkable however is the appearance of an NTNH-like sequence downstream of the OrfX1 locus. The protein (NCBI Protein accession number WP_137842862.1) contains matches to the BoNT translocase and peptidase domains, but lacks the catalytic HExxH motif necessary for zinc peptidase activity. The same property is found in the non-toxic non-hemagglutinin genes typical of BoNT gene clusters. As far as we know, this is the first identification of an NTNH-related sequence in a strain of *Bacillus*. Assuming this OrfX gene cluster is truly a part of the *Bacillus* sp. 2SH genome (and not contaminating DNA from another organism), there are two simple and plausible explanations for this: that other NTNH-containing OrfX gene clusters exist in *Bacillus* species which have yet to be detected, or else that an NTNH-positive OrfX gene cluster was laterally transferred into this strain of *Bacillus*. In either case, the presence of an NTNH gene in *Bacillus* is noteworthy because the *ntnh* gene is only known to be located adjacent to *bont* genes, raising the possibility of *Bacillus* species containing neurotoxin-related sequences.

2.4.4 Discussion and Conclusions

To date, BoNT proteins have been discovered within two types of gene clusters containing either hemagglutinin (*ha*) or *orfX* genes. Little is known about the function of the OrfX proteins, as the neurotoxin protein alone is sufficient for toxicity [287]. As seen among gene clusters encoding BoNT serotypes with many traditional BoNTs, BoNT-related proteins from a wide variety of organisms are encoded in OrfX gene clusters. The broad conservation of OrfX clusters in many species suggests that their function may have an important contribution to the fitness of their associated neurotoxin genes.

Our searches against currently available genomes suggest that OrfX clusters are taxonomically widespread, much more than neurotoxin genes (Figure 2.13). The OrfX gene clusters most similar to BoNT-type clusters are found in various species of the order Bacillales, particularly within species of *Paenibacillus* and *Brevibacillus*. Outside of the phylum Firmicutes, OrfX clusters can be found in species of *Pseudomonas*, *Streptomyces*, *Erwinia*, and *Arsenophonus*. Remarkably, nearly all of the organisms with OrfX gene clusters are associated with pathogenicity in either plants or insects. Many specific OrfX clusters include toxin-related sequences (Figure 2.14), including homologs of anthrax toxins in *Paenibacillus* and *Brevibacillus*, RHS-related sequences, and homologs of the δ-endotoxins and Cry toxins of *Bacillus thuringiensis*. Together, these results demonstrate a strong association between OrfX gene clusters and insect pathogenicity.

Even when considering only *Clostridium*, it is clear that lateral gene transfer events have helped to shape BoNT and OrfX gene clusters [152, 235, 146, 288]. This observation is consistent with phylogenetic analysis of OrfX proteins across many taxa (Figure 2.15). In *Clostridium*, the OrfX1, OrfX2, OrfX3, and P47 lineages form well-supported clades which each have homologous representatives among members of the order Bacillales (that is, *Paenibacillus*, *Brevibacillus*, and *Bacillus* spp.). The phylogeny also implies several independent lateral gene transfer events consisting of different *orfx* gene combinations into species of *Pseudomonas*, *Erwinia*, and *Streptomyces*. As well, the examples of OrfX gene clusters in *C. botulinum* str. AM1195 and *Bacillus* sp. 2SH provide evidence of lateral gene transfers of potentially non-BoNT-type OrfX clusters into *Clostridium* as well as the transfer of a potentially BoNT-type OrfX cluster into *Bacillus*. The production of Cry toxins has been described in other strains of *C. botulinum* [289], but their association with an OrfX gene cluster is surprising. The source of the gene cluster in *Bacillus* sp. 2SH is unknown, but with currently available evidence seems most likely to be a species of *Clostridium*, considering the limited taxonomic distribution of NTNH sequences. The genome of *Bacillus* sp. 2SH was assembled from a thermal spring metagenome [286] with temperatures varying from 20-37 °C, meaning interaction with toxigenic clostridia is at least possible.

Numerous avenues support a link between OrfX proteins and insect pathogenicity. OrfX gene clusters are found in disparate, distantly related taxa, and one of the few common threads among them is a direct association with insect pathogenicity or a more broad association with plants (and, therefore, herbivorous insects). As such, we propose that the OrfX gene cluster is a modular toxin gene cluster related to insect pathogenicity. By extension, the presence of OrfX proteins encoded in BoNT-type OrfX gene clusters reflects either an ancestral or an ongoing adaptation for targeting insects. The observation that OrfX proteins enhance mosquitocidal toxicity corroborates the second claim, at least for the BoNT-related protein PMP1 [260]. The large differences between OrfX proteins associated with BoNT clusters compared to those in BoNT-related clusters like the one encoding PMP1 (∼70% amino acid identity) suggests that the two groups may perform different functions; this is perhaps analogous to the toxin complexes of *Yersinia* species, where the complexes of different *Yersinia* species appear to have specific adaptations for different host types [290]. Although comparative genomics is insufficient to explain the function of OrfX proteins by itself, the analyses presented here strongly suggest that OrfX gene clusters contribute to insect pathogenicity, which may also apply to OrfX gene clusters containing *bont* genes.

Chapter 3

Discovery of diphtheria toxin homologs outside of *Corynebacterium*

Material in this chapter has been published or is currently in preparation for publication. The published manuscript is available from the following source:

1. Mansfield, M. J., Sugiman-Marangos, S. N., Melnyk, R. A., & Doxey, A. C. (2018). Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus. FEBS letters, 592(16), 2693-2705. [5].

   https://doi.org/10.1002/1873-3468.13208

3.1 Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus

3.1.1 Introduction

The history of bacterial toxins began in the 1880s with the isolation of a "diphtheretic poison" [now termed diphtheria toxin (DT)] by Émile Roux and Alexandre Yersin in filtrates of the organism *Corynebacterium diphtheriae* [29]. Roux and Yersin demonstrated that filtrates of this "diphtheria bacillus" exhibited toxicity in guinea pigs, similar to that of urine samples taken from children with diphtheria. From this landmark study, the field of bacterial toxinology was born as it was demonstrated for the first time that a bacterial poison (or "toxin" from Greek: τοξικός, poison) was the causative agent of a human infectious disease.

Although *C. diphtheriae* was first discovered by Klebs and Löffler in the early 1880s, the history of diphtheria disease dates back much further. Early accounts of diphtheria can be found in 2500-year-old writings by Hippocrates, and diphtheria-like epidemics were reported in writings from the 1500s [291]. By the late 1800s, diphtheria was one of the most devastating infectious diseases, and was one of the leading causes of childhood death in the prevaccine era [292]. After the introduction of the diphtheria vaccine in the early 1920s, however, diphtheria-related mortalities began to drop dramatically, and since their introduction vaccines are estimated to have prevented over 40 million cases of diphtheria in the US alone [293]. The name "diphtheria" originates from the Greek word *diphthera* (leather), relating to the leathery pseudomembranous films formed in throats of diphtheria patients. The disease is characterized by an infection of the upper respiratory tract by toxigenic *C. diphtheriae*, leading to ulceration of the mucosa and formation of inflammatory pseudomembranous lesions. The diphtheria exotoxin is secreted by *C. diphtheriae* and its adsorption into the bloodstream and transmission into other organs can induce cytoxicity and ultimately lead to death. The mode of action of DT has been characterized in depth through decades of work [7, 294], and involves binding and entry of DT into host cells and subsequent inhibition of protein synthesis through the inactivation of eukaryotic elongation factor 2 (eEF-2). DT is a 60 kDa polypeptide chain, about 535 amino acids in length, possessing a characteristic AB toxin architecture consisting of two subunits linked by a disulfide bond. The B subunit is responsible for cell receptor binding and membrane translocation, and the A subunit is responsible for intracellular catalytic activity.

Toxicity is achieved through the following steps. Following secretion, the DT C-terminal receptor-binding (R) domain (immunoglobulin-like beta-sandwich fold) in the B subunit binds to human heparin-binding epidermal growth factor (HB-EGF) on host cell membranes. DT is then taken up by an endosome, and the middle translocation (T) domain (colicin-like fold) of the B subunit undergoes pH-driven conformational changes in the endosomal membrane, which facilitates translocation of the catalytic domain (C domain, or A subunit) cargo into the host cytosol [7, 294]. For the C-terminal domain to be released, a disulfide bond linking the A and B fragment must be reduced as well as proteolytically nicked by furin and other proteases [78]. At this point, the C domain, which adopts a beta-alpha structure and possesses ADP-ribosylatransferase activity, transfers an ADP-ribose moiety from NAD to a unique post-translationally modified histidine residue termed diphthamide on eEF-2. ADP-ribosylation of eEF-2 results in blockage of protein synthesis and cell death.

In 2003, the first genome sequence of *C. diphtheriae* (strain NCTC13129) was published, following an ongoing outbreak of re-emerging diphtheria in Europe, revealing genomic insights into its pathogenicity [295]. DT is encoded on a mobile, temperate bacteriophage that integrates into the *Corynebacterium* genome at tRNA$^{\mathrm{Arg}}$ loci [296, 297]. The *tox* gene encoding DT is localized at the 30 side of the integrated corynephage, is flanked by an *att* integration site, and is found in a region with relatively low C+G content (43%), which has been suggested to reflect a potentially recent insertion event into the phage [298]. An important regulator of *tox* gene expression is an iron-dependent transcriptional repressor (DtxR), which not only represses *tox* expression but also acts more globally in the regulation of other iron-sensitive genes [299]. Iron limitation is a common strategy for bacterial growth suppression by hosts, and in turn is exploited by numerous pathogens (including *C. diphtheriae*) to upregulate their virulence genes.

In addition to *C. diphtheriae*, two related species of *Corynebacterium* (*C. pseudotuberculosis*, and *C. ulcerans*) have been identified with *tox*-carrying corynephages. These corynephages possess unique differences from those in *C. diphtheriae* [297], and encode DT variants that exhibit high (95%) sequence identity to classic *C. diphtheriae* DT. Both *C. ulcerans* and to a lesser extent *C. pseudotuberculosis* have been shown to be capable of producing diphtheria-like symptoms in humans, but are primarily associated with zoonotic infections [300, 301].

Despite considerable knowledge on DT structure, function, and mechanism, very little is known about the relationship of DT sequences to other proteins. Although the mechanism of the DT catalytic domain is shared with other ADP-ribosylating toxins such as *Pseudomonas* ExoA and cholix toxin [302, 51], no homologs of DT have been identified to date that possess multiple DT domains indicative of common evolutionary ancestry. Therefore, the evolutionary origin and molecular ancestry of DT is unclear. Based on similar work exploring botulinum neurotoxin evolution and diversity [99, 2, 1, 4], we hypothesized that there may exist additional homologs of DT beyond the *Corynebacterium* genus, which if identified could provide insights into its evolutionary relationships and history. In this study, we mined available genomes for DT homologs including 134,749 prokaryotic species in the current Genbank release. We report the bioinformatic discovery of the first homologs of DT outside of *Corynebacterium*. These putative DT-like toxins possess low identity but conserve DT's three domain architecture and key active sites, and are the closest phylogenetic relatives to DT in the current database. DT homologs are most unique in their R binding domains, suggesting that the specificity of DT toward humans may have evolved in part through fine-tuning and accelerated evolution of receptor binding. Based on the taxonomic sources of these homologs and patterns of evolutionary diversification, we propose that DT evolved from an ancient family of DT-like toxins that is concentrated within the phylum Actinobacteria.

3.1.2 Methods

Homolog detection and sequence analysis

A data set of DT and DT-like protein sequences was obtained by first performing BLASTp search against the NCBI nr database (Mar. 29, 2018). Identical sequences and synthetic constructs were removed from the data set. Domain architectures were predicted using the CDD and verified by comparison to other databases using InterProScan. Only proteins containing more than one predicted DT domain were retained, resulting in a final data set of 27 DT and DT-like proteins.

A notable sequence outlier was the *Austwickia chelonae* DT-like protein (NCBI accession number WP_040322835.1), which contains a portion of a diphtheria T domain and a full R domain. Analysis of this genomic region revealed a frameshift mutation, that when corrected resulted in a full-length DT-like gene also containing the C domain (sequence translated from NCBI accession number NZ_BAGZ01000024, region 41461-42285, complement). The concatenated full-length sequence was used for phylogenetic analysis, but is potentially a pseudogene that has been interrupted by a frameshift mutation. It is unclear whether this represents a true frameshift mutation or a sequencing error.

Phylogenetic analysis

A multiple sequence alignment of the data set was created using CLUSTAL-OMEGA (version 1.2.1, [130]) using default parameters. Phylogenetic trees were inferred using maximum likelihood (RAXML version 8.2.4; 1000 rapid bootstraps, automatic model selection with gamma-distributed rates across sites, and a thorough ML search [135]), Bayesian inference (MRBAYES version 3.2.6 [185]; gamma-distributed rates across sites, WAG model of evolution, 1,000,000 generations with Metropolis-coupled Markov chain Monte Carlo sampling) and neighbour-joining (BioNJ [303] as implemented in SEAVIEW version 4.5.4 [304]; default parameters) methods. All tree-building methods were congruent in high-level topology. The statistical overrepresentation within the Actinobacteria was quantified by computing the probability that randomly selected taxa from the GTDB bacterial tree of life [305] could result in six or more actinobacterial species by chance (permutation test, 100,000 iterations).

Sequence and structural modelling

Transmembrane regions were predicted using TMHMM [306]. Structural models for all DT-like proteins were generated using I-TASSER [187]. Each putative domain was modeled separately, with domain regions defined based on alignment to the DT reference structure (PDB accession number 1MDT, [307]). Sequence regions and structure prediction results are available in Table B.1. Structural models for each DT-like domain were structurally aligned to the corresponding domain in DT using PYMOL (`http://pymol.org`). PyMol's align function was used to perform a sequence-guided structural alignment, and the fit of the superimposed structures was evaluated by root mean squared deviation. Solvent-accessible surface area for 1MDT was calculated using FreeSASA [308].

Analysis of evolutionary conservation

Per-residue sequence conservation was computed and mapped to the crystal structure of DT using ConSurf [309] with the Bayesian method for calculating conservation scores. Thirty-one sites were invariant across the alignment and therefore received the lowest normalized scores (most conserved) according to the ConSurf algorithm. Sequence conservation was also visualized using WebLogo [310] with default parameters.

Genomic context analysis

Genome sequences for three corynephage genomes and the genomes of all organisms containing predicted DT-like genes were downloaded from the NCBI Genome database. Regions 15 kb up- and downstream of DT-like genes were selected for the plot used in Figure 3.9, depicted in R using genoPlotR version 0.1 [163]. The genomes containing DT-like genes were searched for the presence of corynephage-like viral DNA using BLASTN version 2.2.31 [112] as well as the phage search tool, PHAST [311].

3.1.3 Results

Genome mining identifies diphtheria toxin homologs outside of the *Corynebacterium* genus

We searched the GenBank database (Mar 29, 2018, 177,666 total genomes) with corynephage beta DT as a query (PDB identifier 1MDT) using BLASTp. The closest full-length matches to DT are variants present in *C. diphtheria*, *C. ulcerans*, and *C. pseudotuberculosis*, with sequence identities ranging from 95 to 100% (Table 3.1). Beyond this group, we identified several homologs with lower identities (23-34%), weaker but significant E-values (1×10^{-36} to 6×10^{-12}), and detectable alignments spanning partial coverage (51-85%) of the full-length DT query sequence (Table 3.1). These detectable DT homologs include predicted protein sequences from the genomes of *Austwickia chelonae*, *Streptomyces albireticuli*, *Streptomyces* sp. TLI_053, *Streptomyces roseoverticillaticus*, *Streptomyces* sp. MBT76, *Streptosporangium nondiastaticum*, and *Seinonella peptonophila*. Reciprocal searches performed with these sequences as queries identified DT with E-values $<1\times10^{-10}$ in all cases, confirming their direct homology to DT.

Table 3.1: Detection of diphtheria toxin-related sequences in the NCBI GenBank database. The percentage of identical and similar residues as well as query coverage (Iden., Sim., and Cov., respectively) are calculated relative to PDB identifier 1MDT.

Accession	**Iden.**	**Sim.**	**Cov.**	***E*-value**	**Phylum**	**Family**	**Genus/Species**
1MDT	100	100	100	0	Actinobacteria	Corynebacteriaceae	*Corynebacterium diphtheriae*
AAN28948.1	95	97	100	0	Actinobacteria	Corynebacteriaceae	*Corynebacterium ulcerans*
WP_014654963.1	95	97	99	0	Actinobacteria	Corynebacteriaceae	*Corynebacterium pseudotuberculosis*
WP_040322835.1	34	51	51	1×10^{-36}	Actinobacteria	Dermatophilaceae	*Austwickia chelonae*
WP_073156187.1	27	43	78	1×10^{-27}	Firmicutes	Thermoactinomycetaceae	*Seinonella peptonophila*
WP_078659863.1	28	47	68	1×10^{-23}	Actinobacteria	Streptomycetaceae	*Streptomyces roseoverticillatus*
PSJ28985.1	27	46	68	1×10^{-22}	Actinobacteria	Streptosporangiaceae	*Streptosporangium nondiastaticum*
WP_079110321.1	29	46	68	2×10^{-22}	Actinobacteria	Streptomycetaceae	*Streptomyces* sp. MBT76
SDT83331.1, WP_093864399.1	25	43	65	2×10^{-21}	Actinobacteria	Streptomycetaceae	*Streptomyces* sp. TLI_053
WP_095582082.1	23	39	85	6×10^{-12}	Actinobacteria	Streptomycetaceae	*Streptomyces albireticuli*

Intriguingly, all of these genomes (with the exception of *Seinonella*) are members of the phylum Actinobacteria, which also contains *Corynebacterium* (Table 3.1). This taxonomic overrepresentation within the Actinobacteria is highly nonrandom (p <0.0001, permutation test, Figure 3.1), further suggesting the common ancestry of these sequences. Although these species all occur within the Actinobacteria, it is important to note that they are phylogenetically scattered among many other lineages that lack DT-like genes. This suggests that DT-like genes are not monophyletically distributed, and that lateral gene transfer and/or gene loss has played a role in the evolution of these genes. Interestingly, the list of species containing DT-like genes includes some known pathogens; for example, *S. albireticuli* is a pathogen of nematodes [312] and has algicidal activity [313], and *A. chelonae* is a pathogen that causes skin infections in reptiles and other wild animals ([314, 315]; see Discussion).

Phylogenetic analysis suggests an ancestral radiation of DT-like lineages

Next, we aligned representative *Corynebacterium* DT sequences with the detected DT-like homologs, and performed phylogenetic analysis using maximum likelihood, Bayesian, and neighbour-joining reconstruction (see Materials and Methods). All three methods produced trees that were identical in clade topology with 98-100% support of all major clades (Figures 3.2 and 3.3). The resulting midpoint-rooted phylogenetic tree reveals an ancestral diversification of DT-like sequences into six main lineages. All *Corynebacterium* DT proteins form a monophyletic lineage with low diversity and 100% clade support. The *Corynebacterium* clade is subdivided into a subclade of DT sequences predominantly from *C. diphtheriae* and a sister group of DT sequences from *C. ulcerans*. The *Corynebacterium* DT group forms a larger group with DT-like sequences from *A. chelonae* with 100% clade support. We have denoted this group $\alpha 1$. Although there is evidence of a frameshift mutation in this gene (see Materials and Methods), which raises questions about its pseudogenicity, these are the most closely related sequences to DT phylogenetically, and also possess the highest sequence identity to DT (34%, Table 3.1). Neighbouring the *Austwickia* + *Corynebacterium* group ($\alpha 1$) is an earlier diverging lineage ($\alpha 2$), which includes various *Streptomyces* spp. and *Streptosporangium nondiastaticum*, which also has 100% bootstrap support. Finally, a group of two divergent DT-like sequences from *Seinonella peptonophila* and *S. albireticuli* (group β) clustered separately from the remaining DT homologs with 100% bootstrap support. The observed incongruence between the gene and species phylogeny (Figures 3.2 and 3.1) is again indicative of lateral transfer of DT-like genes (e.g., a possible transfer between *Streptomyces* and *Streptosporangium* in lineage $\alpha 2$).

DT homologs conserve protein domain architecture but have variable R domains

Next, we performed sequence and structural modelling to assess the conservation of protein domain architecture in DT homologs. At the level of protein domains, all DT homologs possess detectable diphtheria-like ADP ribosyltransferase (catalytic, C) and translocation (T) domains (Figure 3.4), and this annotation was supported by four separate domain databases/prediction tools (Pfam [117], Conserved Domain Database (CDD) [184], Phyre

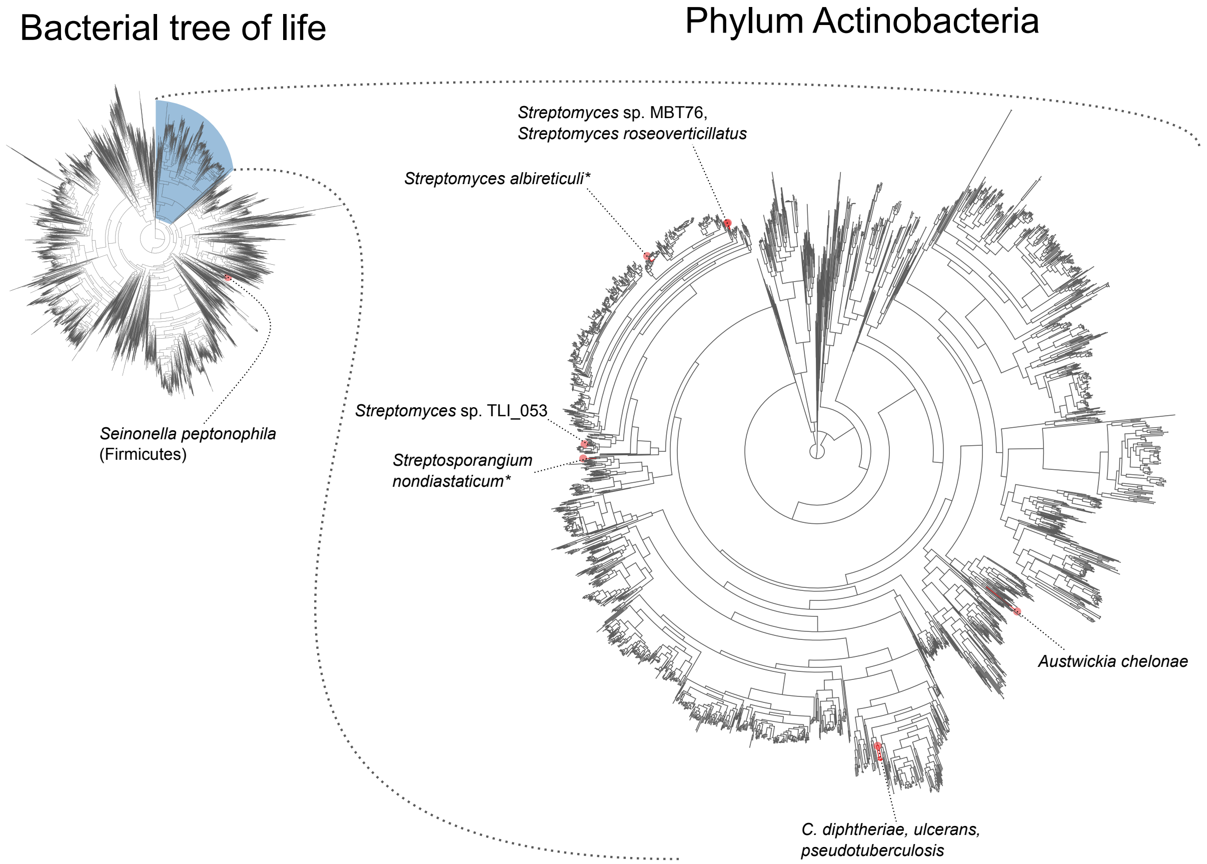


Figure 3.1: Phylogenomic distribution of taxa encoding diphtheria toxin (DT) and diphtheria toxin-like proteins. Gene presence/absence was mapped onto the bacterial tree of life using the Genome Taxonomy Database (GTDB) tree (http://gtdb.ecogenomic.org/) using AnnoTree [1]. The full bacterial tree is shown on the left and highlights the phylum Actinobacteria, which contains all DT-like sequences identified in this study except the example in *Seinonella peptonophila*. The right panel shows the distribution of DT-like sequences within the phylum Actinobacteria, demonstrating the scattered distribution of DT-like sequences across bacteria.

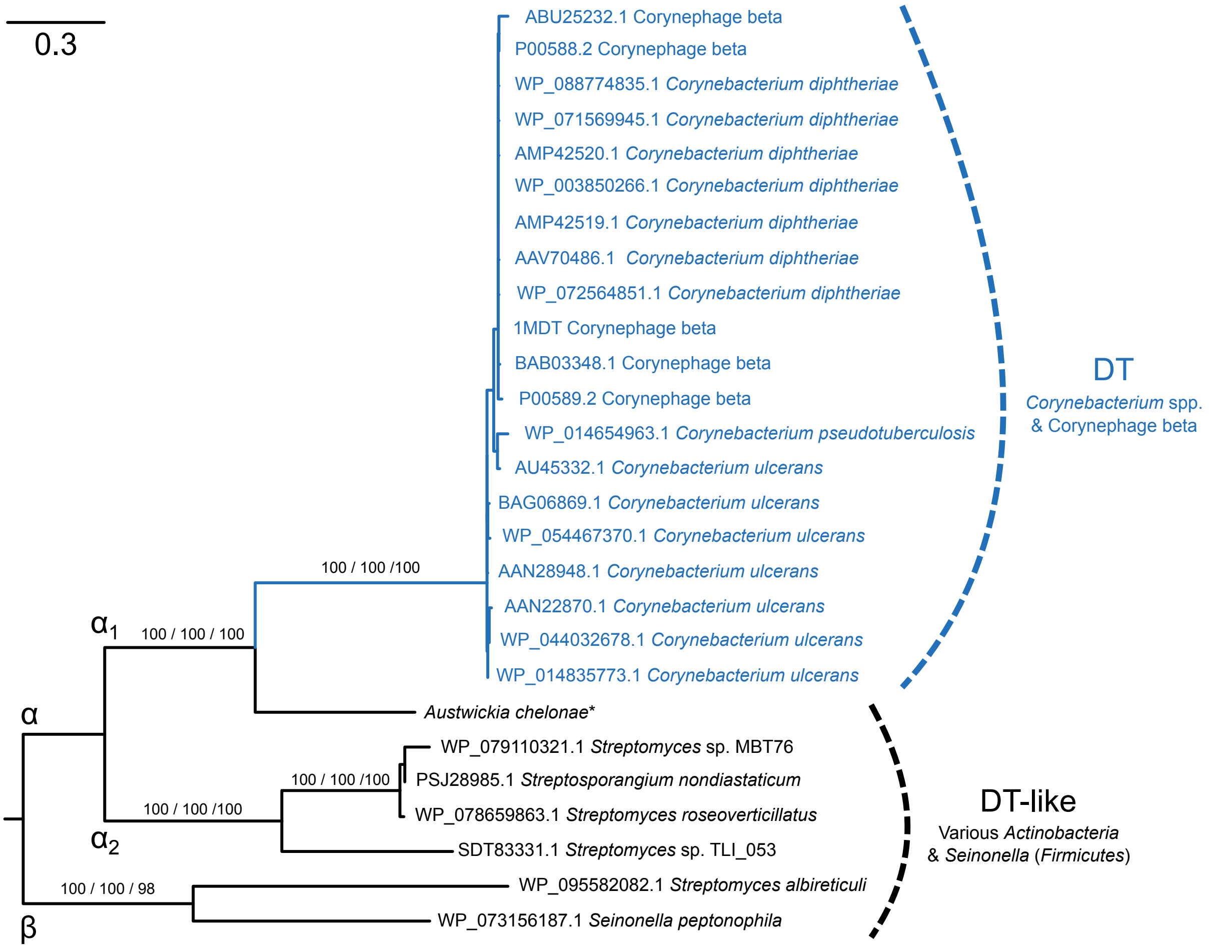

0.3
ABU25232.1 Corynephage beta
P00588.2 Corynephage beta
WP_088774835.1 *Corynebacterium diphtheriae*
WP_071569945.1 *Corynebacterium diphtheriae*
AMP42520.1 *Corynebacterium diphtheriae*
WP_003850266.1 *Corynebacterium diphtheriae*
AMP42519.1 *Corynebacterium diphtheriae*
AAV70486.1 *Corynebacterium diphtheriae*
WP_072564851.1 *Corynebacterium diphtheriae*
1MDT Corynephage beta
BAB03348.1 Corynephage beta
P00589.2 Corynephage beta
WP_014654963.1 *Corynebacterium pseudotuberculosis*
AU45332.1 *Corynebacterium ulcerans*
BAG06869.1 *Corynebacterium ulcerans*
WP_054467370.1 *Corynebacterium ulcerans*
AAN28948.1 *Corynebacterium ulcerans*
AAN22870.1 *Corynebacterium ulcerans*
WP_044032678.1 *Corynebacterium ulcerans*
WP_014835773.1 *Corynebacterium ulcerans*
DT
Corynebacterium spp.
& Corynephage beta
100 / 100 /100
α1
100 / 100 / 100
α
*Austwickia chelonae**
WP_079110321.1 *Streptomyces* sp. MBT76
100 / 100 /100
PSJ28985.1 *Streptosporangium nondiastaticum*
100 / 100 /100
WP_078659863.1 *Streptomyces roseoverticillatus*
α2
SDT83331.1 *Streptomyces* sp. TLI_053
100 / 100 / 98
WP_095582082.1 *Streptomyces albireticuli*
β
WP_073156187.1 *Seinonella peptonophila*
DT-like
Various *Actinobacteria*
& *Seinonella* (*Firmicutes*)

Figure 3.2: Maximum likelihood phylogenetic tree of amino acid sequences from DT and identified DT-like homologs. Clade support values for all three methods are shown above each node (maximum likelihood/Bayesian/neighbour-joining). Bayesian and neighbour-joining methods were identical in high-order topology (Figure 2.10). *This protein sequence was encoded by two separate open reading frames in the *Austwickia chelonae* genome. See Methods for further details.

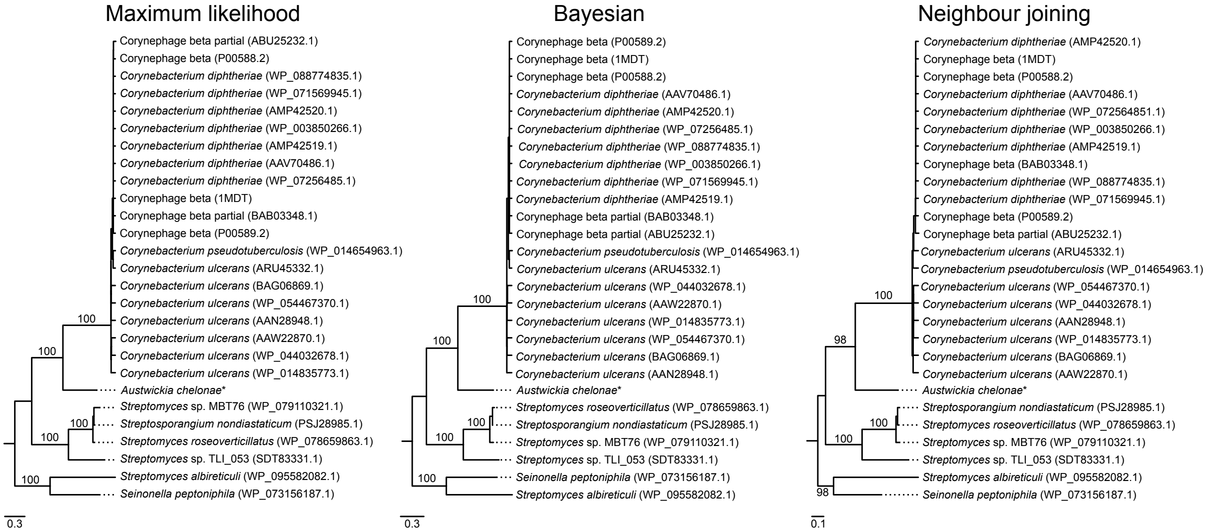

Maximum likelihood
Corynephage beta partial (ABU25232.1)
Corynephage beta (P00588.2)
Corynebacterium diphtheriae (WP_088774835.1)
Corynebacterium diphtheriae (WP_071569945.1)
Corynebacterium diphtheriae (AMP42520.1)
Corynebacterium diphtheriae (WP_003850266.1)
Corynebacterium diphtheriae (AMP42519.1)
Corynebacterium diphtheriae (AAV70486.1)
Corynebacterium diphtheriae (WP_07256485.1)
Corynephage beta (1MDT)
Corynephage beta partial (BAB03348.1)
Corynephage beta (P00589.2)
Corynebacterium pseudotuberculosis (WP_014654963.1)
Corynebacterium ulcerans (ARU45332.1)
Corynebacterium ulcerans (BAG06869.1)
Corynebacterium ulcerans (WP_054467370.1)
Corynebacterium ulcerans (AAN28948.1)
Corynebacterium ulcerans (AAW22870.1)
Corynebacterium ulcerans (WP_044032678.1)
Corynebacterium ulcerans (WP_014835773.1)
Austwickia chelonae*
Streptomyces sp. MBT76 (WP_079110321.1)
Streptosporangium nondiastaticum (PSJ28985.1)
Streptomyces roseoverticillatus (WP_078659863.1)
Streptomyces sp. TLI_053 (SDT83331.1)
Streptomyces albireticuli (WP_095582082.1)
Seinonella peptoniphila (WP_073156187.1)
100
100
100
100
100
0.3
Bayesian
Corynephage beta (P00589.2)
Corynephage beta (1MDT)
Corynephage beta (P00588.2)
Corynebacterium diphtheriae (AAV70486.1)
Corynebacterium diphtheriae (AMP42520.1)
Corynebacterium diphtheriae (WP_07256485.1)
Corynebacterium diphtheriae (WP_088774835.1)
Corynebacterium diphtheriae (WP_003850266.1)
Corynebacterium diphtheriae (WP_071569945.1)
Corynebacterium diphtheriae (AMP42519.1)
Corynephage beta partial (BAB03348.1)
Corynephage beta partial (ABU25232.1)
Corynebacterium pseudotuberculosis (WP_014654963.1)
Corynebacterium ulcerans (ARU45332.1)
Corynebacterium ulcerans (WP_044032678.1)
Corynebacterium ulcerans (AAW22870.1)
Corynebacterium ulcerans (WP_014835773.1)
Corynebacterium ulcerans (WP_054467370.1)
Corynebacterium ulcerans (BAG06869.1)
Corynebacterium ulcerans (AAN28948.1)
Austwickia chelonae*
Streptomyces roseoverticillatus (WP_078659863.1)
Streptosporangium nondiastaticum (PSJ28985.1)
Streptomyces sp. MBT76 (WP_079110321.1)
Streptomyces sp. TLI_053 (SDT83331.1)
Seinonella peptoniphila (WP_073156187.1)
Streptomyces albireticuli (WP_095582082.1)
100
100
100
100
100
0.3
Neighbour joining
Corynebacterium diphtheriae (AMP42520.1)
Corynephage beta (1MDT)
Corynephage beta (P00588.2)
Corynebacterium diphtheriae (AAV70486.1)
Corynebacterium diphtheriae (WP_072564851.1)
Corynebacterium diphtheriae (WP_003850266.1)
Corynebacterium diphtheriae (AMP42519.1)
Corynephage beta (BAB03348.1)
Corynebacterium diphtheriae (WP_088774835.1)
Corynebacterium diphtheriae (WP_071569945.1)
Corynephage beta (P00589.2)
Corynephage beta partial (ABU25232.1)
Corynebacterium ulcerans (ARU45332.1)
Corynebacterium pseudotuberculosis (WP_014654963.1)
Corynebacterium ulcerans (WP_054467370.1)
Corynebacterium ulcerans (WP_044032678.1)
Corynebacterium ulcerans (AAN28948.1)
Corynebacterium ulcerans (WP_014835773.1)
Corynebacterium ulcerans (BAG06869.1)
Corynebacterium ulcerans (AAW22870.1)
Austwickia chelonae*
Streptosporangium nondiastaticum (PSJ28985.1)
Streptomyces roseoverticillatus (WP_078659863.1)
Streptomyces sp. MBT76 (WP_079110321.1)
Streptomyces sp. TLI_053 (SDT83331.1)
Streptomyces albireticuli (WP_095582082.1)
Seinonella peptoniphila (WP_073156187.1)
100
98
100
100
98
0.1

Figure 3.3: Phylogenetic tree of diphtheria toxin and diphtheria toxin-like proteins inferred using three different methods (maximum likelihood, Bayesian, and neighbour joining). For all three methods, branch order remained largely consistent.

[116], and HHpred [114]). Interestingly, although C-terminal regions similar in length to the DT receptor binding (R) domain are present in all of the homologs, these regions are highly divergent in sequence and were not recognizable using existing DT-R domain models with the exception of the *A. chelonae* C-terminal fragment (Figure 3.4). Notably, the C-terminal domain from *S. peptonophila* had a statistically significant alignment score with DT-R despite low identity (26.7%, E-value = 0.031 using SSEARCH36 [316]).

We then predicted structures for the three regions of each DT-like toxin using I-TASSER [255]. Consistent with sequence-based analysis, structure predictions for the C and T domains exhibited high similarity to DT, whereas the predicted structures of the R domains exhibited limited similarity (Figure 3.4). However, importantly, the DT-R domain was detected as the top scoring structural template for *A. chelonae*, *S. albireticuli*, and *S. peptonophila* (Table B.1), respectively, suggesting structural similarity. Consistent with the identification of T domains in DT-like homologs, predicted transmembrane helix propensities using TMHMM [306] resulted in a characteristic, conserved pattern over this region, which supports previous models of T domain-mediated translocation (Figure 3.4). Overall, these results suggest that the identified DT homologs possess DT-like C and T domains, but possess C-terminal binding domains that are divergent in sequence but likely related in structure.

DT homologs conserve key functional sites

We then examined the multiple alignment (Figure 3.5) to assess whether key residues in DT are conserved in DT homologs (Table B.2). Functionally important residues in the C domain include His-21 [317], Tyr-65 [318], and the critical Glu-148 [319], which participate in NAD+ binding and catalysis (with residue numbers relative to diphtheria toxin PDB identifier 1MDT). Glu-148 is completely conserved in all DT homologs, Tyr-65 in 7/8 homologs, and His-21 in 6/8 homologs (missing only in group β) (Figures 3.5 and 3.6). Additional residues that participate in NAD-binding are Thr-23, Tyr-27, and Tyr-54. Tyr-54 is completely conserved but Thr-23 and Tyr-27 are substituted in all homologs. Following the C domain in DT is the disulfide bond formed by Cys-186 and Cys-201. A pair of cysteine residues are conserved in all DT-like homologs, and located either in the same alignment position or in the adjacent (± 1) position (Figures 3.5 and 3.6). Within the disulfide linker region, a furin cleavage site (RxxR) is conserved in 6/8 DT homologs, and again is present in group α but missing in group β.

Within the T domain, several key residues have been identified as important for membrane binding and translocation activity, including the critical Pro-345 [321, 322], Glu-349 [323], and Asp-352 [323, 324]. Pro-345, along with a nearby Gly-348 is completely conserved in all homologs, while E-349 and D-352 are conserved among group α (Figure 3.6). The presence of these key residues, the similarity of transmembrane propensity (Figure 3.4), and T domain homology, strongly suggests conservation of translocation function.

Finally, a co-crystal structure of DT complexed with heparin-binding epidermal growth factor (HB-EGF) has revealed the residues that form the DT-receptor interface [325] (PDB identifier 1XDT). DT residues within 3 Å of HB-EGF are S381, H384, H391, R462, D465, S506, D507, and K526 (Figure 3.7). Although in part due to a poor alignment in the C-terminal region, none of these residues are conserved, suggesting the possibility of altered binding specificity.

The presence/absence of these key residues and motifs mapped onto the phylogenetic tree reveals an interesting evolutionary pattern (Figure 3.6). That is, almost all features important for DT function are conserved within lineage α, which not only supports the common ancestry of this group, but suggests the emergence of DT-specific functionality in the ancestor of this lineage. The absence of key catalytic sites, the furin cleavage site, and other motifs in group β suggests a different biochemical specificity/function for these sequences.

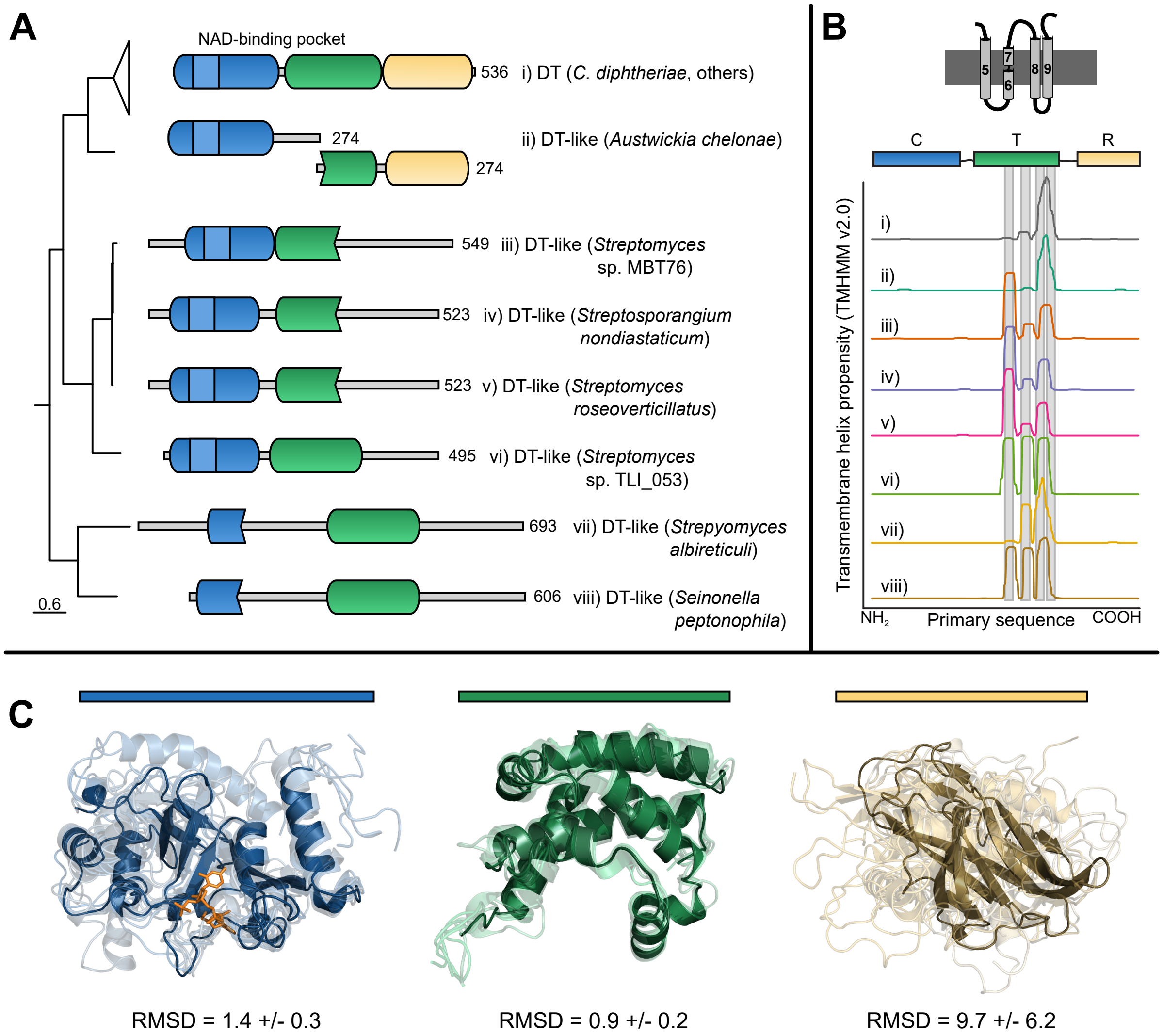

A
NAD-binding pocket
536 i) DT (C. diphtheriae, others)
274
ii) DT-like (Austwickia chelonae)
274
549 iii) DT-like (Streptomyces sp. MBT76)
523 iv) DT-like (Streptosporangium nondiastaticum)
523 v) DT-like (Streptomyces roseoverticillatus)
495 vi) DT-like (Streptomyces sp. TLI_053)
693 vii) DT-like (Strepyomyces albireticuli)
606 viii) DT-like (Seinonella peptonophila)
0.6
B
5
7
6
8
9
C
T
R
Transmembrane helix propensity (TMHMM v2.0)
i)
ii)
iii)
iv)
v)
vi)
vii)
viii)
NH₂
Primary sequence
COOH
C
RMSD = 1.4 +/- 0.3
RMSD = 0.9 +/- 0.2
RMSD = 9.7 +/- 6.2

Figure 3.4: Predicted domain architectures, sequence and structural modeling of diphtheria toxin (DT) homologs. (A) Domain architectures of DT homologs inferred using the NCBI CDD search tool with an E-value threshold of 0.01. The domains are colored as follows: catalytic (C) domain, blue; translocase (T) domain, green; and the receptor binding (R) domain, yellow. (B) Transmembrane helix propensities predicted via TMHMM scores. All T domains of DT-like proteins show a characteristic pattern over the membrane-inserting alpha helical region. (C) Structural models of C, T, and R domains predicted using I-TASSER. Structural alignment was performed using PyMol (http://pymol.org/), and the average RMSD of the models' fit to DT (PDB identifier 1MDT) is indicated.

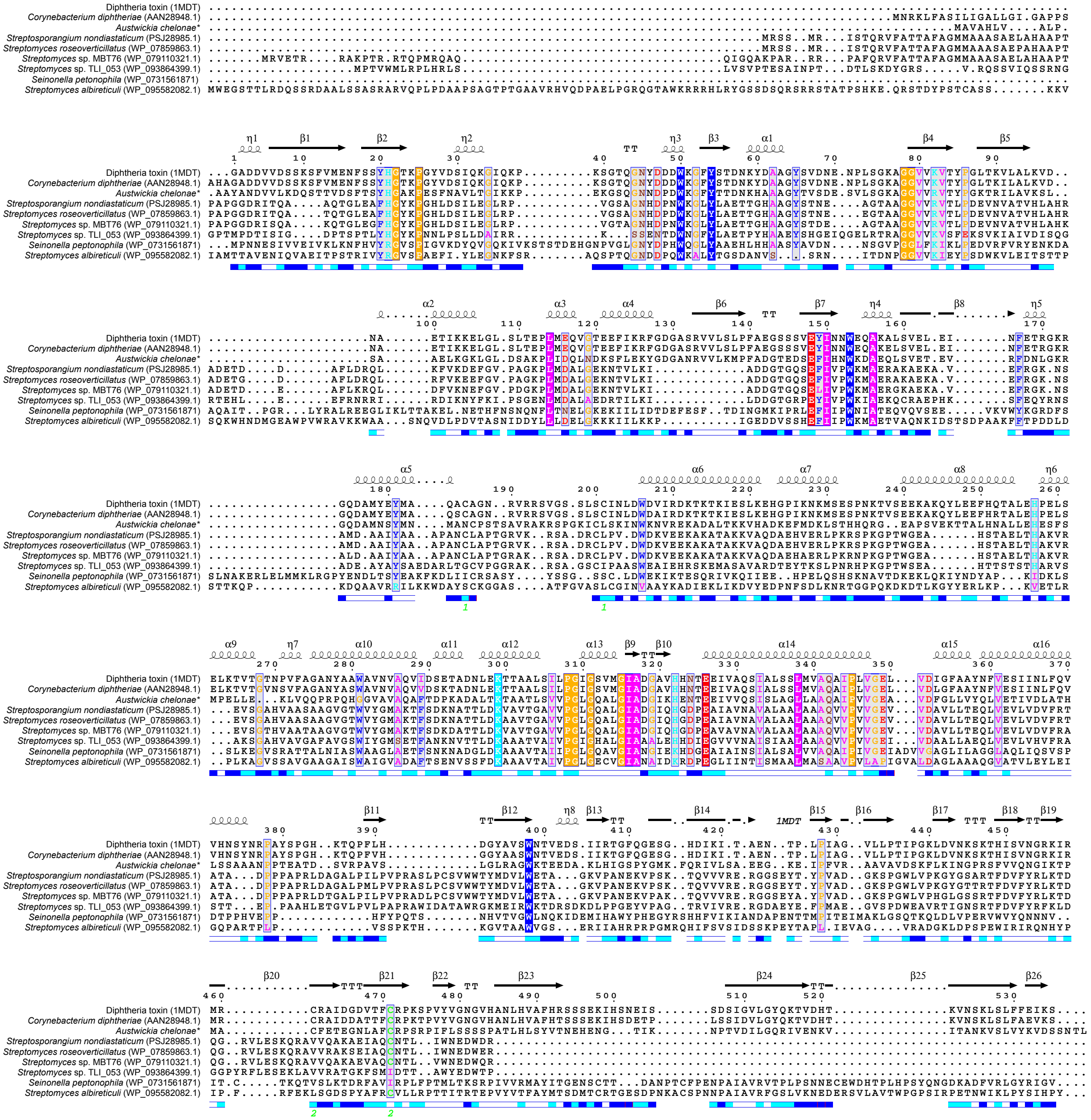
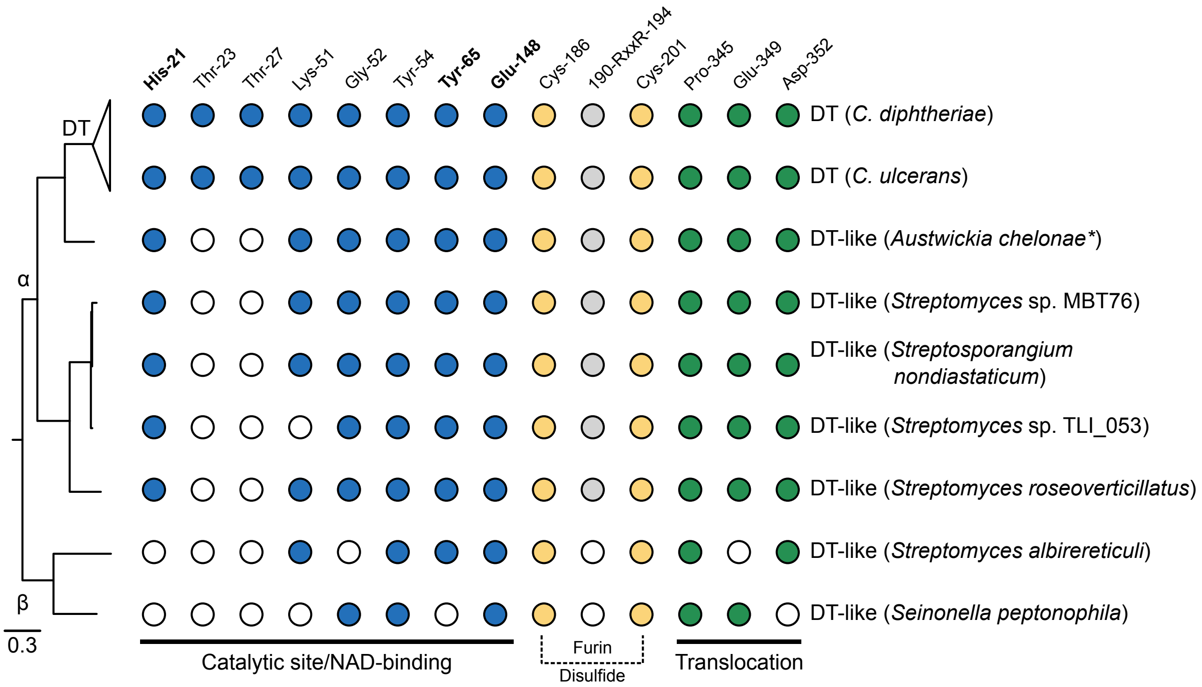

Diphtheria toxin (1MDT)
Corynebacterium diphtheriae (AAN28948.1)
Austwickia chelonae*
Streptosporangium nondiastaticum (PSJ28985.1)
Streptomyces roseoverticillatus (WP_07859863.1)
Streptomyces sp. MBT76 (WP_079110321.1)
Streptomyces sp. TLI_053 (WP_093864399.1)
Seinonella peptonophila (WP_0731561871)
Streptomyces albireticuli (WP_095582082.1)

Figure 3.5: Multiple sequence alignment of DT and DT-like protein sequences. Conserved positions are highlighted, and the secondary structure based on the structure of DT (PDB identifier 1MDT) is plotted above the alignment. Below the alignment, the per-residue relative accessibility is depicted. Image produced using ENDscript [320].



	His-21	Thr-23	Thr-27	Lys-51	Gly-52	Tyr-54	**Tyr-65**	**Glu-148**	Cys-186	190-RxxR-194	Cys-201	Pro-345	Glu-349	Asp-352
DT (*C. diphtheriae*)	●	●	●	●	●	●	●	●	●	●	●	●	●	●
DT (*C. ulcerans*)	●	●	●	●	●	●	●	●	●	●	●	●	●	●
DT-like (*Austwickia chelonae**)	●	○	○	●	●	●	●	●	●	●	●	●	●	●
DT-like (*Streptomyces* sp. MBT76)	●	○	○	●	●	●	●	●	●	●	●	●	●	●
DT-like (*Streptosporangium nondiastaticum*)	●	○	○	●	●	●	●	●	●	●	●	●	●	●
DT-like (*Streptomyces* sp. TLI_053)	●	○	○	○	●	●	●	●	●	●	●	●	●	●
DT-like (*Streptomyces roseoverticillatus*)	●	○	○	●	●	●	●	●	●	●	●	●	●	●
DT-like (*Streptomyces albirereticuli*)	○	○	○	●	○	●	●	●	●	○	●	●	○	●
DT-like (*Seinonella peptonophila*)	○	○	○	○	●	●	○	●	●	○	●	●	●	○

Figure 3.6: Presence/absence of key functional sites and motifs in DT and DT-like homologs. Key functional sites in DT were defined based on previous literature (see text), and their presence/absence pattern has been arranged based on the phylogenetic topology from Figure 3.2. *This protein sequence was encoded by two separate open reading frames in the *Austwickia chelonae* genome. See Methods for further details.

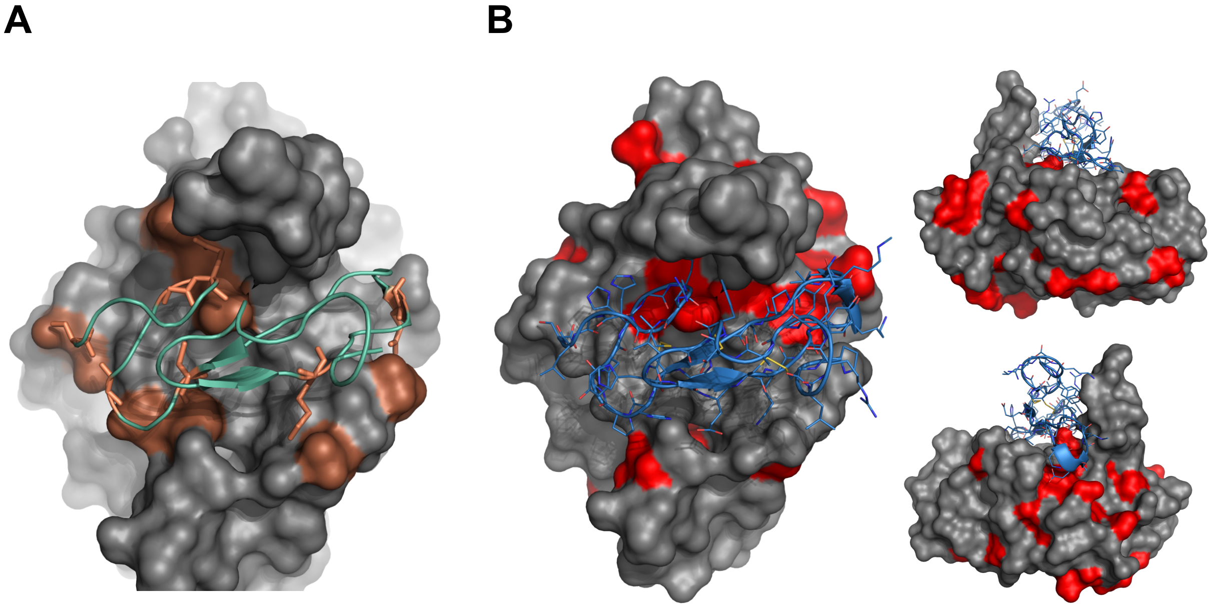


Figure 3.7: (A) A surface model of diphtheria toxin (DT, grey surface; PDB identifier 1XDT) complexed with heparin-binding epidermal growth factor (HB-EGF, green cartoon). Residues that contact HB-EGF are shown in orange, with additional stick representations. (B) All identical residues between the putative receptor binding domain of the *Austwickia chelonae* DT-like protein and DT are shown in red. Residues near the HB-EGF (blue cartoon with lines) interaction pocket are conserved, as well as several more distant residues.

DT homologs reveal purifying selection on known sites and suggest novel sites of functional importance

The MSA of DT with its newly identified homologs is important to provide insights into the potential for conserved functionality among these homologs, but it also facilitates "evolutionary footprinting" of the MSA, which has not been possible previously due to limited sequence variation in the DT family. Evolutionary footprinting allows one to measure the degree of purifying selection that has acted on individual residues in the MSA, which may not only detect sites of known importance but novel sites as well.

A total of 31 DT residues show evidence of strong purifying selection (Figure 3.8 see residues highlighted in red). Included among these residues are the known functional sites, Tyr-54, Tyr-65, Glu-148, and Pro-345. Mapping of evolutionary conservation onto the DT structure (Figure 3.8) reveals conservation of the key residues centered around the NAD-binding/catalytic site in the C domain, strong general conservation of the translocation domain, and weak conservation of the R domain. Evidence of purifying selection can also be seen in sequence logos (Figure 3.8), which shows elevated conservation of key residues in DT compared to neighbouring positions. Aside from residues of known functional importance in DT, conservation footprinting predicted 27 additional residues under strong purifying selection. Of these, 13 occur in the C domain, 10 of these occur in the T domain, and four of these occur in the R domain. Seven of these are exposed sites on the DT surface (residues with a solvent-accessible surface area greater than the mean for 1MDT), and six of these are buried sites and are therefore likely to reflect structural constraints (residues with zero solvent-accessible surface area). An example predicted residue of functional importance based on conservation analysis is Lys-299, which is invariant among all of the DT-like sequences. It is tempting to speculate that this surface-exposed residue and the neighbouring Glu-298, which is substituted to Asp in some homologs, may aid in pH-driven DT translocation.

DT homologs have novel genomic contexts and do not appear to reside in corynephage loci

Finally, one important question concerning the identified homologs is their degree of similarity at the level of genomic architecture and content. Do the identified DT-like homologs reside within genomic regions suggestive of a corynephage origin similar to DT, or do they occur in novel loci? Can other important genes in DT regulation be identified within the strains containing DT-like genes?

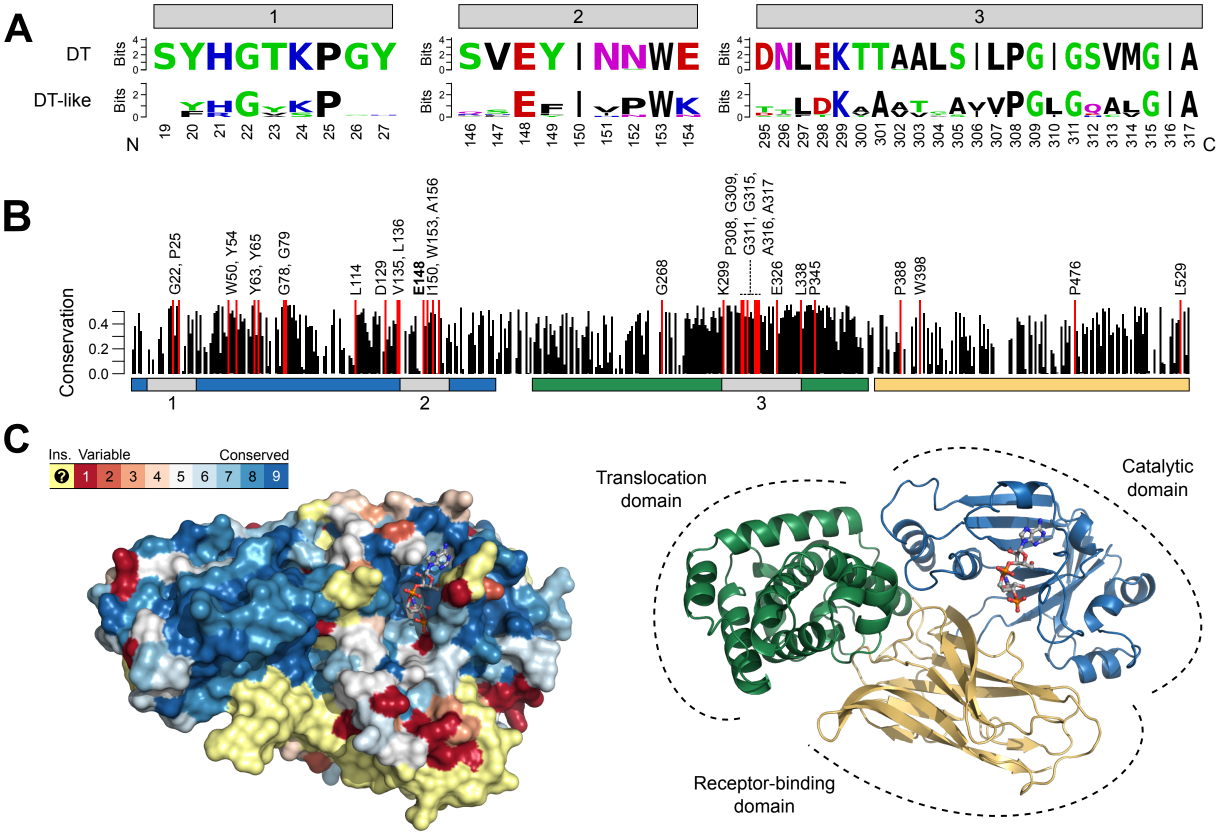


Figure 3.8: Conservation footprinting of the DT sequence family. (A) Sequence logos for three key segments of conservation highlighted in B. These three regions contain many of the most highly conserved residues between DT and DT-like proteins, including several of the residues required for catalytic and translocation activity in DT. (B) Per-residue sequence conservation scores based on the DT family sequence alignment (residue numbering based on PDB identifier 1MDT). Conservation scores were computed using ConSurf. The domain boundaries of DT are indicated below the conservation scores as follows: catalytic (C) domain, blue; translocase (T) domain, green; and the receptor-binding (R) domain, yellow. (C) Evolutionary conservation scores mapped to the DT protein structure (1MDT).

Shown in Figure 3.9 is a plot of the genomic architectures surrounding each of the homologs, which has been arranged based on the evolutionary relationships of the DT-like genes themselves. Except for an orthologous region shared between *Streptomyces* sp. MBT76, *S. nondiastaticum*, and *S. roseoverticillatus* (Figure 3.9), DT-like genes share little similarity overall, with virtually no similarity between corynephage loci within *Corynebacterium* and the DT-like group. Thus, it is unlikely that these toxins have simply been transferred to these species via a standard corynephage, and more likely that they have been acquired through an independent mechanism, or perhaps lost genomic evidence of mobile acquisition over time.

We also searched the genomes containing the DT-like genes for homologs of DtxR, the iron-mediated DT transcriptional repressor responsible for *tox* expression only under iron-limiting growth conditions [326]. Only one genome, *Streptomyces* sp. TLI_053, had a significant match (chromosome: I; 4,423,119 to 4,423,805; $E = 3{\times}10^{-66}$; 51% identity), suggesting that DtxR may not be involved in the regulation of these DT-like genes. Other regulatory genes are present adjacent to some of the DT homologs, including an FNR-type transcription factor gene adjacent to the *S. roseoverticillatus* DT-like gene, which is known to regulate genes under anaerobic conditions and is essential for virulence in some pathogens [327]. Also present is a TetR-family gene, which is a common family of transcription factors in *Streptomyces* genomes [328]. Thus, the conditions necessary for expression of DT-like genes are currently unclear and appear different from those used by *C. diphtheriae* to regulate DT.

3.1.4 Discussion and Conclusions

In this work, we have identified the first homologs of DT in bacterial species outside of *Corynebacterium* genus. These homologs are the closest relatives of DT (within currently available genomic databases), possess a DT-like domain architecture and conserve key functional sites. Since these toxins form lineages that cluster outside of the DT group, it is likely that they reflect early-diverging DT-like lineages that predate the emergence of *Corynebacterium* DT. This model is supported by their occurrence within phylogenetically neighbouring species of Actinobacteria, a phylum that includes *Corynebacterium*, *Austwickia*, and *Streptomyces* as *descendant* lineages. Also consistent with this model is the finding that DT-like genes do not appear to be carried by corynephages, but rather have an independent evolutionary history and mode of acquisition.

A major question concerning these homologs is whether they are functional toxins. Although biochemical studies of these proteins are necessary to answer this question, based

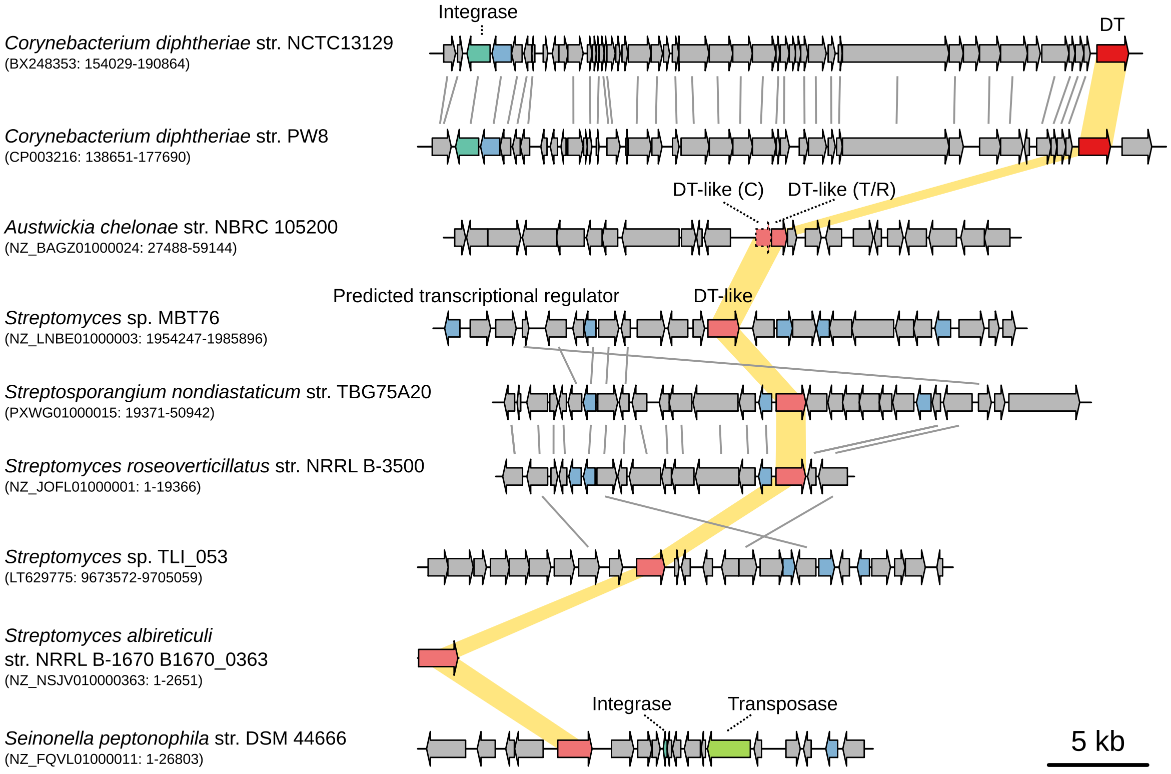

Integrase
DT
Corynebacterium diphtheriae str. NCTC13129
(BX248353: 154029-190864)
Corynebacterium diphtheriae str. PW8
(CP003216: 138651-177690)
DT-like (C)
DT-like (T/R)
Austwickia chelonae str. NBRC 105200
(NZ_BAGZ01000024: 27488-59144)
Predicted transcriptional regulator
DT-like
Streptomyces sp. MBT76
(NZ_LNBE01000003: 1954247-1985896)
Streptosporangium nondiastaticum str. TBG75A20
(PXWG01000015: 19371-50942)
Streptomyces roseoverticillatus str. NRRL B-3500
(NZ_JOFL01000001: 1-19366)
Streptomyces sp. TLI_053
(LT629775: 9673572-9705059)
Streptomyces albireticuli
str. NRRL B-1670 B1670_0363
(NZ_NSJV010000363: 1-2651)
Integrase
Transposase
Seinonella peptonophila str. DSM 44666
(NZ_FQVL01000011: 1-26803)
5 kb

Figure 3.9: Genomic context surrounding DT and identified DT-like genes. *Corynebacterium* DT genes are highlighted in dark red, and DT-like genes are shown in lighter red. Detectable orthology between genes based on top reciprocal BLAST matches is indicated by gray lines. *Corynebacterium* DT genes are located within corynephages and there is clear synteny between both strains. However, identified DT-like genes are associated with novel genomic architectures that share some synteny between each other but not to *tox*-carrying corynephages. The *S. albireticuli* DT-like gene occurs as an isolated contig and therefore genomic context is lacking. Putative transcriptional regulators (blue), transposases (green), and integrases (teal) have been highlighted.

on the bioinformatic analyses presented here, we hypothesize that these genes encode A-B toxins with a DT-like mechanism. Their domain structure, combined with the conservation of catalytic sites and key residues, suggests that DT-like proteins may be capable of host-receptor binding (although only *A. chelonae* exhibits strong similarity to the DT-R domain), translocation into the cytosol, furin-mediated cleavage and disulfide linkage, and binding and transfer of ADP to a target protein. However, the identity of the target organism, cell type, and even target protein is unknown at this point, which is a recurring theme and disadvantage of bioinformatically identified versus protein toxins identified from clinical isolates [1].

Although DT-like toxins share numerous features in common with DT, they are also distinct from DT in several respects, and these differences provide potential insights into their functionality. They cluster outside of the *Corynebacterium* DT lineage in phylogenetic analysis, their putative binding domains are divergent in sequence from DT and lack numerous human HB-EGF binding residues, and importantly, they have never been associated with human infection before. Based on these differences, it is possible that the DT-like toxins target different (nonhuman) hosts. Consistent with this idea, the closest homolog of DT is the DT-like protein from *A. chelonae*, which was originally named *Dermatophilus chelonae* but later reclassified [329]. Interestingly, *A. chelonae* was first isolated from a nose scab of a snapping turtle and shown to have low infectivity for mammals [314]. Subsequently, *A. chelonae* has been associated with skin diseases and lesions in snakes and reptiles [315, 330] similar to *C. ulcerans* in mammals. Although the DT-like gene is likely a pseudogene in the sequenced *Austwickia* genome, it is possible that a functional copy may exist in related strains, and the appearance of a DT-like gene in a skin pathogen is nonetheless noteworthy. It is tempting to speculate that the DT-like toxin plays a role in the pathogenesis of *A. chelonae*, and indeed may be responsible for the tissue death observed in *A. chelonae*-associated infections.

It is unclear from sequence analysis alone whether DT-like proteins are capable of binding DT's canonical target HB-EGF. Although the presence of HB-EGF is conserved across vertebrates, it displays sequence variation among different groups. For example, HB-EGF exhibits 97-99% identity across primates, and quickly drops among other vertebrates, from 88% in cows (human vs. cow), to 81% in mice (human vs. mouse), 71% in chickens (human vs. chicken), and 56% identity in lizards (human vs. green anole). On the other hand, the enzymatic target of the DT C domain, eEF-2, displays extreme sequence conservation among vertebrates (95-100% identity for the organisms listed above). Based on this, an intriguing possibility is that DT-like proteins have maintained enzymatic function but diversified in terms of receptor binding. The conservation observed in the C and T domains could reflect stricter functional constraints in these domains and smaller variation among their target molecules, while greater divergence in the R domain could be the product of co-evolutionary pressure or potentially specialization for different target receptors or different hosts. If this model is correct, this would implicate the DT R domain as the evolutionary determinant of DT specificity and toxicity in humans.

3.2 Structural characterization of diphtheria toxin homologs

3.2.1 Introduction

As demonstrated in the preceding chapter, diphtheria toxin homolog sequences possess notable similarities to diphtheria toxin (DT), and conserve several functional features [5]. These properties suggest that diphtheria toxin homologs also adopt a similar three-dimensional structure. To investigate this, as part of ongoing work, Sugiman-Marangos et al. have recombinantly expressed, purified, and structurally characterized the DT homologs from *Seinonella peptonophila* (SP) and *Streptomyces albireticuli* (SA). These sequences are compared below in Table 3.2, which demonstrates that the sequences of SA and SP are distinct from DT and each other.

Table 3.2: Pairwise amino acid sequence identities between DT and the DT homologs from *Seinonella peptonophila* (SP) and *Streptomyces albireticuli* (SA).

	DT	**SP**	**SA**
DT 1MDT	100	-	-
SP WP_073156187.1	28	100	-
SA WP_095582082.1	23.4	30.3	100

3.2.2 Methods

Chimera generation and protein purification

E. coli codon-optimized gBlocks® gene fragments for full-length DT, SA and SA were purchased from Integrated DNA Technologies (IDT). Two strains were used: *E. coli* str. BL21(DE3) for SP, and B834(DE3) for SA. Primary amino sequences were obtained from NCBI database entries WP_072564851, WP_095582082, and WP_073156187, respectively. Strain B834(DE3) was chosen in order to supplement with heavy atom derivativized selenomethionine. Individual domains were amplified by PCR, and domain-swapped chimeras were generated with the NEBuilder® HiFi DNA Assembly Cloning Kit (New England BioLabs). All toxins and chimeras were expressed and purified with cleavable N-terminal 6His-SUMO fusions as lBioLabs), grown to an OD600 of 0.8 in LB media, and induced for 18 hours at 18 °C using 0.1 mM IPTG. Cell pellets were pelleted by centrifugation and then re-suspended in lysis buffer (20 mM Tris pH 8.0, 500 mM NaCl) and lysed by 3-passes through an Emulsiflex C3 (Avestin) at 15,000 psi. Cell lysates were clarified by centrifugation (20 minutes at 18,000×g) and bound to a 5 mL HisTrap^{TM} FF Crude column (GE Healthcare) and eluted with 50-75 mM imidazole. Eluted protein was diluted to ∼150 mM NaCl and incubated overnight at 4 °C with SUMO protease to cleave the affinity tag. Cleaved protein was separated from SUMO, SUMO protease and uncleaved protein with Ni-NTA resin, concentrated by centrifugation and exchanged into 20 mM Tris pH 7.5, 150 mM NaCl by dialysis.

Crystallization

All crystals were grown by hanging-drop vapor diffusion at 20 °C. For SA, 1 µL of selenomethionine derivatized SA (11.8 mg mL^{-1}) was mixed with 1 µL of mother liquor (10 mM Tris-HCl pH 7.0, 200 mM calcium acetate hydrate, 20% PEG3000) and dehydrated over 200 µL of mother liquor. Diffraction quality SA crystals were obtained following successive rounds of micro-seeding and flash frozen in liquid nitrogen without any additional cryoprotectant. SP crystals were grown from drops containing 1 µL SP (15.8 mg mL^{-1}) and 1 µL mother liquor (100 mM potassium iodide, 22% PEG3350) dehydrated over 200 µL of mother liquor. SP crystals used for data collection were grown following successive rounds of micro-seeding in drops dehydrated over 550 mM ammonium sulfate, and flash frozen in liquid nitrogen without any additional cryoprotectant. Data sets were collected remotely at a wavelength of 0.979 Å for SA crystals and 2.0 Å for SP crystals on the AMX beamline at NSLSII (Brookhaven National Labs).

Structure solution

All diffraction data was processed with XDS using the AutoPROC package [331]. Structure solution of SA by SAD phasing was carried using the CRANK2 pipeline [332] in CPP4 [333]. The initial solution yielded 39 heavy atom sites (selenium) with an FOM of 56.8, producing a partial model with an Rfactor of 43.37% (Buccaneer). A single monomer from the partial model was used to perform molecular replacement using the PHENIX software package [334], which located 4 copies of SA in the asymmetric unit, which were then rebuilt with PHENIX-AutoBuild. SAD phasing of SP was performed with PHENIX-AutoSol, yielding 18 heavy atom sites (iodine) with an FOM of 35.2 and a partial model (monomer) with an Rfactor of 48.6%. Model building and refinement was carried out through multiple iterations of manual building in Coot[335] and PHENIX-Refine until R and R_{free} values converged and geometry statistics reached suitable ranges.

Sequence and structural comparisons

Global pairwise sequence alignments used in Table 3.2 were calculated using needle from the EMBOSS package [106] with a gap opening penalty of 10, gap extension penalty of 0.5, and the BLOSUM30 scoring matrix to improve alignment between distantly-related sequences. Structural alignments used in Table 3.3 and Figure 3.11 were calculated and visualized using PyMol. The crystal structures of SP and SA were separated into domains based on alignment with the structure of diphtheria toxin (PDB identifier 1MDT, [307]). Each SP and SA domain was subsequently used as a search query against sets of non-redundant protein structures from the PDB using the DALI [336] and VAST [337] web servers on July 30, 2019. The results of these structural comparisons were visualized using R. Protein cartoons were made using the Pro-origami web server [338].

3.2.3 Results

SA and SP are structurally similar to DT

Similar to DT, the crystal structures of the DT homologs (Figure 3.10a) possess a 3-domain structure, with homologous catalytic (C), translocase (T), and receptor-binding (R) domains. SP and SA both possess an $\alpha + \beta$ catalytic domain, a hydrophobic α-helix rich T domain, and a jelly-roll-like R domain. The largest structural displacements occur in loop regions, which contributes to their relatively high RMSD after superposition (Figure 3.10a; Table 3.3); additionally, the structural alignment algorithm used to calculate RMSD considers only aligned positions, and as such these RMSD values actually underestimate the overall topological differences between DT and SP or SA. Nevertheless, when the individual domains of SP and SA are compared to all other structures in the PDB, DT is consistently the best-scoring match (Figure 3.10b) using two different structural comparison algorithms (DALI and VAST [336, 337]). For the SP and SA C domains, the top-scoring match after DT was cholix toxin, an ADP-ribosyltransferase toxin from *Vibrio cholerae* with known similarity to DT [94].

Table 3.3: Per-domain structural alignments of DT homologs with DT and associated RMSD values. The RMSD of the structural alignments for every domain were calculated by PyMol using PDB 1MDT as a reference. The high RMSD of SP's C domain can be partially explained by the presence of non-conserved flexible loops, as well as an extended unique α-helix.

	C domain	**T domain**	**R domain**
SP	10.190	4.322	5.184
SA	2.229	5.635	4.634

The homologs' structures contain several unique structural features. The secondary structure elements for each domain are visualized in cartoon representations in Figure 3.11. The C domains of SP and SA feature generally elongated α-helices compared to DT. The C terminus of the SP C domain also contains a unique, extended α-helix of 28 amino acids (residues 208-236) that partially occludes the NAD+ binding site. The corresponding α-helix in DT is comprised of only 12 amino acids. Further, the SP α-helix contains a slight preference for positively charged residues (containing a total of six K or R residues), with K216, K232 and R239 having externally facing side chains, representing potential interactors. The topology of the T domains from DT, SP, and SA are largely the same, and all are predicted to possess an overall negative charge at neutral pH (pI of 4.94, 5.31, 5.17, respectively). The R domains SP and SA are larger than DT (by 28 and 23 residues), and notably contain unique loops and a short C-terminal helical turn that sit within DT's EGF binding pocket [325].

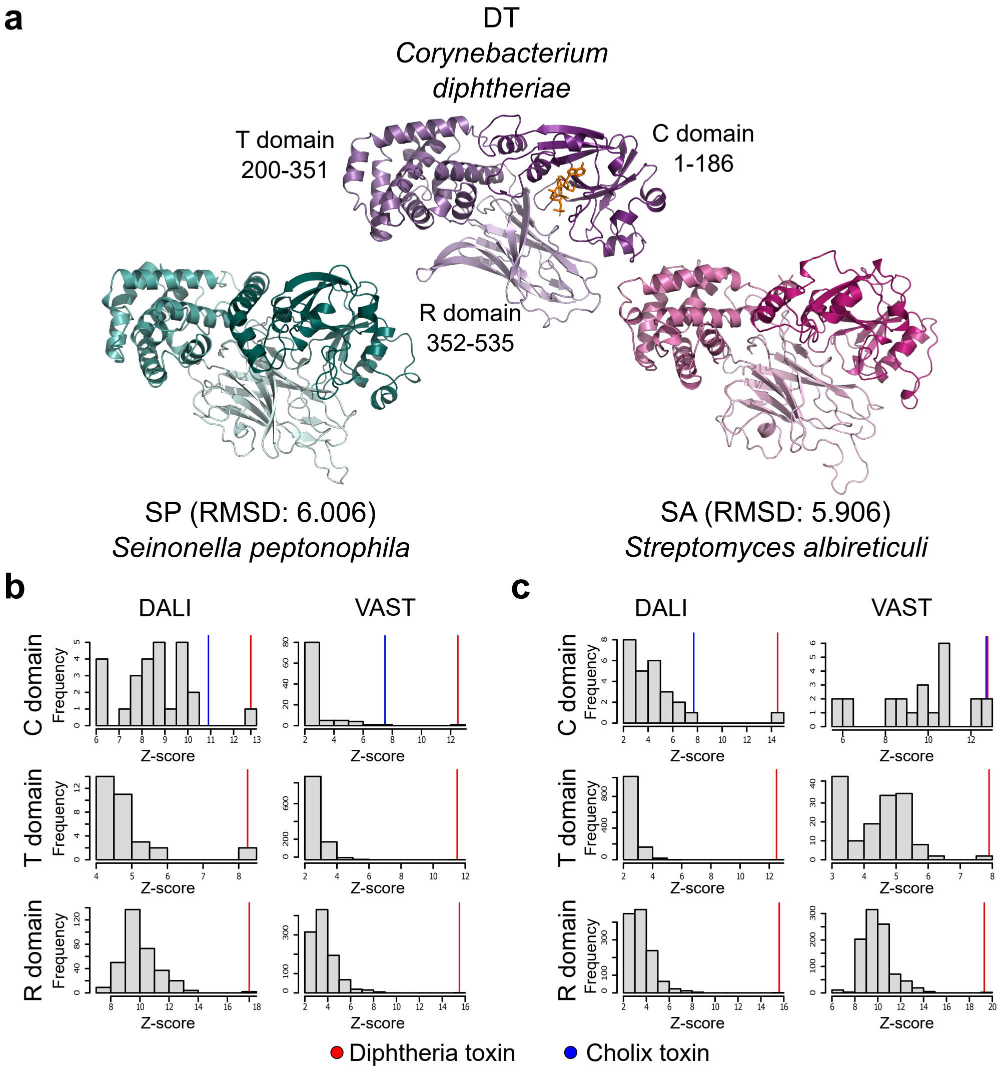


Figure 3.10: Crystal structures of two diphtheria toxin (DT) homologs. (a) DT (PDB identifier 1MDT) and the DT-related proteins from *Seinonella peptonophila* (SP) and *Streptomyces albireticuli* (SA) adopt a three-domain structure with similar topology. The RMSD of the structural alignment of full-length SP and SA to DT (calculated using PyMol, with PDB identifier 1MDT as a reference) is indicated. In (b) and (c), two structure-based search algorithms (DALI and VAST) were used to compare the domains of SP and SA to sets of non-redundant structures in the PDB. For both SP (b) and SA (c), for every domain (C, T, and R, labels indicated vertically), DT is the highest-ranked structural match by Z-score (red lines). The second-highest match for the SA and SP C domains is cholix toxin, excepting SA's C domain when searched by VAST, which produces an equal Z-score to DT's C domain.

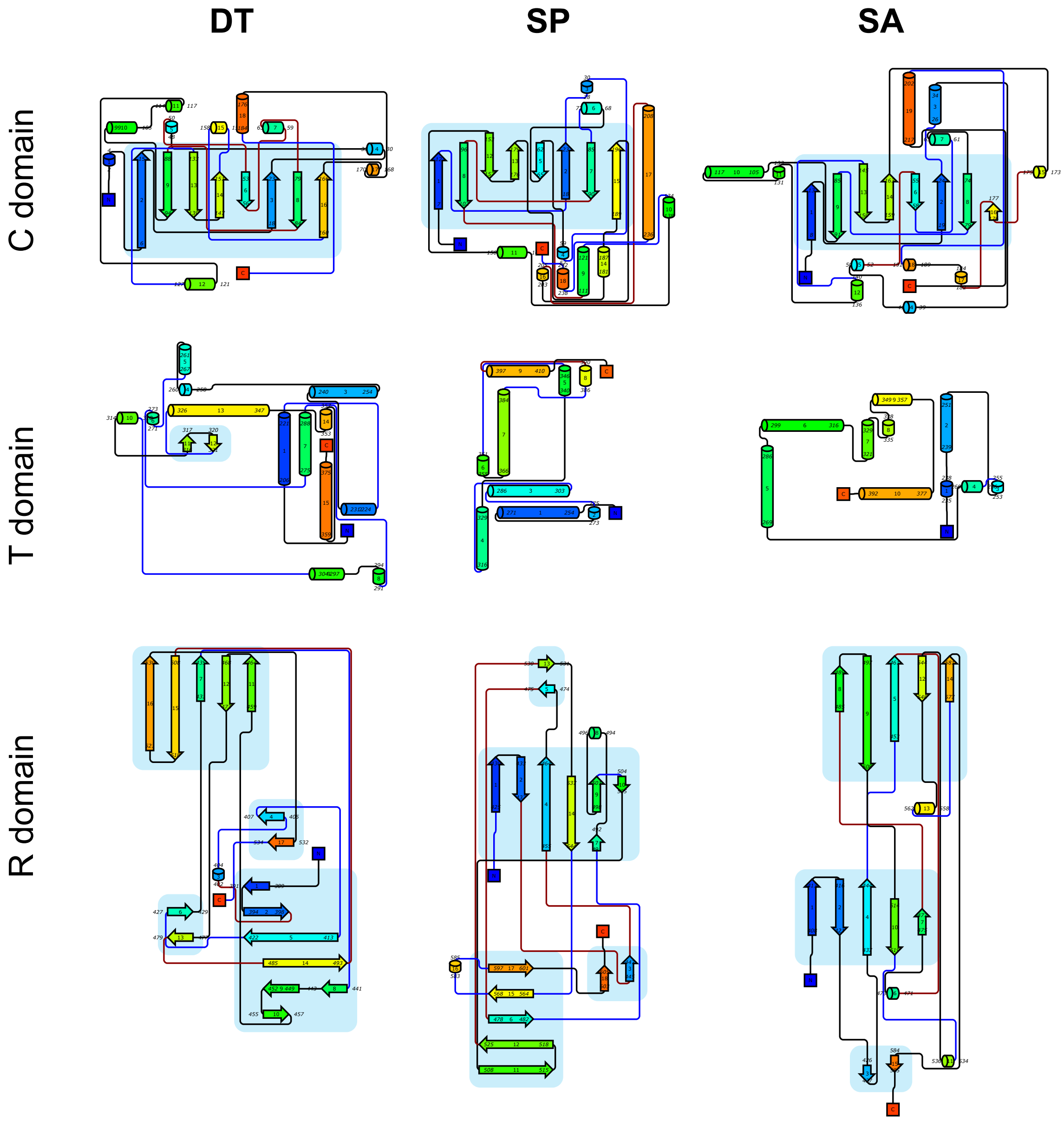


Figure 3.11: Two-dimensional topology diagrams of DT (PDB identifier 1MDT) and the newly solved structure of DT homologs SP and SA. Each query was separated into distinct domains (see Methods) and used as a query to the Pro-origami web server. Each secondary structural element is rainbow-coloured from N to C. The topologies of each domain are similar to DT in SP and SA, although they possess several unique features. In particular, SP contains an elongated α-helix within its C domain, and the region of the DT R domain corresponding to the EGF interface is different in SP and SA.

3.2.4 Discussion and Conclusions

In additional to sequence conservation, SP and SA possess significant structural similarities to DT. For both SP and SA, and for each of their C, T, and R domains, DT is respectively the closest structural relative. After DT, the closest relative of the homologs' C domains is cholix toxin, an ADP-ribosyltransferase toxin from *Vibrio cholerae* [94]. The DT homologs also feature unique structural elements, such as the large C-terminal α-helix in the SP C domain. Consistent with patterns in sequence conservation, the T domains of SP, SA, and DT possess nearly identical structural folds. The R domains of SP and SA are the most divergent in sequence, but maintain structures similar to DT's R domain. A notable feature unique to the R domains of SP and SA is the apparent insertion of a short helical turn corresponding to the EGF-binding region of the DT R domain.

Additional characterization is required to determine how the structural differences unique to SP and SA affect their function. In the solved structures, neither homologs' C domain appeared to have bound an NAD+ moiety, and both lack electron density within the C domain binding pocket that might correspond to the ligand. As well, a unique C-terminal α-helical element in the SP C domain may occlude the NAD+ binding pocket. It is possible that ligand-induced conformational changes alter the structure of the C domain to prevent occlusion in SP, and ligand-induced conformational changes may also be possible for the SA C domain. Nonetheless, the interaction between ADP-ribosyltransferase toxins with eEF-2 seems to be strongly conserved, since the interaction is mechanistically similar to eEF2's interaction with the eukaryotic ribosome [339]. This makes the appearance of structural novelties like the SP C-terminal α-helix more surprising. These outstanding questions make it desirable to obtain structures with an NAD+ ligand or an appropriate structural analog to confirm these C domains do indeed bind NAD+.

A similarly surprising difference is seen in the SP and SA R domains, which appear to contain loops and a short turn that interfere with the binding interface between DT and HB-EGF [325]. If the R domains of SP and SA are capable of binding human HB-EGF, it might utilize a different binding mechanism, or binding potentially involves conformational change in the R domains to better accommodate the receptor protein. An intriguing alternative explanation is that the R domains of SP and SA may simply be optimized for a different receptor.

It remains to be seen whether SP and SA are capable of performing the same functions as DT: binding a surface receptor, translocating into host cells, and ADP-ribosylating a host protein. Beyond that, the particular biological role of these sequences in virulence must be examined *in vivo*. Although *Streptomyces albireticuli* is known to invade the epithelium of *Caenorhabditis elegans* [312], relatively little is known about the pathogenicity or ecology of either organism, which makes pursuing an *in vivo* approach more difficult. It is unlikely that SP and SA have no function *in vivo*. Even if SP and SA were recent acquisitions through lateral gene transfer, they must presumably have performed some function in the source bacterium. Given that SP and SA have significant sequence and structural similarity to DT, it seems probable that this original function was to intoxicate host cells. A logical goal is therefore to elucidate which host types are susceptible, and further, to determine the relevance of this intoxication to the toxin producer's life cycle.

Chapter 4

Evolution of the large clostridial toxin translocase domain

Material in this chapter has been accepted for publication. When published, the manuscript will be available from the following source:

1. Orrell, K. E., Mansfield, M. J., Doxey, A. C., & Melnyk, R. A. (2019). The *C. difficile* toxin translocase is an evolutionarily conserved apparatus for bacterial protein delivery into host cells. Nature Communications (accepted).

4.1 The *C. difficile* toxin translocase is a conserved apparatus for bacterial protein delivery into host cells

4.1.1 Introduction

Large Clostridial Toxins (LCTs) are a family of bacterial toxins comprised of six proteins (TcdA, TcdB, TcsL, TcsH, TpeL, TcnA) [340, 341], defined first by their similar biochemical, immunological and pharmacological effects [342], and later differentiated by their clinical phenotype. TcdA and TcdB are the major causative agents of *C. difficile* infection, the leading cause of hospital-acquired diarrhea in developed countries [288], while other LCTs are implicated in gas gangrene, enterocolitis and toxic shock syndrome [343, 344, 345, 346].

Although LCTs vary in their clinical manifestation, they all have highly similar structure and function. LCTs are high molecular weight (>200 kDa) single-chain polypeptides, sharing between 36-90% sequence identity [340] and inactivate GTPases in the Ras superfamily by glycosylation [347]. In order to gain entry into cells and access cytosolic GTPases, LCTs utilize their multi-domain architecture [348], much like other AB toxin families, including diphtheria toxin (DT) [7] and botulinum neurotoxin (BoNT) [349]. In brief, using their central translocation and receptor-binding domain (herein referred to as T-domain), LCTs bind cell surface receptors and undergo receptor-mediated endocytosis. Low-pH mediated conformational changes in acidified vesicles culminates in insertion of regions of the T-domain into the endosomal membrane, resulting in formation of a translocation pore. The translocation pore facilitates passage of the LCT glycosyltransferase (GTD) and cysteine protease (CPD) into the cytosol, where the GTD is proteolytically released.

While much is known about the enzymatically active LCT domains, the function(s) of the LCT T-domain have remained much more elusive [348]. The LCT T-domain is much larger than the T-domain of other similar toxins (LCT: >100 kDa [350]; BoNT: ∼50 kDa [349]; DT: ∼20 kDa [7]), and has a unique structural fold at high pH [350]. The LCT T-domain at high pH is mostly composed of extended β-sheets, with a hydrophobic α-helical region that extends and wraps around the β- sheet structures. Within the β-sheet enriched region of the T-domain, four different LCT receptors have been identified (Tcdb: CSPG4 [351], Fzd [63, 352], PVRL3 [353]; TcdA: LDLR [64]; TpeL: LRP1 [354]) that all bind within the C-terminal region of the T-domain, with one receptor (CSPG4) also binding partially to the C-terminal repeating region (CROPS) of TcdB [355]. The dual functionality of the LCT T-domain to bind receptors and facilitate translocation has made it difficult to disentangle receptor-binding from translocation, although several studies have concluded that the N-terminal region of the T-domain is important for pore formation and translocation. We and others have identified a pore-forming region between residues 956-1115 [356, 357, 358], which maps to the hydrophobic α-helical stretch in the T-domain, and important pore formation and translocation residues clustered between residues 1035-1107 [357]. Recently, the structure of full length TcdB was solved at endosomal pH with 3 neutralizing VHHs [359]. Conformational changes can be observed within the pore-forming region, although binding of a VHH within the pore-forming region and lack of a membrane prevent a complete understanding of the toxin structure at low pH and in the membrane. Outside 956-1115, the functional significance of the N-terminal region of the T-domain remains unclear. Comparison of the six LCT T-domain sequences does not reveal any striking patterns in conservation or hydropathy, and by extension, obvious clues into important functional regions [348].

In the past five years, genomics-driven approaches have facilitated the discovery of hundreds of bacterial toxin homologs, providing fundamental insights into toxin evolution and diversity [1]. Although homologs of major AB-toxins such as BoNT [99, 2, 181, 3], DT [5], and others [96, 51] have been identified using bioinformatic approaches, there have been no genomics-driven approaches to uncover and characterize LCT homologs. For BoNT and DT, most studies have focused on identification and characterization of homologs conserving the full toxin architecture, such as the BoNT-like toxin in a commensal strain of *Enterococcus faecium* [2, 181]. Among multidomain homologs, the receptor (which in other toxins is distinct from the T-domain) and effector domains are the most extensively analyzed, since these domains contain well-characterized functionally important residues. Compared to receptor and effector domains, less is known about AB toxin translocation, and to date, no study has used genomics to elucidate AB toxin translocation. Since homologs have been critical to understand the function of countless other proteins, we speculate that T-domain homologs have the potential to make significant strides in our understanding of toxin translocation. Understanding the process of translocation is not only critical for a complete understanding of toxin entry and uptake into cells, but also has numerous applications, both in therapeutic interventions of toxin-mediated diseases, and in biotechnology applications, such as bacterial toxin-mediated drug delivery [360, 361, 362].

In order to unravel the function of the elusive LCT T-domain and to gain insights into translocation, we took a genomics-driven approach to uncover distant homologs of the LCT T domain. Here, we use the hundreds of newly uncovered LCT-T containing proteins to shed light on LCT-T domain diversity and distribution, and to pinpoint an evolutionarily conserved determinant of translocation present in hundreds of proteins. Our results provide fundamental insights into translocation of a group of medically and biotechnologically relevant toxins.

4.1.2 Methods

Detection of proteins containing LCT-like T domains, data set curation, and annotation

The TcdB T-domain (UniProt ID P18177.3, residues 800-1814) was used as a query for two iterations of PSI-BLAST [112] (with default parameters: BLOSUM62 substitution matrix, gap existence 11, gap extension 1) against the NCBI non-redundant protein database (nr) on June 13, 2019. A total of 1,573 protein sequences were retrieved. Proteins labeled as "partial" or otherwise truncated, as well as any proteins with $<$100 amino acids upstream of the translocase were removed from the data set. The T-domains from this set of sequences were further reduced to a final set of 203 non-redundant sequences by clustering with USEARCH (v10.0.240, [107]) at 90% identity. These sequences were aligned with the L-INS-I algorithm of the MAFFT package (v7.407, [131]), and a maximum likelihood tree was inferred using RAxML (v8.2.4, [135]) with automatic evolutionary model selection (LG), 4 gamma-distributed rate categories, automatic bootstopping with autoMRE, and a thorough ML search. This was visualized using the ape package in R [363]. This tree and alignment were used as inputs to the Consurf web server and related to the structure of TcdA [309], depicted using PyMol (available from `https://pymol.org`). Protein domains were annotated using InterProScan (v5.33-72, [126, 364, 184, 118, 121]) and transmembrane helix prediction was performed with TMHMM2.0 [306]. AnnoTree was used for phylogenomic visualization of species containing LCT-T homologs across the bacterial tree of life [365]. Proteome-level Pfam annotations from the AnnoTree database were used to determine the domains most correlated with an annotated LCT pore-forming domain model (Pfam model PF12920), the top 100 of which are available in Table C.1. Metagenomic surveys were performed using EBI's MGnify server (`http://www.ebi.ac.uk/metagenomics`) using the TcdB translocase (UniProt ID P18177.3, residues 800-1814) as the query. Genomic context plots were generated using genoPlotR (version 0.8.9, [163]).

Pathogenicity analysis of LCT-T homologs

The association of bacterial organisms with pathogenicity was assessed based on where the organism was isolated and reviewing the literature, where possible. Broadly, an organism's level of pathogenicity was categorized into one of four possibilities: no known pathogenicity or host association, host-associated with no known pathogenicity, known pathogen of non-human hosts, and known pathogen of humans.

Comparison of effector diversity from different AB toxin families

The effector domains from different toxin families were retrieved by searching with each toxin's translocase domain as a query (BoNT: PDB identifier 3BTA, residues 548-865; DT: SwissProt identifier P00588.2, residues 232-383) against the NCBI non-redundant protein database with two iterations of PSI-BLAST. For DT and BoNT, the entire portion upstream of the translocase hit region was extracted and treated as the effector region. For the LCT family, the effector region was more difficult to define because it contains the glycosyltransferase domain as well as the autoproteolytic cysteine peptidase domain, and not all LCT-T homologs have detectable peptidase domains. Thus, the entire region upstream of translocases in proteins lacking a peptidase, and the regions upstream of peptidases in peptidase-containing sequences, were extracted separately to yield the set of

LCT effectors. The putative effector regions from BoNT, DT, and the LCTs were clustered at increments of 5% cluster sequence identity between 50% and 100% using USEARCH. Effector types were assigned using InterProScan.

Generation of toxin chimeras and TcdB 851-1473

Regions of the TcdB T-domain were amplified from a codon-optimized TcdB gene for expression in *E. coli*, and fused into a pET28a vector using the In-Fusion HD Cloning (Clontech). Vectors for ADPR-[truncated TcdB T-domain]-DTR chimeras contained the diphtheria toxin (DT) ADP-ribosyltransferase (ADPR) (defined here as residues 1-201) with an intact furin cleavage site and DT receptor binding region (DTR) (defined as in DT as amino acids 378-535), and the vector for GTD-CPD-[TcdB 851-1473]-DTR chimera contained the TcdB glucosyltransferase (GTD) and cysteine protease domain (CPD) (defined here in TcdB as residues 1-543 and 544-799, respectively) and the DT receptor binding region (DTR). For regions of the TcdB T-domain truncation beginning at 851 or 881, a short linker (four glycine followed by one serine (G_4S)) was added between the truncated TcdB T-domain, and the CPD or ADPR.

Expression and purification of recombinant toxin chimeras and TcdB 851-1473

Toxin chimeras and TcdB 851-1473 were transformed into *E. coli* BL21 DE3 competent cells and expressed with N-terminal H_6-SUMO and C-terminal Strep tags or an N-terminal H_6-SUMO tag, respectively. A total of 20 mL of overnight culture was inoculated into 1.0 L of TB with 50 µg mL^{-1} kanamycin and induced at OD_{600} ∼0.6-0.8 with 1 mM isopropyl-β-thiogalactopyranoside (IPTG) at 25 °C for 4 hrs or 18 °C overnight. Cells were harvested by centrifugation and resuspended with lysis buffer (20 mM Tris pH 8.0, 500 mM NaCl) and lysed by an EmusliFlex C3 microfluidizer (Avestin) at 15,000 psi. Whole cell lysates were then centrifuged at 15,000g for 20 min and filtered through a 0.2 µm filter. Proteins were purified by Strep-Tactin affinity chromatography using a Strep-Tactin column (GE Healthcare) and eluted in 20 mM Tris pH 8.0, 150 mM NaCl, 1 mM D-desthiobiotin and 5% glycerol. For TcdB 851-1473, the H_6-SUMO tag was removed by adding 1U of Sumo protease (Life Sensor) to purified protein in 20 mM Tris-HCl pH 8.0 containing 150 mM NaCl and 1 mM DTT. The cleavage reaction mixture was incubated at 25 °C for 2hrs followed by purification using His-Pure Ni-NTA resin (Thermo Scientific) to remove the His-Sumo protease and His-Sumo tag from the purified protein samples. All protein were verified by SDS-PAGE, concentrated with a 30,000 MWCO ultracentrifugation device. Protein concentration was calculated by densitometry using ImageJ software.

HPTS/DPX dye release from liposomes

The HPTS/DPX dye release assay was based on protocols from Genisyuerek et. al. 2011 [358]. In brief, liposomes were prepared with 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC) (Avanti Polar Lipids), with 0.8% 1,2-dioleoyl-sn-glycero-3-[(N-(5-amino-1-carboxypentyl)iminodiacetic acid)succinyl] (nickel salt) (DGS-NTA[Ni]) (Avanti Polar Lipids). After drying down with N_2, the lipid film was resuspended in 20 mM Tris, 150 mM NaCl, pH 8.0, 35 mM 8-Hydroxypyrene- 1,3,6-trisulfonic acid (HPTS) and 50 mM (p-xylene-bis-pyridinium bromide) (DPX) (Thermo Fischer). Lipid vesicles were subjected to 10x freeze-thaw cycles and extruded using a 200 µm filter. To get rid of un-encapsulated dye, the lipid vesicles were then subjected to gel filtration and eluted in 20 mM Tris, 150 mM NaCl. To assess fluorophore leakage, protein were added in a ratio of 1:10,000 with liposomes, such that the final liposome concentration was ∼400 µM. The fluorescence was monitored in a 96-well opaque plate (Corning) (excitation 403 nm, emission 510 nm) in high pH buffer (20 mM Tris, 150 mM NaCl, pH 8.0) or low pH buffer (20 mM Na-acetate, 150 mM NaCl, pH 4.5) (toxin chimeras), or with citrate-phosphate buffers ranging from pH 4.0-pH 7.5 in 0.5 pH increments (TcdB 851-1473). To determine total HPTS fluorescence, Triton X-100 was added to each well to a final concentration of 0.3%. All spectra were normalized to 100% dye release by 0.3% triton.

Cell viability

Vero cells were cultured in DMEM (Wisent) with 10% FBS (Wisent). Vero cells were seeded at a density of 4,000 cells per well in 96-well plates (Corning) and cultivated at 37 °C and 5% CO_2 overnight. Toxin chimeras were added to Vero cells in a serial dilution of 1/3 (ranging from ∼50-100 nM; Figure C.2) and incubated at 37 °C and 5% CO_2. After 48 hrs, cell viability was assessed by PrestoBlue Cell Viability Reagent (Life Technologies). Fluorescence was read on a Spectramax M5 plate reader (Molecular devices). For each toxin condition, the data were blank subtracted (i.e. buffer only, no cells) and normalized (cells, untreated, representing 100% viable cells) and converted to fraction viable cells. From these data, dose response curves and half-maximal effective concentrations (EC_{50} values) were generated and calculated using Prism software.

Protein synthesis inhibition

Vero cells stably expressing NanoLuc (Nluc) Luciferase (Promega) were cultured in DMEM (Wisent) with 10% FBS (Wisent) and 1% penicillin/ streptomycin (Wisent). Vero Nluc cells were seeded at a density of 4,000 cells per well in 96-well plates (Corning) and cultivated at 37 °C and 5% CO_2 overnight. Toxin chimeras were added to Vero NLuc cells in a serial dilution of 1/3 (ranging from ∼50-100 nM; Figure C.2) and incubated at 37 °C and 5% CO_2 for 24 hrs. Nano-Glo Luciferase Assay substrate and buffer (Promega) were added to cells as per the manufacturer's instructions, and luminescence was read on a Spectramax M5 plate reader at (Molecular Devices). For each toxin condition, the data were blank subtracted (i.e. buffer only, no cells) and normalized (cells, untreated, representing 0% protein synthesis inhibition) and converted to fraction protein synthesis inhibition. From these data, dose response curves and half-maximal effective concentrations (EC_{50} values) were generated and calculated using Prism software.

Cell rounding

Vero cells were cultured in DMEM (Wisent) with 10% FBS (Wisent). Vero cells were seeded at a density of 8,000 cells per well in 96-well plates (Corning) and cultivated at 37 °C and 5% CO_2 overnight. The next day, media was exchanged with serum-free media and cells were intoxicated by adding toxin chimeras at 1 nM. After 3 hrs, light microscope images were taken at 10x magnification to assess rounding of cells.

Rac1 glucosylation

Vero cells (ATCC, Cat #CCL-81) were cultured in DMEM (Wisent) with 10% FBS (Wisent) and 1% penicillin/streptomycin (Wisent). Vero cells were seeded at a density of 100,000 cells per well in 6-well plates (Corning) and cultivated at 37 °C and 5% CO_2 overnight. The next day, media was exchanged with serum-free media and cells were intoxicated by adding GTD-CPD-TcdB(851-1473)-DTR or TcdB at 1 nM. After 1 hr, media was aspirated from cells, cells were washed with PBS and lysed by addition of Laemmli loading buffer with beta-mercaptoethanol (Bio-Rad) to each well. Samples were heated to 90 °C before immediately loading on an SDS polyacrylamide gel. Following electrophoresis, samples were transferred to nitrocellulose using standard wet transfer protocols, blocked with 5% w/v non-fat powdered milk (Biobasics) in Trisbuffered saline (TBS; 5 g/100 mL, with a final concentration of 0.1 M Tris HCl, 0.15 M NaCl) and probed for total Rac1 (1:1,000 dilution) with Anti-Rac1 antibody 23A8 (Millipore Signma, Cat #05-389(or for non-glucosylated Rac1 (1:1,000 dilution) with Anti-Rac Mab102 (BD Biosciences, Cat #610651). Anti-α-tubulin (1:5,000 dilution; Sigma, Cat #T5168) was used as the loading control. Following overnight incubation with the primary antibody, the blot was washed with TBS/0.1% Tween20 and incubated with (1:10,000 dilution) with Anti-mouse conjugated horseradish peroxidase (GE Healthcare, Cat #NXA931V) for 60 min. After the final washes in Tris-buffered saline with Tween20, chemiluminescent detection was carried out using Clarity Western ECL Substrate (Bio-Rat) and exposing to BioMax MR film (Kodak).

Stability studies

TcdB 851-1473 (5-10 µM) was incubated in citrate-phosphate buffers ranging from pH 4.0-pH 7.5 in 0.5 pH increments at room temperature in presence and absence of 20 mM dodecylphosphocholine (DPC). After 30 minutes, samples were spun at 5,000 × g for 5 minutes to pellet aggregates (but not detergent). The supernatant was removed from each sample, and mixed 1:1 with Laemmli loading buffer with beta-mercaptoethanol (Bio-Rad) and boiled for 2 min. Samples were then loaded onto an SDS-PAGE gel and stained using GelCode Blue (Thermo Fisher).

Circular dichroism (CD) spectroscopy

Far-UV CD spectra were recorded at room temperature using a J-810 spectropolarimeter (Jasco) with 0.1 cm path length cuvettes. Protein was added to a final concentration of 5-10 µM in presence or absence of 20 mM dodecylphosphocholine (DPC) in citrate-phosphate buffers ranging from pH 4.0-pH 7.5 in 0.5 pH increments. After 30 minutes at room temperature, all samples were spun down at 5,000 × g for 5 minutes to pellet aggregates but not detergent. The supernatant was removed for each sample, and CD spectra were acquired from 250 to 190 nm at 50 nm min^{-1}, with a data pitch of 0.1 nM and three accumulations. Spectra were then averaged, blank subtracted and converted to mean residue ellipticity using standard formulas.

4.1.3 Results

LCT-T containing proteins are widespread in pathogenic bacteria outside of clostridia

To begin to explore the distribution, diversity, and function of the LCT T-domain, we searched 200,270 available genomes (8,141 eukaryotes, 192,129 prokaryotes) within the Genbank database, and retrieved all sequences containing an LCT-like T-domain. To this end, a PSI-BLAST search was performed using the putative TcdB T-domain as a query

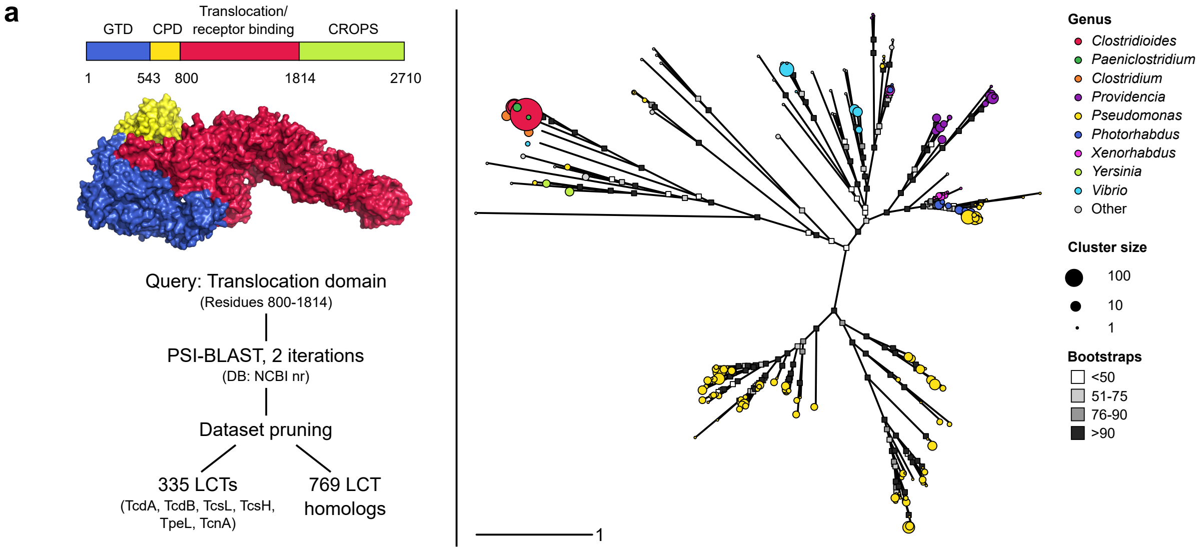


Figure 4.1: Discovery of LCT-like translocases in diverse species. (a) The search strategy for discovering proteins with LCT-like translocases began by searching the NCBI non-redundant protein database (NR) using the TcdB T-domain as a query (UniProt P18177.3), followed by two iterations of PSI-BLAST searches. After removal of partial and poorly aligning sequences, a total of 1,104 LCT and LCT-T homologs were retrieved. (b) A redundancy-removed and pruned alignment of translocase sequences was used to generate a maximum likelihood phylogeny. Each tip represents a sequence cluster centroid, and coloured according to their genus (genera that represent more than 1% of the total data set are coloured). Tip radius is proportional to the size of the cluster.

(UniProt ID P18177.3, residues 800-1814). After removing partial and truncated hits a remarkable 1,104 sequences were uncovered, including 335 LCT sequences found in various clostridia (*C. difficile*, *C. perfringens*, *C. novyi* and *Paeniclostridium sordellii*, previously *C. sordellii*), and an unprecedented 769 sequences in species outside of clostridia (hereafter referred to as LCT-T homologs; Figure 4.1a). LCT-T homologs share an average of 18.6% amino acid identity with the TcdB translocase, reflecting remote homology, and are mostly found distributed among the class Gammaproteobacteria (688 sequences across 32 genera); of the 32 genera, sequences are mostly found in *Pseudomonas* (419 sequences), followed by *Vibrio* (72 sequences) and *Providencia* (67 sequences) (Figure 4.1b). The patchy distribution of LCT-T homolog sequences across the bacterial tree of life indicates that evolution by lateral gene transfer has likely played a strong role in the family's evolution, a pattern that has been observed in other toxin families (Figure 4.2) [365].

Consistent with their wide taxonomic distribution, analysis of metagenomes revealed

LCT-T homologs in a broad distribution of environments (Figure 4.3). We detected LCT-T homologs in human gut, soil, wastewater, marine and aquatic environments, where the T-domains of other AB toxins (i.e. BoNT, DT) were conspicuously absent (Figure 4.3). Notably, LCT-T homologs were not encoded within analogous LCT pathogenicity loci (PaLoc) [366], with genes for toxin regulation (*tcdR*, *tcdC*) or toxin export (*tcdE*). Despite their occurrence in a wide variety of genomic contexts (Figure 4.4), many LCT-T homolog genes were located near components of type I, III, IV, and VI secretion systems, and through proteomic association, the top co-occurring protein families with the LCT-T domain included secretion systems, along with many other virulence genes and mobile genes, including transposons and insertion elements (Table C.1). Together with their phylogenetic distribution, these data suggest that LCT-T homologs may function as putative toxins, many of which utilize non-LCT modes of bacterial secretion and export.

In line with their toxin-like genomic signatures, 74% of LCT-T containing proteins were found in organisms with evidence of pathogenicity. In addition to species with known pathogenic potential, homologs were found in a range of host-associated microbes, which may be suggestive of cryptic pathogenic potential. Of the known pathogenic species, 22% were human pathogens, including *Pseudomonas fluorescens*, *Photorhabdus asymbiotica*, *Serratia mascerens* and several species of *Providencia* (*P. alcalifaciens*, *P. rettgeri*, *P. stuartii*); these bacteria are generally opportunistic pathogens, and are associated with severe diseases in immunocompromised individuals [367, 368, 369, 370, 371, 372]. Interestingly, the majority of remaining LCT-T homolog sequences occurred in species associated with pathogenicity in non-human hosts, including species of *Vibrio*, *Pseudomonas*, *Xenorhabdus* and *Photorhabdus*, which are known pathogens of aquatic organisms, insects, and fungi [373, 374]. Notably, species of *Pseudomonas*, *Xenorhabdus*, and *Photorhabdus* produce insecticidal toxins FitD and Mcf, which have with previously noted homology to the TcdA/TcdB T-domain [375]. Although the association of LCT-T containing proteins with disease or infection is not known, their presence in pathogenic species strengthens the claim of a putative toxin functionality.

LCT-T homologs occur in putative toxins with diverse effector types

We next annotated the individual domains within all 1,104 LCT-T containing proteins, to determine the types of effectors upstream of (and therefore, potentially translocated by) LCT-T homologs. In support of a putative function as a toxin, most LCT-T containing proteins have upstream "LCT toxin domains", with ∼30% of sequences containing a glycosyltransferase (GTD-containing), ∼20% with a glycosyltransferase and cysteine protease (GTD-CPD, or "LCT-like") and ∼10% with a cysteine protease (CPD-containing) (Figure

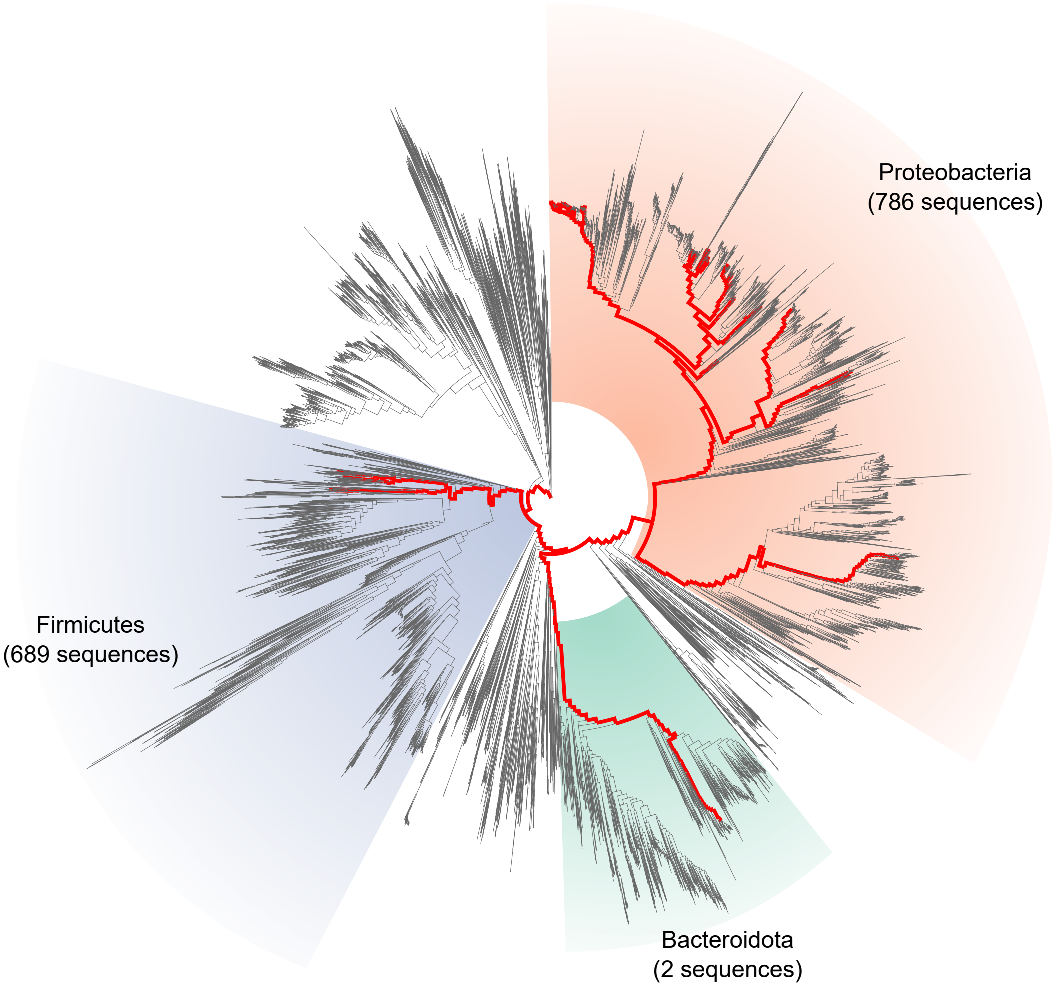


Figure 4.2: Phylogenomic distribution of organisms encoding proteins with LCT-like T domains. Organisms encoding at least one protein with an LCT-like T domain (red highlighting) are mostly found within the Firmicutes and Proteobacteria. Visualization of phylogenomic distributions was performed using AnnoTree [365].

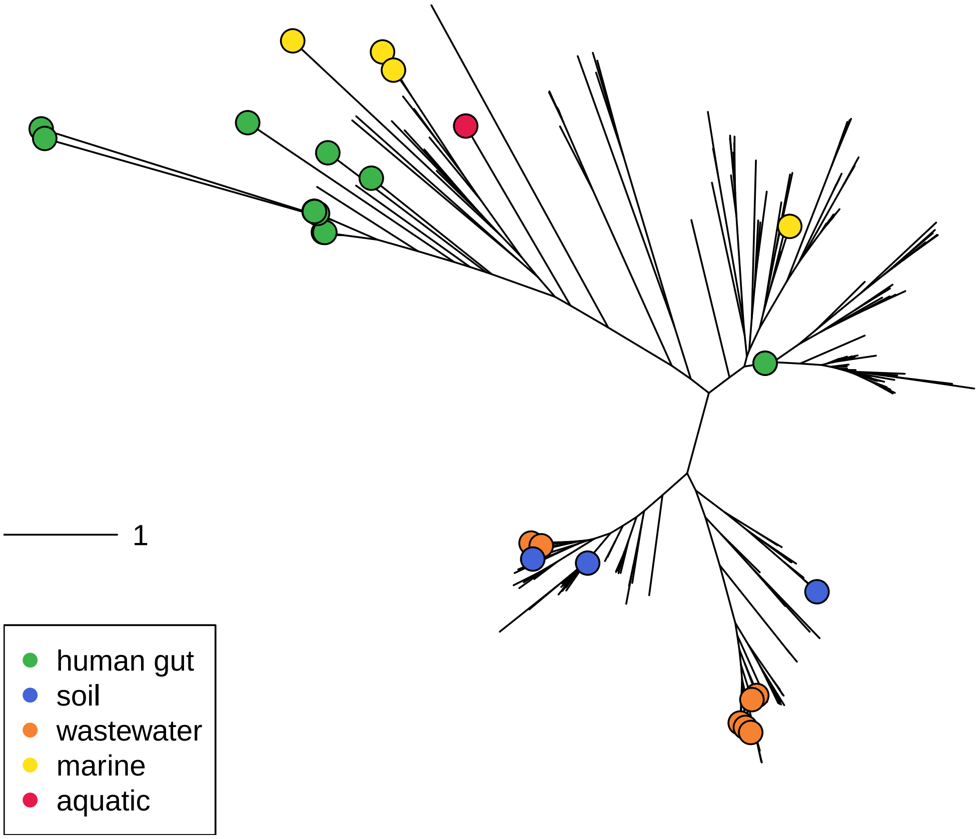


Figure 4.3: Phylogenetic placement of metagenomic LCT-like translocases onto the tree of LCT translocases from Figure 4.1. Coloured circles represent metagenomically-derived sequences with significant similarity to the TcdB translocase. Sequence fragments from human gut metagenomes include nearly identical matches to segments of the TcdA and TcdB translocases. Metagenomic sequences from other environments generally have lower identity to the TcdB translocase, and place more closely to sequences from *Pseudomonas* and *Vibrio* spp.

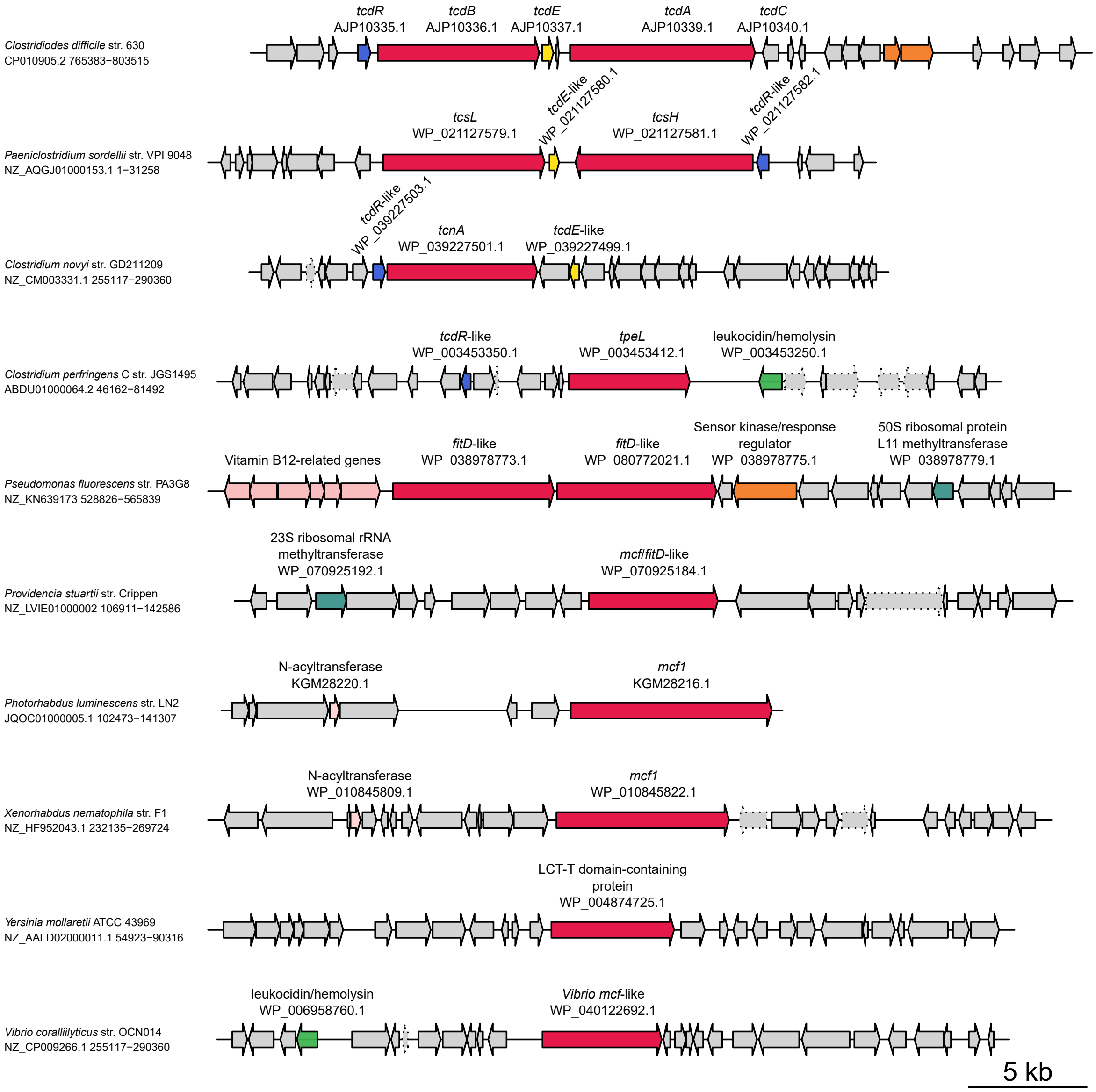


Figure 4.4: Genomic architecture of genes with LCT-like translocases. In *C. difficile*, the LCT genes *tcdA* and *tcdB* are found in a pathogenicity locus (PaLoc) that also contains the regulator *tcdR* and the holin gene *tcdE*, responsible for exporting the toxin genes. Other LCTs are found in loci with some shared features. Outside of the clostridia, proteins with LCT-like translocase domains are found in a wider variety of genomic contexts; a small subset is shown here. These proteins are associated with the presence of secretion systems, but otherwise have few commonalities. One apparently conserved locus, shared between many species of *Pseudomonas* and highlighted in pink, is related to B12 metabolism.

4.5a). We also identified proteins containing different toxin families upstream of LCT-T domains (e.g., a homolog of anthrax toxin lethal factor in WP_102423241.1 from *Vibrio* sp. 10N.261.52.A1). Interestingly, ∼40% of sequences have upstream sequences that are not annotated, containing protein(s) either falling below detection thresholds and/or are novel, uncharacterized proteins. By comparison, through the same type of analysis, the putative effectors of DT-T and BoNT-T homologs were predominantly ADP ribosyltransferases (ADPR) and peptidases, respectively, the well-known effectors of DT and BoNT (Figure 4.5b).

With respect to the size of the upstream translocated cargo, we found that the LCT-Ts on average translocate much larger proteins: 965 amino acids, compared to 479 amino acids for BoNT-T homologs: and 218 amino acids DT-T homologs (Figure 4.5b). Moreover we found that the translocated cargo had greater sequence diversity, regardless of clustering at any sequence identity (Figure 4.5c). These data suggest that although each AB toxin-translocase may be fine-tuned for translocating particular types of effector, LCT-T translocases may be capable of translocating more diverse effectors with a wider range of sequence and size diversity.

Evolutionary footprinting identifies a conserved region across LCT-T homologs

Next, we leveraged the greatly expanded number and diversity of newly identified LCT-T homologs to uncover conserved molecular features of the T-domain of LCT-T homologs. Alignment of the TcdB T-domain and LCT-T homologs revealed a shared core region, with a distribution of start and end sites at amino acids 815 (±6 residues) and 1514 (±99 residues), respectively (Figure 4.6a). In the context of the best characterized homolog, TcdB, this region encompasses regions previously implicated in pore-formation and translocation [356, 357]. Within this stretch are three distinct regions (region i: 956-1019; region ii: 1029-1078; region iii: 1090-1110) that share a remarkably similar pattern of hydropathy - one small peak, followed by two larger peaks - that map to putative membrane-insertion regions (Figure 4.6b) [356]. Furthermore, many of residues that were found to be highly conserved within region ii and region iii among the LCT-T homologs (TcdB residues: I1035, D1037, L1041, P1095, G1098, I1099, L1106 and V1107) correspond to residues in TcdB that were previously shown to be implicated in pore formation and/or translocation (Figure 4.6c) [357]. The conservation of important translocation features is consistent with the LCT-T homologs functioning as protein translocases in putative toxins.

The identification of an evolutionarily-conserved region in LCT-T homologs led us to hypothesize that such smaller-sized forms of the larger T-domain might comprise the core machinery that is necessary and sufficient for pore-formation and translocation. To address

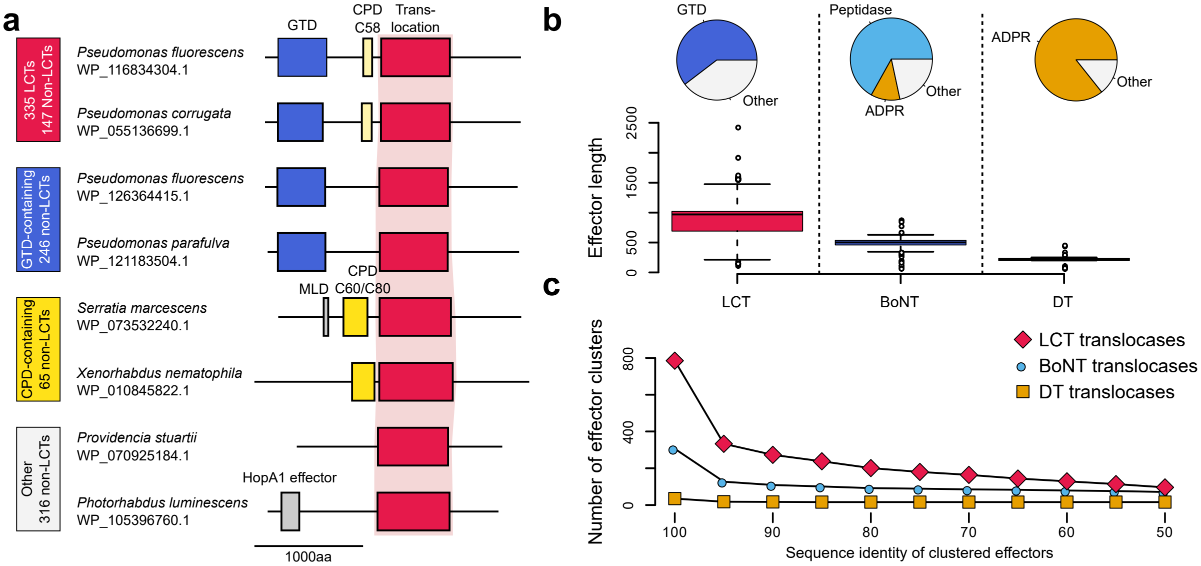


Figure 4.5: Analysis of proteins with LCT-like T-domains. (a) Proteins with LCT-like T-domains can be found in three domain architecture types: LCT-like (red; 147 sequences), GTD-containing (blue; 246 sequences), CPD-containing (yellow; 65 sequences), and variable (other) composition (grey; 316 sequences). Within each domain architecture, two examples are shown. (b) By clustering the effectors at different levels of sequence identity, the diversity of effectors between the translocation domain families can be compared. At all levels of identity, the LCT family yields the largest number of effector sequence clusters, suggesting the LCT T-domain transports more diverse cargo than BoNT or DT translocases. (c) Comparison of effector lengths and types between BoNTs, DTs, and LCTs. The DT family possesses the smallest effectors, most of which are ADP-ribosyltransferases (ADPRs). The BoNT effector family is comprised mostly of metallopeptidases (peptidases) and ADPRs. The LCT family possesses the largest effectors, which are mostly GTDs.

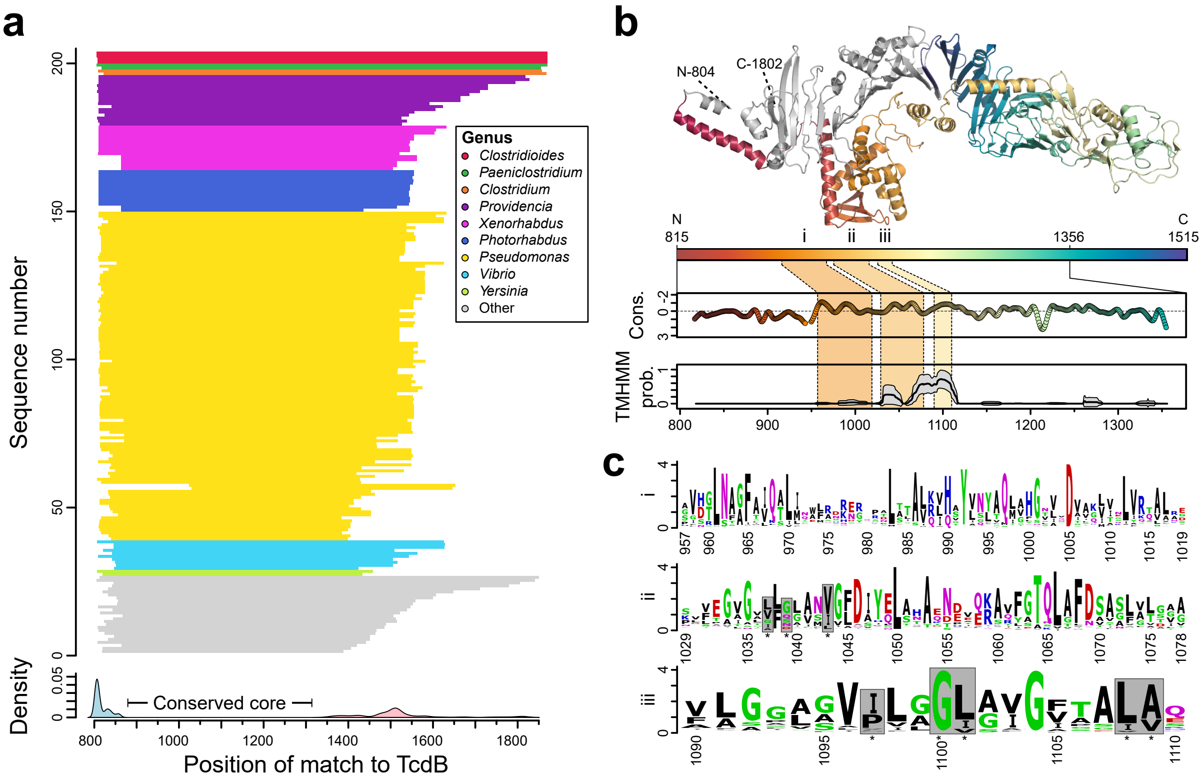


Figure 4.6: Patterns of sequence conservation among LCT-like T-domains. (a) The distribution of BLAST matches to the TcdB T-domain demonstrates that the region common to all LCT-like T domains is limited to within the first 700 amino acids of the N terminus. (b) The structural context of the mean BLAST start (815) and end (1515) sites, as mapped on to the TcdA T domain (PDB 4R04). Normalized sequence conservation for the region spanning residues 815 to 1356 shows three main peaks of sequence conservation (i, ii, and iii). Two of these conserved regions correspond to an average increase in predicted transmembrane helix propensity. (c) Within the three sequence conservation peaks, several sites known to result in the loss of translocation activity in TcdB (underlighted with black boxes and denoted with an asterix) are conserved among LCT-like T-domains. However, several strongly conserved residues are not associated with a loss of translocation function, suggesting they are related to some other key function in LCT-like T-domains (including Y991 and D1005).

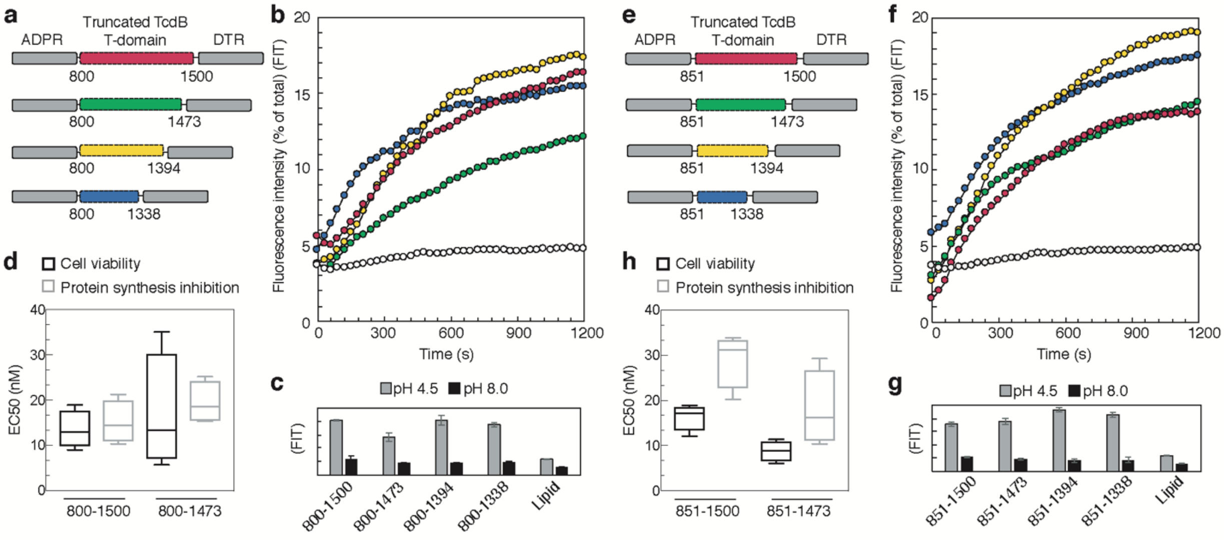


Figure 4.7: Defining the evolutionarily conserved and minimal LCT translocase region. (a) Construct design of ADPR-[truncated TcdB T-domain]-DTR chimeras. (b) Dye release from HPTS/DPX loaded liposomes at pH 4.5, with truncated TcdB T-domains coloured as follows: 800-1500 (red), 800-1473 (green), 800-1394 (yellow), 800-1338 (blue). (c) Quantification of dye release from HPTS/DPX loaded liposomes after 20 minutes, at pH 4.5 (grey) and pH 8.0 (black) (N=3). (d) EC50 values from cell viability (black) and protein synthesis inhibition (grey) of Vero cells (N=4). Panels e, f, g, h the same as a, b, c, d, respectively, but with truncated TcdB T domains 851-1500 (red), 851-1473 (green), 851-1394 (yellow), 851-1338 (blue).

this directly, we used the relative distribution frequency of start and end sequence coverage sites as guides to design a series of T-domain truncations in the most well characterized LCT homolog, TcdB, generating TcdB truncations with two different N-terminal start sites (*viz.*, residues 800 and 851), and a variable C-terminus (X = 1500, 1473, 1394, 1338) (Figure 4.7, Figure C.1). In order to assess translocation in cell-based assays, we developed a pore formation/translocation platform using the DT ADP-ribosyltransferase (ADPR) and the DT receptor-binding domain (DTR) as a scaffold, such that test chimeras would have the general ADPR-[truncated TcdB T-domain]-DTR architecture. We used DT because of the well-established and facile readout of the ADPR (protein synthesis) and the robust binding of DTR to the ubiquitous HB-EGF receptor, present on many cell lines. Practically, the ADPR and DTR domains are amenable to greater levels of expression in *E. coli* over the GTD, CPD of LCTs, making a DT-TcdB chimera a more feasible platform to screen a large number of constructs.

We subjected the chimeras to two rounds of experimental testing, first testing their ability to form pores, and then, their ability to translocate. To probe pore formation, we pre-loaded liposomes with the quenched dye-pair HPTS/DPX, and monitored dye release (fluorescence) over time, with dye release as a surrogate for pore formation. To assess translocation, we evaluated intoxication (indirect measure of ADPR translocation into the cytosol) and protein synthesis inhibition (direct measure of ADPR translocation into the cytosol). All TcdB truncations formed pores in our dye release assay (Figure 4.7b, c); however only two truncations were able to facilitate translocation (i.e. 800-1500, 800-1473) (Figure 4.7d); constructs were considered non-toxic if unable to intoxicate cells at concentrations up to 50 nM, the highest concentration we could test all chimeras (Figure C.2). In line with the above data, we found that all TcdB T domain truncations starting at residue 851 formed pores (Figure 4.7f, g), while only 851-1500, 851-1473 were able to facilitate translocation (Figure 4.7h, (Figure C.2). Further truncation to 881 (i.e. 881-1473); however, abrogated translocation (Figure C.2). Taken together, these results indicate that residues 851-1473 comprise all of the components needed for pore-formation and translocation.

To further interrogate the evolutionarily conserved (and minimal TcdB) translocase, we also produced TcdB 851-1473 in a hybrid TcdB-DT system, with the GTD and CPD of TcdB and the receptor-binding domain of DT, such that the chimera was GTD-CPD-[TcdB(851-1473)]-DTR (Figure 4.8a, (Figure C.1). To evaluate translocation, we assessed cell rounding and Rac1 glucosylation (both direct measure of GTD translocation into the cytosol) and intoxication (indirect measure of GTD translocation into the cytosol). TcdB 851-1473 caused cells to round (Figure 4.8), glucosylated Rac1 (Figure 4.8c) and intoxicated cells (Figure 4.8d), while GTD-CPD-DTR lacking any of the TcdB T-domain did not cause cells to round and was non-toxic, reinforcing TcdB 851-1473 as a functional translocase region.

The evolutionarily conserved translocase retains its form and function

With the discovery of an evolutionarily conserved translocase shared among distant LCT-T homologs, and a minimal LCT translocation region, we aimed to further investigate the properties of TcdB 851-1473. We were able to recombinantly produce soluble TcdB 851-1473 to high levels of purity for characterization. TcdB 851-1473 exhibited pH-dependent pore formation, with maximal dye release at pH 4.0 and minimal dye release above pH 5.0 (Figure C.3). To further gauge the behaviour of TcdB 851-1473 as a function of pH, we assessed stability and conformational changes of TcdB 851-1473 in the presence and absence of a membrane mimetic, dodecylphosphocholine (DPC). At low pH (i.e. pH 4.0, 4.5, 5.0), TcdB 851-1473 rapidly aggregated out of solution (Figure C.3d), but remained in solution

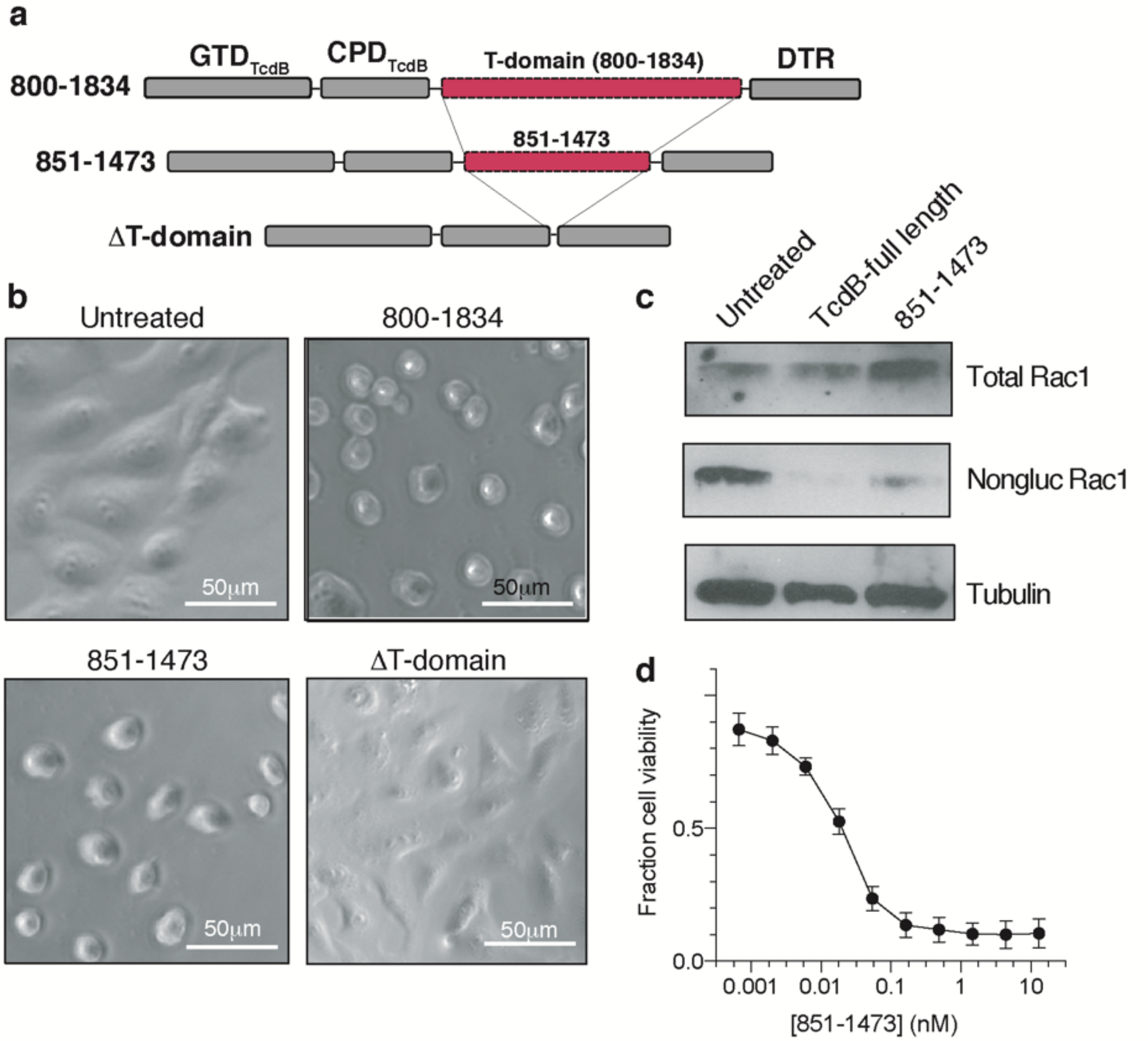


Figure 4.8: TcdB 851-1473 functions as a translocase in a hybrid TcdB-DT context. (a) Construct design of GTD-CPD-[TcdB(851-1473)]-DTR and GTD-CPD-DTR (i.e. ΔT-domain). (b) Light microscopy images of Vero cell rounding, assessed after 3 hrs. Representative images are shown for untreated cells, and for cells treated with TcdB, ΔT-domain and GTD-CPD-[TcdB(851-1473)]-DTR. All protein were added to cells at a final concentration of 1 nM. (c) Fraction cell viability of Vero cells incubated with GTD-CPD-[TcdB(851-1473)]-DTR, with 1.0 indicating 100% viable cells and 0 indicating 0% viable cells (N=4). (d) Western blot to detect total Rac1, non-glucosylated Rac1 and tubulin in untreated Vero cells, and for cells treated with TcdB or GTD-CPD-TcdB(851-1473)-DTR. All protein were added to cells at a final concentration of 1 nM.

and soluble in presence of DPC (Figure C.3e), suggesting that low pH results in exposure of hydrophobic surfaces of the translocase, which readily insert into the hydrophobic interior of the DPC micelle to prevent aggregation. By circular dichroism (CD) spectroscopy, at high pH (pH 7.5 to 6.0), TcdB 851-1473 had the characteristic spectra of a helical protein, with minima at 208 nm and 222 nm (Figure C.3f). Upon lowering of pH and in the presence of DPC, TcdB 851-1473 underwent structural changes, transforming into a protein with a mixture of both helical and β-sheet content, with a minimum at 218 nm (Figure C.3g). Taken together, our data indicated that TcdB 851-1473, and by extrapolation, the evolutionarily conserved region, when isolated is autonomous as it retained proper folding and function, and more specifically, pore formation and translocation activity.

4.1.4 Discussion and Conclusions

In this work, we conducted a targeted search to identify proteins that have homology to the T-domain of TcdB. TcdB is the best characterized member of the small LCT family, of which there was previously only 6 total members (TcdA, TcdB, TcsL, TcsH, TpeL, TcnA). Querying just the T-domain of TcdB, rather than the entire toxin, enabled identification of proteins with little or no homology in any of the other three domains, that would have otherwise gone undetected using other bioinformatic strategies. This approach facilitated the identification of hundreds of LCT-T containing proteins distributed widely throughout the bacterial kingdom, including in pathogenic species outside of clostridia. Conservation of hydropathy and key residues (clustered in conserved regions ii and iii) supports the function of these regions as pore-forming elements, suggesting that LCT-T homologs have highly similar membrane-inserted structures of LCTs, indicating that homologs could be utilized to gain insights into the translocation pore, and broadly, translocation.

There are very few studies of LCT translocation [356, 357, 376, 377], which may reflect the difficulty in studying the process. Homologs (especially for membrane proteins) have been used countless times to gain structural and functional insights of more relevant family members, helping to circumvent numerous experimental obstacles, and we hope that LCT-T homologs could be exploited for this function. In addition to functioning as an experimental proxy for LCTs, LCT-T homologs revealed strongly conserved residues that were not 100% conserved in LCTs (such as TcdB D1103), suggesting that some critically important residues may only be apparent through analysis of hundreds of distant homologs. In addition to important single residues, it is intriguing that, similar to BoNT and DT [3], a PxxG (more specifically, PxxGL) motif was identified as being strongly conserved in LCT-T homologs. We hope this study provides the platform to begin to interrogate these novel and uncharacterized translocation features.

One of our most striking findings was that of an evolutionarily conserved translocation region, which we show acts as a functional and minimal translocase in TcdB, and in isolation, retains low-pH mediated pore formation activity; notably, our studies indicated that pore formation alone is not sufficient for translocation, as many LCT regions could form pores but not translocate. On its own, the identified minimal conserved translocase may not be the most efficient or even abundant translocator among LCT-T homologs, but instead reflects the minimal necessary and sufficient requirements for translocation. Further, in the specific context of LCTs, although TcdB 851-1473 may not be as efficient as the entire T-domain to facilitate translocation, due to its smaller size (70 kDa vs. 100 kDa full length TcdB T-domain), favorable expression and conservation of pH-dependent pore formation and translocation activities, the protein is an ideal candidate for mechanistic studies of pore formation and translocation.

In comparing the T-domains of BoNT and DT, we found that the T-domains of AB toxins are predominantly associated with one type of effector: ADPR (DT), peptidases (BoNT) and GTD (LCT). We speculate that the association of a translocase family with an effector family may be indicative of translocase-effector co-evolution, implying that AB toxin T-domains are fine-tuned to translocate a particular effector type. This perhaps clarifies why three AB toxin translocase domains exist, instead of just one. Although each translocase may be best suited to translocate a particular effector type, LCT T-homologs seem to accommodate effectors of greater sequence diversity, length, and potentially type. We speculate that LCTs may in turn be more amendable as translocation scaffolds for protein cargo that is larger and more diverse in sequence. The potential permissiveness of LCT-T homologs may have direct applications in bacterial toxin drug delivery, where bacterial toxins are harnessed to deliver therapeutic cargo (namely proteins) into cells [360]. We think an interesting application of this work is to explore the potential of LCTs as a delivery platform, and we hope that this study motivates future research into this application.

With the identification of hundreds of new LCT-T homologs, our work raises the intriguing question of the ecological role(s) of LCT-T containing proteins. Bioinformatics is a powerful starting point for toxin identification, and provides clues into the putative toxin functionality of LCT-T homologs. Further guidelines are still required for definitive toxin identification, such as Falkow's molecular Koch's postulates [89], which require that toxins (or virulence factors) exist only in a pathogenic strain, with mutation or deletion of the virulence factor resulting in loss of pathogenicity. We hope our work provides a starting point and framework to further interrogate the function and ecological significance of these hundreds of new putative toxins.

Chapter 5

Conclusions

> [...] we have to remember that what we observe is not nature in itself but nature exposed to our method of questioning.
>
> *Physics and Philosophy: the Revolution in Modern Science*
> Werner Heisenberg

Research into bacterial toxins began nearly 150 years ago, just as microbiology had begun to develop as a discipline in its own right. Then and now, studies of toxigenic pathogens have spurred much research, primarily in medicine and epidemiology. Beyond the considerable impact toxin research has had on human health, toxins have proven useful for a wide variety of purposes. The discovery that botulinum neurotoxins cleave SNAREs and thereby inhibit neurotransmitter release established the role of SNAREs in exocytosis, and at the same time provided strong evidence in favour of the quantal release hypothesis [52, 378, 379, 380]. The fact that the adenylate cyclase toxin of *Bordetella pertussis* is only active when injected intracellularly has been exploited to discover effector proteins from pathogens like *Salmonella*, *Yersinia*, and *Xanthomonas* [54, 381, 382, 383]. Diphtheria toxin - the cause of mortality rates as high as 10% in diphtheria disease, which in some areas once afflicted as many as 1 in 20 children [384] - has now been re-engineered as an immunotoxin to kill cancer cells [385]. In a field that has already proven fruitful, the research and research applications made possible by genomics seem especially promising. In the following chapter, I briefly describe the contributions of the research presented in this thesis, discuss some of its limitations, and conclude with a prospective for the field.

5.1 Summary of major findings

In Chapter 2, several analyses the botulinum neurotoxin family were presented, focusing on the identification of novel variants from genomic data. Chapter 3 presents a similar genomic survey for diphtheria toxin, presenting the first sequences directly homologous to the toxin outside of *Corynebacterium*. Chapter 4 utilizes bioinformatic methods to hone in on the process of toxin translocation in large clostridial toxins. In each case, genome sequencing has enabled the discovery of novel and diverse toxin sequences that may have gone unnoticed through traditional toxin identification methods. The main differences between each chapter are that they highlight different aspects of toxin evolution. In botulinum neurotoxins, novel toxins and toxin gene cluster sequences appear to be much more widespread than previously thought, and further, genomically-identified neurotoxin homologs appear to possess functional novelties compared to currently described lineages. Similarly, genomically-identified diphtheria toxin homologs greatly expand its protein family and known taxonomic distribution, but also display significant structural conservation. In large clostridial toxins, the evolutionary analysis of distantly related sequences facilitates detailed functional analysis of the TcdB translocase domain. In short, it is possible to gain many insights into toxin families when they are examined from an evolutionary perspective.

5.1.1 Evolution of botulinum neurotoxins

The discovery of a protein related to botulinum neurotoxins (BoNTs) in *Enterococcus faecium* provides evidence of the first closely-related BoNT sequence outside of *Clostridium*, far closer than the homolog in *Weissella oryzae* (or, BoNT/Wo [101]) [99]. Moreover, it provides the first evidence of BoNT-related sequences in *Enterococcus*. The strain observed in this experiment was not associated with disease, and appeared to be most closely related to commensal strains [1], but nonetheless reveals the apparent ability of *Enterococcus* to acquire and potentially disseminate *bont* genes. Since enterococci continue to be a major source of hospital-acquired infection [156, 157], the acquisition of a *bont*-like gene is certainly noteworthy. Additionally, preliminary functional characterization results (see Zhang et al. [2]) suggest that BoNT/En is capable of inducing botulism-like paralysis making BoNT/En a potential health concern, although toxicity is only observed at unusually high concentrations for BoNTs. The closest phylogenetic relative of BoNT/En is

[1] The thesis includes only the bioinformatic portions of the published paper. For details on results and methods not included in this thesis, see the publication by Zhang et al. [2].

the recently-described BoNT/X [155]. As the two group together and separately from the rest of the BoNT family, it is possible that they represent a large, early-branching clade of BoNTs that has remained undiscovered. Considering there is no specific mechanism known for the transfer of a clostridial gene cluster in to *Enterococcus*, it is possible that BoNT/En was transferred from a different, unknown host bacterium.

A genomic screen for additional BoNT-related sequences revealed partial and complete BoNT homologs in a variety of other species, including bacteria and fungi. The identified sequences have many conserved BoNT characteristics, particularly in their metalloprotease domains. Surprisingly, BoNT-related sequences with metalloprotease domains provided a link between the BoNT protease and a family of type III secretion system (T3SS) metalloprotease effectors (NleD in *Escherichia coli*, and HopH1 in *Pseudomonas syringae*). The relationship between the two is distant, but BoNT, BoNT-like proteases, and the T3SS effectors clearly have large segments in common, including catalytic residues. Two BoNT-specific insertions clearly differentiate the two families, and given the extended interaction of BoNT with its substrate, it is likely that these regions play a role in determining substrate specificity. Analysis of the translocase regions from BoNT-related sequences revealed similarity both to BoNT translocases, fungi sequences of unknown function, and diphtheria toxin (DT) translocases. Among the BoNTs, residues that play a role in membrane insertion in DT are strongly conserved. This finding is remarkable because the roles of particular residues in BoNT translocation remain largely unknown; recently, a structure-based study of BoNT translocation provided strong evidence that these conserved residues do indeed participate in translocation [386].

A family of BoNT-related sequences was discovered in *Chryseobacterium piperi*. One *C. piperi* sequence, designated Cp1, was found to induce necrotic cell death in human cells using a metalloprotease-dependent mechanism. Its cellular target(s) remains unknown, since inducing expression of the Cp1 protease causes death too rapidly to yield samples suitable for proteomic analysis. However, since the canonical BoNT target proteins VAMP, SNAP25, and syntaxin were not cleaved by Cp1, its activity must be due to cleavage of some other host substrate(s). This observation underscores another evolutionary facet of the BoNT family. Although it is possible to rationally design a BoNT protease capable of cleaving targets beyond traditional BoNT substrates [387], it seems that Cp1 does so naturally. Thus, genomically-identified BoNT-related sequences demonstrate both sequence and functional diversity within the BoNT family. It is likely that genomic investigations will continue to uncover more novel BoNTs with novel functions [388].

Between Cp1 and BoNT/En, it seems that the BoNT family is not restricted to *Clostridium* species [389]. The distribution of OrfX proteins across distantly related bacteria suggests that the OrfX cluster is similarly not restricted to clostridia. The OrfX gene clusters with the greatest similarity to BoNT gene clusters are encoded by the various genomically-identified BoNT-related sequences, including the clusters containing BoNT/X in *C. botulinum* str. 111, BoNT/En in *E. faecium*, and the recently identified PMP1 in *Paraclostridium bifermentans* [260]. Outside of these neurotoxin-containing OrfX clusters, the OrfX clusters found in *Bacillus thuringiensis*, *Bacillus* sp. 2SH, *Brevibacillus*, and *Paenibacillus* are the most similar by analysis of sequence content and phylogenetics. Notably, *B. thuringiensis*, *Brevibacillus* spp., and *Paenibacillus* spp. are broadly associated with insect pathogenicity. Indeed, many OrfX gene clusters are found in the genomes of distantly related insect pathogens, and these OrfX clusters typically contain non-BoNT, toxin-related sequences.

The nature of the relationship between OrfX proteins and insect pathogenicity is unknown. In BoNTs, the analogous genes of hemagglutinin clusters are known to form toxin complexes with NTNH proteins. It has thus been widely assumed that OrfX proteins participate in a similar complex, although the evidence to support this is relatively sparse. It is possible that the lack of evidence for OrfX-BoNT complex formation is related to insect-specific functions or conditions, whereas most BoNT experiments have been performed with mammalian cells. Although this is speculative, what is clear based on comparative genomics is that the OrfX gene cluster appears to be a flexible component or scaffold of many different toxin clusters, and also that *orfX* genes undergo lateral gene transfer independently of *bont* genes. As such, while it cannot be ruled out that OrfX proteins do indeed function analogously to HA proteins by complexing with toxins and assisting uptake through the gut [149], based on the structural similarity of OrfX proteins to membrane permeability-increasing proteins [261] and the association of OrfX-related genes with diverse toxin-related genes, an additional possibility is that OrfX proteins may participate more generally in toxin release and export. The relatively broad taxonomic distribution of *orfX* genes and their association with a variety of host types means that significant characterization efforts will be required to address these outstanding questions about OrfX functionality.

5.1.2 Discovery and analysis of diphtheria toxin homologs

Homologs of diphtheria toxin (DT) are distributed broadly among the phylum Actinobacteria, with many diverse variants found in species of *Streptomyces*. In addition to significant sequence similarity (E-values «1×10^{-10}), several residues that are known to contribute to DT function are strongly conserved in DT homologs, notably including the catalytic E148 residue. Overall, the catalytic (C) and translocase (T) domains of DT homologs are more similar to DT than the receptor binding (R) domain. This might reflect stricter functional constraints in the C and T domains compared to the R domain [390, 391], but is also potentially related to the strong conservation seen among orthologs of DT's catalytic target (eEF-2) compared to its receptor (HB-EGF). The identification of diphtheria toxin-related sequences outside of *Corynebacterium* has further implications for the gene family. In *Corynebacterium*, the *tox* gene encoding diphtheria toxin is located within a corynephage [296, 297], implicating the phage as the major mechanism for lateral gene transfer. While this may be true within *Corynebacterium*, none of the DT-related sequences are predicted to reside within phages. In fact, only one of the *Streptomyces* genomes containing a DT-related sequence had evidence of any phage, and the phage locus did not contain the toxin gene. This suggests that the corynephages are not the sole mechanism for lateral gene transfer of *tox*-related genes. Furthermore, functional characterization of the DT homologs is necessary to determine whether or not these proteins are capable of DT-like functionality, although the lack of a known host suitable for *in vivo* experimentation complicates matters.

In addition to sequence similarity, the DT homologs from *Seinonella peptonophila* (SP) and *Streptomyces albireticuli* (SA) exhibit structures highly similar to DT. SP and SA are both larger than DT, and in general possess larger α-helices, but their overall topologies are similar to DT, and DT's domains are the best structural matches currently in the Protein Data Bank. It is worth noting that neither structure was solved with a bound NAD+ ligand. Since the SA and SP positions aligning to DT's NAD+-binding residues are not conserved, it is necessary in the future to verify their C domains are capable of coordinating an NAD+ ligand or an appropriate structural analog. As well, the R domains of SP and SA both contain extraneous loop regions within the binding interface of DT to HB-EGF. Further research is clearly needed to determine the effects of these structural differences on the proteins' functions, whatever those functions may be.

5.1.3 Conservation of the large clostridial toxin translocase domain

The large clostridial toxin (LCT) family currently consists of a set of six proteins found in several clostridia: TpeL in *Clostridium perfringens*, TcsH and TcsL in *Paeniclostridium sordellii*, TcnA in *Clostridium novyi*, and TcdA and TcdB in *Clostridioides difficile*. TcdA and TcdB are particularly important because of their role in hospital-acquired *C. difficile* infections. Between the two, TcdB appears to play a more important role, making the function of TcdA less clear [392, 393]. At any rate, LCTs can be considered functionally modular, containing regions that perform specific functions. In other toxin families, the functions of binding a target receptor, translocating into the cell, and intoxicating the host are essentially independent. The LCTs are unique in that the translocation domain seems to be intimately related to receptor binding, since the C-terminal region of the translocase appear to contribute to receptor binding, which involves binding multiple distinct receptors [63, 351, 352, 353, 64, 354, 355]. The task of differentiating between the functions of translocation and receptor binding in TcdB is not possible through comparison of the six canonical LCTs. However, by including more distantly related sequences within the LCT translocase family, it is possible to the narrow down the regions related to the basal function of translocation, common to all members of the family. Within this broader family, the translocase N-terminal region is much more strongly conserved than the C terminus. Indeed, these results suggest the conserved N-terminal region (residues 851-1473) encodes a minimal translocase capable of delivering cargo into cells.

Proteins containing an LCT-like translocase can be found in a wide variety of bacteria. Many species with LCT-like translocases are associated with pathogenicity or infections in a variety of hosts, and some proteins containing an LCT-like translocase are in fact toxins, including *Pseudomonas* FitD and *Photorhabdus* Mcf [375]. Most proteins with LCT-like translocases have not been functionally characterized, although many are annotated with glucosyltransferase domains. Considering the only characterized members of the LCT-like translocase family are toxins, and that the LCT translocase tends to associate with known effector domains, it seems likely that many of these sequences of unknown function are toxins. The potential effector regions associated with LCT-like translocases possess more sequence diversity than comparable toxin families, suggesting that the LCT translocase potentially permits more diverse effector cargo. If so, it is possible that the LCT translocase will be a useful scaffold for future immunotoxin therapeutics.

5.2 The limitations of bioinformatic predictions

All biological techniques require a set of assumptions to hold true in order for their conclusions to be valid, including any *in vivo*, *in vitro*, *in silico*, or statistical tests. While the approximations made by most methods simplify data interpretation, the reality is that their assumptions may not reflected in the data. Although some techniques are robust to deviation (for example, that Student's t-test can be considered valid for non-normal data given sufficient sample size [394, 395]), models can suffer a loss of predictive power or be vitiated entirely when their assumptions are violated [396, 397]. The question, then, is how severely these deviations impact the interpretation of results. In this section, I discuss some of the factors that confound the interpretation of the experiments presented here. For the sake of clarity, it is worth distinguishing between technical and biological sources of error, although the two are generally related.

5.2.1 Sources of error in bioinformatic analyses

Homology detection

Many experiments in bioinformatics begin by identifying a set of homologous sequences, which usually relies on sequence alignment. When these homologous sequences are derived from a large variety of distantly related organisms, the sequence identities of the resulting alignments are likely to be low. As the identity of the alignment decreases, it becomes increasingly difficult to distinguish statistically significant alignment scores (and by extension, homology) from random alignments, particular for alignments below 30% identity (the so-called twilight zone of homology) [110, 111]. The toxin alignments for the BoNTs, DTs, and LCT translocases and their respective homologs approach or exceed this cutoff. Further, not all toxin-related sequences have matches to families or domains in protein functional annotation databases, which might be related to the limitations of annotated sequence databases, or perhaps more fundamentally reflects that annotation methods are not equally applicable across protein families [398]. Another more specific error in detecting homologs is homologous over-extension by PSI-BLAST [399], which is used several times in the thesis. Homologous over-extension introduces two types of error; in the first, repeated search iterations produce progressively worse alignments and eventually introduce non-homologous sequences. In the second, repeated PSI-BLAST iterations produce alignments that extend past the boundaries of homologous regions, which degrades the quality of the position-specific scoring matrix [399].

The issues related to homologous sequence detection are unlikely to significantly affect the results in any of the three chapters for several reasons. First, the phenomenon of homologous over-extension was originally noted by repeated PSI-BLAST iterations with Pfam domains, becoming most apparent after 4 iterations [399]. The homologous sequence sets used here were retrieved after a maximum of three PSI-BLAST iterations, which has been deliberately selected as a conservative cutoff to decrease the chance of spurious alignments. Pfam domains, which enabled the detection of homologous over-extension through their well-defined domain boundaries, were used here to annotate the detected sequences separately from the PSI-BLAST result. Generally, the PSI-BLAST-detected sequences share at least one Pfam domain in common with the query toxin sequence, which supports a homologous relationship. These domain annotations are also capable of distinguishing between types of partial homologs seen for a given query; this differentiates, for example, BoNT-metalloprotease homologs from BoNT-translocase homologs. Domain annotations were therefore used to partition sets of sequences in order to limit analysis to homologous regions, which diminishes the second effect of homologous over-extension. Second, potentially homologous sequences were consistently subjected to further analysis by multiple sequence alignment (using several multiple alignment algorithms including MUSCLE [129], ClustalO [130], and MAFFT [131]), inferring phylogenies (using several phylogenetic methods including maximum likelihood [135] and Bayesian estimation [185]), and examination of conserved functional residues and motifs. Each of these lines of evidence increases or decreases support for a homologous relationship independently of the method used to detect them. Finally, the fundamental assumption about homology employed by statistical sequence alignment methods is that the simplest explanation for a statistically significant alignment score is homology. As stated by Pearson and Sierk [400], this approach makes no assumptions about alignments with scores below statistical significance, which may or may not be related. One can easily imagine a scenario where a particular sequence in a homologous family diverges such that alignment with any of the other family members produces a non-significant alignment score. If that circumstance does arise (regularly or rarely), methods for homology detection based on alignment scores actually represent a consistent *underestimation* of a homologous family's true diversity.

Homology is a property that must be inferred on the basis of evidence. For each of the BoNTs, DTs, and LCTs, the evidence strongly supports a homologous relationship between the toxin-like sequences presented here and the toxin sequences used to detect them. It is not possible to quantify homology, which is a binary characteristic - either a set of sequences are related or they are not. However, by themselves, alignment statistics like E-values do give a measure of the statistical significance of an alignment by comparing it to randomly sampled alignments. In all cases, the BoNT-, DT- and LCT-like sequences do produce statistically significant alignments (E-values «1×10^{-3}). The detected alignments span several hundred residues up to the length of a full-length toxin query, and the alignments are detected within databases containing hundreds of millions of sequences. In other words, the chance that these alignments are spurious is very low. Beyond that, the detected toxin-like sequences often have multiple annotated domains in common with their query toxin and preserve key motifs for toxin function. As an example, the exhaustive pairwise comparison between the BoNT/A1 protease with all protease sequences in MEROPS (at the time of the analysis, the database consisted of 1,103,662 sequences) [124] corroborated that the M91 peptidases are the BoNT peptidases' next closest family. This result was congruent with phylogenetic trees that distinguished BoNT, divergent BoNT homologs, and M91 peptidases, which was again consistent with features like domain annotations. To summarize, the analyses of BoNT-, DT-, and LCT-like sequences all provided positive evidence for homology, which went well beyond statistically significant alignment scores.

Genomes as a source of predicted proteins

Although there are now hundreds of thousands of sequenced bacterial genomes, decidedly few of them are complete, closed genomes. Based on statistics from the NCBI Genome database at the time of writing (plotted in Figure 1.2), the number of complete prokaryotic genomes is ∼2,600 out of ∼210,000 total, or slightly above 1%. Most bacterial genome sequences instead consist of a number of loosely connected contiguous sequences (contigs) or scaffolds, ranging from 1 to 2746 sequence fragments (with a mean of 1420 and standard deviation of 857.75). This discontiguity is partly a consequence of sequencing platforms that produce relatively short read lengths, which are then stitched together to create assembled genome sequences (using for example de Bruijn graphs [401]). Increasing sequencing depth can provide better coverage of a genome, but repetitive regions will continue to be a problem if read lengths are smaller than the length of a genomic repeat regardless of sequencing depth [402, 403]. Long-read sequencing platforms have the potential to address this - read lengths as long as 882kb were reported in a recent effort to sequence the human genome by nanopore sequencing [404] - but comparatively high per-base error rates create additional problems, such as errors in protein prediction [405]. As well, platform-independent biases will continue to be a problem, including general patterns in resolving G+C biases and homopolymer repeats [80, 406, 81]. As it currently stands, a hybrid approach that takes advantage of multiple sequencing platforms seems to be optimal, but continued improvements in long-read sequencing may yet rival the accuracy of short-read or hybrid approaches [407].

For the BoNT-related toxins in *E. faecium* and *C. piperi*, long-read sequencing was necessary to close gaps in short-read assemblies. In the case of *C. piperi*, in an initial draft genome assembly, homopolymer errors yielded apparent frameshift mutations in some of the *bont*-related genes that were only resolved by combining short- and long-read sequencing data with Pilon [182, 408]. The problem of fragmentary genome assemblies is increased in fungal genomes, which appear to contain partial homologs of BoNT and LCTs [409], since eukaryote genomes tend to be much larger and more repetitive than prokaryotes'. Another potential example can be seen in the *Austwickia chelonae* diphtheria homologs, where the C and T domains appear to be separated by a premature stop codon. A read pile-up generated by aligning the sequencing reads to the assembled genome did not reveal obvious homopolymer or similar common sequencing errors. However, Jiang et al. reported the genome sequence of another *A. chelonae* strain (LK16-18) [410] associated with serious skin infections in crocodile lizards [411] which appears to contain an intact copy of the DT-like gene (NCBI protein accession WP_116115734.1; NCBI assembly number GCF_003391095.1). Notably, the assembly reported by Jiang et al. used a combination of single-molecule real time (Pacific Biosciences) and solid-phase amplification (Illumina) sequencing. While it is possible that the originally-identified *Austwickia* DT-like gene did in fact contain a frameshift mutation, the identification of a strain containing an intact copy suggests this may have simply reflected a sequencing error. The LCTs pose a unique problem in that their C-terminal receptor-binding domains [355] contain oligopeptide repeats that may impede accurate genome assembly. In general, many BoNT, DT, and LCT homologs are encoded within short contigs, which decreases the informative value of genomic context analysis.

The accuracy of the toxin homolog sequences derived from genomes is limited by current sequencing technologies. Unfortunately, it is currently infeasible to re-sequence and re-annotate the thousands of genomes containing toxin-related sequences within the course of a doctoral program. Nonetheless, the recent evidence of an uninterrupted full-length DT homolog in *Austwickia*, as predicted in Chapter 3, is cause for cautious optimism. Assuming the sequenced DNA originated from a single clonal population, toxin-related genes predicted from fragmented and error-containing assemblies are more likely to be slightly improved over time (for example, by more accurately predicting gene sequences and improved contig scaffolding) than completely invalidated. In other words, the genes predicted from fragmented genome assemblies likely represent real encoded genes, notwithstanding sequencing and assembly errors. As sequencing technology continues to improve, the quality of genome assemblies will also improve, and the existence of toxin-related genes will be substantiated or refuted.

Phylogenetic inference

The objective of phylogenetic methods is to generate a tree describing the ancestry of a set of homologous sequences. Many factors affect the accuracy of phylogenetic inference, the first of which is alignment quality. The accuracy of tree topology decreases with decreasing alignment quality; unbalanced alignments and trees containing clades with long branch lengths have a pronounced effect [412]. Unbalanced multiple alignments can be caused by inadequate taxon sampling (which is addressed more thoroughly in section 5.2.2). Long branch artifacts were first noted as a problem for early phylogenetic methods [413] but have persisted among model-based methods [414, 415]. Long branch artifacts form spurious phylogenetic groupings because of homoplasy or degradation of the phylogenetic signal in deep-branching trees [416]. A problem more specific to Bayesian inference is the general inflation of support values compared to maximum likelihood bootstraps [417, 418].

The analyses presented here possess some of these problematic characteristics and other properties that negatively affect phylogenetic inference. As an example, I focus on phylogenetic issues associated with the analysis of BoNT sequences, although similar issues apply to the DT and LCT families. In the BoNT family, there are clearly identifiable recombination events that have produced chimeric or mosaic toxin sequences (serotypes C/D, D/C, and more recently H/A [419]). Ignoring the effects of recombination tends to produce inaccuracies, especially in determining rate heterogeneity between sites [420, 421]. Phylogenetic networks, rather than phylogenetic trees, can potentially alleviate this by demonstrating the support for multiple tree topologies at once (for example using SplitsTree [422]).

So far, the issue of recombination is restricted to canonical BoNT sequences that clearly reside within the BoNT family. A more pressing issue for the BoNT-like toxins (BoNT/En, BoNT/X, BoNT/Wo, and more distantly Cp1) is related to the long branches they appear to reside on and the inequality of sampling between serotypes. The eight canonical BoNT serotypes mostly contain more than one subtype, giving a measure of within-serotype sequence diversity while providing more balance to the input set. Between-serotype differences vary by up to 70% amino acid identity, thereby yielding deeply branched phylogenies because more substitutions must have occurred since their divergence. Together, uneven sampling and inherently long branch lengths make BoNT phylogenies more prone to long branch artifacts. The identification of additional sequence variants could conceivably help to expand poorly represented serotypes and decrease between-serotype distances. To that effect, a BoNT-related toxin denoted PMP1 was recently discovered in the genome of a strain of *Paraclostridium bifermentans* and appears to fall within the BoNT/En-BoNT/X clade [260]. However, it bears mentioning that some phylogenetic complexities cannot be remedied by the addition of more sequences [415, 423, 424, 425] or inclusion of additional sites [415], meaning that the phylogenetic issues within the BoNTs or any toxin family may not be easily resolved.

In spite of these phylogenetic limitations, several features of the phylogenies in preceding chapters are informative and reveal a few generalizable principles. In every toxin family examined, the branches associated with human-specific toxins form monophyletic groups. This is true for the BoNTs and the BoNT-like sequences, which can be clearly distinguished from NleD-like M91 peptidases, as well as the DT sequences from *Corynebacterium* species, which form a clear monophyletic lineage distinct from DT-related sequences from other members of the phylum Actinobacteria. In both of these cases, the toxin-related sequences from genomic data sets sit on long branches that aggregate at the base of the tree. For BoNTs and DTs, these long branch lengths correlate with an important functional property: the canonical BoNTs and DTs are associated with human disease, while the toxin-related sequences are not. The BoNT-related protein BoNT/En causes mild botulism-like paralysis in mice at high concentrations but is not associated with any human disease, and none of the early-diverging DT-related sequences have been implicated in disease. Therefore, the evolutionary changes that must have occurred between the toxin-related lineages and canonical toxin family members must be sufficient to explain the emergence of human specificity. Although the potential for long branch artifacts makes it difficult to ascertain their position on the tree, the toxin-related sequences appear to consistently form early-branching clades in the toxin families. Considering the conservation of toxin features in these early-branching lineages and their clear distinction from human-specific clades, it is tempting to speculate that most of these toxin-related sequences represent early-diverging toxin family members adapted for non-human hosts.

5.2.2 Biological factors affecting results

Inadequate and uneven sampling of pathogens, toxin sequences, and toxigenic microorganisms

Inadequate taxon sampling creates issues for phylogenetic inference, but it is only one aspect of a broader predicament. Pathogens and their toxins are chiefly discovered and described because of an association with disease, and human disease in particular. This produces a bias toward the kinds of microbes that are already known to produce diseases, while the types and strains of microbes that do not produce disease or display attenuated virulence go under-sampled or undetected. This anthropocentric bias is reflected in sequencing efforts and databases [2]. As a result, considerably less is known about obligate or opportunistic pathogens with non-human hosts. Toxin-related sequences (BoNT-like, DT-like, and LCT-like) are mostly found in bacteria only loosely associated with pathogenicity, particularly in non-human hosts. This includes the association of BoNTs and BoNT gene clusters with insects, DTs in a reptile skin pathogen [329, 410, 411] and environmental isolates of *Streptomyces* spp., and LCTs in insect pathogens and environmental isolates

[2] The NCBI genome database currently contains 532 *Clostridium* genome assemblies (including all sequenced species and strains of the genus) compared to 10,882 genome assemblies for *Salmonella enterica*. More broadly, the NCBI Sequence Run Archive (SRA) contains raw sequencing information for different kinds of sequencing experiments (including whole genome sequencing, microarray data, and many others), which provides a rough estimate of potential taxonomic biases in sequencing efforts. The SRA currently contains 11,432 samples matching the text query *"Clostridium"* versus 260,461 samples matching *"Salmonella enterica"*. By comparison, *"Austwickia chelonae"*, an organism not commonly implicated in human disease, has 3 genome assemblies and matches 3 samples in the SRA. These numbers are not intended to make any statement about the relative value of research in any organism - any research with the potential to alleviate the burden of suffering caused by the pathogens of humans, other animals, or plants, is valuable in the view of the author - but instead to highlight the limitations of currently available sequencing information.

of *Pseudomonas* species. These environments and organisms are simply not studied to the same level of depth or scrutiny as others more clearly associated with human disease. Many toxin-related sequences are clearly distinct from other members of their family (∼30% amino acid identity), but are represented by a single known sequence that is likely a part of a larger, undiscovered subfamily; ideally, sequencing efforts in these uncharacterized environments and the identification of additional organisms with toxin-related sequences will supply a more complete description of these subfamilies.

Most toxin-related sequences are derived from organisms represented by a single genome sequence. A single sequenced genome from a given strain of bacteria can provide an adequate representation of the species, or it might represent only a fraction of the genomic diversity present in the species [426]. In *Bacillus anthracis*, fewer than five genomes might provide an adequate description of its pangenome, whereas a similar level of pangenomic completeness for *Streptococcus agalactiae* might require hundreds of sequenced genomes [426]. On the other hand, the complete pangenomic description of a species is likely not feasible given that the processes of conjugation, transduction, transposition, and transformation generate new genotypes constantly, between closely-related organisms and by lateral gene transfer between distantly related taxa. The rates of lateral gene transfer and thereby "pangenomic evolution" remain mostly unknown, but are surely related to many properties, some of which are likely taxon-specific (for example, the rate of lateral transfer in to *Streptomyces* spp. appears to be relatively slow [427]). There is some evidence that the rate of lateral gene transfer is much higher than rates of sequence evolution [428]. Like rates of lateral gene transfer, an enormous number of factors affect substitution rates [429], and new alleles can also be generated through recombination, insertions, and deletions. Between changes at the level of the genome and changes at the level of particular genes, any given sequenced genome represents a snapshot of an organism rather than a complete description of its genome, which is, in all likelihood, a dynamic construction. This presents two contrary problems for genomes as a source of toxin sequences: first, that toxin genes can be a part of the pangenome but go undetected because of insufficient sampling, and second, that a correctly identified toxin-related sequence may be be a transient addition to a genome rather than a permanent acquisition. Although the issues outlined above will persist, sufficient genome sampling approaching pangenomic completeness would help to minimize the negative effects of these two problems.

Are genomically-derived toxin-related sequences actually toxins?

The relationship between sequence, structure, and function is complex. That the toxin-related sequences of BoNTs, DTs, and LCTs display significant sequence similarity has been thoroughly demonstrated. The toxin-related sequences also exhibit structural similarity, exemplified by the light chains of BoNT/X [430], BoNT/Wo [102], and the DT-related sequences of *Seinonella peptonophila* and *Streptomyces albireticuli*. However, determining the structure of a protein does not always clarify what functions they perform, as seen for OrfX proteins [261] or the structures of DT-related sequences reported here. Therefore, in spite of observed sequence and structural similarities, establishing the relationships between these features and function remains a challenge. Some limited examples of shared toxin functionality have been demonstrated here - as seen in the BoNT-like cleavage of SNARE proteins by BoNT/En [2] and potentially BoNT/Wo [101], or the conserved activity of the evolutionarily conserved minimal TcdB translocase - but the functions of toxin-related sequences are for the most part unknown.

Unlike traditional methods for toxin identification, where a known phenotype is caused by an unknown toxin, the goal for genomically-predicted toxin-related sequences is its inverse: identifying a phenotype associated with a toxin genotype. One major hindrance to estimating a toxin's phenotype is that it is intrinsically linked to host specificity. Attempting to associate a toxin-related sequence with a phenotype is analogous to attempting to find a lock after being given a key when one does not even know which door to check. Many toxin-related sequences are found in the genomes of organisms lacking any appreciable description of their ecological context, and thus their interactions with different hosts are unknown. The fact that most toxin-related sequences are not associated with significant diseases in humans or other organisms therefore has many plausible explanations. The toxin-related sequence might produce a significant and obvious disease phenotype in a narrow host range that has yet to be examined; or, the toxin acts in a way to explicitly attenuate virulence, as may be the case for certhrax [97, 98] and typhoid toxin [92]; or, the proteins could have some other function. It is not trivial to distinguish between these possibilities and the range of possible hosts for a given toxin sequence is vast.

Assuming one knows the host target of toxin-related sequence, an important aspect of further characterization is the development of *in vivo* models. While *in silico* methods can help to identify and characterize toxin sequences and *in vitro* approaches can be used to evaluate their functions and mechanisms, many questions about toxin biology can only be answered in through *in vivo* experimentation. The use of *in vivo* models made it possible to establish anthrax toxin as the cause of death in systemic anthrax infections [37, 38]. De's isolation of cholera toxin [39] was made possible by his earlier development of an *ex vivo* rabbit ileal loop model [431]. These key insights into toxin biology were only possible by examining toxin function in the context of living host organisms. As an example from this work, it would be valuable to determine the contribution of the *Austwickia chelonae* DT-like sequence to virulence in crocodile lizards, assuming the DT-like sequence is functional

[410, 411]. More generally, appropriate *in vivo* models will help to assess the questions of how and when toxin-related genes are expressed, as well as how toxin-related genes influence their microbes' lifestyle during and apart from host association. *In vivo* models are capable of contributing a greater understanding of toxin biology from the perspective of the toxin producer as well as the affected host.

In order to understand what roles toxin-related sequences have, it is necessary to gain a greater understanding of the ecology of the microbes that encode them. This requires an understanding of the microbes' life cycle, how the microbes interact with different types of hosts, and how the presence or absence of toxin genes affects these properties. It is important to consider also the diversity of the microbe, including pathogenic and non-pathogenic varieties. Pathogenic and non-pathogenic strains of related organisms are still capable of interaction and lateral gene transfer, even though pathogenic varieties are more likely to have received attention [432]. A comprehensive understanding of a pathogen's diversity must also include varieties that are resistant to culturing methods, which for some taxa may represent a considerable fraction (as an example, the spore-forming members of the phylum Firmicutes are consistently undersampled in metagenomic data sets [433, 434]). These factors and others contribute to an underestimation of pathogen diversity, which necessarily results in gaps of knowledge about even basic aspects of pathogen biology. It is perhaps counterintuitive, but an improved understanding of pathogens and their toxins in part requires an improved understanding of their non-pathogenic and non-toxic relatives.

5.3 Final remarks

In spite of the aforementioned limitations, a few key conclusions can be gleaned from this thesis. The most generalizable principle can be summarized in Liebniz's quote that "nature does not make leaps" [435] - an idea popularized by Linnaeus [436] and invoked in Heisenberg's essay that introduced this chapter [437]. In other words, all families of bacterial toxins are the product of a continuous evolutionary process, and the bacterial toxins that afflict humans can be understood as branches within broader toxin families. Examining toxins from a broad evolutionary perspective allows for the discovery of novel toxins with unique properties, improves our understanding of the evolutionary shifts that yield highly potent and host-specific toxins, and is able to grant a deeper understanding of toxin functionality. It is a testament to the predictive power of evolution that this knowledge can be gained from naturally noisy and incomplete biological data, consistent with the dictum that nothing in biology makes sense except in the light of evolution [438].

This thesis presents the first relatives of historically significant and medically relevant bacterial protein toxins. The toxin families of clostridial neurotoxins and diphtheria toxins, initially described more than a century ago [8, 29, 27, 33, 34, 35, 36], can now be contextualized within their gene families using data derived directly from whole bacterial genomes. Genomic data also yields greater insights into the behaviours of toxins affiliated with more modern afflictions, including those related to large clostridial toxins. Although it is difficult to ascertain the functions of toxin-related sequences, the evidence strongly suggests that many of them are themselves toxins. In fact, the conservation of toxin features, their early-diverging branches in their respective phylogenetic trees, and their lack of association with human disease together imply that these sequences represent toxins adapted for non-human hosts, where they may elicit disease pathologies similar to those found in humans.

Many toxin-related sequences are highly dissimilar from other members of their families, suggesting they represent larger, uncharacterized subfamilies. As such, continued efforts to study the diversity of microbes and their hosts through genomics and metagenomics will deliver valuable insights into the diversity of pathogens and toxin sequences [388]. This diversity is paramount to understanding evolutionary dynamics within toxin families, but also reaps additional benefits. The identification of novel toxin sequences with unique functions has the potential to develop novel and improved toxin-based therapeutics. Furthermore, the improved understanding of toxin function facilitated by evolutionary studies could contribute valuable insights to immunotoxin development or the design of inhibitors, vaccines, and antibodies. In doing so, it is possible to harness the agents of disease and affliction and direct them toward more humane goals.

Copyright Permissions

The published articles included in this thesis are made with the permission of all publishers. At the time of writing (August 2019), the policies of the publishers MDPI, Nature Publishing Group, and Oxford University Press do not require additional permissions for the reproduction of articles in theses. The relevant policies can be found at the following links:

1. MDPI

   https://www.mdpi.com/authors/rights

2. Nature Publishing Group

   https://www.nature.com/nature-research/reprints-and-permissions/permissions-requests

3. Oxford University Press

   https://global.oup.com/academic/rights/permissions/autperm/?lang=en&cc=us

The inclusion of the following article is reproduced with the permission of the publisher John Wiley and Sons, published in FEBS Letters as a part of the Wiley Online Library.

1. Mansfield, M. J., Sugiman-Marangos, S. N., Melnyk, R. A., & Doxey, A. C. (2018). Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus. FEBS letters, 592(16), 2693-2705. [5].

   https://doi.org/10.1002/1873-3468.13208

A copy of the RightsLink ® agreement is included below.

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Oct 01, 2019

This Agreement between Michael Mansfield ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number: 4640450189588

License date: Aug 01, 2019

Licensed Content Publisher: John Wiley and Sons

Licensed Content Publication: FEBS Letters

Licensed Content Title: Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus

Licensed Content Author: Michael J. Mansfield, Seiji N. Sugiman-Marangos, Roman A. Melnyk, et al

Licensed Content Date: Aug 16, 2018

Licensed Content Volume: 592

Licensed Content Issue: 16

Licensed Content Pages: 13

Type of use: Dissertation/Thesis

Requestor type: Author of this Wiley article

Format: Print and electronic

Portion: Full article

Will you be translating? No

Title of your thesis / dissertation: Bioinformatic insights into the diversity and evolutionary ancestry of bacterial toxins

Expected completion date: Nov 2019

Michael Mansfield
200 University Ave. N
Waterloo, ON N2L3G1
Canada
Expected size (number of pages) Requestor Location: Attn: Michael Mansfield

Publisher Tax ID EU826007151

Total 0.00 CAD

References

[1] A. C. Doxey, M. J. Mansfield, and C. Montecucco, “Discovery of novel bacterial toxins by genomics and computational biology,” *Toxicon*, vol. 147, pp. 2–12, 2018.

[2] S. Zhang, F. Lebreton, M. J. Mansfield, S.-I. Miyashita, J. Zhang, J. A. Schwartzman, L. Tao, G. Masuyer, M. Martínez-Carranza, P. Stenmark, *et al.*, “Identification of a botulinum neurotoxin-like toxin in a commensal strain of *Enterococcus faecium*,” *Cell host & microbe*, vol. 23, no. 2, pp. 169–176, 2018.

[3] M. J. Mansfield, T. G. Wentz, S. Zhang, E. J. Lee, M. Dong, S. K. Sharma, and A. C. Doxey, “Bioinformatic discovery of a toxin family in *Chryseobacterium piperi* with sequence similarity to botulinum neurotoxins,” *Scientific reports*, vol. 9, 2019.

[4] M. J. Mansfield and A. C. Doxey, “Genomic insights into the evolution and ecology of botulinum neurotoxins,” *Pathogens and disease*, vol. 76, no. 4, p. fty040, 2018.

[5] M. J. Mansfield, S. N. Sugiman-Marangos, R. A. Melnyk, and A. C. Doxey, “Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus,” *FEBS letters*, vol. 592, no. 16, pp. 2693–2705, 2018.

[6] A. Casadevall and L.-A. Pirofski, “What is a pathogen?,” *Annals of medicine*, vol. 34, no. 1, pp. 2–4, 2002.

[7] R. Collier, “Understanding the mode of action of diphtheria toxin: a perspective on progress during the 20th century,” *Toxicon*, vol. 39, no. 11, pp. 1793–1803, 2001.

[8] F. Löffler, *Untersuchungen über die Bedeutung der Mikroorganismen für die Entstehung der Diphtherie beim Menschen, bei der Taube und beim Kalbe.* 1884.

[9] R. Koch, “Die Aetiologie der Tuberkulose,” 1884. Reprinted in 2010 by the Robert Koch-Institut.

[10] A. Sakula, “Robert Koch: centenary of the discovery of the tubercle bacillus, 1882.,” *Thorax*, vol. 37, no. 4, pp. 246–251, 1982.

[11] L.-a. Pirofski and A. Casadevall, “Q&A: What is a pathogen? A question that begs the point,” *BMC biology*, vol. 10, no. 1, p. 6, 2012.

[12] P.-O. Méthot and S. Alizon, “What is a pathogen? Toward a process view of host-parasite interactions,” *Virulence*, vol. 5, no. 8, pp. 775–785, 2014.

[13] A. Taylor-Brown, L. Vaughan, G. Greub, P. Timms, and A. Polkinghorne, "Twenty years of research into *Chlamydia*-like organisms: a revolution in our understanding of the biology and pathogenicity of members of the phylum Chlamydiae," *Pathogens and disease*, vol. 73, no. 1, pp. 1–15, 2015.

[14] O. Lukjancenko, T. M. Wassenaar, and D. W. Ussery, "Comparison of 61 sequenced *Escherichia coli* genomes," *Microbial ecology*, vol. 60, no. 4, pp. 708–720, 2010.

[15] M. Otto, "*Staphylococcus epidermidis*—the 'accidental' pathogen," *Nature reviews microbiology*, vol. 7, no. 8, p. 555, 2009.

[16] G. D. Ehrlich, N. L. Hiller, and F. Z. Hu, "What makes pathogens pathogenic," *Genome biology*, vol. 9, no. 6, p. 225, 2008.

[17] A. Casadevall and L.-a. Pirofski, "Host-pathogen interactions: the attributes of virulence," *The Journal of infectious diseases*, vol. 184, no. 3, pp. 337–344, 2001.

[18] T. M. Wassenaar and W. Gaastra, "Bacterial virulence: can we draw the line?," *FEMS microbiology letters*, vol. 201, no. 1, pp. 1–7, 2001.

[19] A. A. Weiss and E. L. Hewlett, "Virulence factors of *Bordetella pertussis*," *Annual Reviews in Microbiology*, vol. 40, no. 1, pp. 661–686, 1986.

[20] S. K. Hoiseth and B. Stocker, "Aromatic-dependent *Salmonella typhimurium* are non-virulent and effective as live vaccines," *Nature*, vol. 291, no. 5812, p. 238, 1981.

[21] R. J. Roantree, "*Salmonella* O antigens and virulence," *Annual Reviews in Microbiology*, vol. 21, no. 1, pp. 443–466, 1967.

[22] B. Liu, D. Zheng, Q. Jin, L. Chen, and J. Yang, "VFDB 2019: a comparative pathogenomic platform with an interactive web interface," *Nucleic acids research*, vol. 47, no. D1, pp. D687–D692, 2018.

[23] D. L. Milton, R. O'Toole, P. Horstedt, and H. Wolf-Watz, "Flagellin A is essential for the virulence of *Vibrio anguillarum.*," *Journal of bacteriology*, vol. 178, no. 5, pp. 1310–1319, 1996.

[24] L. J. Pettit, H. P. Browne, L. Yu, W. K. Smits, R. P. Fagan, L. Barquist, M. J. Martin, D. Goulding, S. H. Duncan, H. J. Flint, *et al.*, "Functional genomics reveals that *Clostridium difficile* Spo0A coordinates sporulation, virulence and metabolism," *BMC genomics*, vol. 15, no. 1, p. 160, 2014.

[25] S. M. Hingley-Wilson, V. K. Sambandamurthy, and W. R. Jacobs Jr, "Survival perspectives from the world's most successful pathogen, *Mycobacterium tuberculosis*," *Nature immunology*, vol. 4, no. 10, p. 949, 2003.

[26] M. R. Wessels, C. E. Rubens, V.-J. Benedi, and D. L. Kasper, "Definition of a bacterial virulence factor: sialylation of the group B streptococcal capsule," *Proceedings of the National Academy of Sciences*, vol. 86, no. 22, pp. 8983–8987, 1989.

[27] L. Brieger, "Zur Kenntniss der Aetiologie des Wundstarrkrampfes nebst Bemerkungen über das Cholerароth," *DMW-Deutsche Medizinische Wochenschrift*, vol. 13, no. 15, pp. 303–305, 1887.

[28] L. Brieger, "Weitere Erfahrungen über Bakteriengifte," *Medical Microbiology and Immunology*, vol. 19, no. 1, pp. 101–112, 1895.

[29] É. Roux and A. Yersin, *Contribution à l'étude de la diphtérie.* Inst. Pasteur, 1889.

[30] A. M. Pappenheimer *et al.*, "Studies in Diphtheria Toxin Production. I: The Effect of Iron and Copper," *British journal of experimental pathology*, vol. 17, no. 5, p. 335, 1936.

[31] J. Murphy, J. Skiver, and G. McBride, "Isolation and partial characterization of a corynebacteriophage beta, tox operator constitutive-like mutant lysogen of *Corynebacterium diphtheriae.*," *Journal of virology*, vol. 18, no. 1, pp. 235–244, 1976.

[32] A. Pappenheimer Jr, "Diphtheria toxin," *Annual review of biochemistry*, vol. 46, no. 1, pp. 69–94, 1977.

[33] K. Faber, "Die pathogenie des tetanus," *Berl klin Wochenschr*, vol. 27, pp. 717–720, 1890.

[34] G. Tizzoni and G. Cattani, "XXIII. Untersuchungen über das Tetanusgift," *Naunyn-Schmiedeberg's Archives of Pharmacology*, vol. 27, no. 6, pp. 432–450, 1890.

[35] S. Kitasato, "Experimentelle Untersuchungen über das Tetanusgift," *Medical microbiology and immunology*, vol. 10, no. 1, pp. 267–305, 1891.

[36] E. v. Ermengem, "Über einen neuen anaeroben *Bacillus* und seine Beziehungen zum Botulismus," *Medical Microbiology and Immunology*, vol. 26, no. 1, pp. 1–56, 1897.

[37] H. Smith and J. Keppie, "Observations on experimental anthrax: demonstration of a specific lethal factor produced *in vivo* by *Bacillus anthracis*," *Nature*, vol. 173, no. 4410, p. 869, 1954.

[38] J. Keppie, H. Smith, and P. W. Harris-Smith, "The chemical basis of the virulence of *Bacillus anthracis*. III: The role of the terminal bacteraemia in death of guinea-pigs from anthrax," *British journal of experimental pathology*, vol. 36, no. 3, p. 315, 1955.

[39] S. N. De, "Enterotoxicity of bacteria-free culture-filtrate of *Vibrio cholerae*," *Nature*, vol. 183, no. 4674, p. 1533, 1959.

[40] T. Katada and M. Ui, "ADP ribosylation of the specific membrane protein of C6 cells by islet-activating protein associated with modification of adenylate cyclase activity.," *Journal of Biological chemistry*, vol. 257, no. 12, pp. 7210–7216, 1982.

[41] F. Okajima and M. Ui, "ADP-ribosylation of the specific membrane protein by islet-activating protein, pertussis toxin, associated with inhibition of a chemotactic peptide-induced arachidonate release in neutrophils. A possible role of the toxin substrate in Ca2+-mobilizing biosignaling.," *Journal of Biological Chemistry*, vol. 259, no. 22, pp. 13863–13871, 1984.

[42] G. Menestrina, G. Schiavo, and C. Montecucco, "Molecular mechanisms of action of bacterial protein toxins," *Molecular aspects of medicine*, vol. 15, no. 2, pp. 79–193, 1994.

[43] G. Schiavo and F. G. van der Goot, "The bacterial toxin toolkit," *Nature Reviews Molecular Cell Biology*, vol. 2, no. 7, p. 530, 2001.

[44] N. Goessweiner-Mohr, K. Arends, W. Keller, and E. Grohmann, "Conjugative type IV secretion systems in Gram-positive bacteria," *Plasmid*, vol. 70, no. 3, pp. 289–302, 2013.

[45] J. C. Madden, N. Ruiz, and M. Caparon, "Cytolysin-mediated translocation (CMT): a functional equivalent of type III secretion in gram-positive bacteria," *Cell*, vol. 104, no. 1, pp. 143–152, 2001.

[46] C. M. Collazo and J. E. Galán, "The invasion-associated type-III protein secretion system in *Salmonella* - a review," *Gene*, vol. 192, no. 1, pp. 51–59, 1997.

[47] S. Cunnac, M. Lindeberg, and A. Collmer, "*Pseudomonas syringae* type III secretion system effectors: repertoires in search of functions," *Current opinion in microbiology*, vol. 12, no. 1, pp. 53–60, 2009.

[48] K. Sekiya, M. Ohishi, T. Ogino, K. Tamano, C. Sasakawa, and A. Abe, "Supermolecular structure of the enteropathogenic *Escherichia coli* type III secretion system and its direct interaction with the EspA-sheath-like structure," *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11638–11643, 2001.

[49] R.-G. Zhang, D. L. Scott, M. L. Westbrook, S. Nance, B. D. Spangler, G. G. Shipley, and E. M. Westbrook, "The three-dimensional crystal structure of cholera toxin," *Journal of molecular biology*, vol. 251, no. 4, pp. 563–573, 1995.

[50] P. E. Stein, A. Boodhoo, G. D. Armstrong, S. A. Cockle, M. H. Klein, and R. J. Read, "The crystal structure of pertussis toxin," *Structure*, vol. 2, no. 1, pp. 45–57, 1994.

[51] N. C. Simon, K. Aktories, and J. T. Barbieri, "Novel bacterial ADP-ribosylating toxins: structure and function," *Nature reviews Microbiology*, vol. 12, no. 9, p. 599, 2014.

[52] G. G. Schiavo, F. Benfenati, B. Poulain, O. Rossetto, P. P. de Laureto, B. R. DasGupta, and C. Montecucco, "Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin," *Nature*, vol. 359, no. 6398, p. 832, 1992.

[53] N. S. Duesbery, C. P. Webb, S. H. Leppla, V. M. Gordon, K. R. Klimpel, T. D. Copeland, N. G. Ahn, M. K. Oskarsson, K. Fukasawa, K. D. Paull, *et al.*, "Proteolytic inactivation of MAP-kinase-kinase by anthrax lethal factor," *Science*, vol. 280, no. 5364, pp. 734–737, 1998.

[54] N. H. Carbonetti, "Pertussis toxin and adenylate cyclase toxin: key virulence factors of *Bordetella pertussis* and cell biology tools," *Future microbiology*, vol. 5, no. 3, pp. 455–469, 2010.

[55] S. H. Leppla, "Anthrax toxin edema factor: a bacterial adenylate cyclase that increases cyclic AMP concentrations of eukaryotic cells," *Proceedings of the National Academy of Sciences*, vol. 79, no. 10, pp. 3162–3166, 1982.

[56] F. Hofmann, C. Busch, U. Prepens, I. Just, and K. Aktories, "Localization of the glucosyltransferase activity of *Clostridium difficile* toxin B to the N-terminal part of the holotoxin," *Journal of Biological Chemistry*, vol. 272, no. 17, pp. 11074–11078, 1997.

[57] T. Frisan, X. Cortes-Bratti, and M. Thelestam, "Cytolethal distending toxins and activation of DNA damage-dependent checkpoint responses," *International journal of medical microbiology*, vol. 291, no. 6-7, pp. 495–499, 2001.

[58] M. Yamaizumi, E. Mekada, T. Uchida, and Y. Okada, "One molecule of diphtheria toxin fragment A introduced into a cell can kill the cell," *Cell*, vol. 15, no. 1, pp. 245–250, 1978.

[59] S. Carle, M. Pirazzini, O. Rossetto, H. Barth, and C. Montecucco, "High conservation of tetanus and botulinum neurotoxins cleavage sites on human SNARE proteins suggests that these pathogens exerted little or no evolutionary pressure on humans," *Toxins*, vol. 9, no. 12, p. 404, 2017.

[60] C. Montecucco and G. Schiavo, "Structure and function of tetanus and botulinum neurotoxins," *Quarterly reviews of biophysics*, vol. 28, no. 4, pp. 423–472, 1995.

[61] K. Bercsenyi, F. Giribaldi, and G. Schiavo, "The elusive compass of clostridial neurotoxins: deciding when and where to go?," in *Botulinum Neurotoxins*, pp. 91–113, Springer, 2012.

[62] C. Montecucco, "How do tetanus and botulinum toxins bind to neuronal membranes?," *Trends in biochemical sciences*, vol. 11, no. 8, pp. 314–317, 1986.

[63] L. Tao, J. Zhang, P. Meraner, A. Tovaglieri, X. Wu, R. Gerhard, X. Zhang, W. B. Stallcup, J. Miao, X. He, *et al.*, "Frizzled proteins are colonic epithelial receptors for *C. difficile* toxin B," *Nature*, vol. 538, no. 7625, p. 350, 2016.

[64] L. Tao, S. Tian, J. Zhang, Z. Liu, L. Robinson-McCarthy, S.-I. Miyashita, D. T. Breault, R. Gerhard, S. Oottamasathien, S. P. Whelan, *et al.*, "Sulfated glycosaminoglycans and low-density lipoprotein receptor contribute to *Clostridium difficile* toxin A entry into cells," *Nature microbiology*, p. 1, 2019.

[65] J. A. Young and R. J. Collier, "Anthrax toxin: receptor binding, internalization, pore formation, and translocation," *Annu. Rev. Biochem.*, vol. 76, pp. 243–265, 2007.

[66] P. E. Stein, A. Boodhoo, G. D. Armstrong, L. D. Heerze, S. A. Cockle, M. H. Klein, and R. J. Read, "Structure of a pertussis toxin–sugar complex as a model for receptor binding," *Nature structural biology*, vol. 1, no. 9, p. 591, 1994.

[67] E. Fan, E. A. Merritt, C. L. Verlinde, and W. G. Hol, "AB_5 toxins: structures and inhibitor design," *Current opinion in structural biology*, vol. 10, no. 6, pp. 680–686, 2000.

[68] T.-i. Nishiki, Y. Kamata, Y. Nemoto, A. Omori, T. Ito, M. Takahashi, and S. Kozaki, "Identification of protein receptor for *Clostridium botulinum* type B neurotoxin in rat brain synaptosomes.," *Journal of Biological Chemistry*, vol. 269, no. 14, pp. 10498–10503, 1994.

[69] M. Dong, H. Liu, W. H. Tepp, E. A. Johnson, R. Janz, and E. R. Chapman, "Glycosylated SV2A and SV2B mediate the entry of botulinum neurotoxin E into neurons," *Molecular biology of the cell*, vol. 19, no. 12, pp. 5226–5237, 2008.

[70] M. Dong, F. Yeh, W. H. Tepp, C. Dean, E. A. Johnson, R. Janz, and E. R. Chapman, "SV2 is the protein receptor for botulinum neurotoxin A," *Science*, vol. 312, no. 5773, pp. 592–596, 2006.

[71] L. Abrami, S. Liu, P. Cosson, S. H. Leppla, and F. G. van der Goot, "Anthrax toxin triggers endocytosis of its receptor via a lipid raft–mediated clathrin-dependent process," *J Cell Biol*, vol. 160, no. 3, pp. 321–328, 2003.

[72] P. Papatheodorou, C. Zamboglou, S. Genisyuerek, G. Guttenberg, and K. Aktories, "Clostridial glucosylating toxins enter cells via clathrin-mediated endocytosis," *PloS one*, vol. 5, no. 5, p. e10673, 2010.

[73] F. Antonucci, C. Rossi, L. Gianfranceschi, O. Rossetto, and M. Caleo, "Long-distance retrograde effects of botulinum neurotoxin A," *Journal of Neuroscience*, vol. 28, no. 14, pp. 3689–3696, 2008.

[74] K. L. Thoren and B. A. Krantz, "The unfolding story of anthrax toxin translocation," *Molecular microbiology*, vol. 80, no. 3, pp. 588–595, 2011.

[75] D. H. Hoch, M. Romero-Mira, B. E. Ehrlich, A. Finkelstein, B. R. DasGupta, and L. L. Simpson, "Channels formed by botulinum, tetanus, and diphtheria toxins in planar lipid bilayers: relevance to translocation of proteins across membranes," *Proceedings of the National Academy of Sciences*, vol. 82, no. 6, pp. 1692–1696, 1985.

[76] R. Ratts, H. Zeng, E. A. Berg, C. Blue, M. E. McComb, C. E. Costello, J. R. Murphy, *et al.*, "The cytosolic entry of diphtheria toxin catalytic domain requires a host cell cytosolic translocation factor complex," *The Journal of cell biology*, vol. 160, no. 7, pp. 1139–1150, 2003.

[77] M. Pirazzini, G. Zanetti, A. Megighian, M. Scorzeto, S. Fillo, C. C. Shone, T. Binz, O. Rossetto, F. Lista, and C. Montecucco, "Thioredoxin and its reductase are present on synaptic vesicles, and their inhibition prevents the paralysis induced by botulinum neurotoxins," *Cell reports*, vol. 8, no. 6, pp. 1870–1878, 2014.

[78] M. Tsuneoka, K. Nakayama, K. Hatsuzawa, M. Komada, N. Kitamura, and E. Mekada, "Evidence for involvement of furin in cleavage and activation of diphtheria toxin.," *Journal of Biological Chemistry*, vol. 268, no. 35, pp. 26461–26465, 1993.

[79] M. Egerer, T. Giesemann, T. Jank, K. J. F. Satchell, and K. Aktories, "Auto-catalytic cleavage of *Clostridium difficile* toxins A and B depends on cysteine protease activity," *Journal of Biological Chemistry*, vol. 282, no. 35, pp. 25314–25321, 2007.

[80] M. L. Metzker, "Sequencing technologies—the next generation," *Nature reviews genetics*, vol. 11, no. 1, p. 31, 2010.

[81] M. Kchouk, J.-F. Gibrat, and M. Elloumi, "Generations of sequencing technologies: From first to next generation," *Biology and Medicine*, vol. 9, no. 3, 2017.

[82] J.-F. Tomb, O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, *et al.*, "The complete genome sequence of the gastric pathogen *Helicobacter pylori*," *Nature*, vol. 388, no. 6642, p. 539, 1997.

[83] C. K. Stover, X. Q. Pham, A. Erwin, S. Mizoguchi, P. Warrener, M. Hickey, F. Brinkman, W. Hufnagle, D. Kowalik, M. Lagrou, *et al.*, "Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen," *Nature*, vol. 406, no. 6799, p. 959, 2000.

[84] J. F. Heidelberg, J. A. Eisen, W. C. Nelson, R. A. Clayton, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, L. Umayam, *et al.*, "DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*," *Nature*, vol. 406, no. 6795, p. 477, 2000.

[85] N. J. Loman and M. J. Pallen, "Twenty years of bacterial genome sequencing," *Nature Reviews Microbiology*, vol. 13, no. 12, p. 787, 2015.

[86] K. Weedmark, D. Lambert, P. Mabon, K. Hayden, C. Urfano, D. Leclair, G. Van Domselaar, J. Austin, and C. Corbett, "Two novel toxin variants revealed by whole-genome sequencing of 175 *Clostridium botulinum* type E strains," *Appl. Environ. Microbiol.*, vol. 80, no. 20, pp. 6334–6345, 2014.

[87] K. Weedmark, P. Mabon, K. Hayden, D. Lambert, G. Van Domselaar, J. Austin, and C. Corbett, "*Clostridium botulinum* group II isolate phylogenomic profiling using whole-genome sequence data," *Appl. Environ. Microbiol.*, vol. 81, no. 17, pp. 5938–5948, 2015.

[88] C. H. Williamson, J. W. Sahl, T. J. Smith, G. Xie, B. T. Foley, L. A. Smith, R. A. Fernández, M. Lindström, H. Korkeala, P. Keim, *et al.*, "Comparative genomic analyses reveal broad diversity in botulinum-toxin-producing Clostridia," *BMC genomics*, vol. 17, no. 1, p. 180, 2016.

[89] S. Falkow, "Molecular Koch's postulates applied to microbial pathogenicity," *Reviews of infectious diseases*, pp. S274–S276, 1988.

[90] S. Falkow, “Molecular Koch’s postulates applied to bacterial pathogenicity—a personal recollection 15 years later,” *Nature Reviews Microbiology*, vol. 2, no. 1, p. 67, 2004.

[91] J. Song, X. Gao, and J. E. Galán, “Structure and function of the *Salmonella* Typhi chimaeric A_2B_5 typhoid toxin,” *Nature*, vol. 499, no. 7458, p. 350, 2013.

[92] L. D. B. Belluz, R. Guidi, I. S. Pateras, L. Levi, B. Mihaljevic, S. F. Rouf, M. Wrande, M. Candela, S. Turroni, C. Nastasi, *et al.*, “The typhoid toxin promotes host survival and the establishment of a persistent asymptomatic infection,” *PLoS pathogens*, vol. 12, no. 4, p. e1005528, 2016.

[93] A. Purdy, F. Rohwer, R. Edwards, F. Azam, and D. H. Bartlett, “A glimpse into the expanded genome content of *Vibrio cholerae* through identification of genes present in environmental strains,” *Journal of bacteriology*, vol. 187, no. 9, pp. 2992–3001, 2005.

[94] R. Jørgensen, A. E. Purdy, R. J. Fieldhouse, M. S. Kimber, D. H. Bartlett, and A. R. Merrill, “Cholix toxin, a novel ADP-ribosylating factor from *Vibrio cholerae*,” *Journal of Biological Chemistry*, vol. 283, no. 16, pp. 10671–10678, 2008.

[95] S. P. Awasthi, M. Asakura, N. Chowdhury, S. B. Neogi, A. Hinenoya, H. M. Golbar, J. Yamate, E. Arakawa, T. Tada, T. Ramamurthy, *et al.*, “Novel cholix toxin variants, ADP-ribosylating toxins in *Vibrio cholerae* non-O1/non-O139 strains, and their pathogenicity,” *Infection and immunity*, vol. 81, no. 2, pp. 531–541, 2013.

[96] D. Visschedyk, A. Rochon, W. Tempel, S. Dimov, H.-W. Park, and A. R. Merrill, “Certhrax toxin, an anthrax-related ADP-ribosyltransferase from *Bacillus cereus*,” *Journal of Biological Chemistry*, vol. 287, no. 49, pp. 41089–41102, 2012.

[97] N. C. Simon and J. T. Barbieri, “*Bacillus cereus* Certhrax ADP-ribosylates vinculin to disrupt focal adhesion complexes and cell adhesion,” *Journal of Biological Chemistry*, vol. 289, no. 15, pp. 10650–10659, 2014.

[98] Y. I. Seldina, C. D. Petro, S. L. Servetas, J. M. Vergis, C. L. Ventura, D. S. Merrell, and A. D. O’Brien, “Certhrax is an antivirulence factor for the anthrax-like organism *Bacillus cereus* strain G9241,” *Infection and immunity*, vol. 86, no. 6, pp. e00207–18, 2018.

[99] M. J. Mansfield, J. B. Adams, and A. C. Doxey, “Botulinum neurotoxin homologs in non-*Clostridium* species,” *FEBS letters*, vol. 589, no. 3, pp. 342–348, 2015.

[100] M. Pirazzini, O. Rossetto, R. Eleopra, and C. Montecucco, “Botulinum neurotoxins: biology, pharmacology, and toxicology,” *Pharmacological reviews*, vol. 69, no. 2, pp. 200–235, 2017.

[101] I. Zornetta, G. Arrigoni, F. Anniballi, L. Bano, O. Leka, G. Zanotti, T. Binz, and C. Montecucco, “The first non-clostridial botulinum-like toxin cleaves VAMP within the juxtamembrane domain,” *Scientific reports*, vol. 6, p. 30257, 2016.

[102] S. Košenina, G. Masuyer, S. Zhang, M. Dong, and P. Stenmark, “Crystal structure of the catalytic domain of the *Weissela oryzae* botulinum-like toxin,” *FEBS letters*, 2019.

[103] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[104] T. F. Smith, M. S. Waterman, *et al.*, “Identification of common molecular subsequences,” *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[105] W. R. Pearson, “Rapid and sensitive sequence comparison with FASTP and FASTA,” 1990.

[106] P. Rice, I. Longden, and A. Bleasby, “EMBOSS: the European molecular biology open software suite,” 2000.

[107] R. C. Edgar, “Search and clustering orders of magnitude faster than BLAST,” *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[108] B. Buchfink, C. Xie, and D. H. Huson, “Fast and sensitive protein alignment using DIAMOND,” *Nature methods*, vol. 12, no. 1, p. 59, 2015.

[109] J. A. Gerlt and P. C. Babbitt, “Can sequence determine function?,” *Genome biology*, vol. 1, no. 5, pp. reviews0005–1, 2000.

[110] B. Rost and A. Valencia, “Pitfalls of protein sequence analysis,” *Current Opinion in Biotechnology*, vol. 7, no. 4, pp. 457–461, 1996.

[111] B. Rost, “Twilight zone of protein sequence alignments,” *Protein engineering*, vol. 12, no. 2, pp. 85–94, 1999.

[112] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[113] S. R. Eddy, “Profile hidden Markov models.,” *Bioinformatics (Oxford, England)*, vol. 14, no. 9, pp. 755–763, 1998.

[114] A. Hildebrand, M. Remmert, A. Biegert, and J. Söding, “Fast and accurate automatic structure prediction with HHpred,” *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. S9, pp. 128–132, 2009.

[115] G. Yona and M. Levitt, “Within the twilight zone: a sensitive profile-profile comparison tool based on information theory,” *Journal of molecular biology*, vol. 315, no. 5, pp. 1257–1275, 2002.

[116] L. A. Kelley and M. J. Sternberg, “Protein structure prediction on the Web: a case study using the Phyre server,” *Nature protocols*, vol. 4, no. 3, p. 363, 2009.

[117] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, *et al.*, “The Pfam protein families database: towards a more sustainable future,” *Nucleic acids research*, vol. 44, no. D1, pp. D279–D285, 2015.

[118] T. E. Lewis, I. Sillitoe, N. Dawson, S. D. Lam, T. Clarke, D. Lee, C. Orengo, and J. Lees, “Gene3D: extensive prediction of globular domains in proteins,” *Nucleic acids research*, vol. 46, no. D1, pp. D435–D439, 2017.

[119] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and A. G. Murzin, “SCOP2 prototype: a new approach to protein structure mining,” *Nucleic acids research*, vol. 42, no. D1, pp. D310–D314, 2013.

[120] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, “The COG database: a tool for genome-scale analysis of protein functions and evolution,” *Nucleic acids research*, vol. 28, no. 1, pp. 33–36, 2000.

[121] H. Mi, S. Poudel, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, "PANTHER version 10: expanded protein families and functions, and analysis tools," *Nucleic acids research*, vol. 44, no. D1, pp. D336–D342, 2015.

[122] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[123] H. M. Berman, P. E. Bourne, J. Westbrook, and C. Zardecki, "The protein data bank," in *Protein Structure*, pp. 394–410, CRC Press, 2003.

[124] N. D. Rawlings, A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn, "The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database," *Nucleic acids research*, vol. 46, no. D1, pp. D624–D632, 2017.

[125] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, *et al.*, "InterPro: the integrative protein signature database," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D211–D215, 2008.

[126] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, *et al.*, "InterProScan 5: genome-scale protein function classification," *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, 2014.

[127] B. Lobb and A. C. Doxey, "Novel function discovery through sequence and structural data mining," *Current opinion in structural biology*, vol. 38, pp. 53–61, 2016.

[128] E. V. Koonin, "Orthologs, paralogs, and evolutionary genomics," *Annu. Rev. Genet.*, vol. 39, pp. 309–338, 2005.

[129] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[130] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, *et al.*, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular systems biology*, vol. 7, no. 1, 2011.

[131] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[132] M. Holder and P. O. Lewis, "Phylogeny estimation: traditional and Bayesian approaches," *Nature reviews genetics*, vol. 4, no. 4, p. 275, 2003.

[133] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Bioinformatics*, vol. 13, no. 5, pp. 555–556, 1997.

[134] Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites," *Genetics*, vol. 155, no. 1, pp. 431–449, 2000.

[135] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, 2014.

[136] W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science*, vol. 284, no. 5423, pp. 2124–2128, 1999.

[137] L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, *et al.*, "A new view of the tree of life," *Nature microbiology*, vol. 1, no. 5, p. 16048, 2016.

[138] H. Philippe, D. Casane, S. Gribaldo, P. Lopez, and J. Meunier, "Heterotachy and functional shift in protein evolution," *IUBMB life*, vol. 55, no. 4-5, pp. 257–265, 2003.

[139] S. S. Arnon, R. Schechter, T. V. Inglesby, D. A. Henderson, J. G. Bartlett, M. S. Ascher, E. Eitzen, A. D. Fine, J. Hauer, M. Layton, *et al.*, "Botulinum toxin as a biological weapon: medical and public health management," *Jama*, vol. 285, no. 8, pp. 1059–1070, 2001.

[140] C. Montecucco and J. Molgó, "Botulinal neurotoxins: revival of an old killer," *Current opinion in pharmacology*, vol. 5, no. 3, pp. 274–279, 2005.

[141] M. Montal, "Botulinum neurotoxin: a marvel of protein design," *Annual review of biochemistry*, vol. 79, pp. 591–617, 2010.

[142] O. Rossetto, M. Pirazzini, and C. Montecucco, "Botulinum neurotoxins: genetic, structural and mechanistic insights," *Nature Reviews Microbiology*, vol. 12, no. 8, p. 535, 2014.

[143] G. Schiavo, M. Matteoli, and C. Montecucco, "Neurotoxins affecting neuroexocytosis," *Physiological reviews*, vol. 80, no. 2, pp. 717–766, 2000.

[144] R. Jahn and R. H. Scheller, "SNAREs—engines for membrane fusion," *Nature reviews Molecular cell biology*, vol. 7, no. 9, p. 631, 2006.

[145] T. C. Südhof and J. E. Rothman, "Membrane fusion: grappling with SNARE and SM proteins," *Science*, vol. 323, no. 5913, pp. 474–477, 2009.

[146] K. K. Hill, G. Xie, B. T. Foley, and T. J. Smith, "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins," *Toxicon*, vol. 107, pp. 2–8, 2015.

[147] S. Gu, S. Rumpel, J. Zhou, J. Strotmeier, H. Bigalke, K. Perry, C. B. Shoemaker, A. Rummel, and R. Jin, "Botulinum neurotoxin is shielded by NTNHA in an interlocked complex," *Science*, vol. 335, no. 6071, pp. 977–981, 2012.

[148] K. Lee, X. Zhong, S. Gu, A. M. Kruel, M. B. Dorner, K. Perry, A. Rummel, M. Dong, and R. Jin, "Molecular basis for disruption of E-cadherin adhesion by botulinum neurotoxin A complex," *Science*, vol. 344, no. 6190, pp. 1405–1410, 2014.

[149] Y. Sugawara, T. Matsumura, Y. Takegahara, Y. Jin, Y. Tsukasaki, M. Takeichi, and Y. Fujinaga, "Botulinum hemagglutinin disrupts the intercellular epithelial barrier by directly binding E-cadherin," *The Journal of cell biology*, vol. 189, no. 4, pp. 691–700, 2010.

[150] J. R. Barash and S. S. Arnon, "A novel strain of *Clostridium botulinum* that produces type B and type H botulinum toxins," *The Journal of infectious diseases*, vol. 209, no. 2, pp. 183–191, 2013.

[151] N. Dover, J. R. Barash, K. K. Hill, G. Xie, and S. S. Arnon, "Molecular characterization of a novel botulinum neurotoxin type H gene," *The Journal of infectious diseases*, vol. 209, no. 2, pp. 192–202, 2013.

[152] K. Hill, T. Smith, C. Helma, L. Ticknor, B. Foley, R. Svensson, J. Brown, E. Johnson, L. Smith, R. Okinaka, *et al.*, "Genetic diversity among botulinum neurotoxin-producing clostridial strains," *Journal of bacteriology*, vol. 189, no. 3, pp. 818–832, 2007.

[153] S. E. Maslanka, C. Lúquez, J. K. Dykes, W. H. Tepp, C. L. Pier, S. Pellett, B. H. Raphael, S. R. Kalb, J. R. Barr, A. Rao, *et al.*, "A novel botulinum neurotoxin, previously reported as serotype H, has a hybrid-like structure with regions of similarity to the structures of serotypes A and F and is neutralized with serotype A antitoxin," *The Journal of infectious diseases*, vol. 213, no. 3, pp. 379–385, 2015.

[154] C. Montecucco and M. B. Rasotto, "On botulinum neurotoxin variability," *MBio*, vol. 6, no. 1, pp. e02131–14, 2015.

[155] S. Zhang, G. Masuyer, J. Zhang, Y. Shen, D. Lundin, L. Henriksson, S.-I. Miyashita, M. Martínez-Carranza, M. Dong, and P. Stenmark, "Identification and characterization of a novel botulinum neurotoxin," *Nature communications*, vol. 8, p. 14130, 2017.

[156] F. Lebreton, W. van Schaik, A. M. McGuire, P. Godfrey, A. Griggs, V. Mazumdar, J. Corander, L. Cheng, S. Saif, S. Young, *et al.*, "Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains," *MBio*, vol. 4, no. 4, pp. e00534–13, 2013.

[157] F. Lebreton, A. L. Manson, J. T. Saavedra, T. J. Straub, A. M. Earl, and M. S. Gilmore, "Tracing the enterococci from Paleozoic origins to the hospital," *Cell*, vol. 169, no. 5, pp. 849–861, 2017.

[158] S. Schloissnig, M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D. R. Mende, J. R. Kultima, J. Martin, *et al.*, "Genomic variation landscape of the human gut microbiome," *Nature*, vol. 493, no. 7430, p. 45, 2013.

[159] D. Van Tyne and M. S. Gilmore, "Friend turned foe: evolution of enterococcal virulence and antibiotic resistance," *Annual review of microbiology*, vol. 68, pp. 337–356, 2014.

[160] C. A. Arias and B. E. Murray, "The rise of the *Enterococcus*: beyond vancomycin resistance," *Nature Reviews Microbiology*, vol. 10, no. 4, p. 266, 2012.

[161] M. S. Gilmore, F. Lebreton, and W. van Schaik, "Genomic transition of enterococci from gut commensals to leading causes of multidrug-resistant hospital infection in the antibiotic era," *Current opinion in microbiology*, vol. 16, no. 1, pp. 10–16, 2013.

[162] P. Courvalin, "Transfer of antibiotic resistance genes between gram-positive and gram-negative bacteria.," *Antimicrobial agents and chemotherapy*, vol. 38, no. 7, p. 1447, 1994.

[163] L. Guy, J. Roat Kultima, and S. G. Andersson, "genoPlotR: comparative gene and genome visualization in R," *Bioinformatics*, vol. 26, no. 18, pp. 2334–2335, 2010.

[164] A. Rummel, S. Mahrhold, H. Bigalke, and T. Binz, "The $\mathrm{HC_C}$-domain of botulinum neurotoxins A and B exhibits a singular ganglioside binding site displaying serotype specific carbohydrate interaction," *Molecular microbiology*, vol. 51, no. 3, pp. 631–643, 2004.

[165] D. M. Gill, "Bacterial toxins: a table of lethal amounts.," *Microbiological reviews*, vol. 46, no. 1, p. 86, 1982.

[166] E. C. Lim and R. C. Seet, "Use of botulinum toxin in the neurology clinic," *Nature Reviews Neurology*, vol. 6, no. 11, p. 624, 2010.

[167] M. Peck, T. Smith, F. Anniballi, J. Austin, L. Bano, M. Bradshaw, P. Cuervo, L. Cheng, Y. Derman, B. Dorner, *et al.*, "Historical perspectives and guidelines for botulinum neurotoxin subtype nomenclature," *Toxins*, vol. 9, no. 1, p. 38, 2017.

[168] L. Peng, W. H. Tepp, E. A. Johnson, and M. Dong, "Botulinum neurotoxin D uses synaptic vesicle protein SV2 and gangliosides as receptors," *PLoS pathogens*, vol. 7, no. 3, p. e1002008, 2011.

[169] M. Dong, D. A. Richards, M. C. Goodnough, W. H. Tepp, E. A. Johnson, and E. R. Chapman, "Synaptotagmins I and II mediate entry of botulinum neurotoxin B into cells," *The Journal of cell biology*, vol. 162, no. 7, pp. 1293–1303, 2003.

[170] A. Rummel, T. Karnath, T. Henke, H. Bigalke, and T. Binz, "Synaptotagmins I and II act as nerve cell receptors for botulinum neurotoxin G," *Journal of Biological Chemistry*, vol. 279, no. 29, pp. 30865–30870, 2004.

[171] S. Mahrhold, A. Rummel, H. Bigalke, B. Davletov, and T. Binz, "The synaptic vesicle protein 2C mediates the uptake of botulinum neurotoxin A into phrenic nerves," *FEBS letters*, vol. 580, no. 8, pp. 2011–2014, 2006.

[172] A. Rummel, K. Häfner, S. Mahrhold, N. Darashchonak, M. Holt, R. Jahn, S. Beermann, T. Karnath, H. Bigalke, and T. Binz, "Botulinum neurotoxins C, E and F bind gangliosides via a conserved binding site prior to stimulation-dependent uptake with botulinum neurotoxin F utilising the three isoforms of SV2 as second receptor," *Journal of neurochemistry*, vol. 110, no. 6, pp. 1942–1954, 2009.

[173] Z. Fu, C. Chen, J. T. Barbieri, J.-J. P. Kim, and M. R. Baldwin, "Glycosylated SV2 and gangliosides as dual receptors for botulinum neurotoxin serotype F," *Biochemistry*, vol. 48, no. 24, pp. 5631–5641, 2009.

[174] R. P. Berntsson, L. Peng, M. Dong, and P. Stenmark, "Structure of dual receptor binding to botulinum neurotoxin B," *Nature communications*, vol. 4, p. 2058, 2013.

[175] Q. Chai, J. W. Arndt, M. Dong, W. H. Tepp, E. A. Johnson, E. R. Chapman, and R. C. Stevens, "Structural basis of cell surface receptor recognition by botulinum neurotoxin B," *Nature*, vol. 444, no. 7122, p. 1096, 2006.

[176] R. Jin, A. Rummel, T. Binz, and A. T. Brunger, "Botulinum neurotoxin B recognizes its protein receptor with high affinity and specificity," *Nature*, vol. 444, no. 7122, p. 1092, 2006.

[177] T. Binz, S. Sikorra, and S. Mahrhold, "Clostridial neurotoxins: mechanism of SNARE cleavage and outlook on potential substrate specificity reengineering," *Toxins*, vol. 2, no. 4, pp. 665–682, 2010.

[178] J. Blasi, E. Chapman, S. Yamasaki, T. Binz, H. Niemann, and R. Jahn, "Botulinum neurotoxin C1 blocks neurotransmitter release by means of cleaving HPC-1/syntaxin.," *The EMBO journal*, vol. 12, no. 12, pp. 4821–4828, 1993.

[179] G. Schiavo, C. C. Shone, M. K. Bennett, R. H. Scheller, and C. Montecucco, "Botulinum neurotoxin type C cleaves a single Lys-Ala bond within the carboxyl-terminal region of syntaxins," *Journal of biological chemistry*, vol. 270, no. 18, pp. 10566–10570, 1995.

[180] S. Pantano and C. Montecucco, "The blockade of the neurotransmitter release apparatus by botulinum neurotoxins," *Cellular and molecular life sciences*, vol. 71, no. 5, pp. 793–811, 2014.

[181] J. Brunt, A. T. Carter, S. C. Stringer, and M. W. Peck, "Identification of a novel botulinum neurotoxin gene cluster in *Enterococcus*," *FEBS letters*, vol. 592, no. 3, pp. 310–317, 2018.

[182] T. G. Wentz, T. Muruvanda, S. Lomonaco, N. Thirunavukkarasu, M. Hoffmann, M. W. Allard, D. R. Hodge, S. P. Pillai, T. S. Hammack, E. W. Brown, *et al.*, "Closed genome sequence of *Chryseobacterium piperi* Strain CTMT/ATCC BAA-1782, a gram-negative bacterium with clostridial neurotoxin-like coding sequences," *Genome Announc.*, vol. 5, no. 48, pp. e01296–17, 2017.

[183] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2—a multiple sequence alignment editor and analysis workbench," *Bioinformatics*, vol. 25, no. 9, pp. 1189–1191, 2009.

[184] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, *et al.*, "CDD: NCBI's conserved domain database," *Nucleic acids research*, vol. 43, no. D1, pp. D222–D226, 2014.

[185] J. P. Huelsenbeck and F. Ronquist, "MRBAYES: Bayesian inference of phylogenetic trees," *Bioinformatics*, vol. 17, no. 8, pp. 754–755, 2001.

[186] S. Wu and Y. Zhang, "LOMETS: a local meta-threading-server for protein structure prediction," *Nucleic acids research*, vol. 35, no. 10, pp. 3375–3382, 2007.

[187] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The I-TASSER Suite: protein structure and function prediction," *Nature methods*, vol. 12, no. 1, p. 7, 2015.

[188] R. Hatherley, D. K. Brown, M. Glenister, and Ö. T. Bishop, "PRIMO: An interactive homology modeling pipeline," *PloS one*, vol. 11, no. 11, p. e0166698, 2016.

[189] F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, and C. Notredame, "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W604–W608, 2006.

[190] B. Webb and A. Sali, "Comparative protein structure modeling using MODELLER," *Current protocols in bioinformatics*, vol. 47, no. 1, pp. 5–6, 2014.

[191] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *Journal of applied crystallography*, vol. 26, no. 2, pp. 283–291, 1993.

[192] C. C. Huang, G. S. Couch, E. F. Pettersen, and T. E. Ferrin, "Chimera: an extensible molecular modeling application constructed using standard components," in *Pacific symposium on biocomputing*, vol. 1, p. 724, World Scientific, 1996.

[193] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: an information aesthetic for comparative genomics," *Genome research*, vol. 19, no. 9, pp. 1639–1645, 2009.

[194] K. Creuzburg, C. Giogha, T. W. F. Lung, N. E. Scott, S. Mühlen, E. L. Hartland, and J. S. Pearson, "The type III effector NleD from enteropathogenic *Escherichia coli* differentiates between host substrates p38 and JNK," *Infection and immunity*, vol. 85, no. 2, pp. e00620–16, 2017.

[195] J.-F. Bernardet, C. Hugo, and B. Bruun, "The genera *Chryseobacterium* and *Elizabethkingia*," *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass*, pp. 638–676, 2006.

[196] L. Zamora, A. I. Vela, M. A. Palacios, L. Domínguez, and J. F. Fernández-Garayzábal, "First isolation and characterization of *Chryseobacterium shigense* from rainbow trout," *BMC veterinary research*, vol. 8, no. 1, p. 77, 2012.

[197] R. Pukall, P. Schumann, C. Schütte, R. Gols, and M. Dicke, "*Acaricomes phytoseiuli* gen. nov., sp. nov., isolated from the predatory mite *Phytoseiulus persimilis*," *International journal of systematic and evolutionary microbiology*, vol. 56, no. 2, pp. 465–469, 2006.

[198] R. J. Wallace Jr, B. A. Brown, and G. O. Onyi, "Skin, soft tissue, and bone infections due to *Mycobacterium chelonae* chelonae: importance of prior corticosteroid therapy, frequency of disseminated infections, and resistance to oral antimicrobials other than clarithromycin," *Journal of Infectious Diseases*, vol. 166, no. 2, pp. 405–412, 1992.

[199] N. D. Rawlings, A. J. Barrett, and R. Finn, "Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors," *Nucleic acids research*, vol. 44, no. D1, pp. D343–D350, 2015.

[200] T. Binz, S. Bade, A. Rummel, A. Kollewe, and J. Alves, "Arg362 and Tyr365 of the botulinum neurotoxin type A light chain are involved in transition state stabilization," *Biochemistry*, vol. 41, no. 6, pp. 1717–1723, 2002.

[201] O. Rossetto, P. Caccin, M. Rigoni, F. Tonello, N. Bortoletto, R. Stevens, and C. Montecucco, "Active-site mutagenesis of tetanus neurotoxin implicates TYR-375 and GLU-271 in metalloproteolytic activity," *Toxicon*, vol. 39, no. 8, pp. 1151–1159, 2001.

[202] N. R. Silvaggi, D. Wilson, S. Tzipori, and K. N. Allen, "Catalytic features of the botulinum neurotoxin A light chain revealed by high resolution structure of an inhibitory peptide complex," *Biochemistry*, vol. 47, no. 21, pp. 5736–5745, 2008.

[203] R. M. Mizanur, V. Frasca, S. Swaminathan, S. Bavari, R. Webb, L. A. Smith, and S. A. Ahmed, "The C terminus of the catalytic domain of type A botulinum neurotoxin may facilitate product release from the active site," *Journal of Biological Chemistry*, vol. 288, no. 33, pp. 24223–24233, 2013.

[204] M. S. Monta, R. Blewitt, J. M. Tomich, and M. Montal, "Identification of an ion channel-forming motif in the primary structure of tetanus and botulinum neurotoxins," *FEBS letters*, vol. 313, no. 1, pp. 12–18, 1992.

[205] F. J. Lebeda and M. A. Olson, "Structural predictions of the channel-forming region of botulinum neurotoxin heavy chain," *Toxicon*, vol. 33, no. 4, pp. 559–567, 1995.

[206] A. Fischer, S. Sambashivan, A. T. Brunger, and M. Montal, "Beltless translocation domain of botulinum neurotoxin A embodies a minimum ion-conductive channel," *Journal of Biological Chemistry*, vol. 287, no. 3, pp. 1657–1661, 2012.

[207] P. K. Kienker, Z. Wu, and A. Finkelstein, "Topography of the TH5 segment in the diphtheria toxin T-domain channel," *The Journal of membrane biology*, vol. 249, no. 1-2, pp. 181–196, 2016.

[208] S. R. Kalb, J. Baudys, B. H. Raphael, J. K. Dykes, C. Lúquez, S. E. Maslanka, and J. R. Barr, "Functional characterization of botulinum neurotoxin serotype H as a hybrid of known serotypes F and A (BoNT F/A)," *Analytical chemistry*, vol. 87, no. 7, pp. 3911–3917, 2015.

[209] M. W. Peck, “Biology and genomic analysis of *Clostridium botulinum*,” *Advances in microbial physiology*, vol. 55, pp. 183–320, 2009.

[210] M. R. Weigand, A. Pena-Gonzalez, T. B. Shirey, R. G. Broeker, M. K. Ishaq, K. T. Konstantinidis, and B. H. Raphael, “Implications of genome-based discrimination between *Clostridium botulinum* group I and *Clostridium sporogenes* strains for bacterial taxonomy,” *Appl. Environ. Microbiol.*, vol. 81, no. 16, pp. 5420–5429, 2015.

[211] S. S. Dineen, M. Bradshaw, and E. A. Johnson, “Neurotoxin gene clusters in *Clostridium botulinum* type A strains: sequence comparison and evolutionary implications,” *Current microbiology*, vol. 46, no. 5, pp. 0345–0352, 2003.

[212] M. Sebaihia, M. W. Peck, N. P. Minton, N. R. Thomson, M. T. Holden, W. J. Mitchell, A. T. Carter, S. D. Bentley, D. R. Mason, L. Crossman, *et al.*, “Genome sequence of a proteolytic (Group I) *Clostridium botulinum* strain Hall A and comparative analysis of the clostridial genomes,” *Genome research*, vol. 17, no. 7, pp. 1082–1092, 2007.

[213] H. Skarin and B. Segerman, “Horizontal gene transfer of toxin genes in *Clostridium botulinum*: involvement of mobile elements and plasmids,” *Mobile genetic elements*, vol. 1, no. 3, pp. 213–215, 2011.

[214] A. T. Carter and M. W. Peck, “Genomes, neurotoxins and biology of *Clostridium botulinum* Group I and Group II,” *Research in microbiology*, vol. 166, no. 4, pp. 303–317, 2015.

[215] H. Brüggemann, E. Brzuszkiewicz, D. Chapeton-Montes, L. Plourde, D. Speck, and M. R. Popoff, “Genomics of *Clostridium tetani*,” *Research in microbiology*, vol. 166, no. 4, pp. 326–331, 2015.

[216] S.-I. Miyashita, Y. Sagane, T. Suzuki, T. Matsumoto, K. Niwa, and T. Watanabe, ““Non-toxic” proteins of the botulinum toxin complex exert in-vivo toxicity,” *Scientific reports*, vol. 6, p. 31043, 2016.

[217] B. R. DasGupta, “Botulinum neurotoxins: perspective on their existence and as polyproteins harboring viral proteases,” *The Journal of general and applied microbiology*, vol. 52, no. 1, pp. 1–8, 2006.

[218] C. H. Williamson, T. J. Smith, B. T. Foley, K. Hill, P. Keim, and J. W. Sahl, “Botulinum-neurotoxin-like sequences identified from an *Enterococcus* sp. genome assembly,” *BioRxiv*, p. 228098, 2017.

[219] M. J. Mansfield, T. G. Wentz, S. Zhang, E. J. Lee, M. Dong, S. K. Sharma, and A. C. Doxey, “Newly identified relatives of botulinum neurotoxins shed light on their molecular evolution,” *bioRxiv*, p. 220806, 2017.

[220] M. Collins and A. East, “Phylogeny and taxonomy of the food-borne pathogen *Clostridium botulinum* and its neurotoxins.,” *Journal of applied microbiology*, vol. 84, no. 1, pp. 5–17, 1998.

[221] C. Montecucco and G. Schiavo, “Mechanism of action of tetanus and botulinum neurotoxins,” *Molecular microbiology*, vol. 13, no. 1, pp. 1–8, 1994.

[222] R. Pellizzari, O. Rossetto, G. Schiavo, and C. Montecucco, “Tetanus and botulinum neurotoxins: mechanism of action and therapeutic uses,” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 354, no. 1381, pp. 259–268, 1999.

[223] J. E. Cohen, R. Wang, R.-F. Shen, W. W. Wu, and J. E. Keller, "Comparative pathogenomics of *Clostridium tetani*," *PloS one*, vol. 12, no. 8, p. e0182909, 2017.

[224] R. Kumar, T.-W. Chang, and B. R. Singh, "Evolutionary traits of toxins," *Biological Toxins and Bioterrorism*, pp. 527–557, 2015.

[225] J. Adams, M. J. Mansfield, D. J. Richard, and A. C. Doxey, "Lineage-specific mutational clustering in protein structures predicts evolutionary shifts in function," *Bioinformatics*, vol. 33, no. 9, pp. 1338–1345, 2017.

[226] D. M. Althoff, K. A. Segraves, and M. T. Johnson, "Testing for coevolutionary diversification: linking pattern with process," *Trends in Ecology & Evolution*, vol. 29, no. 2, pp. 82–89, 2014.

[227] R. Kumar and B. R. Singh, "Evolution of toxin," in *Protein Toxins in Modeling Biochemistry*, pp. 113–134, Springer, 2016.

[228] M. E. Woolhouse, J. P. Webster, E. Domingo, B. Charlesworth, and B. R. Levin, "Biological and biomedical implications of the co-evolution of pathogens and their hosts," *Nature genetics*, vol. 32, no. 4, p. 569, 2002.

[229] D. Ebert, "Host-parasite coevolution: insights from the *Daphnia*-parasite model system," *Current opinion in microbiology*, vol. 11, no. 3, pp. 290–301, 2008.

[230] S. Paterson, T. Vogwill, A. Buckling, R. Benmayor, A. J. Spiers, N. R. Thomson, M. Quail, F. Smith, D. Walker, B. Libberton, *et al.*, "Antagonistic coevolution accelerates molecular evolution," *Nature*, vol. 464, no. 7286, p. 275, 2010.

[231] M. F. Dybdahl, C. E. Jenkins, and S. L. Nuismer, "Identifying the molecular basis of host-parasite coevolution: merging models and mechanisms," *The American Naturalist*, vol. 184, no. 1, pp. 1–13, 2014.

[232] L. Masri, A. Branca, A. E. Sheppard, A. Papkou, D. Laehnemann, P. S. Guenther, S. Prahl, M. Saebelfeld, J. Hollensteiner, H. Liesegang, *et al.*, "Host-pathogen coevolution: the selective advantage of *Bacillus thuringiensis* virulence and its cry toxin genes," *PLoS biology*, vol. 13, no. 6, p. e1002169, 2015.

[233] M. Takeda, K. Tsukamoto, T. Kohda, M. Matsui, M. Mukamoto, and S. Kozaki, "Characterization of the neurotoxin produced by isolates associated with avian botulism," *Avian diseases*, vol. 49, no. 3, pp. 376–381, 2005.

[234] A. M. Yule, I. K. Barker, J. W. Austin, and R. D. Moccia, "Toxicity of *Clostridium botulinum* type E neurotoxin to Great Lakes fish: implications for avian botulism," *Journal of wildlife diseases*, vol. 42, no. 3, pp. 479–493, 2006.

[235] T. J. Smith, K. K. Hill, B. T. Foley, J. C. Detter, A. C. Munk, D. C. Bruce, N. A. Doggett, L. A. Smith, J. D. Marks, G. Xie, *et al.*, "Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3,/Ba4 and/B1 clusters are located within plasmids," *PloS one*, vol. 2, no. 12, p. e1271, 2007.

[236] G. E. Hannett, W. B. Stone, S. W. Davis, and D. Wroblewski, "Biodiversity of *Clostridium botulinum* type E associated with a large outbreak of botulism in wildlife from Lake Erie and Lake Ontario," *Appl. Environ. Microbiol.*, vol. 77, no. 3, pp. 1061–1068, 2011.

[237] L. Tao, L. Peng, R. P.-A. Berntsson, S. M. Liu, S. Park, F. Yu, C. Boone, S. Palan, M. Beard, P.-E. Chabrier, *et al.*, "Engineered botulinum neurotoxin B with improved efficacy for targeting human receptors," *Nature communications*, vol. 8, no. 1, p. 53, 2017.

[238] L. Peng, M. Adler, A. Demogines, A. Borrell, H. Liu, L. Tao, W. H. Tepp, S.-C. Zhang, E. A. Johnson, S. L. Sawyer, *et al.*, "Widespread sequence variations in VAMP1 across vertebrates suggest a potential selective pressure from botulinum neurotoxins," *PLoS pathogens*, vol. 10, no. 7, p. e1004177, 2014.

[239] V. V. Vaidyanathan, K.-i. Yoshino, M. Jahnz, C. Dörries, S. Bade, S. Nauenburg, H. Niemann, and T. Binz, "Proteolysis of SNAP-25 isoforms by botulinum neurotoxin types A, C, and E: domains and amino acid residues controlling the formation of enzyme-substrate complexes and cleavage," *Journal of neurochemistry*, vol. 72, no. 1, pp. 327–337, 1999.

[240] J. W. Arndt, W. Yu, F. Bi, and R. C. Stevens, "Crystal structure of botulinum neurotoxin type G light chain: serotype divergence in substrate recognition," *Biochemistry*, vol. 44, no. 28, pp. 9574–9580, 2005.

[241] S. Chen, J.-J. P. Kim, and J. T. Barbieri, "Mechanism of substrate recognition by botulinum neurotoxin serotype A," *Journal of biological chemistry*, vol. 282, no. 13, pp. 9621–9627, 2007.

[242] A. C. Doxey, M. D. Lynch, K. M. Müller, E. M. Meiering, and B. J. McConkey, "Insights into the evolutionary origins of clostridial neurotoxins from analysis of the *Clostridium botulinum* strain A neurotoxin gene cluster," *BMC evolutionary biology*, vol. 8, no. 1, p. 316, 2008.

[243] S. Amatsu, Y. Sugawara, T. Matsumura, K. Kitadokoro, and Y. Fujinaga, "Crystal structure of *Clostridium botulinum* whole hemagglutinin reveals a huge triskelion-shaped molecular complex," *Journal of Biological Chemistry*, vol. 288, no. 49, pp. 35617–35625, 2013.

[244] J. W. Arndt, J. Gu, L. Jaroszewski, R. Schwarzenbacher, M. A. Hanson, F. J. Lebeda, and R. C. Stevens, "The structure of the neurotoxin-associated protein HA33/A from *Clostridium botulinum* suggests a reoccurring β-trefoil fold in the progenitor toxin complex," *Journal of molecular biology*, vol. 346, no. 4, pp. 1083–1093, 2005.

[245] C. V. Jongeneel, J. Bouvier, and A. Bairoch, "A unique signature identifies a family of zinc-dependent metallopeptidases," *FEBS letters*, vol. 242, no. 2, pp. 211–214, 1989.

[246] K. Baruch, L. Gur-Arie, C. Nadler, S. Koby, G. Yerushalmi, Y. Ben-Neriah, O. Yogev, E. Shaulian, C. Guttman, R. Zarivach, *et al.*, "Metalloprotease type III effectors that specifically cleave JNK and NF-κB," *The EMBO journal*, vol. 30, no. 1, pp. 221–231, 2011.

[247] V. Cavalli, F. Vilbois, M. Corti, M. J. Marcote, K. Tamura, M. Karin, S. Arkinstall, and J. Gruenberg, "The stress-induced MAP kinase p38 regulates endocytic trafficking via the GDI: Rab5 complex," *Molecular cell*, vol. 7, no. 2, pp. 421–432, 2001.

[248] M. R. Baldwin, M. Bradshaw, E. A. Johnson, and J. T. Barbieri, "The C-terminus of botulinum neurotoxin type A light chain contributes to solubility, catalysis, and stability," *Protein expression and purification*, vol. 37, no. 1, pp. 187–195, 2004.

[249] D. Kumaran, R. Rawat, M. L. Ludivico, S. A. Ahmed, and S. Swaminathan, "Structure-and substrate-based inhibitor design for *Clostridium botulinum* neurotoxin serotype A," *Journal of Biological Chemistry*, vol. 283, no. 27, pp. 18883–18891, 2008.

[250] V. E. Bychkova, R. H. Pain, and O. B. Ptitsyn, “The ‘molten globule’ state is involved in the translocation of proteins across membranes?,” *FEBS letters*, vol. 238, no. 2, pp. 231–234, 1988.

[251] F. Van der Goot, J. Gonzalez-Manas, J. Lakey, and F. Pattus, “A 'molten-globule' membrane-insertion intermediate of the pore-forming domain of colicin A,” *Nature*, vol. 354, no. 6352, p. 408, 1991.

[252] A. Puhar, E. Johnson, O. Rossetto, and C. Montecucco, “Comparison of the pH-induced conformational change of different clostridial neurotoxins,” *Biochemical and biophysical research communications*, vol. 319, no. 1, pp. 66–71, 2004.

[253] R. Kukreja and B. Singh, “Biologically active novel conformational state of botulinum, the most poisonous poison,” *Journal of Biological Chemistry*, vol. 280, no. 47, pp. 39346–39352, 2005.

[254] L. Muraro, S. Tosatto, L. Motterlini, O. Rossetto, and C. Montecucco, “The N-terminal half of the receptor domain of botulinum neurotoxin A binds to microdomains of the plasma membrane,” *Biochemical and biophysical research communications*, vol. 380, no. 1, pp. 76–80, 2009.

[255] Y. Zhang, A. S. Gardberg, T. E. Edwards, B. Sankaran, H. Robinson, S. M. Varnum, and G. W. Buchko, “Structural insights into the functional role of the Hcn sub-domain of the receptor-binding domain of the botulinum neurotoxin mosaic serotype C/D,” *Biochimie*, vol. 95, no. 7, pp. 1379–1385, 2013.

[256] J. A. Payne, “A summer carrion study of the baby pig *Sus scrofa* Linnaeus,” *Ecology*, vol. 46, no. 5, pp. 592–602, 1965.

[257] I. Joseph, D. G. Mathew, P. Sathyan, and G. Vargheese, “The use of insects in forensic investigations: An overview on the scope of forensic entomology,” *Journal of forensic dental sciences*, vol. 3, no. 2, p. 89, 2011.

[258] Z. Hubálek and J. Halouzka, “Persistence of *Clostridium botulinum* type C toxin in blow fly (Calliphoridae) larvae as a possible cause of avian botulism in spring,” *Journal of wildlife diseases*, vol. 27, no. 1, pp. 81–85, 1991.

[259] I. Anza, D. Vidal, and R. Mateo, “New insight in the epidemiology of avian botulism outbreaks: necrophagous flies as vectors of *Clostridium botulinum* type C/D,” *Environmental microbiology reports*, vol. 6, no. 6, pp. 738–743, 2014.

[260] E. Contreras, G. Masuyer, N. Qureshi, S. Chawla, H. S. Dhillon, H. L. Lee, J. Chen, P. Stenmark, and S. S. Gill, “A neurotoxin that specifically targets *Anopheles* mosquitoes,” *Nature communications*, vol. 10, no. 1, p. 2869, 2019.

[261] R. Gustafsson, R. P.-A. Berntsson, M. Martínez-Carranza, G. El Tekle, R. Odegrip, E. A. Johnson, and P. Stenmark, “Crystal structures of OrfX2 and P47 from a Botulinum neurotoxin OrfX-type gene cluster,” *FEBS letters*, vol. 591, no. 22, pp. 3781–3792, 2017.

[262] K. ho Lam, R. Qi, S. Liu, A. Kroh, G. Yao, K. Perry, A. Rummel, and R. Jin, “The hypothetical protein p47 of *Clostridium botulinum* e1 strain beluga has a structural topology similar to bactericidal/permeability-increasing protein,” *Toxicon*, vol. 147, pp. 19 – 26, 2018. Basic science and clinical aspects of botulinum and other toxins.

[263] D. Chapeton-Montes, L. Plourde, C. Bouchier, L. Ma, L. Diancourt, A. Criscuolo, M. R. Popoff, and H. Brüggemann, "The population structure of *Clostridium tetani* deduced from its pan-genome," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[264] U. Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2018.

[265] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," *BMC bioinformatics*, vol. 10, no. 1, p. 421, 2009.

[266] R. Kolde, "Pheatmap: Pretty Heatmaps (version 1.0. 12)," 2019.

[267] G. Csardi, T. Nepusz, *et al.*, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[268] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell, "NCBI prokaryotic genome annotation pipeline," *Nucleic acids research*, vol. 44, no. 14, pp. 6614–6624, 2016.

[269] A. L. Mitchell, T. K. Attwood, P. C. Babbitt, M. Blum, P. Bork, A. Bridge, S. D. Brown, H.-Y. Chang, S. El-Gebali, M. I. Fraser, *et al.*, "InterPro in 2019: improving coverage, classification and access to protein sequence annotations," *Nucleic acids research*, vol. 47, no. D1, pp. D351–D360, 2018.

[270] A.-M. Grenier, G. Duport, S. Pagès, G. Condemine, and Y. Rahbé, "The phytopathogen *Dickeya dadantii* (*Erwinia chrysanthemi* 3937) is a pathogen of the pea aphid," *Appl. Environ. Microbiol.*, vol. 72, no. 3, pp. 1956–1965, 2006.

[271] N. Piqué, D. Miñana-Galbis, S. Merino, and J. M. Tomás, "Virulence factors of *Erwinia amylovora*: a review," *International journal of molecular sciences*, vol. 16, no. 6, pp. 12836–12854, 2015.

[272] D. P. Labeda, "Multilocus sequence analysis of phytopathogenic species of the genus *Streptomyces*," *International journal of systematic and evolutionary microbiology*, vol. 61, no. 10, pp. 2525–2531, 2011.

[273] C. B. Michielse and M. Rep, "Pathogen profile update: *Fusarium oxysporum*," *Molecular plant pathology*, vol. 10, no. 3, pp. 311–324, 2009.

[274] M. Fernández, M. Porcel, J. de la Torre, M. A. Molina-Henares, A. Daddaoua, M. A. Llamas, A. Roca, V. Carriel, I. Garzón, J. L. Ramos, *et al.*, "Analysis of the pathogenic potential of nosocomial *Pseudomonas putida* strains," *Frontiers in microbiology*, vol. 6, p. 871, 2015.

[275] E. N. Grady, J. MacDonald, L. Liu, A. Richman, and Z.-C. Yuan, "Current knowledge and perspectives of *Paenibacillus*: a review," *Microbial cell factories*, vol. 15, no. 1, p. 203, 2016.

[276] A. K. Panda, S. S. Bisht, S. DeMondal, N. S. Kumar, G. Gurusubramanian, and A. K. Panigrahi, "*Brevibacillus* as a biological tool: a short review," *Antonie Van Leeuwenhoek*, vol. 105, no. 4, pp. 623–639, 2014.

[277] R. L. Gherna, J. H. Werren, W. Weisburg, R. Cote, C. R. Woese, L. Mandelco, and D. J. Brenner, "*Arsenophonus nasoniae* gen. nov., sp. nov., the causative agent of the son-killer trait in the parasitic wasp *Nasonia vitripennis*," *International Journal of Systematic and Evolutionary Microbiology*, vol. 41, no. 4, pp. 563–565, 1991.

[278] E. Nováková, V. Hypša, and N. A. Moran, "*Arsenophonus*, an emerging clade of intracellular symbionts with a broad host distribution," *BMC microbiology*, vol. 9, no. 1, p. 143, 2009.

[279] L. Palma, D. Muñoz, C. Berry, J. Murillo, and P. Caballero, "*Bacillus thuringiensis* toxins: an overview of their biocidal activity," *Toxins*, vol. 6, no. 12, pp. 3296–3325, 2014.

[280] C. M. Sieber, W. Lee, P. Wong, M. Münsterkötter, H.-W. Mewes, C. Schmeitzl, E. Varga, F. Berthiller, G. Adam, and U. Güldener, "The *Fusarium graminearum* genome reveals more secondary metabolite gene clusters and hints of horizontal gene transfer," *PLoS One*, vol. 9, no. 10, p. e110311, 2014.

[281] K.-L. Sheahan, C. L. Cordero, and K. J. F. Satchell, "Identification of a domain within the multifunctional *Vibrio cholerae* RTX toxin that covalently cross-links actin," *Proceedings of the National Academy of Sciences*, vol. 101, no. 26, pp. 9798–9803, 2004.

[282] D. Zhang, R. F. de Souza, V. Anantharaman, L. M. Iyer, and L. Aravind, "Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics," *Biology direct*, vol. 7, no. 1, p. 18, 2012.

[283] S. Koskiniemi, J. G. Lamoureux, K. C. Nikolakakis, C. t. de Roodenbeke, M. D. Kaplan, D. A. Low, and C. S. Hayes, "Rhs proteins from diverse bacteria mediate intercellular competition," *Proceedings of the National Academy of Sciences*, vol. 110, no. 17, pp. 7032–7037, 2013.

[284] A. Jamet and X. Nassif, "New players in the toxin field: polymorphic toxin systems in bacteria," *MBio*, vol. 6, no. 3, pp. e00285–15, 2015.

[285] K. Sandvig and B. van Deurs, "Entry of ricin and Shiga toxin into cells: molecular mechanisms and medical perspectives," *The EMBO journal*, vol. 19, no. 22, pp. 5943–5950, 2000.

[286] R. Pedron, A. Esposito, I. Bianconi, E. Pasolli, A. Tett, F. Asnicar, M. Cristofolini, N. Segata, and O. Jousson, "Genomic and metagenomic insights into the microbial community of a thermal spring," *Microbiome*, vol. 7, no. 1, p. 8, 2019.

[287] H. Bigalke, "Botulinum toxin: application, safety, and limitations," in *Botulinum Neurotoxins*, pp. 307–317, Springer, 2012.

[288] W. K. Smits, D. Lyras, D. B. Lacy, M. H. Wilcox, and E. J. Kuijper, "*Clostridium difficile* infection," *Nature reviews Disease primers*, vol. 2, p. 16020, 2016.

[289] F. Barloy, M.-M. Lecadet, and A. Delécluse, "Distribution of clostridial cry-like genes among *Bacillus thuringiensis* and *Clostridium* strains," *Current microbiology*, vol. 36, no. 4, pp. 232–237, 1998.

[290] M. C. Hares, S. J. Hinchliffe, P. C. Strong, I. Eleftherianos, A. J. Dowling, N. Waterfield, *et al.*, "The *Yersinia pseudotuberculosis* and *Yersinia pestis* toxin complex is active against cultured mammalian cells," *Microbiology*, vol. 154, no. 11, pp. 3503–3517, 2008.

[291] R. Henry, "Etymologia: Diphtheria," *Emerging infectious diseases*, vol. 19, no. 11, p. 1838, 2013.

[292] K. Zakikhany and A. Efstratiou, "Diphtheria in Europe: current problems and new challenges," *Future microbiology*, vol. 7, no. 5, pp. 595–607, 2012.

[293] W. G. Van Panhuis, J. Grefenstette, S. Y. Jung, N. S. Chok, A. Cross, H. Eng, B. Y. Lee, V. Zadorozhny, S. Brown, D. Cummings, *et al.*, "Contagious diseases in the United States from 1888 to the present," *The New England journal of medicine*, vol. 369, no. 22, p. 2152, 2013.

[294] R. J. Collier, "Diphtheria toxin: mode of action and structure.," *Bacteriological reviews*, vol. 39, no. 1, p. 54, 1975.

[295] A. Cerdeno-Tarraga, A. Efstratiou, L. Dover, M. Holden, M. Pallen, S. Bentley, G. Besra, C. Churcher, K. James, A. De Zoysa, *et al.*, "The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129," *Nucleic acids research*, vol. 31, no. 22, pp. 6516–6523, 2003.

[296] D. Leong and J. R. Murphy, "Characterization of the diphtheria tox transcript in *Corynebacterium diphtheriae* and *Escherichia coli.*," *Journal of bacteriology*, vol. 163, no. 3, pp. 1114–1119, 1985.

[297] T. Sekizuka, A. Yamamoto, T. Komiya, T. Kenri, F. Takeuchi, K. Shibayama, M. Takahashi, M. Kuroda, and M. Iwaki, "*Corynebacterium ulcerans* 0102 carries the gene encoding diphtheria toxin on a prophage different from the *C. diphtheriae* NCTC 13129 prophage," *BMC microbiology*, vol. 12, no. 1, p. 72, 2012.

[298] L. Crossman, A. Cerdeño-Tárraga, S. Bentley, and J. Parkhill, "Pathogenomics," *Nature Reviews Microbiology*, vol. 1, no. 3, p. 176, 2003.

[299] A. De Zoysa, A. Efstratiou, and P. M. Hawkey, "Molecular characterization of diphtheria toxin repressor (*dtxR*) genes present in nontoxigenic *Corynebacterium diphtheriae* strains isolated in the United Kingdom," *Journal of clinical microbiology*, vol. 43, no. 1, pp. 223–228, 2005.

[300] A. Sing, S. Bierschenk, and J. Heesemann, "Classical diphtheria caused by *Corynebacterium ulcerans* in Germany: amino acid sequence differences between diphtheria toxins from *Corynebacterium diphtheriae* and *C. ulcerans*," *Clinical infectious diseases*, vol. 40, no. 2, pp. 325–326, 2005.

[301] R. Hogg, J. Wessels, J. Hart, A. Efstratiou, A. De Zoysa, G. Mann, T. Allen, and G. Pritchard, "Possible zoonotic transmission of toxigenic *Corynebacterium ulcerans* from companion animals in a human case of fatal diphtheria," 2009.

[302] M. S. Cohen and P. Chang, "Insights into the biogenesis, function, and regulation of ADP-ribosylation," *Nature chemical biology*, vol. 14, no. 3, p. 236, 2018.

[303] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.," *Molecular biology and evolution*, vol. 14, no. 7, pp. 685–695, 1997.

[304] M. Gouy, S. Guindon, and O. Gascuel, "SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building," *Molecular biology and evolution*, vol. 27, no. 2, pp. 221–224, 2009.

[305] D. Parks, M. Chuvochina, D. Waite, C. Rinke, A. Skarshewski, P. Chaumeil, *et al.*, "A proposal for a standardized bacterial taxonomy based on genome phylogeny. bioRxiv," 2018.

[306] E. L. Sonnhammer, G. Von Heijne, A. Krogh, *et al.*, "A hidden Markov model for predicting transmembrane helices in protein sequences.," in *Ismb*, vol. 6, pp. 175–182, 1998.

[307] M. Bennett and D. Eisenberg, "Refined structure of monomelic diphtheria toxin at 2.3 {aa resolution," *Protein Science*, vol. 3, no. 9, pp. 1464–1475, 1994.

[308] S. Mitternacht, "FreeSASA: An open source C library for solvent accessible surface area calculations," *F1000Research*, vol. 5, 2016.

[309] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids," *Nucleic acids research*, vol. 38, no. suppl_2, pp. W529–W533, 2010.

[310] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[311] Y. Zhou, Y. Liang, K. H. Lynch, J. J. Dennis, and D. S. Wishart, "PHAST: a fast phage search tool," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W347–W352, 2011.

[312] J.-O. Park, K. El-Tarabily, E. Ghisalberti, and K. Sivasithamparam, "Pathogenesis of *Streptoverticillium albireticuli* on *Caenorhabditis elegans* and its antagonism to soil-borne fungal pathogens," *Letters in Applied Microbiology*, vol. 35, no. 5, pp. 361–365, 2002.

[313] B.-H. Zhang, Z.-G. Ding, H.-Q. Li, X.-Z. Mou, Y.-Q. Zhang, J.-Y. Yang, E.-M. Zhou, and W.-J. Li, "Algicidal activity of *Streptomyces eurocidicus* JXJ-0089 metabolites and their effects on *Microcystis* physiology," *Appl. Environ. Microbiol.*, vol. 82, no. 17, pp. 5132–5143, 2016.

[314] A. Masters, T. Ellis, J. Carson, S. Sutherland, and A. Gregory, "*Dermatophilus chelonae* sp. nov., isolated from chelonids in Australia," *International Journal of Systematic and Evolutionary Microbiology*, vol. 45, no. 1, pp. 50–56, 1995.

[315] J. F. Wellehan, C. Turenne, D. J. Heard, C. J. Detrisac, and J. J. O'Kelley, "*Dermatophilus chelonae* in a king cobra (*Ophiophagus hannah*)," *Journal of Zoo and Wildlife Medicine*, vol. 35, no. 4, pp. 553–557, 2004.

[316] W. R. Pearson, "Flexible sequence similarity searching with the FASTA3 program package," in *Bioinformatics methods and protocols*, pp. 185–219, Springer, 2000.

[317] E. Papini, G. Schiavo, D. Sandona, R. Rappuoli, and C. Montecucco, "Histidine 21 is at the NAD+ binding site of diphtheria toxin.," *Journal of Biological Chemistry*, vol. 264, no. 21, pp. 12385–12388, 1989.

[318] E. Papini, A. Santucci, G. Schiavo, M. Domenighini, P. Neri, R. Rappuoli, and C. Montecucco, "Tyrosine 65 is photolabeled by 8-azidoadenine and 8-azidoadenosine at the NAD binding site of diphtheria toxin.," *Journal of Biological Chemistry*, vol. 266, no. 4, pp. 2494–2498, 1991.

[319] S. F. Carroll and R. J. Collier, "NAD binding site of diphtheria toxin: identification of a residue within the nicotinamide subsite by photochemical modification with NAD," *Proceedings of the National Academy of Sciences*, vol. 81, no. 11, pp. 3307–3311, 1984.

[320] X. Robert and P. Gouet, "Deciphering key features in protein structures with the new ENDscript server," *Nucleic acids research*, vol. 42, no. W1, pp. W320–W324, 2014.

[321] H. Zhan, J. Elliott, W. Shen, P. Huynh, A. Finkelstein, and R. Collier, "Effects of mutations in proline 345 on insertion of diphtheria toxin into model membranes," *The Journal of membrane biology*, vol. 167, no. 2, pp. 173–181, 1999.

[322] V. G. Johnson, P. Nicholls, W. Habig, and R. Youle, "The role of proline 345 in diphtheria toxin translocation.," *Journal of Biological Chemistry*, vol. 268, no. 5, pp. 3514–3519, 1993.

[323] P. Kaul, J. Silverman, W. H. Shen, S. R. Blanke, P. D. Huynh, A. Finkelstein, and R. John Collier, "Roles of Glu 349 and Asp 352 in membrane insertion and translocation by diphtheria toxin," *Protein science*, vol. 5, no. 4, pp. 687–692, 1996.

[324] J. A. Silverman, J. A. Mindell, A. Finkelstein, W. H. Shen, and R. J. Collier, "Mutational analysis of the helical hairpin region of diphtheria toxin transmembrane domain.," *Journal of Biological Chemistry*, vol. 269, no. 36, pp. 22524–22532, 1994.

[325] G. V. Louie, W. Yang, M. E. Bowman, and S. Choe, "Crystal structure of the complex of diphtheria toxin with an extracellular fragment of its receptor," *Molecular cell*, vol. 1, no. 1, pp. 67–78, 1997.

[326] J. Boyd, M. N. Oza, and J. R. Murphy, "Molecular cloning and DNA sequence analysis of a diphtheria tox iron-dependent regulatory element (*dtxR*) from *Corynebacterium diphtheriae*.," *Proceedings of the National Academy of Sciences*, vol. 87, no. 15, pp. 5968–5972, 1990.

[327] R. C. Fink, M. R. Evans, S. Porwollik, A. Vazquez-Torres, J. Jones-Carson, B. Troxell, S. J. Libby, M. McClelland, and H. M. Hassan, "FNR is a global regulator of virulence and anaerobic metabolism in *Salmonella enterica* serovar Typhimurium (ATCC 14028s)," *Journal of bacteriology*, vol. 189, no. 6, pp. 2262–2273, 2007.

[328] S. D. Bentley, K. F. Chater, A.-M. Cerdeño-Tárraga, G. L. Challis, N. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, *et al.*, "Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3 (2)," *Nature*, vol. 417, no. 6885, p. 141, 2002.

[329] M. Hamada, T. Iino, T. Iwami, S. Harayama, T. Tamura, and K.-i. Suzuki, "*Mobilicoccus pelagius* gen. nov., sp. nov. and *Piscicoccus intestinali*s gen. nov., sp. nov., two new members of the family Dermatophilaceae, and reclassification of *Dermatophilus chelonae* (Masters et al. 1995) as *Austwickia chelonae* gen. nov., comb. nov.," *The Journal of general and applied microbiology*, vol. 56, no. 6, pp. 427–436, 2010.

[330] K. Tamukai, T. Tokiwa, H. Kobayashi, and Y. Une, "Ranavirus in an outbreak of dermatophilosis in captive inland bearded dragons (Pogona vitticeps)," *Veterinary dermatology*, vol. 27, no. 2, pp. 99–e28, 2016.

[331] C. Vonrhein, C. Flensburg, P. Keller, A. Sharff, O. Smart, W. Paciorek, T. Womack, and G. Bricogne, "Data processing and analysis with the autoPROC toolbox," *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. 4, pp. 293–302, 2011.

[332] N. S. Pannu, W.-J. Waterreus, P. Skubák, I. Sikharulidze, J. P. Abrahams, and R. A. de Graaff, "Recent advances in the CRANK software suite for experimental phasing," *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. 4, pp. 331–337, 2011.

[333] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. Leslie, A. McCoy, *et al.*, "Overview of the CCP4 suite and current developments," *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. 4, pp. 235–242, 2011.

[334] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, *et al.*, "PHENIX: a comprehensive Python-based system for macromolecular structure solution," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 2, pp. 213–221, 2010.

[335] P. Emsley and K. Cowtan, "Coot: model-building tools for molecular graphics," *Acta Crystallographica Section D: Biological Crystallography*, vol. 60, no. 12, pp. 2126–2132, 2004.

[336] L. Holm and L. M. Laakso, "Dali server update," *Nucleic acids research*, vol. 44, no. W1, pp. W351–W355, 2016.

[337] J.-F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison," *Current opinion in structural biology*, vol. 6, no. 3, pp. 377–385, 1996.

[338] A. Stivala, M. Wybrow, A. Wirth, J. C. Whisstock, and P. J. Stuckey, "Automatic generation of protein structure cartoons with Pro-origami," *Bioinformatics*, vol. 27, no. 23, pp. 3315–3316, 2011.

[339] R. Jørgensen, A. R. Merrill, S. P. Yates, V. E. Marquez, A. L. Schwan, T. Boesen, and G. R. Andersen, "Exotoxin A–eEF2 complex structure indicates ADP ribosylation by ribosome mimicry," *Nature*, vol. 436, no. 7053, p. 979, 2005.

[340] C. von Eichel-Streiber, P. Boquet, M. Sauerborn, and M. Thelestam, "Large clostridial cytotoxins — a family of glycosyltransferases modifying small GTP-binding proteins," *Trends in microbiology*, vol. 4, no. 10, pp. 375–382, 1996.

[341] K. Amimoto, T. Noro, E. Oishi, and M. Shimizu, "A novel toxin homologous to large clostridial cytotoxins found in culture supernatant of *Clostridium perfringens* type C," *Microbiology*, vol. 153, no. 4, pp. 1198–1206, 2007.

[342] P. Bette, A. Oksche, F. Mauler, C. Eichel-Streiber, M. Popoff, and E. Habermann, "A comparative biochemical, pharmacological and immunological study of *Clostridium novyi* α-toxin, *C. difficile* toxin B and *C. sordellii* lethal toxin," *Toxicon*, vol. 29, no. 7, pp. 877–887, 1991.

[343] D. M. Aronoff and P. H. Kazanjian, "Historical and contemporary features of infections due to *Clostridium novyi*," *Anaerobe*, vol. 50, pp. 80–84, 2018.

[344] M. Aldape, A. Bryant, and D. Stevens, "*Clostridium sordellii* infection: epidemiology, clinical findings, and current perspectives on diagnosis and treatment," *Clinical Infectious Diseases*, vol. 43, no. 11, pp. 1436–1446, 2006.

[345] Y. Shindo, Y. Dobashi, T. Sakai, C. Monma, H. Miyatani, and Y. Yoshida, "Epidemiological and pathobiological profiles of *Clostridium perfringens* infections: review of consecutive series of 33 cases over a 13-year period," *International journal of clinical and experimental pathology*, vol. 8, no. 1, p. 569, 2015.

[346] M. R. Popoff and P. Bouvet, "Clostridial toxins," *Future microbiology*, vol. 4, no. 8, pp. 1021–1064, 2009.

[347] J. Schirmer and K. Aktories, "Large clostridial cytotoxins: cellular biology of Rho/Ras-glucosylating toxins," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1673, no. 1-2, pp. 66–74, 2004.

[348] K. E. Orrell, Z. Zhang, S. N. Sugiman-Marangos, and R. A. Melnyk, "*Clostridium difficile* toxins A and B: Receptors, pores, and translocation into cells," *Critical reviews in biochemistry and molecular biology*, vol. 52, no. 4, pp. 461–473, 2017.

[349] M. Pirazzini, O. Leka, G. Zanetti, O. Rossetto, and C. Montecucco, "On the translocation of botulinum and tetanus neurotoxins across the membrane of acidic intracellular compartments," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1858, no. 3, pp. 467–474, 2016.

[350] N. M. Chumbler, S. A. Rutherford, Z. Zhang, M. A. Farrow, J. P. Lisher, E. Farquhar, D. P. Giedroc, B. W. Spiller, R. A. Melnyk, and D. B. Lacy, "Crystal structure of *Clostridium difficile* toxin A," *Nature microbiology*, vol. 1, no. 1, p. 15002, 2016.

[351] P. Yuan, H. Zhang, C. Cai, S. Zhu, Y. Zhou, X. Yang, R. He, C. Li, S. Guo, S. Li, *et al.*, "Chondroitin sulfate proteoglycan 4 functions as the cellular receptor for *Clostridium difficile* toxin B," *Cell research*, vol. 25, no. 2, p. 157, 2015.

[352] P. Chen, L. Tao, T. Wang, J. Zhang, A. He, K.-h. Lam, Z. Liu, X. He, K. Perry, M. Dong, *et al.*, "Structural basis for recognition of frizzled proteins by *Clostridium difficile* toxin B," *Science*, vol. 360, no. 6389, pp. 664–669, 2018.

[353] M. E. LaFrance, M. A. Farrow, R. Chandrasekaran, J. Sheng, D. H. Rubin, and D. B. Lacy, "Identification of an epithelial cell receptor responsible for *Clostridium difficile* TcdB-induced cytotoxicity," *Proceedings of the National Academy of Sciences*, vol. 112, no. 22, pp. 7073–7078, 2015.

[354] B. Schorch, S. Song, F. R. Van Diemen, H. H. Bock, P. May, J. Herz, T. R. Brummelkamp, P. Papatheodorou, and K. Aktories, "LRP1 is a receptor for *Clostridium perfringens* TpeL toxin indicating a two-receptor model of clostridial glycosylating toxins," *Proceedings of the National Academy of Sciences*, vol. 111, no. 17, pp. 6431–6436, 2014.

[355] P. Gupta, Z. Zhang, S. N. Sugiman-Marangos, J. Tam, S. Raman, J.-P. Julien, H. K. Kroh, D. B. Lacy, N. Murgolo, K. Bekkari, *et al.*, "Functional defects in *Clostridium difficile* TcdB toxin uptake identify CSPG4 receptor-binding determinants," *Journal of Biological Chemistry*, vol. 292, no. 42, pp. 17290–17301, 2017.

[356] K. E. Orrell, Å. Tellgren-Roth, M. Di Bernardo, Z. Zhang, F. Cuviello, J. Lundqvist, G. von Heijne, I. Nilsson, and R. A. Melnyk, "Direct Detection of Membrane-Inserting Fragments Defines the Translocation Pores of a Family of Pathogenic Toxins," *Journal of molecular biology*, vol. 430, no. 18, pp. 3190–3199, 2018.

[357] Z. Zhang, M. Park, J. Tam, A. Auger, G. L. Beilhartz, D. B. Lacy, and R. A. Melnyk, "Translocation domain mutations affecting cellular toxicity identify the *Clostridium difficile* toxin B pore," *Proceedings of the National Academy of Sciences*, vol. 111, no. 10, pp. 3721–3726, 2014.

[358] S. Genisyuerek, P. Papatheodorou, G. Guttenberg, R. Schubert, R. Benz, and K. Aktories, "Structural determinants for membrane insertion, pore formation and translocation of *Clostridium difficile* toxin B," *Molecular microbiology*, vol. 79, no. 6, pp. 1643–1654, 2011.

[359] P. Chen, K.-h. Lam, Z. Liu, F. A. Mindlin, B. Chen, C. B. Gutierrez, L. Huang, Y. Zhang, T. Hamza, H. Feng, *et al.*, "Structure of the full-length *Clostridium difficile* toxin B," *Nature structural & molecular biology*, vol. 26, no. 8, pp. 712–719, 2019.

[360] G. L. Beilhartz, S. N. Sugiman-Marangos, and R. A. Melnyk, "Repurposing bacterial toxins for intracellular delivery of therapeutic proteins," *Biochemical pharmacology*, vol. 142, pp. 13–20, 2017.

[361] A. E. Rabideau and B. L. Pentelute, "Delivery of non-native cargo into mammalian cells using anthrax lethal toxin," *ACS chemical biology*, vol. 11, no. 6, pp. 1490–1501, 2016.

[362] K. Sandvig and B. van Deurs, "Delivery into cells: lessons learned from plant and bacterial toxins," *Gene therapy*, vol. 12, no. 11, p. 865, 2005.

[363] E. Paradis and K. Schliep, "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R," *Bioinformatics*, vol. 35, no. 3, pp. 526–528, 2018.

[364] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, *et al.*, "The Pfam protein families database in 2019," *Nucleic acids research*, vol. 47, no. D1, pp. D427–D432, 2018.

[365] K. Mendler, H. Chen, D. H. Parks, B. Lobb, L. A. Hug, and A. C. Doxey, "AnnoTree: visualization and exploration of a functionally annotated microbial tree of life," *Nucleic acids research*, vol. 47, no. 9, pp. 4442–4448, 2019.

[366] M. Rupnik and S. Janezic, "An update on *Clostridium difficile* toxinotyping," *Journal of clinical microbiology*, vol. 54, no. 1, pp. 13–18, 2016.

[367] M. D. Gershman, D. J. Kennedy, J. Noble-Wang, C. Kim, J. Gullion, M. Kacica, B. Jensen, N. Pascoe, L. Saiman, J. McHale, *et al.*, "Multistate outbreak of *Pseudomonas fluorescens* bloodstream infection after exposure to contaminated heparinized saline flush prepared by a compounding pharmacy," *Clinical Infectious Diseases*, vol. 47, no. 11, pp. 1372–1379, 2008.

[368] G. Mulley, M. L. Beeton, P. Wilkinson, I. Vlisidou, N. Ockendon-Powell, A. Hapeshi, N. J. Tobias, F. I. Nollmann, H. B. Bode, J. Van Den Elsen, *et al.*, "From insect to man: *Photorhabdus* sheds light on the emergence of human pathogenicity," *PLoS One*, vol. 10, no. 12, p. e0144937, 2015.

[369] S. D. Mahlen, "*Serratia* infections: from military experiments to current practice," *Clinical microbiology reviews*, vol. 24, no. 4, pp. 755–791, 2011.

[370] M. M. Shah, E. Odoyo, P. S. Larson, E. Apondi, C. Kathiiko, G. Miringu, M. Nakashima, and Y. Ichinose, "First report of a foodborne *Providencia alcalifaciens* outbreak in Kenya," *The American journal of tropical medicine and hygiene*, vol. 93, no. 3, pp. 497–500, 2015.

[371] S. Sagar, N. Narasimhaswamy, and J. d'Souza, "*Providencia rettgeri*: An emerging nosocomial uropathogen in an indwelling urinary catheterised patient," *Journal of clinical and diagnostic research: JCDR*, vol. 11, no. 6, p. DD01, 2017.

[372] M. H. Patel, G. R. Trivedi, S. M. Patel, and M. M. Vegad, "Antibiotic susceptibility pattern in urinary isolates of gram negative bacilli with special reference to AmpC β-lactamase in a tertiary care hospital," *Urology annals*, vol. 2, no. 1, p. 7, 2010.

[373] C. Osunla and A. Okoh, "*Vibrio* pathogens: A public health concern in rural water resources in sub-Saharan Africa," *International journal of environmental research and public health*, vol. 14, no. 10, p. 1188, 2017.

[374] J. M. Chaston, G. Suen, S. L. Tucker, A. W. Andersen, A. Bhasin, E. Bode, H. B. Bode, A. O. Brachmann, C. E. Cowles, K. N. Cowles, *et al.*, "The entomopathogenic bacterial endosymbionts *Xenorhabdus* and *Photorhabdus*: convergent lifestyles from divergent genomes," *PloS one*, vol. 6, no. 11, p. e27909, 2011.

[375] B. Ruffner, M. Péchy-Tarr, M. Höfte, G. Bloemberg, J. Grunder, C. Keel, and M. Maurhofer, "Evolutionary patchwork of an insecticidal toxin shared between plant-associated pseudomonads and the insect pathogens *Photorhabdus* and *Xenorhabdus*," *BMC genomics*, vol. 16, no. 1, p. 609, 2015.

[376] T. Giesemann, T. Jank, R. Gerhard, E. Maier, I. Just, R. Benz, and K. Aktories, "Cholesterol-dependent pore formation of *Clostridium difficil*e toxin A," *Journal of biological chemistry*, vol. 281, no. 16, pp. 10808–10815, 2006.

[377] H. Barth, G. Pfeifer, F. Hofmann, E. Maier, R. Benz, and K. Aktories, "Low pH-induced formation of ion channels by *Clostridium difficile* toxin B in target cells," *Journal of biological chemistry*, vol. 276, no. 14, pp. 10670–10676, 2001.

[378] D. Cutler, "Progress by poisoning," *Nature*, vol. 359, no. 6398, pp. 773–773, 1992.

[379] W. B. Huttner, "Cell biology. Snappy exocytoxins.," *Nature*, vol. 365, no. 6442, p. 104, 1993.

[380] B. Katz, "Quantal mechanism of neural transmitter release," *Science*, vol. 173, no. 3992, pp. 123–126, 1971.

[381] K. Geddes, M. Worley, G. Niemann, and F. Heffron, "Identification of new secreted effectors in *Salmonella enterica* serovar Typhimurium," *Infection and immunity*, vol. 73, no. 10, pp. 6260–6271, 2005.

[382] I. Gendlina, K. G. Held, S. Schesser Bartra, B. M. Gallis, C. E. Doneanu, D. R. Goodlett, G. V. Plano, and C. M. Collins, "Identification and type III-dependent secretion of the *Yersinia pestis* insecticidal-like proteins," *Molecular microbiology*, vol. 64, no. 5, pp. 1214–1227, 2007.

[383] A. Furutani, M. Takaoka, H. Sanada, Y. Noguchi, T. Oku, K. Tsuno, H. Ochiai, and S. Tsuge, "Identification of novel type III secretion effectors in *Xanthomonas oryzae pv. oryzae*," *Molecular plant-microbe interactions*, vol. 22, no. 1, pp. 96–106, 2009.

[384] R. W. Byard, "Diphtheria - 'the strangling angel' of children," *Journal of forensic and legal medicine*, vol. 20, no. 2, pp. 65–68, 2013.

[385] S. Choudhary, M. Mathew, and R. S. Verma, "Therapeutic potential of anticancer immunotoxins," *Drug discovery today*, vol. 16, no. 11-12, pp. 495–503, 2011.

[386] K.-h. Lam, Z. Guo, N. Krez, T. Matsui, K. Perry, J. Weisemann, A. Rummel, M. E. Bowen, and R. Jin, "A viral-fusion-peptide-like molecular switch drives membrane insertion of botulinum neurotoxin A1," *Nature communications*, vol. 9, no. 1, p. 5367, 2018.

[387] S. Sikorra, C. Litschko, C. Müller, N. Thiel, T. Galli, T. Eichner, and T. Binz, "Identification and characterization of botulinum neurotoxin A substrate binding pockets and their re-engineering for human SNAP-23," *Journal of molecular biology*, vol. 428, no. 2, pp. 372–384, 2016.

[388] M. Pirazzini, "Novel botulinum neurotoxins: Exploring underneath the iceberg tip," *Toxins*, vol. 10, no. 5, p. 190, 2018.

[389] M. R. Popoff, "Botulinum Neurotoxins: Still a Privilege of Clostridia?," *Cell host & microbe*, vol. 23, no. 2, pp. 145–146, 2018.

[390] A. del Sol, H. Fujihashi, D. Amoros, and R. Nussinov, "Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families," *Protein Science*, vol. 15, no. 9, pp. 2120–2128, 2006.

[391] J. M. Thornton, C. A. Orengo, A. E. Todd, and F. M. Pearl, "Protein folds, functions and evolution," *Journal of molecular biology*, vol. 293, no. 2, pp. 333–342, 1999.

[392] G. P. Carter, M. M. Awad, M. L. Kelly, J. I. Rood, and D. Lyras, "TcdB or not TcdB: a tale of two *Clostridium difficile* toxins," *Future microbiology*, vol. 6, no. 2, pp. 121–123, 2011.

[393] J. Steele, J. Mukherjee, N. Parry, and S. Tzipori, "Antibody against TcdB, but not TcdA, prevents development of gastrointestinal and systemic *Clostridium difficile* disease," *The Journal of infectious diseases*, vol. 207, no. 2, pp. 323–330, 2012.

[394] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance," *Review of educational research*, vol. 42, no. 3, pp. 237–288, 1972.

[395] D. G. Altman and J. M. Bland, "Statistics notes: the normal distribution," *Bmj*, vol. 310, no. 6975, p. 298, 1995.

[396] D. W. Zimmerman, "Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions," *The Journal of experimental education*, vol. 67, no. 1, pp. 55–68, 1998.

[397] R. Wilcox, M. Carlson, S. Azen, and F. Clark, "Avoid lost discoveries, because of violations of standard assumptions, by using modern robust statistical methods," *Journal of clinical epidemiology*, vol. 66, no. 3, pp. 319–329, 2013.

[398] D. Petrey and B. Honig, "Is protein classification necessary? Toward alternative approaches to function annotation," *Current opinion in structural biology*, vol. 19, no. 3, pp. 363–368, 2009.

[399] M. W. Gonzalez and W. R. Pearson, "Homologous over-extension: a challenge for iterative similarity searches," *Nucleic acids research*, vol. 38, no. 7, pp. 2177–2189, 2010.

[400] W. R. Pearson and M. L. Sierk, "The limits of protein sequence comparison?," *Current opinion in structural biology*, vol. 15, no. 3, pp. 254–260, 2005.

[401] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, *et al.*, "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing," *Journal of computational biology*, vol. 19, no. 5, pp. 455–477, 2012.

[402] M. Pop, "Genome assembly reborn: recent computational challenges," *Briefings in bioinformatics*, vol. 10, no. 4, pp. 354–366, 2009.

[403] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage: key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, p. 121, 2014.

[404] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, *et al.*, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nature biotechnology*, vol. 36, no. 4, p. 338, 2018.

[405] M. Watson and A. Warr, "Errors in long-read assemblies can critically affect protein prediction," *Nature biotechnology*, vol. 37, no. 2, p. 124, 2019.

[406] N. Rieber, M. Zapatka, B. Lasitschka, D. Jones, P. Northcott, B. Hutter, N. Jäger, M. Kool, M. Taylor, P. Lichter, *et al.*, "Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies," *PloS one*, vol. 8, no. 6, p. e66621, 2013.

[407] S. Koren, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose, "Reply to 'Errors in long-read assemblies can critically affect protein prediction'," *Nature biotechnology*, vol. 37, no. 2, p. 127, 2019.

[408] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, *et al.*, "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement," *PloS one*, vol. 9, no. 11, p. e112963, 2014.

[409] K. V. Ambrose, A. M. Koppenhöfer, and F. C. Belanger, "Horizontal gene transfer of a bacterial insect toxin gene into the Epichloë fungal symbionts of grasses," *Scientific Reports*, vol. 4, p. 5562, 2014.

[410] H.-Y. Jiang, M.-W. Huang, L.-B. Lin, N. He, and J.-P. Chen, "Complete genome sequence of *Austwickia chelonae* LK16-18, isolated from crocodile lizards," *Microbiol Resour Announc*, vol. 7, no. 16, pp. e01140–18, 2018.

[411] H. Jiang, X. Zhang, L. Li, J. Ma, N. He, H. Liu, R. Han, H. Li, Z. Wu, and J. Chen, "Identification of *Austwickia chelonae* as cause of cutaneous granuloma in endangered crocodile lizards using metataxonomics," *PeerJ*, vol. 7, p. e6574, 2019.

[412] T. H. Ogden and M. S. Rosenberg, "Multiple sequence alignment accuracy and phylogenetic inference," *Systematic biology*, vol. 55, no. 2, pp. 314–328, 2006.

[413] J. Felsenstein, "Cases in which parsimony or compatibility methods will be positively misleading," *Systematic zoology*, vol. 27, no. 4, pp. 401–410, 1978.

[414] B. Kolaczkowski and J. W. Thornton, "Long-branch attraction bias and inconsistency in Bayesian phylogenetics," *PloS one*, vol. 4, no. 12, p. e7891, 2009.

[415] P. Kück, C. Mayer, J.-W. Wägele, and B. Misof, "Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model," *PLoS One*, vol. 7, no. 5, p. e36593, 2012.

[416] E. Mossel and M. Steel, "How much can evolved characters tell us about the tree that generated them?," *arXiv preprint q-bio/0406048*, 2004.

[417] Y. Suzuki, G. V. Glazko, and M. Nei, "Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics," *Proceedings of the National Academy of Sciences*, vol. 99, no. 25, pp. 16138–16143, 2002.

[418] M. P. Simmons, K. M. Pickett, and M. Miya, "How meaningful are Bayesian support values?," *Molecular Biology and Evolution*, vol. 21, no. 1, pp. 188–199, 2004.

[419] K.-h. Lam, S. Sikorra, J. Weisemann, H. Maatsch, K. Perry, A. Rummel, T. Binz, and R. Jin, "Structural and biochemical characterization of the protease domain of the mosaic botulinum neurotoxin type HA," *Pathogens and disease*, vol. 76, no. 4, p. fty044, 2018.

[420] M. H. Schierup and J. Hein, "Consequences of recombination on traditional phylogenetic analysis," *Genetics*, vol. 156, no. 2, pp. 879–891, 2000.

[421] M. Arenas and D. Posada, "The effect of recombination on the reconstruction of ancestral sequences," *Genetics*, vol. 184, no. 4, pp. 1133–1139, 2010.

[422] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Molecular biology and evolution*, vol. 23, no. 2, pp. 254–267, 2005.

[423] B. Schierwater, M. Eitel, W. Jakob, H.-J. Osigus, H. Hadrys, S. L. Dellaporta, S.-O. Kolokotronis, and R. DeSalle, "Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis," *PLoS biology*, vol. 7, no. 1, p. e1000020, 2009.

[424] H. Philippe, R. Derelle, P. Lopez, K. Pick, C. Borchiellini, N. Boury-Esnault, J. Vacelet, E. Renard, E. Houliston, E. Quéinnec, *et al.*, "Phylogenomics revives traditional views on deep animal relationships," *Current Biology*, vol. 19, no. 8, pp. 706–712, 2009.

[425] C. W. Dunn, A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, *et al.*, "Broad phylogenomic sampling improves resolution of the animal tree of life," *Nature*, vol. 452, no. 7188, p. 745, 2008.

[426] H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, *et al.*, "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"," *Proceedings of the National Academy of Sciences*, vol. 102, no. 39, pp. 13950–13955, 2005.

[427] B. R. McDonald and C. R. Currie, "Lateral gene transfer dynamics in the ancient bacterial genus *Streptomyces*," *MBio*, vol. 8, no. 3, pp. e00644–17, 2017.

[428] M. Vos, M. C. Hesselman, T. A. te Beek, M. W. van Passel, and A. Eyre-Walker, "Rates of lateral gene transfer in prokaryotes: high but why?," *Trends in microbiology*, vol. 23, no. 10, pp. 598–605, 2015.

[429] H. Maughan, "Rates of molecular evolution in bacteria are relatively constant despite spore dormancy," *Evolution*, vol. 61, no. 2, pp. 280–288, 2007.

[430] G. Masuyer, S. Zhang, S. Barkho, Y. Shen, L. Henriksson, S. Košenina, M. Dong, and P. Stenmark, "Structural characterisation of the catalytic domain of botulinum neurotoxin X-high activity and unique substrate specificity," *Scientific reports*, vol. 8, no. 1, p. 4518, 2018.

[431] S. N. De and D. Chatterje, "An experimental study of the mechanism of action of *Vibrio cholerae* on the intestinal mucous membrane," *The Journal of pathology and bacteriology*, vol. 66, no. 2, pp. 559–562, 1953.

[432] T. J. Treangen and M. Pop, "You can't always sequence your way out of a tight spot," *EMBO reports*, vol. 19, no. 12, p. e47036, 2018.

[433] S. Filippidou, T. Junier, T. Wunderlin, C.-C. Lo, P.-E. Li, P. S. Chain, and P. Junier, "Underdetection of endospore-forming Firmicutes in metagenomic data," *Computational and structural biotechnology journal*, vol. 13, pp. 299–306, 2015.

[434] C. Von Mering, P. Hugenholtz, J. Raes, S. Tringe, T. Doerks, L. Jensen, N. Ward, and P. Bork, "Quantitative phylogenetic assessment of microbial communities in diverse environments," *science*, vol. 315, no. 5815, pp. 1126–1130, 2007.

[435] G. W. Leibniz, "New Essays on Human Understanding (1646-1716), translated by P. REM-NANT, J. BENNETT," 1981.

[436] S. Freer, "Linnaeus' *Philosophia Botanica*," 2004.

[437] W. Heisenberg, "Physics and philosophy," 1958.

[438] T. Dobzhansky, “Nothing in biology makes sense except in the light of evolution,” *The american biology teacher*, vol. 75, no. 2, pp. 87–92, 2013.

[439] G. Segal, J. J. Russo, and H. A. Shuman, “Relationships between a new type IV secretion system and the *icm/dot* virulence system of *Legionella pneumophila*,” *Molecular microbiology*, vol. 34, no. 4, pp. 799–809, 1999.

[440] D. Burstein, F. Amaro, T. Zusman, Z. Lifshitz, O. Cohen, J. A. Gilbert, T. Pupko, H. A. Shuman, and G. Segal, “Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires,” *Nature genetics*, vol. 48, no. 2, p. 167, 2016.

[441] M. Pirazzini, T. Henke, O. Rossetto, S. Mahrhold, N. Krez, A. Rummel, C. Montecucco, and T. Binz, “Neutralisation of specific surface carboxylates speeds up translocation of botulinum neurotoxin type B enzymatic domain,” *FEBS letters*, vol. 587, no. 23, pp. 3831–3836, 2013.

[442] L. Chen, J. Yang, J. Yu, Z. Yao, L. Sun, Y. Shen, and Q. Jin, “VFDB: a reference database for bacterial virulence factors,” *Nucleic acids research*, vol. 33, no. suppl_1, pp. D325–D328, 2005.

[443] B. Lobb, D. A. Kurtz, G. Moreno-Hagelsieb, and A. C. Doxey, “Remote homology and the functions of metagenomic dark matter,” *Frontiers in genetics*, vol. 6, p. 234, 2015.

[444] K. P. Schliep, “phangorn: phylogenetic analysis in R,” *Bioinformatics*, vol. 27, no. 4, pp. 592–593, 2010.

[445] T. S. Walker, H. P. Bais, E. Déziel, H. P. Schweizer, L. G. Rahme, R. Fall, and J. M. Vivanco, “*Pseudomonas aeruginosa*-plant root interactions. Pathogenicity, biofilm formation, and root exudation,” *Plant physiology*, vol. 134, no. 1, pp. 320–331, 2004.

[446] E. P. Rocha, “Evolutionary patterns in prokaryotic genomes,” *Current opinion in microbiology*, vol. 11, no. 5, pp. 454–460, 2008.

[447] B. L. Strahan, K. C. Failor, A. M. Batties, P. S. Hayes, K. M. Cicconi, C. T. Mason, and J. D. Newman, “*Chryseobacterium piperi* sp. nov., isolated from a freshwater creek,” *International journal of systematic and evolutionary microbiology*, vol. 61, no. 9, pp. 2162–2166, 2011.

[448] H. Philippe, H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain, “Resolving difficult phylogenetic questions: why more sequences are not enough,” *PLoS biology*, vol. 9, no. 3, p. e1000602, 2011.

APPENDICES

Appendix A

Supplementary Material: Chapter 2

A.1 Supplementary Table 1

Table A.1: Sequence identifiers, categories and protein lengths for each sequence used in this study. All sequences were retrieved from the NCBI *nr* database and are publicly available.

Accession	**Species**	**Classification**	**Length**
WP_055473237.1	*Streptomyces pathocidini*	Actinobacteria	1066
WP_083906476.1	*Acaricomes phytoseiuli*	Actinobacteria	1370
SDT83331.1	*Streptomyces* sp. TLI 053	Actinobacteria	495
WP_058043469.1	*Streptomyces* sp. MBT76	Actinobacteria	512
EFL04418.1	*Streptomyces* sp. AA4	Actinobacteria	2761
WP_030364034.1	*Streptomyces roseoverticillatus*	Actinobacteria	519
SDS61334.1	*Streptomyces* sp. TLI 053	Actinobacteria	231
GAO13068.1	*Streptomyces* sp. NBRC 110027	Actinobacteria	841
BAF91946.1	*Clostridium botulinum* str. Osaka 05 BoNT B6	BoNT	1291
ABM73981.1	*Clostridium botulinum* BoNT E2	BoNT	1252
AFV91339.1	*Clostridium botulinum* str. CDC66177 BoNT E9	BoNT	1251
ACQ51417.1	*Clostridium botulinum* Ba4 str. 657 BoNT BvA4	BoNT	1296
KEI05265.1	*Clostridium botulinum* CD str. BKT2873 BoNT CD	BoNT	1291
(continued on next page)			

Table A.1 – *(continued from previous page)*

Accession	Species	Classification	Length
AEN25581.1	*Clostridium botulinum* BoNT B7	BoNT	1291
ADA79573.1	*Clostridium botulinum* BoNT F5	BoNT	1277
ABM73977.1	*Clostridium botulinum* str. CDC 795 BoNT B3	BoNT	1291
ACA46990.1	*Clostridium botulinum* B1 str. Okra BoNT Ba4	BoNT	1291
CAA44558.1	*Clostridium botulinum* BoNT E1	BoNT	1252
BAP25804.1	*Clostridium botulinum* BoNT B2	BoNT	1291
CAA38175.1	*Clostridium botulinum* BoNT D	BoNT	1276
ADA79566.1	*Clostridium botulinum* BoNT F3	BoNT	1279
AER11392.1	*Clostridium botulinum* str. E134 BoNT E8	BoNT	1252
EEP52948.1	*Clostridium butyricum* E4 str. BoNT E BL5262 BoNT E4	BoNT	1252
ABS38337.1	*Clostridium botulinum* A str. Hall BoNT A1	BoNT	1296
EDT74844.1	*Clostridium butyricum* 5521 BoNT E	BoNT	1252
CBA17654.1	*Clostridium botulinum* BoNT CD	BoNT	1280
ABS41202.1	*Clostridium botulinum* F str. Langeland BoNT F1	BoNT	1278
ACD14195.1	*Clostridium botulinum* B str. Eklund 17B BoNT B4	BoNT	1291
ADU57954.1	*Clostridium botulinum* BoNT F6	BoNT	1275
ABM73985.1	*Clostridium botulinum* BoNT B2	BoNT	1291
ACT33194.1	*Clostridium botulinum* BoNT A5	BoNT	1296
EES49627.1	*Clostridium botulinum* E1 str. BoNT E Beluga BoNT E1	BoNT	1252
BAH84879.1	*Clostridium botulinum* BoNT DC	BoNT	1285
ACA57525.1	*Clostridium botulinum* A3 str. Loch Maree BoNT A3	BoNT	1292
EDS76240.1	*Clostridium botulinum* C str. Eklund BoNT C	BoNT	1280
CAA52275.1	*Clostridium botulinum* BoNT G	BoNT	1297
BAD90567.1	*Clostridium botulinum* BoNT C	BoNT	1291
			(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	Species	Classification	Length
BAB03522.1	*Clostridium butyricum* str. LCL 095 BoNT E5	BoNT	1251
ADA79579.1	*Clostridium baratii* BoNT F7	BoNT	1268
ACO83782.1	*Clostridium botulinum* A2 str. Kyoto BoNT A2	BoNT	1296
BAQ12790.1	*Clostridium botulinum* str 111 BoNT X	BoNT	1306
KGO12225.1	*Clostridium botulinum* BoNT H BoNT B	BoNT	1291
ADA79562.1	*Clostridium botulinum* BoNT F4	BoNT	1277
ADA79557.1	*Clostridium botulinum* BoNT F2	BoNT	1280
AER11391.1	*Clostridium botulinum* str. IBCA97 BoNT E7	BoNT	1252
CAM91125.1	*Clostridium botulinum* E str. K35 BoNT E6	BoNT	1252
KGO15617.1	*Clostridium botulinum* BoNT H BoNT Fa	BoNT	1288
KEH96501.1	*Clostridium botulinum* D str. 16868 BoNT D	BoNT	1287
ACQ51206.1	*Clostridium botulinum* Ba4 str. 657 BoNT BvBb5	BoNT	1291
ACD53549.1	*Clostridium botulinum* E3 str. Alaska E43 BoNT E3	BoNT	1252
OTO22244.1	*Enterococcus* sp. 3G1_DIV0629	BoNT	1280
WP_034687877.1	*Chryseobacterium piperi*	*Chryseobacterium*	388
KFF17709.1	*Chryseobacterium piperi* partial	*Chryseobacterium*	2112
WP_034687879.1	*Chryseobacterium piperi*	*Chryseobacterium*	695
WP_034681279.1	*Chryseobacterium piperi*	*Chryseobacterium*	1747
WP_034687193.1	*Chryseobacterium piperi*	*Chryseobacterium*	1496
WP_034687872.1	*Chryseobacterium piperi*	*Chryseobacterium*	1467
WP_034681281.1	*Chryseobacterium piperi*	*Chryseobacterium*	1617
WP_034687874.1	*Chryseobacterium piperi*	*Chryseobacterium*	1116
NZ_BAGZ01000024:3993143386	*Austwickia chelonae*	Diphtheria	260
WP_040322835.1	*Austwickia chelonae*	Diphtheria	274
WP_073156187.1	*Seinonella peptoniphila*	Diphtheria	606
			(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	**Species**	**Classification**	**Length**
ABU25232.1	Corynephage beta	Diphtheria	563
4AE0.1	*Corynebacterium diphtheria*	Diphtheria	535
WP_071569945.1	*Corynebacterium diphtheriae*	Diphtheria	560
BAG06869.1	*Corynebacterium ulcerans*	Diphtheria	560
WP_029975703.1	*Corynebacterium ulcerans*	Diphtheria	560
WP_038617330.1	*Corynebacterium ulcerans*	Diphtheria	560
WP_044032678.1	*Corynebacterium ulcerans*	Diphtheria	560
AAV70486.1	*Corynebacterium diphtheriae*	Diphtheria	536
AND74674.1	*Corynebacterium diphtheriae* bv. mitis	Diphtheria	560
AGT63319.1	*Corynebacterium ulcerans*	Diphtheria	296
WP_014654963.1	Corynebacterium pseudotuberculosis	Diphtheria	560
BAB03348.1	Corynephage beta	Diphtheria	535
AMP42519.1	*Corynebacterium diphtheriae*	Diphtheria	560
AMP42520.1	*Corynebacterium diphtheriae*	Diphtheria	560
WP_014835773.1	*Corynebacterium ulcerans*	Diphtheria	560
AAW22870.1	*Corynebacterium ulcerans*	Diphtheria	560
P00588.2	*Corynebacterium diphtheria*	Diphtheria	567
WP_054467370.1	*Corynebacterium ulcerans*	Diphtheria	560
WP_003850266.1	*Corynebacterium diphtheria*	Diphtheria	560
KHN96101.1	*Metarhizium album* ARSEF 1941	Fungal	730
XP_018137534.1	*Pochonia chlamydosporia* 170	Fungal	874
XP_007815863.1	*Metarhizium acridum* CQMa 102	Fungal	905
KID82327.1	*Metarhizium guizhouense* ARSEF 977	Fungal	994
KID81771.1	*Metarhizium guizhouense* ARSEF 977	Fungal	734
KFG79709.1	*Metarhizium anisopliae*	Fungal	704
XP_007806665.1	*Metarhizium acridum* CQMa 102	Fungal	795
KJZ74539.1	*Hirsutella minnesotensis* 3608	Fungal	951
KHN97968.1	*Metarhizium album* ARSEF 1941	Fungal	707
XP_007825170.1	*Metarhizium robertsii* ARSEF 23	Fungal	1082
KOM18317.1	*Ophiocordyceps unilateralis*	Fungal	723
KJK83643.1	*Metarhizium anisopliae* BRIP 53293	Fungal	1048
KFG84249.1	*Metarhizium anisopliae*	Fungal	909

(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	**Species**	**Classification**	**Length**
KID84774.1	*Metarhizium guizhouense* ARSEF 977	Fungal	1081
KID84474.1	*Metarhizium guizhouense* ARSEF 977	Fungal	290
XP_018138178.1	*Pochonia chlamydosporia* 170	Fungal	1069
KFG84771.1	*Metarhizium anisopliae*	Fungal	994
XP_014548865.1	*Metarhizium brunneum* ARSEF 3297	Fungal	908
EQL01877.1	*Ophiocordyceps sinensis* CO18	Fungal	615
XP_014541277.1	*Metarhizium brunneum* ARSEF 3297	Fungal	1082
KOM20214.1	*Ophiocordyceps unilateralis*	Fungal	798
XP_008600141.1	*Beauveria bassiana* ARSEF 2860	Fungal	172
XP_006673829.1	*Cordyceps militaris* CM01	Fungal	990
KOM18061.1	*Ophiocordyceps unilateralis*	Fungal	1006
KYK53973.1	*Drechmeria coniospora*	Fungal	753
OAR01471.1	*Cordyceps confragosa*	Fungal	901
XP_006672770.1	*Cordyceps militaris* CM01	Fungal	1046
KOM19981.1	*Ophiocordyceps unilateralis*	Fungal	676
XP_018701747.1	*Isaria fumosorosea* ARSEF 2679	Fungal	966
KJZ72439.1	*Hirsutella minnesotensis* 3608	Fungal	732
KJK78274.1	*Metarhizium anisopliae* BRIP 53293	Fungal	994
KOM17722.1	*Ophiocordyceps unilateralis*	Fungal	692
KID81025.1	*Metarhizium guizhouense* ARSEF 977	Fungal	1026
KFG78556.1	*Metarhizium anisopliae*	Fungal	1082
KID81342.1	*Metarhizium guizhouense* ARSEF 977	Fungal	1051
KHN96509.1	*Metarhizium album* ARSEF 1941	Fungal	799
XP_007813862.1	*Metarhizium acridum* CQMa 102	Fungal	731
KHN94110.1	*Metarhizium album* ARSEF 1941	Fungal	850
OAA45953.1	*Cordyceps brongniartii* RCEF 3172	Fungal	896
XP_014576605.1	*Metarhizium majus* ARSEF 297	Fungal	707
XP_007821069.1	*Metarhizium robertsii* ARSEF 23	Fungal	680
KOM19459.1	*Ophiocordyceps unilateralis*	Fungal	1134
KJK74733.1	*Metarhizium anisopliae* BRIP 53293	Fungal	777

(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	Species	Classification	Length
XP_008597266.1	*Beauveria bassiana* ARSEF 2860	Fungal	905
XP_007823153.1	*Metarhizium robertsii* ARSEF 23	Fungal	994
KHN96007.1	*Metarhizium album* ARSEF 1941	Fungal	182
XP_014574393.1	*Metarhizium majus* ARSEF 297	Fungal	1052
KOM22197.1	*Ophiocordyceps unilateralis*	Fungal	872
KFG79783.1	*Metarhizium anisopliae*	Fungal	790
KFG84514.1	*Metarhizium anisopliae*	Fungal	680
OAA46382.1	*Metarhizium rileyi* RCEF 4871	Fungal	1144
KGQ06896.1	*Beauveria bassiana* D1	Fungal	958
XP_007815132.1	*Metarhizium acridum* CQMa 102	Fungal	700
KZZ89413.1	*Aschersonia aleyrodis* RCEF 2490	Fungal	775
OAA38271.1	*Cordyceps brongniartii* RCEF 3172	Fungal	957
KFG81441.1	*Metarhizium anisopliae*	Fungal	961
KID85357.1	*Metarhizium guizhouense* ARSEF 977	Fungal	908
KOM22373.1	*Ophiocordyceps unilateralis*	Fungal	1051
ODA78204.1	*Drechmeria coniospora*	Fungal	795
KJK91804.1	*Metarhizium anisopliae* BRIP 53284	Fungal	1041
XP_008593942.1	*Beauveria bassiana* ARSEF 2860	Fungal	892
XP_018701218.1	*Isaria fumosorosea* ARSEF 2679	Fungal	864
KJK74458.1	*Metarhizium anisopliae* BRIP 53293	Fungal	300
KJZ74925.1	*Hirsutella minnesotensis* 3608	Fungal	779
KOM17891.1	*Ophiocordyceps unilateralis*	Fungal	800
KGQ09460.1	*Beauveria bassiana* D1	Fungal	892
KJZ75232.1	*Hirsutella minnesotensis* 3608	Fungal	967
OAA33269.1	*Aschersonia aleyrodis* RCEF 2490	Fungal	776
KOM19891.1	*Ophiocordyceps unilateralis*	Fungal	684
KJK77338.1	*Metarhizium anisopliae* BRIP 53293	Fungal	909
XP_018141808.1	*Pochonia chlamydosporia* 170	Fungal	836
KJZ77068.1	*Hirsutella minnesotensis* 3608	Fungal	651
KZZ93685.1	*Aschersonia aleyrodis* RCEF 2490	Fungal	882
KOM20843.1	*Ophiocordyceps unilateralis*	Fungal	667
EQL03202.1	*Ophiocordyceps sinensis* CO18	Fungal	707
XP_018702425.1	*Isaria fumosorosea* ARSEF 2679	Fungal	970
OAA33955.1	*Cordyceps brongniartii* RCEF 3172	Fungal	984
KZZ97314.1	*Aschersonia aleyrodis* RCEF 2490	Fungal	1053

(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	Species	Classification	Length
XP_014543958.1	*Metarhizium brunneum* ARSEF 3297	Fungal	1055
KOM18203.1	*Ophiocordyceps unilateralis*	Fungal	759
OAA38924.1	*Cordyceps brongniartii* RCEF 3172	Fungal	835
KID83117.1	*Metarhizium guizhouense* ARSEF 977	Fungal	802
KHN95172.1	*Metarhizium album* ARSEF 1941	Fungal	686
KZZ88788.1	*Aschersonia aleyrodis* RCEF 2490	Fungal	634
KHO00998.1	*Metarhizium album* ARSEF 1941	Fungal	846
XP_006671066.1	*Cordyceps militaris* CM01	Fungal	1164
OAA70995.1	*Cordyceps confragosa* RCEF 1005	Fungal	953
KOM19467.1	*Ophiocordyceps unilateralis*	Fungal	620
EQL04075.1	*Ophiocordyceps sinensis* CO18	Fungal	965
KFG86514.1	*Metarhizium anisopliae*	Fungal	1048
KHO02117.1	*Metarhizium album* ARSEF 1941	Fungal	907
XP_007825853.1	*Metarhizium robertsii* ARSEF 23	Fungal	908
KJK76584.1	*Metarhizium anisopliae* BRIP 53293	Fungal	782
XP_014544711.1	*Metarhizium brunneum* ARSEF 3297	Fungal	680
KID86267.1	*Metarhizium guizhouense* ARSEF 977	Fungal	467
KID82428.1	*Metarhizium guizhouense* ARSEF 977	Fungal	644
KID61384.1	*Metarhizium anisopliae* ARSEF 549	Fungal	782
EXU95574.1	*Metarhizium robertsii*	Fungal	631
ODA76424.1	*Drechmeria coniospora*	Fungal	969
XP_014581715.1	*Metarhizium majus* ARSEF 297	Fungal	903
KZZ92552.1	*Aschersonia aleyrodis* RCEF 2490	Fungal	1024
KJK78185.1	*Metarhizium anisopliae* BRIP 53293	Fungal	1082
KJZ71660.1	*Hirsutella minnesotensis* 3608	Fungal	986
KID82801.1	*Metarhizium guizhouense* ARSEF 977	Fungal	819
KYK58955.1	*Drechmeria coniospora*	Fungal	969
KID82824.1	*Metarhizium guizhouense* ARSEF 977	Fungal	463
XP_007824532.1	*Metarhizium robertsii* ARSEF 23	Fungal	704

(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	**Species**	**Classification**	**Length**
KJK84451.1	*Metarhizium anisopliae* BRIP 53293	Fungal	680
KHN94021.1	*Metarhizium album* ARSEF 1941	Fungal	673
XP_018700026.1	*Isaria fumosorosea* ARSEF 2679	Fungal	830
KGQ03580.1	*Beauveria bassiana* D1	Fungal	905
XP_014540776.1	*Metarhizium brunneum* ARSEF 3297	Fungal	994
KOM19284.1	*Ophiocordyceps unilateralis*	Fungal	1163
XP_014539926.1	*Metarhizium brunneum* ARSEF 3297	Fungal	790
OAA52226.1	*Cordyceps brongniartii* RCEF 3172	Fungal	945
KJK74970.1	*Metarhizium anisopliae* BRIP 53293	Fungal	704
KJK83642.1	*Metarhizium anisopliae* BRIP 53293	Fungal	305
XP_011411449.1	*Metarhizium robertsii* ARSEF 23	Fungal	827
KOM17955.1	*Ophiocordyceps unilateralis*	Fungal	960
XP_014576580.1	*Metarhizium majus* ARSEF 297	Fungal	759
ODA80573.1	*Drechmeria coniospora*	Fungal	857
XP_008599538.1	*Beauveria bassiana* ARSEF 2860	Fungal	219
KGQ05150.1	*Beauveria bassiana* D1	Fungal	458
KJZ71838.1	*Hirsutella minnesotensis* 3608	Fungal	756
KOM19347.1	*Ophiocordyceps unilateralis*	Fungal	1135
XP_014575725.1	*Metarhizium majus* ARSEF 297	Fungal	704
KYK56264.1	*Drechmeria coniospora*	Fungal	912
KJZ74384.1	*Hirsutella minnesotensis* 3608	Fungal	788
XP_008602550.1	*Beauveria bassiana* ARSEF 2860	Fungal	836
OAA52037.1	*Metarhizium rileyi* RCEF 4871	Fungal	711
XP_006674368.1	*Cordyceps militaris* CM01	Fungal	1035
XP_014574395.1	*Metarhizium majus* ARSEF 297	Fungal	637
EQL00592.1	*Ophiocordyceps sinensis* CO18	Fungal	940
XP_007816336.1	*Metarhizium robertsii* ARSEF 23	Fungal	1052
XP_008601493.1	*Beauveria bassiana* ARSEF 2860	Fungal	970
OAA75559.1	*Cordyceps confragosa* RCEF 1005	Fungal	931
KJZ70085.1	*Hirsutella minnesotensis* 3608	Fungal	471
OAA39888.1	*Cordyceps brongniartii* RCEF 3172	Fungal	902
KJZ78148.1	*Hirsutella minnesotensis* 3608	Fungal	775
WP_011267300.1	*Pseudomonas syringae*	M91	198
CUV34250.1	*Ralstonia solanacearum*	M91	245

(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	**Species**	**Classification**	**Length**
AEX33777.1	*Xanthomonas arboricola* pv. *pruni*	M91	146
WP_077142976.1	*Pseudomonas syringae*	M91	215
WP_061231638.1	*Leptospira noguchii*	M91	334
WP_074375462.1	*Xanthomonas translucens*	M91	204
GAE50625.1	*Xanthomonas arboricola* pv. *pruni* str. MAFF 311562	M91	230
ADX47300.1	*Acidovorax avenae* subsp. *avenae* ATCC 19860	M91	212
WP_026053328.1	*Leptospira santarosai*	M91	245
WP_069342860.1	*Pandoraea* sp. ISTKB	M91	222
WP_074052393.1	*Xanthomonas vesicatoria*	M91	213
WP_071011456.1	*Ralstonia solanacearum*	M91	219
WP_024689564.1	*Pseudomonas syringae* group	M91	212
CAP52016.1	*Xanthomonas campestris* pv. *campestris*	M91	249
CTP91498.1	*Xanthomonas translucens* pv. *poae*	M91	221
WP_020738886.1	*Sorangium cellulosum*	M91	260
WP_004471278.1	*Leptospira santarosai*	M91	361
WP_057176762.1	*Paraburkholderia caribensis*	M91	206
AMV47536.1	*Paraburkholderia caribensis*	M91	284
AEG71857.1	*Ralstonia solanacearum* Po82	M91	225
WP_074812007.1	*Pseudomonas syringae*	M91	230
WP_064048191.1	*Ralstonia solanacearum*	M91	219
SEI45716.1	*Pseudomonas* sp. NFR16	M91	234
AIL29259.1	*Pseudomonas syringae* pv. *actinidiae*	M91	201
WP_019994798.1	*Aureimonas ureilytica*	M91	179
WP_075251160.1	*Xanthomonas oryzae*	M91	168
WP_006453000.1	*Xanthomonas gardneri*	M91	218
WP_064297220.1	*Ralstonia solanacearum*	M91	221
AIE45643.1	*Acidovorax citrulli*	M91	197
WP_035542658.1	*Burkholderia* sp. UYPR1.413	M91	238
WP_006073522.1	*Vibrio*	M91	214
WP_039558791.1	*Ralstonia solanacearum*	M91	186
WP_011409781.1	*Xanthomonas oryzae*	M91	155
WP_036949404.1	*Providencia alcalifaciens*	M91	240

(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	**Species**	**Classification**	**Length**
WP_063885037.1	*Pseudomonas syringae*	M91	183
KPY01823.1	*Pseudomonas amygdali* pv. *mori*	M91	212
KPY92388.1	*Pseudomonas syringae* pv. *tomato*	M91	236
WP_019702312.1	*Acidovorax*	M91	213
WP_069191536.1	*Escherichia coli*	M91	231
GAE54985.1	*Xanthomonas arboricola* pv. *pruni* MAFF 301420	M91	219
WP_043897868.1	*Ralstonia solanacearum*	M91	174
WP_051781314.1	*Janthinobacterium agaricidamnosum*	M91	223
WP_076037893.1	*Xanthomonas campestris*	M91	207
WP_004771984.1	*Leptospira kirschneri*	M91	283
WP_017115174.1	*Xanthomonas vasicola*	M91	206
WP_016971449.1	*Pseudomonas tolaasii*	M91	184
WP_074686786.1	*Acidovorax citrulli*	M91	216
KFA31217.1	*Xanthomonas vasicola* pv. *vasculorum* NCPPB 1326	M91	202
KPY56004.1	*Pseudomonas amygdali* pv. *sesami*	M91	226
KPW50752.1	*Pseudomonas syringae* pv. *berberidis*	M91	221
CUV20876.1	*Ralstonia solanacearum*	M91	234
WP_075241300.1	*Xanthomonas oryzae*	M91	171
WP_071615581.1	*Ralstonia solanacearum*	M91	218
WP_061944074.1	*Collimonas pratensis*	M91	265
AMP07033.1	*Collimonas pratensis*	M91	304
KPX28235.1	*Pseudomonas coronafaciens* pv. *garcae*	M91	159
WP_075242222.1	*Xanthomonas oryzae*	M91	167
KTB84937.1	*Pseudomonas syringae* pv. *syringae* PD2774	M91	196
WP_011003174.1	*Ralstonia solanacearum*	M91	217
WP_071895715.1	*Ralstonia solanacearum*	M91	183
WP_070931164.1	*Mycobacterium chelonae*	*Mycobacterium*	990
WP_070931163.1	*Mycobacterium chelonae*	*Mycobacterium*	391
EES49602.1	*Clostridium botulinum* E1 str. BoNT E Beluga NTNH E1	NTNH	1163

(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	Species	Classification	Length
ACD14165.1	*Clostridium botulinum* B str. Eklund 17B NTNH B4	NTNH	1196
KGO15578.1	*Clostridium botulinum* BoNT H NTNH A	NTNH	1164
ACD52603.1	*Clostridium botulinum* E3 str. Alaska E43 NTNH E3	NTNH	1163
ACA57431.1	*Clostridium botulinum* A3 str. Loch Maree NTNH A3	NTNH	1159
BAQ12789.1	*Clostridium botulinum* str 111 NTNH X	NTNH	1174
CAA61228.1	*Clostridium botulinum* NTNH G	NTNH	1198
BAF91945.1	*Clostridium botulinum* str. Osaka 05 NTNH B6	NTNH	1197
BAP25803.1	*Clostridium botulinum* str. Prevot 25 NTNH B2	NTNH	1197
KEH96500.1	*Clostridium botulinum* D str. 16868 NTNH D	NTNH	1196
KGO12234.1	*Clostridium botulinum* BoNT H NTNH B	NTNH	1197
EDS76246.1	*Clostridium botulinum* C str. Eklund NTNH C	NTNH	1196
EEP54802.1	*Clostridium butyricum* E4 str. BoNT E BL5262 NTNH E4	NTNH	1163
ACQ51342.1	*Clostridium botulinum* Ba4 str. 657 NTNH BivA4	NTNH	1159
CAM91124.1	*Clostridium botulinum* E NTNH E6	NTNH	1163
EDT74767.1	*Clostridium butyricum* 5521 NTNH E	NTNH	1163
KEI05264.1	*Clostridium botulinum* CD str. BKT2873 NTNH CD	NTNH	1196
ACT33193.1	*Clostridium botulinum* NTNH A5	NTNH	1193
ABS37375.1	*Clostridium botulinum* A str. Hall NTNH A1	NTNH	1193
ACA47084.1	*Clostridium botulinum* B1 str. Okra NTNH B1	NTNH	1197

(continued on next page)

Table A.1 – *(continued from previous page)*

Accession	**Species**	**Classification**	**Length**
ACO85717.1	*Clostridium botulinum* A2 str. Kyoto NTNH A2	NTNH	1159
ADU57953.1	*Clostridium botulinum* NTNH F6	NTNH	1165
ABS40335.1	*Clostridium botulinum* F str. Langeland NTNH F1	NTNH	1163
AGR53839.1	*Clostridium baratii* NTNH F7	NTNH	1162
ACQ51274.1	*Clostridium botulinum* Ba4 str. 657 NTNH BvB	NTNH	1197
OTO22243.1	*Enterococcus* sp. 3G1_DIV0629	NTNH-like	1192
AAO37454.1	*Clostridium tetani* E88 TeNT tetani	TeNT	1315
WP_027699549.1	*Weissella oryzae* SG25 WoNT	Weissella	1296
WP_027699548.1	*Weissella oryzae* SG25 WoNTNH	Weissella	1437

A.2 Supplementary Table 2

Table A.2: Gene annotations for toxin gene clusters in *Chryseobacterium piperi*. Homologs were inferred from top 100 results of a BLASTn (nucleotide) or BLASTp (protein) search, respectively.

Locus	Cluster	Type	Annotation	DNA homology	Protein homology	Notes
CJF12_06270	1	DS Flank	Fatty Acid Hydroxylase			
CJF12_06275	1	DS Flank	Short-Chain Dehydrogenase	Broad distribution in *Chryseobacterium*	Full length hit, <1e-131, Genera: *Chryseobacterium*, *Pedobacter*, *Flavobacterium*,	
CJF12_06280	1	DS Flank	Hypothetical Protein (pseudo)	Broad distribution in *Chryseobacterium*	Full length hit, <4e-16, Genera: *Chryseobacterium*, *Elizabethkingia*, *Cruoricaptor*	Peptidase M15 domain protein, Endolysin, Phage-like protein, Van-Y like
CJF12_06285	1	DS Flank	Transposase (IS21)	Broad distribution in *Chryseobacterium*	Full length hit, <1e-161, Genera: *Chryseobacterium*, *Elizabethkingia*, *Flavobacterium*,	IS21
CJF12_06290	1	DS Flank	ATP-Binding Protein (IS21 Transposase)	Broad distribution in *Chryseobacterium*	Full length hit, <1e-94, Genera: *Chryseobacterium*, *Elizabethkingia*, *Flavobacterium*,	IS21 AG
CJF12_06295	1	DS Flank	BoNT_GC1_4			
CJF12_06300	1	DS Flank	Hypothetical Protein	Broad distribution in *Chryseobacterium*	Full length hit,<7e-25, Genera: *Chryseobacterium*, *Elizabethkingia*, *Cruoricaptor*,	Partial Domain Match D-alanyl-D-alanine carboxypeptidase
CJF12_06305	1	BoNT-Like	BoNT_GC1_3			
CJF12_06310	1	BoNT-Like	BoNT_GC1_3			
CJF12_06315	1	BoNT-Like	BoNT_GC1_3			
CJF12_06320	1	NTNH-Like	NTNH_GC1_2			
CJF12_06325	1	NTNH-Like	NTNH_GC1_2			
CJF12_06330	1	BoNT-Like	BoNT_GC1_2			
CJF12_06335	1	BoNT-Like	BoNT_GC1_2			
CJF12_06340	1	BoNT-Like	BoNT_GC1_2			
CJF12_06345	1	NTNH-Like	NTNH_GC1_1			
CJF12_06350	1	BoNT-Like	BoNT_GC1_1			
CJF12_06355	1	BoNT-Like	BoNT_GC1_1			
CJF12_06360	1	US Flank	Response Regulator	Broad distribution in *Chryseobacterium*	Full length hit, <1e-47, Genera: *Chryseobacterium*, *Sphingobacterium*	

(continued on next page)

Table A.2 – *(continued from previous page)*

Locus	Cluster	Type	Annotation	DNA homology	Protein homology	Notes
CJF12_06365	1	US Flank	Chemotaxis Protein CheB	Broad distribution in *Chryseobacterium*	Full length hit, <2e-60, Genera: *Chryseobacterium*, *Sphingobacterium*, *Pedobacter*,	
CJF12_06370	1	US Flank	Chemotaxis Protein CheR	Broad distribution in *Chryseobacterium*	Full length hit, <2e-136, Genera: *Chryseobacterium*, *Sphingobacterium*, *Pedobacter*,	
CJF12_06375	1	US Flank	Response Regulator	Broad distribution in *Chryseobacterium*	Full length hit, <2e-44, Genera: *Chryseobacterium*, *Sphingobacterium*, *Flavobacterium*	
CJF12_06380	1	US Flank	Histidine Kinase	Broad distribution in *Chryseobacterium*	Full length hit, <2e-44, Genera: *Chryseobacterium*, *Sphingobacterium*, *Flavobacterium*	
CJF12_14515	2	Alcohol Dehydrogenase	Broad distribution in *Chryseobacterium*	Full length hit, <1e-127, Genera: *Chryseobacterium*, Chitinophaga, *Flavobacterium*,		
CJF12_14520	2	Hypothetical Protein	Broad distribution in *Chryseobacterium*	Full length hit, = 0.0, Genus: *Chryseobacterium*	Annotation: Pos Tetraricopeptide repeat-containing protein	
CJF12_14525	2	IS1595 Family Transposase ISChpi1	Nucleotide homology to *Elizabethkingia*, Draconibacterium, *Chryseobacterium*, Myroides,	Full length hit, <2e-87, Broad Bacteroidetes distribution	IS1595, similar to CJF12_14620	
CJF12_14530	2	Hypothetical Protein	Low homology region	Low homology protein		
CJF12_14535	2	Hypothetical Protein	Low homology region	Full length hit, 2.6e-30, uncultured Bacteroidetes; partial C-term hits, <8e-20, Broad Bacteroidetes distribution	Annotation: Pos TMF family protein domain in hit *Spirosoma luteum*	
CJF12_14540	2	Hypothetical Protein	Low homology region	Hits (6e-18 to 9e-06) to N-terminus of Glycerophosphodiester phosphordiesterase domain proteins.	Annotation: Pos Glycerophosphodiester phosphordiesterase domain of *Agrobacterium tumefaciens* and similar proteins.	
CJF12_14545	2		N			
CJF12_14550	2		N			

(continued on next page)

Table A.2 – *(continued from previous page)*

Locus	**Cluster**	**Type**	**Annotation**	**DNA homology**	**Protein homology**	**Notes**
CJF12_14555	2	IS1982 Family Transposase	Nucleotide homology to Genera: *Prevotella*, *Clostridium*, *Barnesiella*,	Degraded, Broad Bacteroidetes distribution		
CJF12_14560	2	Alpha/Beta Hydrolase	Nucleotide homology to *Mycobacterium chelonae* CCUG 47445; Genera: *Debaryomyces*, *Dickeya*, *Candida*, *Agarobacterium*, *Mycobacterium*, *Rhodococcus*	Full length hit, <2e-97, Genera: *Chryseobacterium*, *Mycobacterium*, *Rhodococcus*, *Variovorax*, *Actinomadura*, *Nocardia*, *Chintinophaga*, *Pararhizobium*, *Ensifer*, *Pseudomonas*, ...		
			Genera: *Debaryomyces*, *Dickeya*, *Candida*, *Agarobacterium*, *Mycobacterium*, *Rhodococcus*			
CJF12_14565	2	tRNA-Leu	Broad distribution in *Chryseobacterium*, Elizabethingia, Riemerella	N/A	tRNA	
CJF12_14570	2	tRNA-Gly	Broad distribution in *Chryseobacterium*	N/A	tRNA	
CJF12_14575	2	tRNA-Leu	Broad distribution in *Chryseobacterium*	N/A	tRNA	
CJF12_14580	2	tRNA-Leu	Broad distribution in *Chryseobacterium*	N/A	tRNA	
CJF12_14585	2	tRNA-Gly	Broad distribution in *Chryseobacterium*	N/A	tRNA	
CJF12_14590	2	tRNA-Leu	Broad distribution in *Chryseobacterium*	N/A	tRNA	
CJF12_14595	2	DUF3127 Domain-Containing Protein	Broad distribution in *Chryseobacterium*	Full length hit, <6e-51, Genera: *Chryseobacterium*, *Flavobacterium*, Riemerella		
CJF12_14600	2	Leucyl/ Phenylalanyl-tRNA-Protein Transferase	Broad distribution in *Chryseobacterium*	Full length hit, <3e-107, Genus: *Chryseobacterium*		

(continued on next page)

Table A.2 – *(continued from previous page)*

Locus	**Cluster**	**Type**	**Annotation**	**DNA homology**	**Protein homology**	**Notes**
CJF12_14605	2	EamA/RhaT Family Transporter	Broad distribution in *Chryseobacterium*	Full length hit, <2e-142, Genus: *Chryseobacterium*		*CJF12_14565 through CJF12_14605 are inverted
CJF12_14610	2	Ankyrin Repeat Domain Containing Protein	Limited Distribution within *Chryseobacterium*, *Elizabethkingia*, *Sphingobacterium*	Full length hit, <1e-37, Genera: *Chryseobacterium*, *Elizabethkingia*, Em*Pedobacter Flavobacterium*,		
CJF12_14615	2	Catalase	Limited Distribution within *Chryseobacterium*, *Elizabethkingia*, *Sphingobacterium*	Full length hit, <7e-19, Genera: *Chryseobacterium*, *Elizabethkingia*, Em*Pedobacter Flavobacterium*,		
CJF12_14620	2	Transposase			IS1595 N-term, truncated, similar to CJF12_14530	

A.3 Supplementary Table 3

Table A.3: Presence and absence of various features of BoNTs and BoNT-like toxins, with associated literature support. This table provides supporting information for the mapped features shown in Figure 2.12.

Protein (Accession)	Genomic location	NTNH	Cluster type	Domains	Motifs	Substrate	Sources
BoNT/A1 (CAL82360.1)	chr	Yes	HA, ORFX	All domains	Disulfide, SxWY	SNAP25 (Q197)	1. Smith, Theresa J., et al. "Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3,/Ba4 and/B1 clusters are located within plasmids." PloS one 2.12 (2007): e1271. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/A2 (CAA51824.1)	chr, plm	Yes	ORFX	All domains	Disulfide, SxWY	SNAP25 (Q197)	1. Smith, Theresa J., et al. "Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3,/Ba4 and/B1 clusters are located within plasmids." PloS one 2.12 (2007): e1271. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/A3 (ACA57525.1)	plm	Yes	ORFX	All domains	Disulfide, SxWY	SNAP25 (Q197)	1. Smith, Theresa J., et al. "Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3,/Ba4 and/B1 clusters are located within plasmids." PloS one 2.12 (2007): e1271.
BoNT/A4 (ACQ51417.1)	plm	Yes	ORFX	All domains	Disulfide, SxWY	SNAP25 (Q197)	1. Smith, Theresa J., et al. "Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3,/Ba4 and/B1 clusters are located within plasmids." PloS one 2.12 (2007): e1271.
BoNT/A5 (ACG50065.1)	chr	Yes	HA	All domains	Disulfide, SxWY	SNAP25 (Q197)	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20.
BoNT/A6 (ACW83608.1)	-	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Lúquez, Carolina, Brian H. Raphael, and Susan E. Maslanka. "Neurotoxin gene clusters in *Clostridium botulinum* type Ab strains." Applied and environmental microbiology 75.19 (2009): 6094-6101. 2. Kull, Skadi, et al. "Isolation and functional characterization of the novel *Clostridium botulinum* neurotoxin A8 subtype." PLoS One 10.2 (2015): e0116381.
BoNT/A7 (AFV13854.1)	-	Inferred	-	All domains	Disulfide, SxWY	Inferred SNAP25	1. Mazuet, Christelle, et al. "Toxin detection in patients' sera by mass spectrometry during two outbreaks of type A botulism in France." Journal of clinical microbiology 50.12 (2012): 4091-4094.

(continued on next page)

Table A.3 – *(continued from previous page)*

Protein (Accession)	Genomic location	NTNH	Cluster type	Domains	Motifs	Substrate	Sources
BoNT/A8 (AJA05787.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	SNAP25 (Q197)	1. Kull, Skadi, et al. "Isolation and functional characterization of the novel *Clostridium botulinum* neurotoxin A8 subtype." PLoS One 10.2 (2015): e0116381.
BoNT/B1 (ACA46990.1)	plm	Yes	HA	All domains	Disulfide, SxWY	VAMP (Q78)	1. Smith, Theresa J., et al. "Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3,/Ba4 and/B1 clusters are located within plasmids." PloS one 2.12 (2007): e1271. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/B2 (BAC22064.1)	chr, plm	Yes	HA	All domains	Disulfide, SxWY	Inferred VAMP	1. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/B3 (ABM73977.1)	-	Inferred	Inferred HA	All domains	Disulfide, SxWY	Inferred VAMP	1. Umeda, Kaoru, et al. "Genetic characterization of *Clostridium botulinum* associated with type B infant botulism in Japan." Journal of clinical microbiology 47.9 (2009): 2720-2728. 2. Dover, Nir, Jason R. Barash, and Stephen S. Arnon. "Novel *Clostridium botulinum* toxin gene arrangement with subtype A5 and partial subtype B3 botulinum neurotoxin genes." Journal of clinical microbiology 47.7 (2009): 2349-2350.
BoNT/B4 (ABM73987.1)	plm	Yes	HA	All domains	Disulfide, SxWY	Inferred VAMP	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/B5 (ACQ51206.1)	plm	Yes	HA	All domains	Disulfide, SxWY	Inferred VAMP	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/B6 (BAF91946.1)	plm	Yes	HA	All domains	Disulfide, SxWY	Inferred VAMP	1. Umeda, Kaoru, et al. "Genetic characterization of *Clostridium botulinum* associated with type B infant botulism in Japan." Journal of clinical microbiology 47.9 (2009): 2720-2728.

(continued on next page)

Table A.3 – *(continued from previous page)*

Protein (Accession)	Genomic location	NTNH	Cluster type	Domains	Motifs	Substrate	Sources
BoNT/B7 (AFD33678.1)	-	Inferred	Inferred HA	All domains	Disulfide, SxWY	Inferred VAMP	1. Kalb, Suzanne R., et al. "De novo subtype and strain identification of botulinum neurotoxin type B through toxin proteomics." Analytical and bioanalytical chemistry 403.1 (2012): 215-226.
BoNT/B8 (AFN61309.1)	-	Inferred	Inferred HA	All domains	Disulfide, SxWY	Inferred VAMP	1. Wangroongsarb, Piyada, et al. "An Outbreak of type B botulism in Chaiyaphum Province, Thailand 2014." Bulletin of Chiang Mai Associated Medical Sciences 48.1 (2015): 49.
BoNT/C (BAA14235.1)	phage	Yes	HA	All domains	Disulfide, SxWY	Syt (K252), SNAP25 (R198)	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20.
BoNT/CD (BAA08418.1)	phage	Yes	HA	All domains	Disulfide, SxWY	Syt (K252), SNAP25 (R198)	1. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/D (EES90380.1)	phage	Yes	HA	All domains	Disulfide, SxWY	VAMP (K61)	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20.
BoNT/DC (ABP48747.1)	phage	Yes	HA	All domains	Disulfide, SxWY	VAMP (K61)	1. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/E1 (CAA43999.1)	chr, plm	Yes	ORFX	All domains	Disulfide, SxWY	SNAP25 (R180)	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8. 3. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.
BoNT/E2 (ABM73981.1)	inferred chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.

(continued on next page)

Table A.3 – *(continued from previous page)*

Protein (Accession)	Genomic location	NTNH	Cluster type	Domains	Motifs	Substrate	Sources
BoNT/E3 (ABM73980.1)	chr, plm	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8. 3. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.
BoNT/E4 (BAC05434.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8. 3. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.
BoNT/E5 (BAB03512.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.
BoNT/E6 (CAM91125.1)	Inferred chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.
BoNT/E7 (AER11391.1)	Inferred chr	Inferred	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555. 2. Chen, Ying, et al. "Sequencing the botulinum neurotoxin gene and related genes in *Clostridium botulinum* type E strains reveals orfx3 and a novel type E neurotoxin subtype." Journal of bacteriology 189.23 (2007): 8643-8650.

(continued on next page)

Table A.3 – *(continued from previous page)*

Protein (Accession)	**Genomic location**	**NTNH**	**Cluster type**	**Domains**	**Motifs**	**Substrate**	**Sources**
BoNT/E8 (AER11392.1)	Inferred chr	Inferred	-	All domains	Disulfide, SxWY	Inferred SNAP25	1. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.
BoNT/E9 (AFV91339.1)	Inferred chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555. 2. Raphael, Brian H., et al. "Analysis of a unique *Clostridium botulinum* strain from the Southern hemisphere producing a novel type E botulinum neurotoxin subtype." BMC microbiology 12.1 (2012): 245.
BoNT/E10 (AII82330.1)	chr, plm	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.
BoNT/E11 (AII82289.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Carter, Andrew T., et al. "Evolution of chromosomal *Clostridium botulinum* type E neurotoxin gene clusters: Evidence provided by their rare plasmid-borne counterparts." Genome biology and evolution 8.3 (2016): 540-555.
BoNT/E12 (AHK10119.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred SNAP25	1. Mazuet, Christelle, et al. "An atypical outbreak of food-borne botulism due to *Clostridium botulinum* types B and E from ham." Journal of clinical microbiology 53.2 (2015): 722-726.
BoNT/F1 (ABS41202.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	VAMP (Q60)	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8.
BoNT/F2 (CAA73972.1)	plm	Yes	ORFX	All domains	Disulfide, SxWY	Inferred VAMP	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20.
BoNT/F3 (ADA79575.1)	-	Yes	ORFX	All domains	Disulfide, SxWY	Inferred VAMP	1. Raphael, Brian H., et al. "Sequence diversity of genes encoding botulinum neurotoxin type F." Applied and environmental microbiology 76.14 (2010): 4805-4812.

(continued on next page)

Table A.3 – *(continued from previous page)*

Protein (Accession)	Genomic location	NTNH	Cluster type	Domains	Motifs	Substrate	Sources
BoNT/F4 (ADA79569.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred VAMP	1. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8. 2. Dover, Nir, et al. "*Clostridium botulinum* strain Af84 contains three neurotoxin gene clusters: bont/A2, bont/F4 and bont/F5." PloS one 8.4 (2013): e61205.
BoNT/F5 (ADA79560.1)	plm	Yes	ORFX	All domains	Disulfide, SxWY	VAMP (L56)	1. Raphael, Brian H., et al. "Sequence diversity of genes encoding botulinum neurotoxin type F." Applied and environmental microbiology 76.14 (2010): 4805-4812. 2. Dover, Nir, et al. "*Clostridium botulinum* strain Af84 contains three neurotoxin gene clusters: bont/A2, bont/F4 and bont/F5." PloS one 8.4 (2013): e61205.
BoNT/F5A (KGO15617.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	VAMP (L56)	1. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8. 2. Dover, Nir, et al. "Molecular characterization of a novel botulinum neurotoxin type H gene." The Journal of infectious diseases 209.2 (2013): 192-202.
BoNT/F6 (AAA23263.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred VAMP	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8. 3. Smith, Theresa J., et al. "Genomic sequences of six botulinum neurotoxin-producing strains representing three clostridial species illustrate the mobility and diversity of botulinum neurotoxin genes." Infection, Genetics and Evolution 30 (2015): 102-113.
BoNT/F7 (ADK48765.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred VAMP	1. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8. 2. Dover, Nir, et al. "Arrangement of the Clostridium baratii F7 toxin gene cluster with identification of a σ factor that recognizes the botulinum toxin gene cluster promoters." PloS one 9.5 (2014): e97983. 3. Smith, Theresa J., et al. "Genomic sequences of six botulinum neurotoxin-producing strains representing three clostridial species illustrate the mobility and diversity of botulinum neurotoxin genes." Infection, Genetics and Evolution 30 (2015): 102-113.

(continued on next page)

Table A.3 – *(continued from previous page)*

Protein (Accession)	Genomic location	NTNH	Cluster type	Domains	Motifs	Substrate	Sources
BoNT/F8 (WP_076177537.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	Inferred VAMP	1. Giordani, Francesco, et al. "Genomic characterization of Italian *Clostridium botulinum* group I strains." Infection, Genetics and Evolution 36 (2015): 62-71.
BoNT/G (KIE44899.1)	chr, plm	Yes	HA	All domains	Disulfide, SxWY	VAMP (A83)	1. Hill, Karen K., and Theresa J. Smith. "Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes." Botulinum neurotoxins. Springer Berlin Heidelberg, 2012. 1-20. 2. Hill, K. K., et al. "Genetic diversity within the botulinum neurotoxin-producing bacteria and their neurotoxins." Toxicon 107 (2015): 2-8. 3. Smith, Theresa J., et al. "Genomic sequences of six botulinum neurotoxin-producing strains representing three clostridial species illustrate the mobility and diversity of botulinum neurotoxin genes." Infection, Genetics and Evolution 30 (2015): 102-113.
TeNT (WP_011100836.1)	plm	No	No	All domains	Disulfide, SxWY	VAMP (Q78)	1. Cohen, Jonathan E., et al. "Comparative pathogenomics of *Clostridium tetani*." PloS one 12.8 (2017): e0182909.
BoNT/En (OTO22244.1)	plm	Yes	ORFX	All domains	Disulfide, SxWY	VAMP (D68)	1. Zhang, Sicai, et al. "Identification of a Botulinum Neurotoxin-like Toxin in a Commensal Strain of *Enterococcus faecium*." Cell host & microbe (2018). 2. Brunt, Jason, et al. "Identification of a novel botulinum neurotoxin gene cluster in *Enterococcus*." FEBS letters (2018). 3. Williamson, Charles HD, et al. "Botulinum-neurotoxin-like sequences identified from an *Enterococcus* sp. genome assembly." bioRxiv (2017): 228098.
BoNT/X (BAQ12790.1)	chr	Yes	ORFX	All domains	Disulfide, SxWY	VAMP (R66)	1. Zhang, Sicai, et al. "Identification and characterization of a novel botulinum neurotoxin." Nature communications 8 (2017): 14130.
BoNT/Wo (WP_027699549.1)	chr	Partial	No	All domains	Disulfide, SxWY	VAMP (W89)	1. Mansfield, Michael J., Jeremy B. Adams, and Andrew C. Doxey. "Botulinum neurotoxin homologs in non-*Clostridium* species." FEBS letters 589.3 (2015): 342-348.
Cp1 (WP_034687872.1)	chr	Partial	No	All domains	Disulfide, SxWY	None	1. Mansfield, Michael J., et al. "Newly identified relatives of botulinum neurotoxins shed light on their molecular evolution." bioRxiv (2017): 220806.

A.4 Supplementary Table 4

Table A.4: Nucleotide and protein identifiers from the NCBI Genome database corresponding to ORFX gene clusters.

Nucleotide ID	**Start**	**End**	**Species**	**Protein IDs**
NOYG01000011	280147	303094	*Rhodococcus* sp. 06-412-2C	OZC91763.1, OZC91764.1
CP012479	1238333	1262714	*Arthrobacter* sp. ERGS1:01	ALE05666.1, ALE05667.1, ALE05668.1
QBUG01000011	106838	131253	*Promicromonospora* sp. AC04	PUB23505.1, PUB23506.1, PUB23507.1
SLVI01000019	92846	108934	*Promicromonospora* sp. CF082 Ga0189745_119	TCM59590.1, TCM59591.1, TCM59592.1
PHUK01000001	1143225	1166293	*Kitasatospora* sp. OK780	PKB46238.1, PKB46239.1
NZ_JOEH01000025	115748	137271	*Streptacidiphilus jeojiense*	WP_084715219.1, WP_084715221.1
NZ_JQMJ01000004	1273963	1297143	*Streptacidiphilus rugosus*	WP_037604336.1, WP_084713504.1
FN554889	6498875	6522337	*Streptomyces scabiei*	CBG72874.1, CBG72875.1
JPPW01000417	10249	33711	*Streptomyces scabiei*	KFG03295.1, KFG03296.1
JPPX01000225	155836	181048	*Streptomyces scabiei*	KFF96595.1, KFF96596.1, KFF96597.1
NC_013929	6498875	6522442	*Streptomyces scabiei*	WP_013003441.1, WP_078580200.1
NZ_KL997441	218139	243531	*Streptomyces scabiei*	WP_037703294.1, WP_037703295.1, WP_079024416.1
NZ_NHOD01000599	1	11925	*Streptomyces scabiei*	WP_107473801.1
BCMM01000016	142609	167824	*Streptomyces scabiei* str. S58	GAQ63334.1, GAQ63335.1, GAQ63336.1
OLMK01000036	1	13313	*Streptomyces* sp. MA5143a	SPF02795.1
NZ_VDMA01000008	168152	191057	*Microbispora* sp. CR1-09	WP_139575493.1, WP_139575494.1
BHFQ01000004	223103	247606	*Capsulimonas corticalis* str. AX-7	GCE53538.1, GCE53539.1, GCE53540.1
NZ_FNVX01000004	236914	259797	*Bacteroides ihuae* str. Marseille-P2824	WP_071145116.1, WP_071145117.1
FUWZ01000001	2268920	2291808	*Chitinophaga eiseniae* str. DSM 22224	SJZ84530.1, SJZ84582.1
QKTW01000018	248073	263266	*Taibaiella soli*	PZF72469.1, PZF72470.1, PZF72471.1
DMST01000004	1	22284	Cytophagales bacterium isolate UBA9432	HAP58003.1, HAP58004.1, HAP58005.1
DLUS01000425	1	12730	Cytophagales bacterium isolate UBA9867	HAA14938.1, HAA14939.1, HAA14940.1
DLUT01000238	1	10234	Cytophagales bacterium isolate UBA9868	HAA20266.1, HAA20267.1, HAA20268.1
NZ_VCEI01000025	860993	886164	*Dyadobacter sediminis* str. Z12	WP_138282329.1, WP_138282330.1, WP_138282331.1
NZ_KB913013	2148379	2174704	*Rudanella lutea*	WP_019988038.1, WP_019988039.1, WP_019988040.1, WP_019988041.1
PVTE01000035	130	25106	*Spirosoma oryzae*	PRY27013.1, PRY27014.1, PRY27015.1
NZ_RPOC01000005	221970	246949	*Flavobacterium* sp. T13	WP_125721298.1, WP_125721300.1, WP_125721302.1
PEIG01000001	295207	318146	Chroococcales cyanobacterium IPPAS B-1203	PIG95165.1, PIG95166.1
LMOM01000040	1	11995	*Deinococcus* sp. Leaf326	KQR18755.1
NZ_SCNA01000023	13521	38984	*Bacillus* sp. 2SH	WP_137842863.1, WP_137842864.1, WP_137842865.1, WP_137842866.1
NVNL01000046	1	20608	*Bacillus thuringiensis* str. AFS089089	PEA87578.1, PEA87579.1, PEA87580.1, PEA87581.1, PEA87582.1
RHPK01000003	1168334	1194701	*Brevibacillus laterosporus* str. 1951	TPG68520.1, TPG68521.1, TPG68522.1, TPG68523.1

(continued on next page)

Table A.4 – *(continued from previous page)*

Nucleotide ID	**Start**	**End**	**Species**	**Protein IDs**
RHPK01000003_2	3078578	3112612	*Brevibacillus laterosporus* str. 1951	TPG70131.1, TPG70132.1, TPG70134.1, TPG70135.1, TPG70136.1, TPG70137.1
QJJD01000035	1	11251	*Brevibacillus laterosporus* str. MG64	RAP23898.1
RHPL01000052	4818	26803	*Brevibacillus laterosporus* str. Rsp	Pseudo_RHPL01000052.1_prot_19, TPG88748.1, TPG88749.1, TPG88750.1, TPG88752.1
RHPL01000080	1	13585	*Brevibacillus laterosporus* str. Rsp	TPG82513.1, TPG82514.1
NZ_SMZW01000009	92757	118846	*Paenibacillus dendritiformis* str. F-A1	WP_133383096.1, WP_133383097.1, WP_133383098.1, WP_133383099.1
MBTG01000018	104803	128867	*Paenibacillus ferrarius* str. CY1	OPH56035.1, OPH56036.1, OPH56037.1
CP019794	1947657	1973284	*Paenibacillus larvae*	AQT84652.1, AQT84653.1, AQT84654.1, Pseudo_CP019794.1_prot_1979
ADZY03000258	31334	48685	*Paenibacillus larvae* subsp. larvae B-3650	PCK70219.1, PCK70220.1, PCK70221.1, PCK70222.1
ADFW01000001	2046728	2070965	*Paenibacillus larvae* subsp. larvae DSM 25719	ETK28712.1, ETK28713.1, ETK28714.1, ETK28715.1, ETK28716.1
CP019655	4044518	4070146	*Paenibacillus larvae* subsp. larvae str. Eric_III	AVF28391.1, AVF28392.1, AVF28393.1, AVF28394.1
CP020327	1618255	1643883	*Paenibacillus larvae* subsp. pulvifaciens str. CCM 38	AQZ46656.1, AQZ46657.1, AQZ46658.1, AQZ46659.1
CP020557	1842644	1868273	*Paenibacillus larvae* subsp. pulvifaciens str. SAG 10367	ARF68073.1, ARF68074.1, ARF68075.1, ARF68076.1
NZ_NDGK01000036	401497	427560	*Paenibacillus thiaminolyticus*	WP_087443774.1, WP_087443775.1, WP_087443776.1, WP_127510999.1
QYZD01000004	139253	165319	*Paenibacillus thiaminolyticus* str. BO5	RJG25203.1, RJG25204.1, RJG25205.1, RJG25206.1
UGRZ01000006	2940140	2966209	*Paenibacillus thiaminolyticus* str. NCTC11027	SUA98986.1, SUA98987.1, SUA98988.1, SUA98989.1
NGLI01000004	148615	173789	*Enterococcus faecium* str. 3G1_DIV0629	OTO22240.1, OTO22241.1, OTO22242.1
JX847735	1	15200	*Clostridium baratii* F7 str. IBCA03-0045	AGR53835.1, AGR53836.1, AGR53837.1, AGR53838.1
JZTY01000005	1	15748	*Clostridium baratii* str. 771-14	KJU72373.1, KJU72374.1, KJU72375.1, KJU72376.1
NZ_LUSO01000011	106	20128	*Clostridium baratii* str. 796-15	WP_039311664.1, WP_079286133.1, WP_079286134.1, WP_079286135.1
NZ_CP014203	1	25614	*Clostridium baratii* str. CDC51267	WP_039311664.1, WP_039311666.1, WP_045724644.1, WP_045724645.1
CP006905	747263	772975	*Clostridium baratii* str. Sullivan	AIY83090.1, AIY83468.1, AIY84033.1, AIY84667.1
KM233166	1	18952	*Clostridium botulinum* A str. Chemnitz	AJA05781.1, AJA05782.1, AJA05783.1, AJA05785.1
DQ310546	1	10655	*Clostridium botulinum* A str. Mascarpone	ABC25997.1, ABC25998.1, ABC26000.1
NZ_AQPU01000165	1	25604	*Clostridium botulinum* A1 str. CFSAN002368 isolate CDC297	WP_003356995.1, WP_003357203.1, WP_012704373.1, WP_021136344.1
NZ_AUYW01000009	361649	388007	*Clostridium botulinum* A2 117	WP_033066626.1, WP_076174538.1, WP_076174540.1, WP_076174544.1
CP001581	935162	962679	*Clostridium botulinum* A2 str. Kyoto	ACO83474.1, ACO84115.1, ACO84705.1, ACO87274.1

(continued on next page)

Table A.4 – *(continued from previous page)*

Nucleotide ID	**Start**	**End**	**Species**	**Protein IDs**
NC_012563	935162	962679	*Clostridium botulinum* A2 str. Kyoto	WP_012703709.1, WP_012704051.1, WP_012704373.1, WP_012705769.1
NZ_AUZA01000014	244956	260807	*Clostridium botulinum* A2B7 92	WP_012300848.1, WP_012300974.1, WP_076042694.1, WP_076042839.1
NZ_AUZB01000012	57383	83738	*Clostridium botulinum* A2B7 92	WP_033066624.1, WP_033066625.1, WP_033066626.1, WP_033066628.1
CP000963	41634	69054	*Clostridium botulinum* A3 str. Loch Maree	ACA57304.1, ACA57457.1, ACA57490.1, ACA57564.1
NC_010418	41634	69054	*Clostridium botulinum* A3 str. Loch Maree	WP_012300848.1, WP_012300974.1, WP_012301001.1, WP_012301067.1
CP001081	536	28048	*Clostridium botulinum* Ba4 str. 657	ACQ51172.1, ACQ51216.1, ACQ51221.1, ACQ51300.1
NC_012654	536	28048	*Clostridium botulinum* Ba4 str. 657	WP_003362717.1, WP_012720169.1, WP_012720205.1, WP_012720209.1
AM695754	1	12908	*Clostridium botulinum* E str. 31-2570	CAM91132.1, CAM91133.1, CAM91134.1, CAM91135.1
AM941719	1	12908	*Clostridium botulinum* E str. CB11/1-1	CAQ03490.1, CAQ03491.1, CAQ03492.1, CAQ03493.1
AM695752	1	12908	*Clostridium botulinum* E str. K35	CAM91120.1, CAM91121.1, CAM91122.1, CAM91123.1
AM695753	1	12908	*Clostridium botulinum* E str. K81	CAM91126.1, CAM91127.1, CAM91128.1, CAM91129.1
ACSC01000002	2654146	2679872	*Clostridium botulinum* E1 str. 'BoNT E Beluga'	EES48201.1, EES49241.1, EES49667.1, EES50826.1
NC_010723	1167942	1193668	*Clostridium botulinum* E3 str. Alaska E43	WP_003369622.1, WP_003372464.1, WP_003374133.1, WP_012450884.1
NZ_AUZC01000009	372653	400188	*Clostridium botulinum* F str. 357	WP_061318234.1, WP_076177532.1, WP_076177533.1, WP_076177535.1
CP000728	862585	890117	*Clostridium botulinum* F str. Langeland	ABS39418.1, ABS40665.1, ABS41517.1, ABS41658.1
NC_009699	862585	890117	*Clostridium botulinum* F str. Langeland	WP_011987704.1, WP_011987705.1, WP_011987706.1, WP_011987708.1
NZ_AP014696	849030	884468	*Clostridium botulinum* str. 111	WP_045538943.1, WP_045538945.1, WP_045538947.1, WP_045538954.1, WP_080316619.1
NZ_CP006903	383803	410010	*Clostridium botulinum* str. 202F	WP_039307656.1, WP_039307660.1, WP_039307662.1, WP_039310025.1
NZ_LIIR01000075	1	16470	*Clostridium botulinum* str. 301-13	WP_079007104.1, WP_079007105.1, WP_079007106.1, WP_079007108.1
EU341305	1	14371	*Clostridium botulinum* str. 5328	ABY56324.1, ABY56325.1, ABY56326.1, ABY56328.1
EU341307	1	14905	*Clostridium botulinum* str. 657Ba	ABY56338.1, ABY56339.1, ABY56340.1, ABY56342.1
NZ_AOSX01000018	1863335	1890754	*Clostridium botulinum* str. Af84	WP_021106990.1, WP_021106991.1, WP_021106992.1, WP_021106994.1
NZ_AOSX01000029	376397	403914	*Clostridium botulinum* str. Af84	WP_012704373.1, WP_021107444.1, WP_021107445.1, WP_021107446.1
CP013701	868030	893649	*Clostridium botulinum* str. AM1195	AUM94647.1, AUM94648.1, AUM94649.1, AUM97440.1
CP013683	1034633	1060989	*Clostridium botulinum* str. AM282	APQ75142.1, APQ75727.1, APQ77619.1, APQ77626.1
LHUM01000023	197191	222917	*Clostridium botulinum* str. ATCC 17786	KOR61896.1, KOR61897.1, KOR61898.1, KOR61899.1

(continued on next page)

Table A.4 – *(continued from previous page)*

Nucleotide ID	Start	End	Species	Protein IDs
LGII01000021	370814	397021	*Clostridium botulinum* str. ATCC 23387	KON12233.1, KON12234.1, KON12235.1, KON13292.1
MWJB01000001	97235	122961	*Clostridium botulinum* str. Bac-01-03998	OSA99876.1, OSA99877.1, OSA99878.1, OSA99879.1
MWJC01000001	145705	171431	*Clostridium botulinum* str. Bac-02-06430	OSA96954.1, OSA96955.1, OSA96956.1, OSA96957.1
ABDP01000023	4825	32334	*Clostridium botulinum* str. Bf	EDT84096.1, EDT84116.1, EDT84123.1, EDT84152.1
NZ_CP014151	896411	923932	*Clostridium botulinum* str. Br-Dura	WP_012300974.1, WP_012704373.1, WP_085322242.1, WP_101528879.1
NZ_LFPC01000018	484763	510489	*Clostridium botulinum* str. CDC 5247	WP_003369622.1, WP_003374133.1, WP_061314990.1, WP_061314992.1
NZ_LFPA01000066	13732	34372	*Clostridium botulinum* str. CDC KA-95B	WP_003369622.1, WP_003372464.1, WP_003374133.1, WP_053342323.1
EU341306	1	14823	*Clostridium botulinum* str. CDC/A3	ABY56331.1, ABY56332.1, ABY56333.1, ABY56335.1
CP006909	213625	239979	*Clostridium botulinum* str. CDC_1436	AJE13190.1, AJE13301.1, AJE13346.1, AJE13355.1
NZ_CP013247	3457966	3485406	*Clostridium botulinum* str. CDC_53174	WP_061318232.1, WP_061318233.1, WP_061318234.1, WP_061318235.1
FJ981696	1	13726	*Clostridium botulinum* str. CDC41370	ACW83602.1, ACW83603.1, ACW83604.1, ACW83606.1
NZ_AZRQ01000040	1	11350	*Clostridium botulinum* str. CDC54085	WP_025775291.1
NZ_AZRQ01000133	1	16032	*Clostridium botulinum* str. CDC54085	WP_025776038.1, WP_025776039.1, WP_025776040.1
NZ_ALYJ01000070	29776	55494	*Clostridium botulinum* str. CDC66177	WP_017352931.1, WP_017352932.1, WP_017352934.1, WP_035789035.1
NZ_JFGA01000189	1	11424	*Clostridium botulinum* str. CDC67190	WP_033050150.1
NZ_JSCF01000006	1	25179	*Clostridium botulinum* str. CFSAN024410	WP_047402751.1, WP_047402754.1, WP_047402756.1, WP_047402757.1
KC516871	9505	35712	*Clostridium botulinum* str. Craig610	AGL45131.1, AGL45132.1, AGL45133.1, AGL45134.1
NZ_CP013705	2097939	2125359	*Clostridium botulinum* str. F160	WP_021106991.1, WP_025776038.1, WP_075141973.1, WP_075141974.1
KT897280	82708	108434	*Clostridium botulinum* str. FI1111E1	ALT05871.1, ALT05872.1, ALT05873.1, ALT05874.1
KT897278	80133	105859	*Clostridium botulinum* str. FWSKR40E1	ALT05666.1, ALT05667.1, ALT05668.1, ALT05669.1
NZ_NPMY01000001	1	16485	*Clostridium botulinum* str. Hazen	WP_003369622.1, WP_003372464.1, WP_100390136.1, WP_100390137.1
KC516872	9505	35712	*Clostridium botulinum* str. HobbsFT10	AGL45171.1, AGL45172.1, AGL45173.1, AGL45174.1
HQ441176	1	13559	*Clostridium botulinum* str. IBCA66-5436	ADU57949.1, ADU57950.1, ADU57951.1, ADU57952.1
KC516868	9505	35712	*Clostridium botulinum* str. IFR 06/001	AGL45011.1, AGL45012.1, AGL45013.1, AGL45014.1
KC516869	9505	35712	*Clostridium botulinum* str. IFR 06/005	AGL45051.1, AGL45052.1, AGL45053.1, AGL45054.1
KC516870	9505	35712	*Clostridium botulinum* str. IFR 06/005	AGL45091.1, AGL45092.1, AGL45093.1, AGL45094.1

(continued on next page)

Table A.4 – *(continued from previous page)*

Nucleotide ID	**Start**	**End**	**Species**	**Protein IDs**
KT897275	77998	103724	*Clostridium botulinum* str. IFR12/29	ALT05361.1, ALT05362.1, ALT05363.1, ALT05364.1
KT897276	73075	98801	*Clostridium botulinum* str. INGR16-02E1	ALT05470.1, ALT05471.1, ALT05472.1, ALT05473.1
NZ_ABDO02000001	798025	825003	*Clostridium botulinum* str. NCTC 2916	WP_003356338.1, WP_003356469.1, WP_003356995.1, WP_003357203.1
KT897277	73075	98801	*Clostridium botulinum* str. ST0210E1	ALT05568.1, ALT05569.1, ALT05570.1, ALT05571.1
KT897279	81269	106995	*Clostridium botulinum* str. SWKR38E2	ALT05768.1, ALT05769.1, ALT05770.1, ALT05771.1
ACOM01000005	1812899	1838625	*Clostridium butyricum* E4 str. BoNT E BL5262	EEP53291.1, EEP53887.1, EEP54138.1, EEP54392.1
NZ_ABDT01000090	74398	100124	*Clostridium butyricum* str. 5521	WP_003369622.1, WP_003372464.1, WP_003409822.1, WP_003409841.1
FOOX01000007	77844	102988	*Desulfallas arcticus* str. DSM17038	SFG63163.1, SFG63181.1, SFG63207.1
MASS01000023	32306	55777	*Desulfosporosinus* sp. BG	ODA40899.1, ODA40900.1, ODA40901.1
MVQL01000134	100534	121761	*Pelotomaculum* sp. PtaB.Bin104	OPX90105.1, OPX90106.1
BGIScaffold9	73024	102439	*Paraclostridium bifermentans*	Pbm-ORFX1, Pbm-ORFX2, Pbm-ORFX3, Pbm-P47
FOKQ01000059	1	15222	*Ruminococcus albus*	SFD29377.1, SFD29411.1, SFD29443.1
PHAA01000014	50956	72701	Firmicutes bacterium HGW-Firmicutes-15	PKM77288.1
RFGF01000135	1	15732	Nitrospirae bacterium	RMH30825.1, RMH30826.1, RMH30827.1
LDPZ01000045	1	14035	*Aureimonas ureilytica* str. NS226	KTQ87253.1, KTQ87254.1
NZ_PPPT01000056	1	14486	*Bradyrhizobium rifense* str. SM32	WP_140977545.1, WP_140977546.1, WP_140977547.1, WP_140977548.1
AAMY01000012	85960	112090	*Nitrobacter* sp. Nb-311A	EAQ35351.1, EAQ35352.1, EAQ35353.1, EAQ35354.1
NC_007406	950829	976957	*Nitrobacter winogradskyi*	WP_011314189.1, WP_011314190.1, WP_011314191.1, WP_011314192.1
CP000115	950829	976957	*Nitrobacter winogradskyi* str. Nb-255	ABA04148.1, ABA04149.1, ABA04150.1, ABA04151.1
SUIN01000010	194	26027	*Mesorhizobium* sp. isolate N.Ce.Tu.015.02.1	Pseudo_SUIN01000010.1_prot_12, TIN49193.1, TIN49194.1, TIN49195.1, TIN49196.1
LMQF01000068	72241	96595	*Rhizobium* sp. Leaf386	KQS83087.1, KQS83088.1, KQS83089.1
LMRG01000028	227687	252041	*Rhizobium* sp. Leaf453	KQT92872.1, KQT92873.1, KQT92874.1
AUSY01000019	68615	93057	*Sinorhizobium* sp. GW3	KSV75547.1, KSV75548.1, KSV75549.1
NZ_AUBC01000023	25134	51553	*Salinarimonas rosea*	WP_029031725.1, WP_029031726.1, WP_029031727.1, WP_084327582.1
LJSX01000015	1	13355	Salinarimonadaceae bacterium HL-109	KPQ10493.1, KPQ10494.1
NZ_FMBM01000001	1428685	1446352	Salinarimonadaceae bacterium HL-109	WP_074444051.1, WP_083204353.1
NZ_UEHD01000031	1	14737	Rhizobiales bacterium	WP_112407914.1, WP_112407915.1, WP_112407916.1
FNEB01000001	71237	95710	*Lutimaribacter saemankumensis* str. DSM28010	SDH93581.1, SDH93615.1, SDH93644.1
NVTR01000010	14795	39287	*Marinosulfonomonas* sp.	PHQ97418.1, PHQ97419.1, PHQ97420.1

(continued on next page)

Table A.4 – *(continued from previous page)*

Nucleotide ID	Start	End	Species	Protein IDs
NZ_LOAS01000024	78286	95359	*Pseudoruegeria sabulilitoris*	WP_068315418.1, WP_068315421.1, WP_068315432.1, WP_082739190.1
QGKU01000038	103607	126664	Rhodobacteraceae bacterium TG-679	PWR02329.1, PWR02330.1
NZ_QOZS01000001	204716	229120	Rhodospirillaceae bacterium SYSU D60015	WP_119417922.1, WP_119417923.1, WP_119417924.1
AFRQ01000031	153885	178436	*Achromobacter insuavis* str. AXX-A	EGP47333.1, EGP47334.1, EGP47335.1
NZ_GL982453	1338131	1362739	*Achromobacter insuavis* str. AXX-A	WP_006391311.1, WP_083827800.1, WP_083827924.1
FKIF01000007	140391	166058	*Bordetella ansorpii* str. H050680373	SAI71066.1, SAI71067.1, SAI71069.1
NZ_NEVP01000004	455202	476641	*Bordetella* genomosp. 5 str. AU14646	WP_094799339.1
NZ_LOWB01000135	8177	22240	*Burkholderia* sp. TSV86	WP_082710301.1, WP_082710302.1
CP030092	1632415	1658579	*Massilia* sp. YMA4	AXA90938.1, AXA90939.1, AXA90940.1, AXA90941.1
NZ_SDAU01000007	1	15672	Enterobacteriaceae bacterium ML5	WP_136374692.1, WP_136374693.1
CAPE01000026	64930	87831	*Erwinia amylovora*	CCP08590.1, CCP08591.1
FR719197	21987	44888	*Erwinia amylovora*	CBX82130.1, CBX82131.1
NC_013961	3322957	3345858	*Erwinia amylovora*	WP_004160292.1, WP_004160293.1
NZ_CAPD01000022	401569	424470	*Erwinia amylovora*	WP_004171016.1, WP_004171017.1
NZ_NQJL01000001	506181	529082	*Erwinia amylovora*	WP_099257802.1, WP_099257804.1
FN434113	3322957	3345858	*Erwinia amylovora* str. CFBP1430	CBA23296.1, CBA23298.1
CAHS01000021	232771	255471	*Erwinia piriflorinigrans* str. CFBP 5888	CCG88684.1, CCG88685.1
NC_012214	3515945	3538978	*Erwinia pyrifoliae*	WP_012669436.1, WP_014539531.1
FN392235	3515882	3538915	*Erwinia pyrifoliae* str. DSM12163	CAY75867.1, CAY75868.1
CP002124	768201	791099	*Erwinia* sp. Ejp617	ADP10405.1, ADP10406.1
NC_017445	768201	791099	*Erwinia* sp. Ejp617	WP_012669437.1, WP_041474302.1
NC_010694	3367112	3390004	*Erwinia tasmaniensis*	WP_012442723.1, WP_012442724.1
CU468135	3367112	3390004	*Erwinia tasmaniensis* str. ET1/99	CAO98072.1, CAO98073.1
OUND01000001	748277	769899	*Arsenophonus* endosymbiont of *Aleurodicus floccissimus*	SPP31287.1
FN545172	1	18646	*Arsenophonus nasoniae*	CBA72105.1, CBA72106.1, Pseudo_FN545172.1_prot_8
NZ_AUCC01000047	1	16297	*Arsenophonus nasoniae* str. DSM 15247	WP_026823089.1, WP_026823090.1, WP_081700659.1
NZ_NGVR01000030	84480	107462	*Proteus columbae*	WP_100160135.1, WP_100160136.1
CVRZ01000013	107917	123129	*Proteus vulgaris*	CRL64650.1, CRL64651.1, CRL64652.1, CRL64653.1
NZ_BCTS01000011	17443	40359	*Serratia ficaria*	WP_061796564.1, WP_061796565.1
NZ_PQGI01000007	4789	27696	*Serratia marcescens*	WP_103681903.1, WP_103681904.1
AAQJ02000001	707439	730371	*Rickettsiella grylli*	EDP46360.1, EDP46936.1
DOTR01000013	19323	44324	*Halomonas campaniensis* isolate UBA11284	HCA01106.1, HCA01107.1, HCA01108.1
CP022286	1048219	1073222	*Halomonas* sp. N3-2A	ASK18679.1, ASK18680.1, ASK18681.1
AFQW01000048	1	20388	*Halomonas* sp. TD01	EGP19388.1, EGP19389.1, EGP19390.1

(continued on next page)

Table A.4 – *(continued from previous page)*

Nucleotide ID	**Start**	**End**	**Species**	**Protein IDs**
NZ_GL949757	168487	193491	*Halomonas* sp. TD01	WP_009723554.1, WP_009723555.1, WP_083817071.1
NZ_NKHL01000056	1	13494	*Pseudomonas bohemica*	WP_110946518.1, WP_110946520.1
PISL01000046	19070	42013	*Pseudomonas hunanensis* str. P11	PKF23524.1, PKF23525.1
NZ_KK214957	7793	30694	*Pseudomonas monteilii*	WP_021784075.1, WP_028698029.1
PJCG01000017	60221	76519	*Pseudomonas monteilii* str. CY06	PKI23724.1, PKI23725.1
RBLH01000001	4923371	4946272	*Pseudomonas plecoglossicida* str. ZKA3	RKS49527.1, RKS49528.1
LDJF01000015	7248	30134	*Pseudomonas putida*	KMY35635.1, KMY35636.1
NC_002947	2265678	2288579	*Pseudomonas putida*	NP_744156.2, NP_744157.1
NZ_JENB01000021	60650	83593	*Pseudomonas putida*	WP_043200446.1, WP_050491898.1
NZ_MINE01000015	2795354	2818255	*Pseudomonas putida*	WP_021784074.1, WP_103443912.1
NZ_NHBB01000022	53155	76098	*Pseudomonas putida*	WP_003247214.1, WP_087535274.1
NZ_PDEH01000001	254282	277225	*Pseudomonas putida*	WP_098089144.1, WP_098089145.1
NZ_AOUR02000091	60177	76432	*Pseudomonas putida* LF54	WP_021784074.1, WP_021784075.1
NZ_ALPV02000001	116469	139412	*Pseudomonas putida* LS46	WP_003247214.1, WP_003247215.1
AE015451	2265678	2288579	*Pseudomonas putida* sp. KT2440	AAN67620.2, AAN67621.1
NNBI01000001	115342	138243	*Pseudomonas putida* str. DPA1	PNG86274.1, PNG86275.1
NBWA01000085	12616	28086	*Pseudomonas putida* str. DZ-F23	ORL61219.1, ORL61220.1
LSUZ01000058	16735	34052	*Pseudomonas putida* str. IN-Sali382	OAS19658.1, OAS19659.1
MING01000019	1078616	1101517	*Pseudomonas putida* str. KH-18-2	POG13814.1, POG13815.1
APBQ01000200	4876	27777	*Pseudomonas putida* str. TRO1	ENY74411.1, ENY74412.1
NHBC01000017	54853	77796	*Pseudomonas putida* str. UV4/95	OUS85781.1, OUS85782.1
NZ_BCAQ01000058	12837	35780	*Pseudomonas* sp. GTC 16473	WP_049586629.1, WP_049586631.1
CP011525	1893002	1915945	*Pseudomonas* sp. JY-Q	ANI33532.1, ANI33533.1
NZ_BCAX01000013	7460	30403	*Pseudomonas* sp. NBRC 111121	WP_049586631.1, WP_060489093.1
NZ_BCBB01000102	36454	51751	*Pseudomonas* sp. NBRC 111125	WP_003247214.1, WP_060539043.1
NZ_BCBP01000041	7595	30538	*Pseudomonas* sp. NBRC 111139	WP_003247214.1, WP_070096344.1
QJOV01000015	7971	30914	*Pseudomonas* sp. SMT-1	PXZ47850.1, PXZ47851.1
CP035952	4078874	4101753	*Pseudomonas* sp. SNU WT1	QBF27609.1, QBF27610.1
NZ_SEIQ01000082	7091	23749	*Pseudomonas* sp. SWI36	WP_020193110.1, WP_129933394.1
CP026332	4204848	4227749	*Pseudomonas* sp. XWY-1	AUZ60466.1, AUZ60467.1
SEZV01000011	104102	128449	*Pseudomonas syringae* str. MWU 13-30316	TFZ34284.1, TFZ34285.1, TFZ34286.1
NZ_PIFD01000008	1	16185	*Vibrio splendidus*	WP_108185305.1, WP_108185306.1
NZ_PIFN01000001	1	17845	*Vibrio splendidus*	WP_108192224.1, WP_108192225.1
NZ_PIFX01000014	1	16185	*Vibrio splendidus*	WP_108192224.1, WP_108202954.1
NC_030990	4413562	4437678	*Fusarium oxysporum*	XP_018236094.1, XP_018236095.1
KB730215	78110	101577	*Fusarium oxysporum* f. sp. cubense race 1	ENH69816.1, ENH69817.1

(continued on next page)

Table A.4 – *(continued from previous page)*

Nucleotide ID	**Start**	**End**	**Species**	**Protein IDs**
DS231697	3533457	3557573	*Fusarium oxysporum* f. sp. lycopersici 4287	KNA98048.1, KNA98049.1
JH659331	43567	67683	*Fusarium oxysporum* f. sp. melonis 26406	EXK40938.1, EXK40939.1
LN649229	430582	452068	*Fusarium venenatum* str. A3/5	CEI63638.1
NBIV01000071	107302	135039	G*racilariopsis chorda* isolate SKKU-2015	PXF45067.1, PXF45068.1, PXF45069.1

Appendix B

Supplementary Material: Chapter 3

B.1 Supplementary Table 5

Table B.1: Summary of I-TASSER structural modelling results. Each diphtheria toxin-like protein was separated into its domains according to its alignment with the crystal structure of diphtheria toxin, which was also used to evaluate the model's RMSD (PDB identifier 1MDT). The I-TASSER C-score varies from -5 to 2, representing low to high confidence in the model. The domain boundaries for each subsection are indicated in the table.
*The C domain (NZ_BAGZ01000024.1, 41687-42285) and the N-terminus of the T domain (NZ_BAGZ01000024.1, 41462-41641) for the sequence from *Austwickia chelonae* were translated from a pseudogene. The T domain was concatenated from the translated pseudogene and residues 2-87 of WP_040322835.1.

Species	**Model**	**C domain**		**T domain**		**R domain**	
		C-score	RMSD	C-score	RMSD	C-score	RMSD
Auswickia chelonae		Pseudogene*		WP_040322835.1*, 2-87		WP_040322835.1, 88-274	
	model1	0.28	1.112	0.97	0.866	1.05	1.173
	model2	0.62	0.617	-3.52	5.298	-5	1.13
	model3	0.26	0.614	-4.03	4.585	-5	1.184
	model4	0.44	1.155	-4.76	12.279	-4.94	1.174
	model5	-1.8	0.582	-5	14.604	-5	1.217
Streptomyces sp. MBT76		WP_079110321.1, 50-251		WP_079110321.1, 252-397		WP_079110321.1, 428-549	
	model1	-0.47	1.112	0.55	1.202	-3.56	13.764
	model2	-0.49	1.072	-3.87	16.269	-4.14	13.482
	model3	-1.44	1.592	-4.52	15.196	-4.6	13.723
	model4	-1.6	1.211	-4.52	8.669	-4.85	14.324
	model5	-3.6	2.636	-4.58	7.573	-5	14.452

(continued on next page)

Table B.1 – *(continued from previous page)*

Species	Model	C domain		T domain		R domain	
		C-score	RMSD	C-score	RMSD	C-score	RMSD
Streptosporangium nondiastaticum		PSJ28985.1, 9-230		PSJ28985.1, 226-371		PSJ28985.1, 372-523	
	model1	-1.02	1.162	0.62	1.032	-2.74	15.253
	model2	-2.42	1.582	-3.44	14.84	-4.97	15.839
	model3	-2.91	1.238	-4.64	13.396	-4.42	15.143
	model4	-1.93	1.086	-4.14	12.725	-5	14.674
	model5	-2.6	1.972	-4.14	13.172	-5	17.923
Streptomyces verticillatus		WP_07859863.1, 24-225		WP_07859863.1, 226-371		WP_07859863.1, 402-523	
	model1	-0.35	1.472	0.78	0.721	-4.14	11.901
	model2	-1.19	1.069	-4.39	10.76	-4.57	12.818
	model3	-0.55	1.361	-3.73	4.944	-4.64	12.267
	model4	-2.83	0.909	-4.39	11.017	-4.36	11.81
	model5	-1.98	0.932	-4	4.739	-4.77	11.501
Streptomyces sp. TLT_053		SDT83331.1, 1-192		SDT83331.1, 193-339		SDT83331.1, 396-495	
	model1	0.29	1.003	0.25	0.727	-3.84	8.88
	model2	-1.78	0.981	-3.78	11.807	-4.51	6.42
	model3	-3.17	1.773	-3.94	11.462	-5	11.232
	model4	-4.59	0.905	-4.66	8.699	-4.77	6.73
	model5	-4.21	13.372	-5	12.315	-5	7.741
Streptomyces albireticuli		WP_095582082.1, 100-329		WP_095582082.1, 330-481		WP_095582082.1, 510-693	
	model1	-2.15	1.392	0.26	1.191	-2.46	1.125
	model2	-4.4	3.229	-4.31	6.075	-4.4	10.061
	model3	-5	3.455	-4.76	6.437	-4.09	2.509
	model4	-3.67	3.246	-5	8.458	-2.9	1.137
	model5	-3.58	1.505	-5	8.318	-5	7.012
Seinonella peptonophila		WP_073156187.1, 1-245		WP_073156187.1, 246-392		WP_073156187.1, 421-606	
	model1	-1.43	1.953	0.32	0.755	-2.7	15.619
	model2	-4.01	10.512	-4.08	9.806	-5	1.875
	model3	-4.02	13.268	-4.45	6.098	-5	13.803
	model4	-4.24	12.012	-5	6.672	-5	10.658
	model5	-4.61	3.745	-5	14.817	-5	12.406

B.2 Supplementary Table 6

Table B.2: Presence/absence of key DT functional sites among DT homologs. In order, the noted sites perform these functions in DT (and references are provided below): H21 maintains the steric structure of the catalytic site; T23 forms hydrogen bonds with the adenosine ribose; Y27 participates in the active site; K51 and G52 participate in the active site loop; Y54 and Y65 bind NAD; E148 is the key catalytic residue; C186 and C201 form the disulfide bond linking A and B fragments; RxxR is the furin cleavage site; P345, E349, and D352 participate in membrane insertion and pore formation.

Species	**Accession**	**Key functional sites**					
Corynebacterium diphtheriae	1MDT	H21	T23	Y27	K51	G52	Y54
Corynebacterium ulcerans	AAN28948.1	H46	T48	Y52	K56	G77	Y79
Austwickia chelonae	Pseudogene*	H33	A35	S39	K63	G64	Y66
Streptomyces sp. MBT76	WP_079110321.1	H81	Y83	H87	K109	G110	Y112
Streptosporangium nondiastaticum	PSJ28985.1	H47	Y49	H53	K75	G76	Y78
Streptomyces roseoverticillatus	WP_078659863.1	H55	Y57	H64	K83	G84	Y86
Streptomyces sp. TLI_053	SDT83331.1	H17	Y19	N23	N42	G43	Y45
Streptomyces albireticuli	WP_095582082.1	R129	V130	E134	K159	A160	Y162
Seinonella peptonophila	WP_073156187.1	R21	V23	G27	Q60	H61	Y63

Corynebacterium diphtheriae	1MDT	Y65	E148	C186	190-RVRR-194	C201
Corynebacterium ulcerans	AAN28948.1	Y90	E173	C211	215-RVRR-218	C226
Austwickia chelonae	Pseudogene*	Y77	E160	C199*	206-RAKR-218	C215*
Streptomyces sp. MBT76	WP_079110321.1	Y123	E160	C199*	248-RAKR-251	C256*
Streptosporangium nondiastaticum	PSJ28985.1	Y89	E170	C208*	214-RVKR-217	C222*

(continued on next page)

Table B.2 – *(continued from previous page)*

Species	**Accession**	**Key functional sites**				
Streptomyces roseoverticillatus	WP_078659863.1	Y97	E178	C216*	222-RVKR-225	C230*
Streptomyces sp. TLI_053	SDT83331.1	Y56	E141	C184	189-RAKR-192	C197*
Streptomyces albireticuli	WP_095582082.1	-	E266	C323	-	C337*
Seinonella peptonophila	WP_073156187.1	Y74	E176	C239	-	C251*

Corynebacterium diphtheriae	1MDT	P345	E349	D352
Corynebacterium ulcerans	AAN28948.1	P370	E374	D377
Austwickia chelonae	Pseudogene*	P355	E359	362
Streptomyces sp. MBT76	WP_079110321.1	P391	E395	D398
Streptosporangium nondiastaticum	PSJ28985.1	P357	E361	D372
Streptomyces roseoverticillatus	WP_078659863.1	P365	E366	D372
Streptomyces sp. TLI_053	SDT83331.1	P333	E337	D340
Streptomyces albireticuli	WP_095582082.1	P475	E479	D485
Seinonella peptonophila	WP_073156187.1	P386	E390	E396

Appendix C

Supplementary Material: Chapter 4

C.1 Supplementary Table 7

Table C.1: Top 100 Pfam domains correlated with the LCT translocase. Proteome-wide Pfam annotations were retrieved from the Genome Taxonomy Database. Domains associated with LCTs are bolded.

Pfam ID	Clan ID	Clan Name	Pfam Name	Pfam Description	Correlation
PF12920.2			**TcdA_TcdB_pore**	**TcdA/TcdB pore forming domain**	**1**
PF12919.2	**CL0110**	**GT-A**	**TcdA_TcdB**	**TcdA/TcdB catalytic glycosyltransferase domain**	**0.506869721**
PF03538.9			VRP1	Salmonella virulence plasmid 28.1kDa A protein	0.418404007
PF06958.7	CL0446	Bacteriocin_TLN	Pyocin_S	S-type Pyocin	0.407896036
PF14564.1			Membrane_bind	Membrane binding	0.345611493
PF03245.8	CL0331	EpsM	Phage_lysis	Bacteriophage Rz lysis protein	0.345518375
PF02413.12	CL0348	Phage_tail	Caudo_TAP	Caudovirales tail fibre assembly protein, lambda gpK	0.345254532
PF07865.6			DUF1652	Protein of unknown function (DUF1652)	0.333345129
PF13503.1			DUF4123	Domain of unknown function (DUF4123)	0.328320429
PF12255.3			TcdB_toxin_midC	Insecticide toxin TcdB middle/C-terminal region	0.325058121
PF09000.5			Cytotoxic	Cytotoxic	0.322464262
PF07119.7			DUF1375	Protein of unknown function (DUF1375)	0.310421646

(continued on next page)

Table C.1 – *(continued from previous page)*

Pfam ID	**Clan ID**	**Clan Name**	**Pfam Name**	**Pfam Description**	**Correlation**
PF11462.3	CL0266	PH	DUF3203	Protein of unknown function (DUF3203)	0.307089453
PF09909.4			DUF2138	Uncharacterized protein conserved in bacteria (DUF2138)	0.30691289
PF06649.7			DUF1161	Protein of unknown function (DUF1161)	0.301704946
PF10109.4	CL0567	Phage_TACs	Phage_TAC_7	Phage tail assembly chaperone proteins, E, or 41 or 14	0.297103359
PF09634.5			DUF2025	Protein of unknown function (DUF2025)	0.294250433
PF07395.6	CL0257	Acetyltrans	Mig-14	Mig-14	0.294003761
PF05954.6	CL0504	Phage_barrel	Phage_GPD	Phage late control gene D protein (GPD)	0.292797043
PF09498.5			DUF2388	Protein of unknown function (DUF2388)	0.287317097
PF03406.8			Phage_fiber_2	Phage tail fibre repeat	0.283954868
PF04676.9			CwfJ_C_2	Protein similar to CwfJ C-terminus 2	0.283373472
PF04717.7			Phage_base_V	Type VI secretion system, phage-baseplate injector	0.282900265
PF04958.7	CL0257	Acetyltrans	AstA	Arginine N-succinyltransferase beta subunit	0.282377849
PF09906.4			DUF2135	Uncharacterized protein conserved in bacteria (DUF2135)	0.282092198
PF03513.9			Cloacin_immun	Cloacin immunity protein	0.281446515
PF06474.7	CL0421	LppaM	MLTD_N	MltD lipid attachment motif	0.281316062
PF11647.3			**MLD**	**Membrane Localization Domain**	**0.27674684**
PF03502.8	CL0193	MBB	Channel_Tsx	Nucleoside-specific channel-forming protein, Tsx	0.273576354
PF12306.3	CL0026	CU_oxidase	PixA	Inclusion body protein	0.272185422
PF01320.13			Colicin_Pyocin	Colicin immunity protein / pyocin immunity protein	0.27177386
PF10062.4			DUF2300	Predicted secreted protein (DUF2300)	0.269069383
PF12021.3			DUF3509	Protein of unknown function (DUF3509)	0.265764007
PF11713.3	**CL0093**	**Peptidase_CD**	**Peptidase_C80**	**Peptidase C80 family**	**0.265413307**
PF08682.5	CL0236	PDDEXK	DUF1780	Putative endonuclease, protein of unknown function (DUF1780)	0.26432034
PF10976.3			DUF2790	Protein of unknown function (DUF2790)	0.263663893

(continued on next page)

Table C.1 – *(continued from previous page)*

Pfam ID	Clan ID	Clan Name	Pfam Name	Pfam Description	Correlation
PF06611.7			DUF1145	Protein of unknown function (DUF1145)	0.260882618
PF05581.7	NA	NA	NA	NA	0.260763361
PF07634.6			RtxA	RtxA repeat	0.260030609
PF03543.9	CL0125	Peptidase_CA	Peptidase_C58	Yersinia/Haemophilus virulence surface antigen	0.259620448
PF10144.4	CL0165	Cache	SMP_2	Bacterial virulence factor haemolysin	0.258000858
PF13652.1			QSregVF	Putative quorum-sensing-regulated virulence factor	0.255362
PF05488.8			PAAR_motif	PAAR motif	0.252857044
PF12571.3			DUF3751	Phage tail-collar fibre protein	0.251902708
PF03873.8			RseA_C	Anti sigma-E protein RseA, C-terminal domain	0.251688075
PF06790.6			UPF0259	Uncharacterised protein family (UPF0259)	0.247096601
PF05736.6	CL0193	MBB	OprF	OprF membrane domain	0.244047727
PF13693.1	CL0123	HTH	HTH_35	Winged helix-turn-helix DNA-binding	0.241707199
PF04320.9			DUF469	Protein with unknown function (DUF469)	0.240266819
PF03889.8			ArfA	Alternative ribosome-rescue factor A	0.237133137
PF05638.7			T6SS_HCP	Type VI secretion system effector, Hcp	0.236267323
PF11293.3			DUF3094	Protein of unknown function (DUF3094)	0.235268321
PF05947.7			T6SS_TssF	Type VI secretion system, TssF	0.234957281
PF12633.2	CL0260	NTP_transf	Adenyl_cycl_N	Adenylate cyclase NT domain	0.234726283
PF09621.5			LcrR	Type III secretion system regulator (LcrR)	0.234023735
PF06812.7			ImpA_N	ImpA, N-terminal, type VI secretion system	0.233925807
PF06942.7	CL0420	GlpM-like	GlpM	GlpM protein	0.23289032
PF01295.13			Adenylate_cycl	Adenylate cyclase, class-I	0.23274483
PF06672.6	CL0125	Peptidase_CA	DUF1175	Protein of unknown function (DUF1175)	0.232508376
PF04984.9			Phage_sheath_1	Phage tail sheath protein subtilisin-like domain	0.232025929
PF04965.9			GPW_gp25	Gene 25-like lysozyme	0.231901984
PF04888.7			SseC	Secretion system effector C (SseC) like family	0.231616006
PF07216.7			LcrG	LcrG protein	0.231505947

(continued on next page)

Table C.1 – *(continued from previous page)*

Pfam ID	Clan ID	Clan Name	Pfam Name	Pfam Description	Correlation
PF04792.7			LcrV	V antigen (LcrV) protein	0.231168544
PF09025.5	CL0646	T3SS	T3SS_needle_reg	YopR, type III needle-polymerisation regulator	0.230224779
PF09477.5			Type_III_YscG	Bacterial type II secretion system chaperone protein (type_III_yscG)	0.22961717
PF11286.3			DUF3087	Protein of unknown function (DUF3087)	0.227387565
PF06693.6			DUF1190	Protein of unknown function (DUF1190)	0.227077986
PF09619.5			YscW	Type III secretion system lipoprotein chaperone (YscW)	0.225796548
PF03573.8	CL0193	MBB	OprD	outer membrane porin, OprD family	0.224232333
PF11862.3			DUF3382	Domain of unknown function (DUF3382)	0.219804078
PF05844.7			YopD	YopD protein	0.217519462
PF10948.3			DUF2635	Protein of unknown function (DUF2635)	0.216077945
PF11661.3			DUF2986	Protein of unknown function (DUF2986)	0.215827375
PF06450.7	CL0182	IT	NhaB	Bacterial Na+/H+ antiporter B (NhaB)	0.215760319
PF07023.7			DUF1315	Protein of unknown function (DUF1315)	0.214862846
PF08468.6			MTS_N	Methyltransferase small domain N-terminal	0.214854346
PF07409.7			GP46	Phage protein GP46	0.213479959
PF14567.1	CL0526	SUKH	SUKH_5	SMI1-KNR4 cell-wall	0.213253186
PF09904.4	CL0123	HTH	HTH_43	Winged helix-turn helix	0.213153058
PF05597.6			Phasin	Poly(hydroxyalcanoate) granule associated protein (phasin)	0.213144599
PF10618.4			Tail_tube	Phage tail tube protein	0.212457126
PF09392.5			T3SS_needle_F	Type III secretion needle MxiH, YscF, SsaG, EprI, PscF, EscF	0.211573532
PF12790.2	CL0287	Transthyretin	T6SS-SciN	Type VI secretion lipoprotein, VasD, EvfM, TssJ, VC_A0113	0.211571946
PF05106.7	CL0564	Holin-III	Phage_holin_3_1	Phage holin family (Lysis protein S)	0.211297161
PF04985.9	CL0569	Phage_TTPs	Phage_tube	Phage tail tube protein FII	0.211096297
PF07201.6	CL0646	T3SS	HrpJ	HrpJ-like domain	0.210468182

(continued on next page)

Table C.1 – *(continued from previous page)*

Pfam ID	Clan ID	Clan Name	Pfam Name	Pfam Description	Correlation
PF06890.7			Phage_Mu_Gp45	Bacteriophage Mu Gp45 protein	0.210308896
PF05936.7			T6SS_VasE	Bacterial Type VI secretion, VC_A0110, EvfL, ImpJ, VasE	0.210257255
PF06476.7			DUF1090	Protein of unknown function (DUF1090)	0.209734486
PF06295.7			DUF1043	Protein of unknown function (DUF1043)	0.209651638
PF11340.3			DUF3142	Protein of unknown function (DUF3142)	0.209617732
PF06995.6			Phage_P2_GpU	Phage P2 GpU	0.209609326
PF11354.3			DUF3156	Protein of unknown function (DUF3156)	0.2095852
PF10679.4			DUF2491	Protein of unknown function (DUF2491)	0.208922421
PF03865.8	CL0193	MBB	ShlB	Haemolysin secretion/activation protein ShlB/FhaC/HecB	0.20888078
PF03974.8			Ecotin	Ecotin	0.208388611
PF06794.7			UPF0270	Uncharacterised protein family (UPF0270)	0.208022718
PF07157.7			DNA_circ_N	DNA circularisation protein N-terminus	0.208012943
PF09500.5	CL0050	HotDog	YiiD_C	Putative thioesterase (yiiD_Cterm)	0.207210421

C.2 Supplementary Figure 1

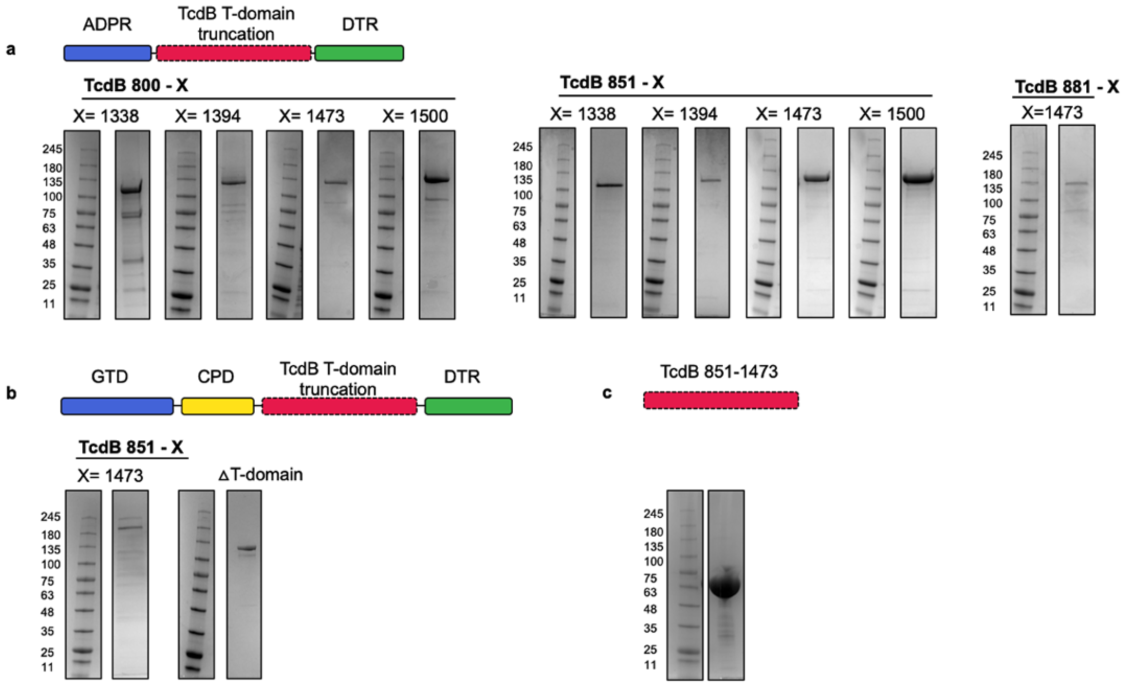


Figure C.1: SDS-PAGE of purified TcdB constructs. (a) ADPR-[truncated TcdB domain]-DTR chimeras, with the truncated T-domain indicated in the figure. (b) GTD-CPD-TcdB(851-1473)-DTR and ΔT-domain. (c) TcdB 851-1473.

C.3 Supplementary Figure 2

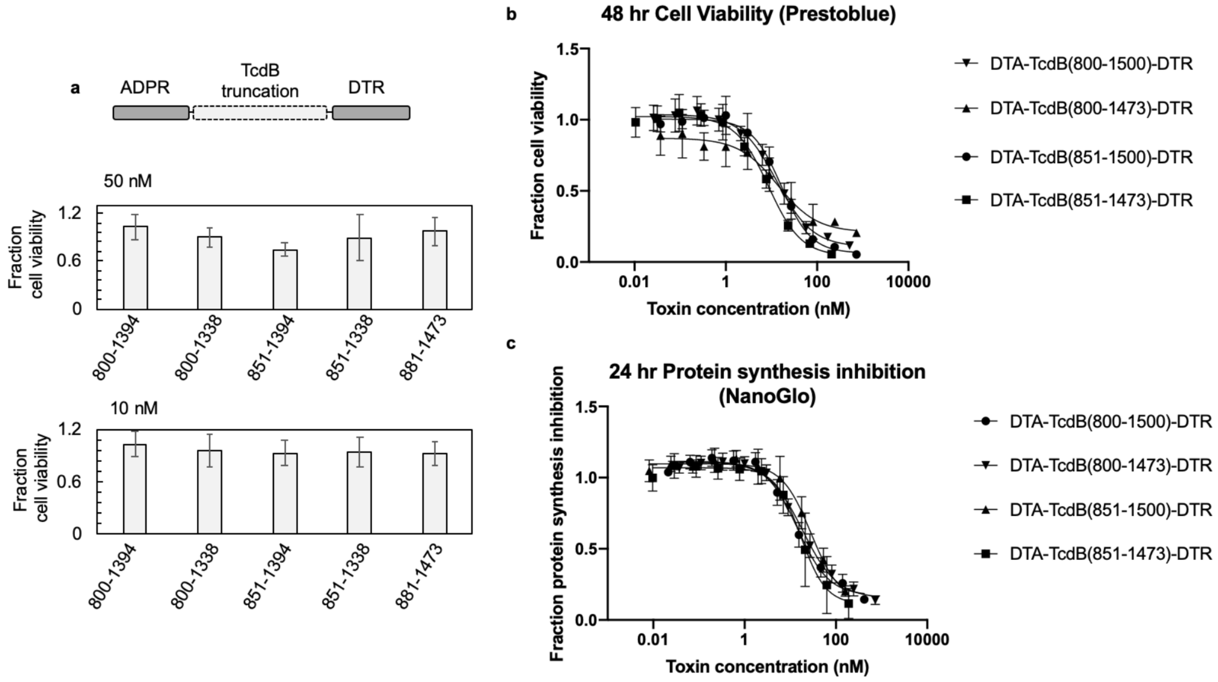


Figure C.2: -
DTR constructs]Cell viability and protein synthesis inhibition ADPR-[truncated TcdB T-domain]-DTR constructs. (a) Fraction cell viability of all non-toxic constructs, tested at both 50nM and 10 nM. The variable TcdB truncation is indicated on the x-axis. (b) Fraction cell viability and (c) fraction protein synthesis inhibition curves of toxic ADPR-[truncated TcdB T-domain]-DTR constructs.

C.4 Supplementary Figure 3

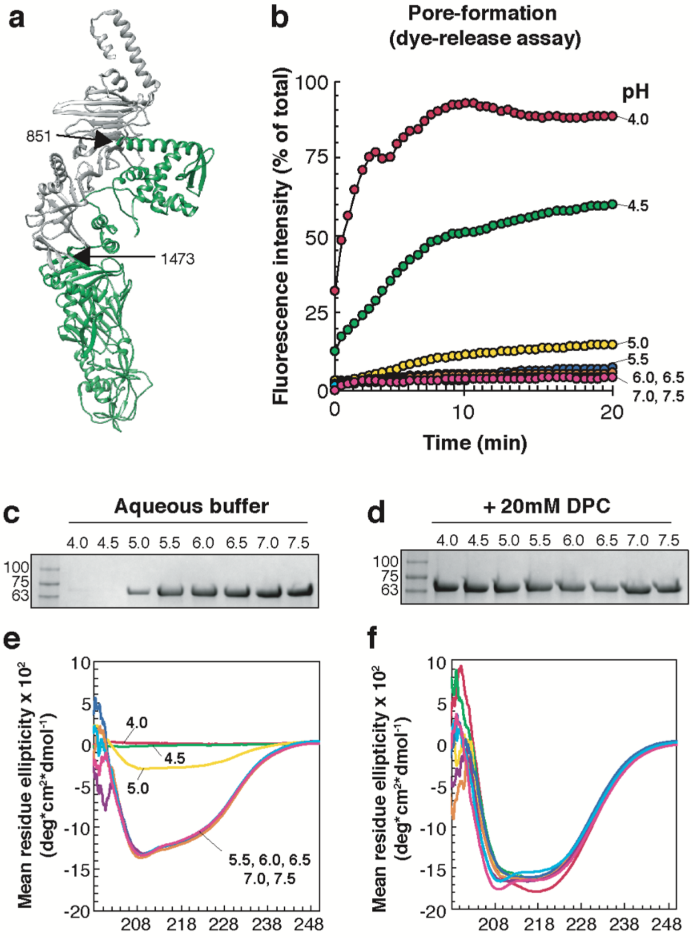


Figure C.3: TcdB 851-1473 retains its form and function. (a) TcdA T-domain (pdb: 4R04), with 851-1473 colored in green. (b) TcdB 851-1473 induced dye release quantification after 20 minutes and (c) a kinetic trace from HPTS/DPX loaded liposomes, from pH 4.0 to pH 7.0 in 0.5 pH increments, with coloring as follows: pH 4.0 (red), pH 4.5 (green), pH 5.0 (yellow), pH 5.5 (blue), pH 6.0 (orange), pH 6.5 (purple), pH 7.0 (aqua), pH 7.5 (magenta) (N=3). (d) Stability of TcdB 851-1473 in aqueous buffer and with (e) 20 mM DPC from pH 4.0 to pH 7.5. (f) Circular dichroism (CD) spectroscopy of TcdB 851-1473 from pH 4.0 to pH 7.5 in 0.5 pH increments (N=3) in aqueous buffer and with (g) 20 mM DPC (N=3). Coloring for each pH is the same as in (b) and (c).