Comparative analysis and visualization of microbial gene neighborhoods: applications to pathogen genomics

by

Xin Wei

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, 2023

© Xin Wei 2023

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. Jesse Shapiro Associate Professor, Department of Microbiology and Immunology, McGill University
Supervisor:	Dr. Andrew Doxey Associate Professor, Department of Biology, University of Waterloo
Internal Member:	Dr. Gabriel Moreno-Hagelsieb Professor, Department of Biology, Wilfrid Laurier University, and Adjunct Associate Professor, Department of Biology, University of Waterloo
Internal Member:	Dr. Trevor Charles Professor, Department of Biology, University of Waterloo
Internal-External Member:	Dr. Valerie Ward Assistant Professor, Department of Chemical Engineering, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this 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.

Abstract

Gene neighborhoods are clusters of genes that are encoded together within the same genomic region, and may share similarity in expression and/or function. Bioinformatic analysis of gene neighborhoods is a powerful approach that provides insights into genome structure, function, and evolution. Despite its importance, existing bioinformatics tools are limited in terms of their ability to analyze and compare gene neighborhoods, especially on a large-scale. In this thesis, I present a novel software tool called AnnoView for large-scale exploration and analysis of microbial gene neighborhoods. I then demonstrate its use by exploring several different bacterial pathogens and their virulence factors, and using genomic context analysis to gain novel insights into pathogen evolution and genome function.

AnnoView is a new gene neighborhood analysis tool that facilitates interactive exploration of microbial gene neighborhoods from 30,238 bacterial genomes and 1,672 archaeal genomes with pre-computed functional annotations derived from KEGG, Pfam, and TIGRFAM. Users also have the flexibility to upload custom datasets in various formats for gene neighborhood visualization.

As a first application of AnnoView, I analyzed the adenylate isopentenyl transferase (*IPT*) gene in bacteria, which encodes a cytokinin-producing enzyme found in plant pathogenic as well as plant growth-promoting bacteria (PGPB). To understand how this gene may function differently between pathogens and PGPB, AnnoView was used to explore and compare its genomic contexts across bacteria. Analysis revealed numerous distinguishing features, including the tendency for pathogen-associated adenylate *IPT* genes tend to occur in predicted virulence loci.

As a second case study utilizing AnnoView, two novel gene clusters were discovered containing putative clostridial neurotoxin (CNT) genes within the genomes of *Paraclostridium ghonii* and *Bacillus toyonensis*. Both gene clusters were analyzed using AnnoView and found to contain unique features not present in other CNT-containing gene clusters. In particular, analysis of the *P. ghonii* toxin gene neighborhood revealed the identification of nearby genes indicating that *P. ghonii* toxin may have a specificity toward insect hosts. This prediction was experimentally validated in collaborative work, revealing *P. ghonii* toxin as a novel insecticidal neurotoxin.

Lastly, as a third case study, I used AnnoView to perform a comparative genomic analysis of a putatively novel species of *Clostridium* known as "clade X" identified from ancient DNA. The identification of conserved, unique gene neighborhood patterns within clade X but absent from closely related *Clostridium* genomes reinforces previous claims that clade X represents a distinct species.

Ultimately, comparative analysis of gene neighborhoods using AnnoView and large-scale genomic information from existing databases (GTDB, AnnoTree, NCBI) is a powerful approach for for microbial genomics studies. Comparative analysis of gene neighborhoods in pathogenic genomes helps to expand our knowledge of virulence factors and pathogen-associated genomic traits, contributing to the characteristic of pathogen biology including even the identification of host species.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Andrew Doxey, for his unwavering support, encouragement, and mentorship. He has been a continuous source of inspiration, guiding the evolution and refinement of my research ideas. This thesis would not have happened without him. I am thankful to my committee members, Dr. Gabriel Moreno-Hagelsieb and Dr. Trevor Charles, for their invaluable insights, expertise, and consistent guidance.

I extend my sincere appreciation to my collaborators: Dr. Bernard Glick for his visionary idea and thoughtful feedback on the chapter about isopentenyl transferase; Dr. Briallen Lobb, Huagang Tan, Zijing Wu, and William Zhen, for their combined efforts on the chapter about AnnoView; Dr. Min Dong's lab for their experimental validation, and Dr. Michael Mansfield for his contribution to the chapter on clostridial neurotoxins.

Additional thanks to all past and current Doxey lab members for always being there in times of need. The casual gatherings after our lab meetings always added a memorable touch. I thank all my friends for bringing vibrant colors to my life in a foreign land. To every educator who has impacted my journey, your wisdom has sculpted my path, shaping who I am today.

Lastly, I would like to thank my family: my parents and my brother, for their unconditional love and support throughout this journey, always believing in me even when I doubted myself.

Dedication

To all who guided and supported me on this journey, to individuals I've met and the experiences shared, to my family who always stand by me, and to my own perseverance and determination for carrying me this far.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Abstract iv

Acknowledgements vi

Dedication vii

List of Figures xi

List of Tables xiii

1 Introduction 1

1.1 Bacterial and archaeal genomics, comparative genomics, and gene neighborhood analysis . . . 1

1.2 Gene neighborhood conservation patterns . . . 3

1.3 Methods for annotating gene neighborhoods . . . 6

1.4 Methods for comparing neighborhoods . . . 9

1.5 Problems with existing gene neighborhood tools . . . 14

1.6 Gene neighborhood analysis in pathogen genomics . . . 15

1.7 Thesis outline . . . 16

2 AnnoView: Gene neighborhood browser for bacterial and archaeal genomes 18

2.1 Introduction . . . 18

2.2 Methods . . . 20

2.3 The AnnoView application . . . 23

2.4 Example workflows . . . 26

2.5 Discussion . . . 30

3 Comparative analysis of adenylate *IPT* genes 32

3.1 Introduction . . . 32

3.2 Methods . . . 37

3.3 Results . . . 41

3.4 Discussion . . . 59

4 Novel Clostridial neurotoxins identified outside *Clostridium* 63

4.1 Introduction . . . 64

4.2 Methods . . . 66

4.3 Results . . . 69

4.4 Experimental characterization of PG toxins . . . 84

4.5 Discussion . . . 89

5 Comparative genomics of a novel *Clostridium* sp. X identified from archaeological samples 92

5.1 Introduction . . . 93

5.2 Methods . . . 96

5.3 Results . . . 97

5.4 Discussion . . . 108

6 Discussion **111**

6.1 Summary of findings 112

6.1.1 Identification of *IPT* gene clusters 112

6.1.2 Identification of two novel clostridial neurotoxins 113

6.1.3 Identification of genomic traits unique to *Clostridium* sp. X 115

6.2 Limitations 116

6.3 Final remarks 121

Letter of Copyright Permission **122**

References **123**

APPENDICES **159**

Supplementary Tables **160**

List of Figures

1.1 The number of sequence records in GenBank and WGS databases 2
1.2 Different genomic context methods 4

2.1 The architecture of AnnoView 21
2.2 Pseudocode for the clustering algorithm 22
2.3 AnnoView gene neighborhood browser functionalities 25
2.4 Search in the AnnoTree database 27
2.5 Upload gene neighborhood dataset to AnnoView 29
2.6 Genome structure of *Betacoronavirus* 30

3.1 General pathway for CK biosynthesis 36
3.2 Phylogenetic distribution of IPT proteins 42
3.3 Phylogenetic tree of bacterial IPT proteins 44
3.4 Phylogenetic tree of IPTs from bacteria and eukaryotes 46
3.5 Comparison of codon usage patterns between *IPT* genes from the pathogen-associated and PGPB-associated clades 48
3.6 Genomic context analysis of bacterial *IPT* genes 51
3.7 Pathogen-associated *IPT* genes tend to reside within predicted virulence loci 53
3.8 Boxplot of the PathFams scores 54
3.9 Selected *IPT* genomic neighborhoods 55
3.10 Analysis of *IPT* promoter sequences 58

4.1 Identification of BoNT-like genes in genomes of two *Paeniclostridium ghonii* strains . . . 70
4.2 Structural and motif analysis of PG-toxins in relation to BoNT . . . 74
4.3 Phylogeny of the BoNT family including PG-toxin . . . 76
4.4 Phylogenetic clustering of *P. ghonii* genomes . . . 78
4.5 Pairwise sequence similarity of BoNT toxins . . . 79
4.6 Visualization of a newly identified BoNT-like gene cluster in *B. toyonensis* 80
4.7 Alignment of *B. toyonensis* BoNT-like toxin with other subtypes . . . 80
4.8 Phylogenetic tree of BoNTs, BTNT, and related toxins . . . 82
4.9 Phylogenetic tree of BoNTs, PG toxins, BTNT and related toxins . . . 83
4.10 PG1 and PG2 injection in adult *Drosophila* induce paralysis and death . . . 86
4.11 Characterization of PG1 and PG2 toxins . . . 89

5.1 Previous work . . . 95
5.2 Venn diagram . . . 98
5.3 The number of clade X specific genes . . . 99
5.4 Unique gene neighborhood structures in clade X and Y acBins . . . 101
5.5 Gene neighborhoods surrounding the *repA* genes . . . 103
5.6 Gene neighborhoods surrounding the *colT* genes . . . 104
5.7 Mapped read coverage to the reference *C. tetani* E88 plasmid and *tent* genes 105
5.8 Gene neighborhoods surrounding the *tent* genes . . . 106
5.9 Identification of a putative deletion event . . . 108
5.10 Evolutionary models of *tent* gene . . . 110

List of Tables

1.1 Bioinformatics tools for gene neighborhood analysis 11

1.2 Summary of thesis chapters 17

5.1 Unique orthologous groups 100

1 AnnoTree results 160

2 IPT protein sequences 168

3 Metadata table of IPT-containing bacteria 173

4 GTDB species name and NCBI species name of the IPT-containing genomes 178

5 Pfam annotations detected five or more than five times in the *IPT* gene neighborhood 180

Chapter 1

Introduction

> Gene order analysis, i.e. the comparative study of genomic sequences that have been scrambled by genome rearrangement events, confers detailed insights into the species' evolutionary past helping to resolve their phylogenetic relationships, the structural organization of ancestral genomes, and functional dependencies of genes, only to name a few applications.
>
> *Comparative Genomics: Methods and Protocols*
> Setubal, J. C., Stoye, J., & Stadler, P. F.[1]

1.1 Bacterial and archaeal genomics, comparative genomics, and gene neighborhood analysis

A genome represents the complete set of DNA sequences of an organism, and is blueprint for the building blocks (i.e., genes) of cells and their regulation. Genomic studies aim to characterize genome function, structure, and evolution, which is becoming increasingly possible through the constant addition of new species to genomic databases. Over recent decades, the surge in genomic data submissions has greatly increased our understanding of prokaryotic genomes. At the time of writing, the whole genome shotgun (WGS) database holds data records for over 2 billion sequences (Figure 1.1). Meanwhile, GenBank contains data for nearly 600,000 distinct prokaryotic organisms (https://ncbi.nlm.nih.gov/genome/browse#!/overview/).

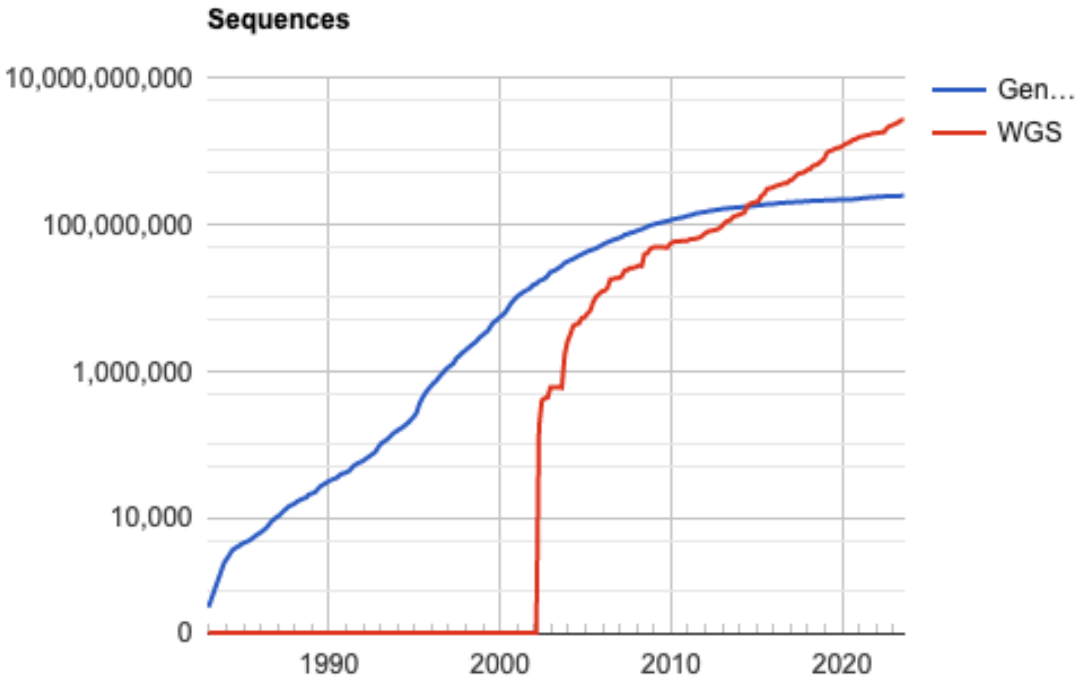


Figure 1.1: The number of sequence records in GenBank and Whole genome shotgun (WGS) databases. Data is obtained from NCBI GenBank (www.ncbi.nlm.nih.gov/genbank/statistics/).

Comparative genomics is a field of study that compares genome structure and function of different organisms to understand their evolutionary relationship, as well as to identify similar or unique genomic features.[1,2] Features that are often examined in comparative genomics include DNA sequence comparison using whole-genome alignment, gene content comparison, pan-genome analysis, functional comparison, and synteny analysis.[1] Synteny analysis refers to the identification of genetic loci conservation pattern across different genomes. In this thesis, gene neighborhood analysis is defined as synteny analysis of gene orders and arrangement on the genomic regions that housing the genes of interests. A gene neighborhood, often termed a "genomic neighborhood" or a "genomic context", denotes the arrangement of genomic features within a specific genomic region, typically surrounding one or more genes of interest. Gene neighborhood analysis involves visualizing and comparing the organization of genes, genetic elements, and other genomic features across different species, as well as identifying conservation patterns and evolutionary changes in gene order. Gene order conservation, or shared synteny patterns can provide functional association clues

between the gene of interest and its neighboring genes, which offers a complementary way of assigning protein functions other than the sequence similarity methods.[3] Changes in gene order combined with phylogenetic data can indicate genomic rearrangement activities that might have occurred during evolution. Generally, gene neighborhood analysis is a crucial component of comparative genomics, and has been demonstrated to be a powerful technique for inferring protein function and tracing gene evolutionary history. This approach has been employed to pinpoint functional associated genes,[4,5] as well as to further refine homology for remote homologs based on the idea that homologs tend to share similar gene neighborhoods.[6] Gene neighborhood analysis also facilitates the discovery and characterization of operon structure[7] or other functional unit made up of a set of homologous genes, such as metabolic pathways,[8,9] biosynthetic gene clusters[10,11] and toxin antitoxin systems.[12]

1.2 Gene neighborhood conservation patterns

Multiple protocols and techniques have been developed to infer protein function associations from various genomic context patterns. These include statistical detection of conserved gene order, phylogenetic co-occurrence, phylogenetic anticorrelation, gene fusion, and conserved divergently transcribed gene pairs (Figure 1.2). While these methods provide valuable clues into functional associations, experimental validation is often necessary to confirm a predicted functional association between gene pairs.

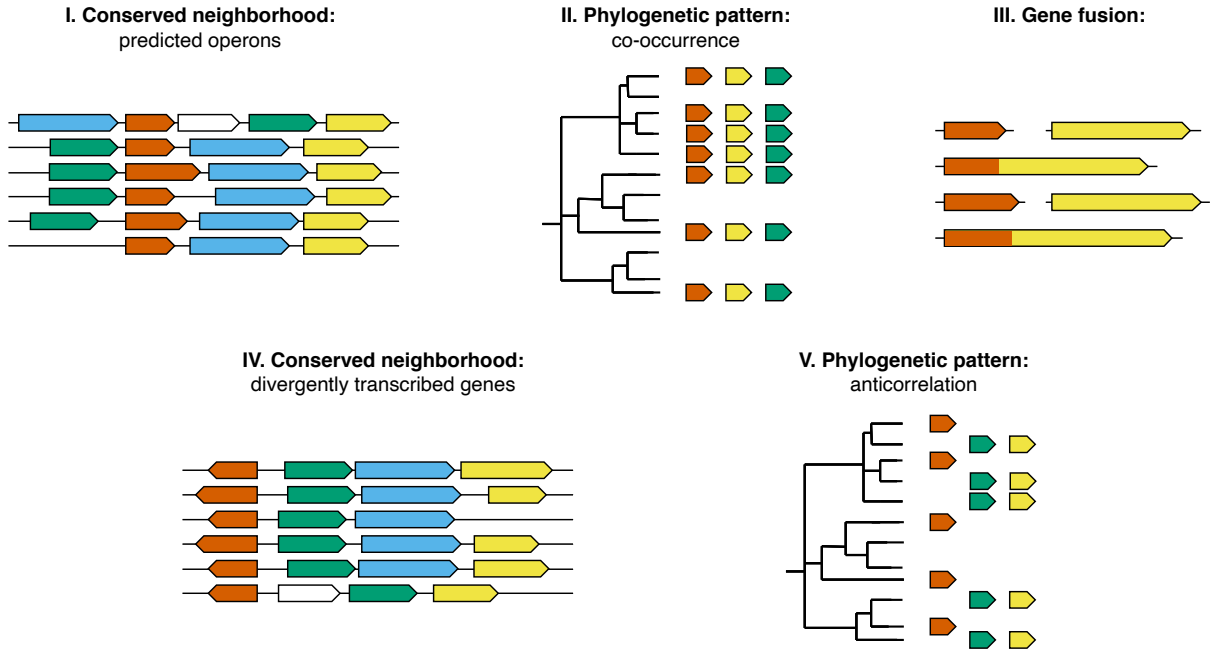


Figure 1.2: Different genomic context methods. Method I examines conserved gene neighborhood and potential operons. Method II focuses on gene co-occurrence patterns in a phylogenetic context. Method III detects gene fusion events. Method IV centers on conserved gene pairs that are divergently transcribed. Method V investigates anti-correlated gene patterns in a phylogenetic context. Figure modified from Korbel *et al.*[13]

The identification of conserved arrangements of two or more genes in multiple genomes is often used to predict a functional similarity or co-regulation.[14] The underlying principle of this technique is that conserved gene order can signify a potential operon structure, in which the genes within the operon are under transcriptional control by common transcription factors or regulatory proteins. However, it is important to recognize that conserved gene order does not always imply an operon structure, but may still imply shared gene functions for other reasons. For instance, adjacent genes may be under the control of common transcription factors and share regulatory elements, but be transcribed into separate mRNAs.[15] A limitation of this approach is that closely related species might naturally display a significant degree of shared synteny due to limited genomic divergence over short evolutionary time scales, rather than conservation due to functional constraints and selection. Hence, this approach should avoid choosing genomes that are too closely related to avoid taxonomic bias, also referred to as "redundancy".

Phylogenetic co-occurrence, or phylogenetic profiling, proposes that two genes are likely to have functional association if they show a correlated pattern of co-occurence across multiple genomes.[16] The difference between this approach and gene order conservation is that it does not require the genes to be in close physical proximity in the genome (i.e., within the same genomic region). The hypothesis of phylogenetic co-occurrence is that functionally interacting proteins, such as those involved in the same biological pathway or protein structure, are likely to evolve in a correlated fashion, as losing either one protein would lose a whole biological function/pathway or a protein complex in which the two proteins form essential components. A pitfall of this approach is that two non-interacting genes can also exhibit similar phylogenetic profiles if sampled from closely related organisms. To address this taxonomic bias, the phylogenetic co-occurrence approach is often combined with a species phylogenetic tree. Furthermore, phylogenetic profiling can provide more accurate prediction of function association when incorporated with gene proximity data.[17] However, it is worth mentioning that this approach might not be able to detect proteins that only co-evolve within certain phylogenetic lineages, known as non-homogeneous coevolution or local co-evolutionary problems.[18,19] In contrast to the phylogenetic co-occurrence approach, phylogenetic anticorrelation describes two genes that have a mutually exclusively occurrence profile across multiple genomes and are likely to have analogous function. This prediction is based on the observation that analogous enzymes can replace each other in different genomes.[20] From an evolutionary perspective, it would be useless for the prokaryotic genomes to encode two proteins with identical functions or functions that can substitute for one another.

Another two genomic context approaches detect patterns of gene fusion and conserved divergently transcribed gene pairs. The gene fusion pattern is based on the principle that when two separate genes (A and B) are occasionally found fused together to form a single gene (AB) in some genomes, this can be evidence of a physical or functional interaction between the individual genes A and B.[21–23] The identification of conserved divergently transcribed ($\leftarrow\rightarrow$) gene pairs indicates another type of functional association. If two adjacent genes are oriented in an opposite direction and transcribed away from each other consistently across genomes, this also suggests coregulation and a functionally association between the genes [17]. This pattern indicates that the adjacent gene pairs may share an overlapping promoter region, and hence suggests a co-regulatory relationship.[24,25] In a previous study, functional characterization of the conserved divergent transcribed (DT) pairs across 101 prokaryotic genomes spanning 47 evolutionary clades reveals that they often consist of one transcriptional regulator.[13] This trend is more clear in *E. coli*, with the DT pairs often composed of one transcriptional regulator and its regulated gene.[13] A recent study observed a similar trend, where researchers investigated "colinear syntenic blocks

(CSBs)" in 1485 prokaryotic genomes covering 31 different phyla, and identified that DT genes are enriched with genes involved in transcription.[26]

1.3 Methods for annotating gene neighborhoods

Gene neighborhood annotation refers to the bioinformatic process for identifying and functionally describing genetic elements that reside within a genomic region. While public databases such as GenBank/EMBL/DDBJ and IMG already offer annotation for these genetic elements, unless they are from non-redundant or reference databases, the annotation method largely depends on the submitter, leading to potential inconsistency.[27] Consequently, applying a consistent annotation pipeline on gene neighborhood dataset is crucial for comparative analysis. There are multiple ways to annotate genes on genomic regions across diverse species and group them into gene families or functional categories for gene order comparison, such as orthology inference, mapping to reference protein databases or precomputed ortholog databases (e.g., COG (Clusters of Orthologous Groups)[28]and eggNOG[29]), and mapping to protein signature databases (e.g., Pfam,[30] KEGG[31] and TIGRFAM[32]). It should be noted that only protein-coding genes and their functional annotation will be explored in this thesis.

Orthology inference

Orthology inference is widely used in comparative studies to study to the functional aspects of genes and genomes across different species. Many orthology inference techniques begin by identifying reciprocal best hits (RBH) from pariwise genome comparison. If gene A in genome X is the highest scoring match for gene B in genome Y, and vice versa, then gene A and gene B are identified as RBH, and together they are considered an orthologous gene pair.[33,34] This is based on the theory that homologous genes that have diverged exclusively due to speciation (orthologs) tend to share more sequence (and functional) similarity than genes that have diverged due to duplication events (paralogs). Of course one exception to this rule is when gene-duplication events occur specifically within a species or lineage (within-species paralogs), which may exhibit greater sequence similarity than they do to orthologs in other species. All-vs-all comparisons using BLAST[35] are one bioinformatic method that is frequently used for RBH identification. This technique performs a pairwise comparison of all genes (or protein sequences) in one genome against all genes in another genome, then generates a similarity score matrix for RBH identification. However, the all-vs-all BLAST technique[35] can be computationally intensive if applied to

multiple genomes. Therefore, some sequence comparison techniques have been emerged for fast protein pairwise alignment, which can be used to more rapidly infer orthology based on RBH, such as DIAMOND[36] and MMseqs2.[37] It is proposed that the use of DIAMOND with the very-sensitive option enabled the detection of a higher number of RBHs, and is therefore the optimal choice for finding RBH in a fast and accurate way.[38]

For large-scale orthology mapping across multiple genomes, specialized tools have been developed. These tools can be categorized into four types, including phylogenetic tree-based approaches, graph-based approaches, hybrid approaches that combine tree-based and graph-based methods, and meta-approaches.[34,39] Tree-based approaches identify orthologs based on phylogenetic knowledge of the type of evolutionary events that could have occurred. It first constructs a gene family tree and a species tree, and then uses phylogenetic tree reconciliation to identify nodes exhibiting incongruence, which may result from gene duplication or losses. Graph-based approaches begin by performing pairwise genome comparisons to identify RBH as orthologous gene pairs. While the identified ortholog relationships are not always limited to one-to-one orthologs, there can be one-to-many or many-to-many orthologs among homologs if additional duplication events have occurred after the speciation event. Therefore, paralogy identification is also crucial. Many graph-based tools identify paralogs within a species by detecting sequences with similarity greater than RBH. The orthologs and paralogs will then be used together to construct a graph using the hierarchical clustering method for orthology inference in multiple genomes.[33,34,39]

Compared to graph-based approaches, tree-based approaches do not require complete proteomes for RBH identification, can also handle complicated evolutionary scenarios and are considered more accurate.[33,34] Nevertheless, orthology inference tools that implement graph-based methods are more computationally efficient and often used nowadays. In recent years, hybrid tools that combine tree-based and graph-based approaches such as orthofinder have gained popularity due to its high accuracy in orthology prediction.[40] Orthofinder also provides gene trees and species tree, allowing the identification of gene duplication and speciation events. Lastly, meta-approaches combine multiple orthology inference tools to enhance orthology prediction accuracy. This can involve merging or intersecting results, or assigning weights to individual predictions to yield a comprehensive result, as well as employing machine learning techniques to derive predictions from various methodologies.[41]

Reference protein databases

Genes in the gene neighborhood dataset can also be functionally annotated by mapping to reference protein sequence database using homology-based search methods. For example,

coding sequences can be functionally annotated by searching (BLAST,[42] MMseqs2,[37] etc.) the translated protein sequences against reference databases, such as NCBI RefSeq,[43] UniProt/Swiss-Prot[44] that is curated with expert knowlege, UniProt/TrEMBL[44] that is annotated by computational pipelines, and Uniref[45] with pre-clustered protein groups that reduce sequence redundancy. RefSeq is a non-redundant protein database that also contains associated genomic and transcript seqeunces, along with related metadata. UniProtKB is a database that combines Swiss-Prot and TrEMBL, offering richer protein sequence annotations compared to RefSeq. Swiss-Prot contains reviewed and manually annotated protein sequences, making it ideal for high-quality protein annotation. As a complement to Swiss-Prot, TrEMBL contains translated protein sequence from EMBL/GenBank/DDBJ with automatic annotations, providing a broader coverage of protein sequences. UniRef clusters proteins from UniProtKB based on sequence similarity are suitable for large-scale protein annotation with reduced redundancy.

Protein signature databases or pre-computed orthologs databases

Gene neighborhoods can be functionally annotated through alignment-based approaches by mapping the translated coding sequence to protein signature databases, such as Pfam,[30] KEGG,[31] and TIGRFAM.[32] Protein signature databases contain information about protein domains, motifs, and conserved sequence regions, which are conserved functional and structural units within proteins. The current Pfam database includes 19,632 protein families (Pfam 35.0) that are represented by multiple sequence alignment and hidden markov models (HMMs).[30] The Pfamscan (at `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/`) and HMMER[46] tools can be used to annotate protein sequence with Pfam domains. Essentially, Pfamscan uses HMMER to search a library of Pfam domains represented by HMMs that match the sequence. TIGRFAM contains 4,488 manually curated protein families (JCVI 15.0), focusing more heavily on those from prokaryotic genomes.[32] TIGRFAM annotations can be mapped to query protein sequences using HMMER to identify matching HMMs.[46] KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database that provides a comprehensive view of biological pathways, molecular interactions, and functional annotations.[31] KEGG orthologs (KO) are integrated with pathway information, allowing the exploration of the biological process and pathway of a genes. KofamScan is a functional annotation tool that assigns a KO term to a query protein sequence.[47] It uses HMMs derived from KO database to identify similar domains in the sequence. The adaptive threshold is used to determine whether the KO identifier that matches the query sequence is more reliable than other matches. Alternatively, BlastKOALA for genomes and GhostKOALA for metagenomes use BLAST and ghostX, respectively, to search homologs in the KEGG

GENES database, and then transfer the KO identifiers to the protein queries.[48]

Gene neighborhoods can also be functionally annotated by leveraging known orthologous relationships to infer the possible function of the unknown protein. Precomputed orthologous groups such as COG[28] and eggNOG clusters[29] provide valuable functional annotation resources for comparing genomic neighborhood. COG is a manually curated orthology database created for functional and comparative genomics.[28] The latest update contains 4,877 COGs from 1,187 bacterial and 122 archaeal species. The eggNOG database is a functionally annotated database of orthologous genes that was created based on the idea of COG, but seeks the identification of orthologs in a much larger range of organisms.[29] The latest update (eggNOG v6.0) contains over 17 million orthologous groups (OGs) from 12,535 reference species, including prokaryotes, eukaryotes, and viruses. One benefit of eggNOG OGs is that proteins with unknown functions were also grouped into OGs, allowing the exploration of functional conservation for these unknown proteins. EggNOG-mapper is a program that assigns OGs to query protein sequences for orthology assignment.[49] The eggNOG OGs are linked with protein signature databases such as Pfam, KEGG orthology, and COG, allowing the transfer of these functional annotations from OGs to the queried protein sequences and thus provides a comprehensive annotation.

Several integrated databases and resources also exist for protein annotation with protein signatures or pre-computed orthologs. For instance, the Interpro consortium integrates protein families and domains, as well as functional sites and other annotation types (e.g., Gene Ontology terms) from 13 member databases.[50] Interproscan assigns these member protein functional signatures to queries.[51] This tool applies a variety of algorithms, including HMMER and BLAST, to compare query sequences against multiple models. Another integrated database, the Conserved Domain Databse (CDD), contains well-annotated protein domain models from multiple resources, including NCBI-curated domains, Pfam, TIGRfAM, COG, SMART, PRK.[52] Batch CD-Search in the CDD uses reverse-position-specific BLAST (RPS-BLAST) to compare quries against the position-specific score matrices (PSSMs) for the identification of conserved domains. Beyond protein domain and orthology mapping tools mentioned previously, several prokaryotic annotation tools and customized pipelines can be used.[53–55] Generally, these programs use similarity search algorithms to search queries against protein models that are curated in various formats.

1.4 Methods for comparing neighborhoods

Once gene neighborhoods from multiple genomes have been annotated, a common next step is to compare them to identify conservation patterns or evolutionary alterations in gene

order. For pairwise or small-scale gene neighborhood comparison, tools such as Artemis Comparison Tool (ACT),[56] Mauve[57] or Jbrowse2[58] may be used. These tools are valuable for visualizing and comparing genomic regions and associated annotations. Synteny mapping tools can also aid in the comparison of gene neighborhoods.[59–62] Additionally, public databases and comparative genomics platforms can provide gene neighborhood visualization within specialized databases.[9,63–68] However, these tools are limited and are not suitable when scaling up to larger scale comparisons. In light of these challenges, specialized tools tailored for larger-scale gene neighborhood analysis have emerged in recent years. Typically, these tools contain a browser or generate figures for visualization and exploring gene neighborhoods, and allow for the analysis of multiple tracks of gene neighborhoods based on a gene of interest. These tools allow for easy comparison and identification of conserved or distinct neighborhood patterns.[64,69–79] For example, some web servers such as Gecoviz facilitates an interactive browser for the visualization and comparison of gene neighborhood within their reference database.[69] On the other hand, some other tools such as WebFlags generates figures for gene neighborhood visualization based on a list of accession numbers from NCBI[77] (See Table 1.1 For a full list of gene neighborhood tools avialable). To compare multiple gene neighborhood tracks, these tracks are usually aligned and sorted based on phylogeny or gene order similarity. Sorting gene neighborhood based on phylogeny offers insights into the evolutionary trajectories of gene clusters and arrangements across species, whereas sorting gene neighborhood based on gene order similarity can uncover conserved patterns that might shed light on functional associations.

Table 1.1: Bioinformatics tools for gene neighborhood analysis.

Tool	Supported genomes	Input	Tracks display limit	Gene annotation method	Sorting method
CAGECAT[80]	unknown	protein sequences or .gbk files or NCBI accessions or HMM profile identifiers	50	identify gene groups based on BLAST similarity	based on the number of matching proteins to the query cluster
EvolClustDB[79]	882 eukaryotic species including fungi, plants, metazoans, insects and protists	protein code linked to NCBI and UniProt or species name or gene cluster name	150	define gene families using all against all BLASTP search	based on species tree obtained from NCBI
fast.genomics[76]	prokaryotes	locus tags or protein identifiers from fast.genomics, or identifiers from UniProt, the NCBI Protein database, RefSeq, or the Protein Data Bank, or protein sequences, or a taxon	200	group similar genes using LAST	rank by descending order of BLAST bit score or gene phylogenetic tree
Flags and WebFlags[77]	designed for prokaryotic and bacteriophage genomes; but can visualize all domains of life	protein seqeunces or NCBI protein accessions	200 for webflags, more for the local version	detect homologous proteins using Jackhmmer	based on protein phylogenetic tree
gcluster[74]	unknown	annotated genomes in .gbk files and a list of interested genes with locus tags	unknown	detect homologs using OrthoMCL; also accept homologs detection results from user	based on phylogenetic diagram provided by user
gcsnap[75]	not restricted to prokaryotic genomes	at least two target sequence identifiers from UniprotKB, UniRef, Ensemble, or Entrez	unknown	identify protein families using all-against-all BLAST searches	based on taxonomic clustering from NCBI, or by user's provided phylogenetic tree
Gecoivz[69]	prokaryotes	gene names or protein sequences or orthologous groups identifiers from eggNOG v5 or KEGG or Pfam databases	250	by COG, KEGG, Pfam, and assign orthology using eggNOG v5	based on species tree obtained from NCBI
genegraphics[73]	unknown	NCBI identifiers or .gbk files or SEED dataset	unknown	NCBI annotations	-
Genespy[71]	all domains of life	NCBI identifiers or BLASTP output files or .gbk files	330	NCBI annotations	based on phylogeny provided by user
Genome Context Viewer[70]	plants, legume and non-legume, metazoan and yeast	gene names or chromosome regions	unknown	functional annotation from their database	use Smith-Waterman and Repeat local alignment algorithms to cluster gene neighborhoods

Continued on next page.

Tool	Supported genomes	Input	Tracks display limit	Gene annotation method	Sorting method
Microbesonline[64]	bacteria, archaea and fungi	protein sequences or gene names	400	identify homologs based on all against all BLAST	based on gene tree
syntax[78]	prokaryotes from NCBI	protein sequences and selected genomes	120	detect homologs by the Smith-Waterman-Gotoh algorithm	rank by normalized sequence similarity scores
TREND[72]	prokaryotes	protein sequences or multiple sequence alignment or protein identifiers from NCBI&MiST or phylogenetic tree	unknown	identify similar genes based on shared domains	based on protein phylogenetic tree

Compare by gene order similarity

Although the tools presented thus far allow for annotation and visualization of gene neighborhoods, it is additionally important to develop quantitative methods for measuring the similarity of gene neighborhood structure and/or gene content. Gene order similarity can be measured by various metrics, such as the number of shared genes and the number of shared contiguous genes that quantifies gene order conservation. Several platforms of gene cluster visualization and detection have developed algorithms for identifying similar gene clusters given a query gene cluster, or grouping similar gene clusters. For example, ClusterBlast[81] calculates an empirical similarity score for identifying similar gene clusters. This score integrates multiple factors such as the total number of shared similar genes, the number of shared genes recognized as significant to the gene cluster according to scientific literature, and the number of conserved gene pairs. MultiGeneBlast[82] generates conservation scores based on the number of shared genes and the number of shared gene pairs to measure shared synteny, as well as incorporating a user-defined weighting factor to determine the importance of synteny. The resulting matrix of conservation scores is then hierarchically clustered for gene neighboring clustering. Clinker[80] employed a modified version of the MultiGeneBlast[82] similarity score, which prioritizes similar gene clusters with higher sequence similarity. Some programs group similar gene clusters using network-based strategies. For instance, Netsyn[83] generates a synteny score based on the ratio of shared genes between two gene neighborhoods, and then constructs a synteny network based on these scores. They also suggest that sequences with similar gene neighborhoods are likely to share similar protein functions, providing an alternative method of detecting homologs beyond sequence similarity. GeneGrouper[84] uses the Jaccard distance between each pair of genomes to compute a similarity matrix for all the genomes. This clustering algorithm facilitates population-level comparison of similar gene clusters by using an unsupervised clustering approach, grouping thousands of gene clusters from various genomes and taxa

into bins, with each bin being represented by a gene cluster. These representative gene clusters will provide an overarching perspective of gene clusters among distant taxa.

Compare by phylogeny

When comparing gene neighborhoods rooted in phylogeny, we can track the presence, absence, or variations of genomic neighborhoods across diverse taxa, as well as infer horizontal gene transfer events based on phylogenetic incongruence. For instance, gene neighborhoods can be organized based on the phylogeny of species retrieved from the NCBI taxonomy, such as GCsnap.[75] The Genome Taxonomy Database (GTDB)[85] provides another option for species phylogeny as it applies a standardized taxonomy classification system.[69,76] Certain tools, like GCsnap, GeneSpy, and Gcluster, allow users to sort gene neighborhoods using a user-defined phylogeny,[71,74,75] while others, such as FlaGs and webFlaGs, utilize the phylogeny of the query sequences.[77] By comparing gene neighborhoods using taxonomic information, users can trace the evolutionary changes and obtain insights into the evolutionary history of the species.

Operon prediction

Operons are a group of genes regulated by the same promoter and transcribed into a single mRNA, which likely participate in the same pathway or function.[86] Conserved gene clusters oriented in the same direction with small intergenic regions across multiple lineages are likely to be involved in an operon structure. In bacteria and archaea, most genes are predicted to be involved in operon structures.[87] It should be noted that most operons are only conserved in closely related species due to frequent genomic rearrangement events occurring in prokaryotes,[88] and only a few operons such as the ribosomal superoperons found to date are conserved in most prokaryotic species.[89]

Although the operon structure needs to be validated through experimental approaches that measure gene expression such as northern blotting, RNA-seq, RT-PCR or gene expression microarrays,[90] operons can be predicted using bioinformatic approaches.[91–93] Several factors derived from gene neighborhood analysis have been used to identify operons, and the most informative ones are intergenic distance, gene order conservation, functional relatedness among genes, regulatory elements, and gene expression profiles.[94,95] Furthermore, gene neighborhood analysis also enables the detection of changes in operon structure and composition. It is worth mentioning that operon-mapper, a tool that uses an artificial neural network (ANN) to predict operons based on genomic sequences, has achieved very high

accuracy, with 94.6 and 93.3% accuracies in *E. coli* and *B. subtilis*, respectively.[91] Some tools also mined for conserved gene clusters across hundreds or thousands of prokaryotic genomes to identify good candidates for operons,[96–98] enabling the discovery of numerous candidate operons that await experimental validation.

1.5 Problems with existing gene neighborhood tools

While microbial genomics data has experienced exponential growth over the past few decades, there are several deficiencies and limitations in available tools and workflows for microbial gene neighborhood visualization and analysis. Aside from using some of the specialized tools described thus far, there are several common strategies that researchers use to analyze and compare gene neighborhoods. One popular method is to explore gene neighborhoods using online genome browsers, such as genome browsers from NCBI,[99] UCSC,[100] Ensembl,[101] as well as standalone tools, such as IGV,[102] Artemis ACT,[56] Mauve,[57] and JBrowse.[58] However, most genome browsers lack comparative analysis features essential for multiple gene neighborhood analysis, such as visualizing related gene neighborhoods across different genomes, coloring similar genes for synteny comparison, arranging and aligning gene neighborhood tracks in order to identify conserved or unique patterns. In addition, gene neighborhood information can be accessed in public genomic databases and comparative genomics platforms,[9,63–68] but integrating information across multiple gene neighborhoods can pose challenges. While synteny detection tools can assist in comparing gene neighborhoods,[59–62] they are typically intended for longer genomic region comparisons (whole contig or chromosome), making them less suitable for shorter regions (10kb to 20kb), especially on a large scale. Importantly, all of the approaches mentioned above are not straightforward and computationally effective for large-scale gene neighborhood comparison. Therefore, several specialized tools tailored for gene neighborhood visualization have been developed in recent years.[64,69–79] These tools provide a more detailed and focused perspective on gene neighborhoods than genome browsers, integrating features such as retrieval and comparison of gene neighborhoods from different genomes, automatic assignment of homologous genes, or visualize gene neighborhood from different functional perspectives beyond standard NCBI annotations. Nonetheless, limitations still persist. Many of these tools have a narrow focus on specific functional aspects, and lack customizability. For example, users may wish to edit their gene neighborhoods manually by deleting genes and annotations, rearranging gene neighborhoods, and sorting tracks based on genes and functions of interest. In addition, while some tools permit analyses within their curated databases for fast gene neighborhood retrieval, they restrict users from exploring gene

neighborhoods from their own data, such as user-generated genomic and metagenomic datasets. In light of these challenges, there is a clear need for a more comprehensive and versatile platform for gene neighborhood search, comparison and visualization.

1.6 Gene neighborhood analysis in pathogen genomics

Gene neighborohood analysis plays a pivotal role in addressing numerous challenges in the pathogen genomics field. It has been applied to investigate the gain or loss of virulence genes, and clarify how they aid in the evolution of specific traits.[103,104] By examining the adjacent genes (e.g. mobile gene elements) of antibiotic resistance genes (ARG), researchers can explore the genetic background of ARGs to better understand ARG transmission patterns and dynamics.[105–107] Additionally, comparing the flanking region of mobile gene elements (MGE) within a phylogenetic framework across multiple genomes can be used to infer potential evolutionary events, such as insertion and mobilization,[108] and characterizing MGE neighboring regions helps infer MGE mechanisms by understanding the functional and regulatory aspects of flanking genes.[109–111] Moreover, analyzing neighboring genes surrounding toxigenic genes helps uncovering the diversity and genomic distribution of toxin gene clusters,[112,113] as well as identify and functionally characterize novel toxin-antitoxin (TA) systems.[114] Next I will provide a general introduction on the biological background of case studies that I explore in this thesis. Additional detailed background information on these topics will be provided in each chapter.

Biological background of case studies

The isopentenyl transferase (*IPT*) gene encodes an enzyme for cytokinin (CK) biosynthesis, with adenylate IPT being a subtype of this enzyme present in both plant pathogenic and plant growth-promoting bacterial (PGPB) strains.[115] Previous studies have shown that PGPB strains containing adenylate IPT generates an optimal amount of CK, but pathogenic strains produce an excessive amount that over-stimulate plant growth.[116–118] It has been hypothesized that there may be differences in *IPT* function between plant pathogens and PGPB. In this study, a comprehensive comparison of *IPT* genes between plant pathogenic and PGPB strains has been performed. By analyzing the genomic context of adenylate *IPT* genes in pathogenic and PGPB strains, I can gain a better understanding of the functional perspective of this gene family, e.g., whether adenylate *IPT* genes involved in different gene clusters and biosynthetic pathways, leading to various amount and type of CK being generated.

Botulinum neurotoxin (BoNT) is a type of clostridial neurotoxins (CNT), which are the most potent toxins known to humankind.[119] BoNT can cause a severe illness called botulism and is primarily produced by species within the *Clostridium* genus. In recent years, BoNT homologs have also been identified in non-*Clostridium* species, such as *Weissella oryzae*,[120] *Paraclostridium bifermentans*[121] and *Enterococcus faecium*.[122,123] The discovery of new clostridial neurotoxins in non-*Clostridium* species improves our understanding of this toxin family, and also raises questions regarding the evolutionary origins of BoNTs and their role outside of the human pathogens with which they are associated. In this study, I identify putative BoNT-like homologs in *Paeniclostridium ghonii* and *Bacillus toyonensis* genomes. However, the target host of these putative toxins is unknown. The central question guiding this study is, can we potentially learn host information from the genomic context of the new toxins? By analyzing neighboring genes that may play a role in *P. ghonii*'s and *Bacillus toyonensis* pathogenesis, it is hypothesized that if we can characterize the target host of adjacent virulence genes, it may allow us to infer that PG toxins and other virulence genes collaboratively target the same host species.

Tetanus neurotoxin (TeNT) is another type of CNT, which is produced by *Clostridium tetani* and can cause the neuroparalytic disease, tetanus. As a wound-associated infectious disease, tetanus is estimated to have plagued *Homo sapiens* throughout history, but it is also associated with animals.[119] Human archaeological samples contain ancient DNA from microbial genomes, some of which potentially belong to extinct or even novel bacterial species. In this study, two new *Clostridium* species, *Clostridium* sp. X and Y, were identified from ancient DNA samples. Phylogenomic analysis suggests that *Clostridium* sp. X is closely related to *Clostridium tetani*. To strengthen the hypothesis that *Clostridium* sp. X representing a distinct lineage, one question is whether gene neighborhood information can be used to clarify phylogenetic relationships.

1.7 Thesis outline

The overarching theme of the work presented in this thesis is the development of a novel platform called AnnoView for gene neighborhood analysis, and its application to several biological case studies of interest. These case studies are comparative genomics studies related to pathogenic organisms and gene families, where the goal is to identify biologically relevant differences in genomic contexts among species/strains.

This thesis is composed of five chapters. This first introductory chapter has provided biological and methodological background for gene neighborhood analysis to aid in the interpretation of the following data chapters (Table 1.2). Chapter 2 presents the development

of the AnnoView platform for prokaryotic gene neighborhood visualization and associated bioinformatic workflows for gene neighborhood analysis. This chapter also provides the methodology of gene neighborhood analysis for the following data chapters. The three data chapters are organized by protein families or genomes studied.

In Chapter 3, I performe comparative studies on the adenylate isopentenyl transferase (*IPT*) gene, with a focus on the genomic context differences that distinguish bacteria phenotypes (plant pathogenic bacteria vs. plant growth-promoting bacteria). In Chapter 4, I identify two putative botulinum neurotoxin (BoNT) sequences from *Paeniclostridium ghonii* and *Bacillus toyonensis* using bioinformatic methods, and then apply the genomic neighborhood analysis to predict the target host of the BoNT homologs in *P. ghonii*. In Chapter 5, I analyze the genomic structures of two novel *Clostridium* species identified in ancient DNA, and perform genome comparisons with closely related *CLostridium* species (i.e., *Clostridium tetani)* to identify key differences, with a special focus on the tetanus neurotoxin (TeNT) gene neighborhood.

Table 1.2: Summary of thesis chapters.

	Gene or genome investigated	**Methods**	**Research question**
Chapter 2		Develop annoview; Create workflows for large-scale neighborhood analysis and functional annotation	
Chapter 3	IPT	Comprehensive analysis of adenylate IPT gene family across bacteria	Do adenylate IPT in pathogenic bacteria and PGPB involved in different biological pathways?
Chapter 4	BoNT	Identify and analyze two new BoNT-like toxin genes in *P. ghonii* and *B. toyonensis*	What may be the target host species for the two putative BoNT toxins?
Chapter 5	Clostridium clade X genomes	Comparative genomic analysis of a new Clostridium species from aDNA	Is species X a new neurotoxigenic (able to express neurotoxin genes) Clostridium? Are there any unique genomic structures in species X?

Chapter 2

AnnoView: Gene neighborhood browser for bacterial and archaeal genomes

This chapter focuses on the development of a novel comparative genomics tool called AnnoView, which is accessible at http://annoview.uwaterloo.ca/. AnnoView facilitates large-scale search, comparison, and visualization of gene neighborhoods across the microbial tree of life, and is powered by the AnnoTree[124]/GTDB[85] databases. AnnoView forms a major methodological framework for further studies presented in the thesis in Chapters 3, 4, and 5.

The work presented in this chapter is a collaborative effort involving several individuals. I was responsible for tool design (along with Dr. Briallen Lobb and Dr. Doxey), analyses and code for bioinformatic workflows. The front-end and back-end application was developed by several programmers including Huagang Tan, William Zhen, and Zijing Wu. This project would not have been possible without the collaborative effort of numerous individuals.

2.1 Introduction

A fundamental aspect of microbial genome analysis is the annotation and visualization of genomic regions or gene neighborhoods. Gene neighborhood-based approaches represent a class of methods for gene function prediction, and typically operate by assigning functions to hypothetical genes based on those of annotatable genes that are found within the same

operon or that commonly co-occur within the same genomic region. Unknown protein functions can be inferred from conserved genomic context patterns in prokaryotic genomes, such as conserved gene orders that suggests an operon structure,[14] frequently co-occurring genes,[16] anticorrelated genes with mutually exclusive occurrence patterns,[20] gene fusions,[21] and bidirectional gene pairs.[13] Additionally, gene neighborhood analysis can be used to unravel genomic rearrangements events and shared synteny patterns, thus providing insights into the evolutionary history of genome structure.

There are multiple ways of visualizing gene neighborhoods for genes of interest. One common way is to explore gene neighborhoods using genome browsers, such as those available from NCBI,[70] UCSC,[100] Ensembl,[101] and standalone tools such as IGV,[102] Artemis ACT,[56] and JBrowse.[58] However, most genome browsers lack comparative analysis features that are often required in gene neighborhood analysis, such as the comparison of gene neighborhoods across different genomes. Therefore, several bioinformatics tools designed specifically for gene neighborhood visualization have been developed.[64,69–79] These gene neighborhood tools provide a more detailed and focused view on gene neighborhoods than genome browsers, with unique features such as the ability to compare gene neighborhoods from different genomes, automatic functional annotation of genes by homology. Moreover, these tools allow for gene neighborhoods to be visualized with different functional annotations, which can reveal different functional perspectives and provide a more holistic view of neighborhood function. However, these tools still have some limitations. Many of them have a narrow focus on specific functional aspects, and lack customizability. For example, users may wish to edit their gene neighborhoods manually by deleting gene neighborhood tracks, rearranging gene neighborhoods, and sorting tracks based on genes and functions of interest. In addition, while some tools permit analyses within their curated databases for fast gene neighborhood retrieval, they restrict users from exploring gene neighborhoods from their own data, such as user-generated genomic and metagenomic datasets.

In this Chapter, I introduce AnnoView (http://annoview.uwaterloo.ca/), an online prokaryotic gene neighborhood browser that enables gene neighborhood exploration on a large scale through different functional annotations, while at the same time also allowing data visualization outside our pre-computed database by enabling users to upload their own gene neighborhood dataset. Our precomputed database is built upon the AnnoTree[124]/GTDB[85] database, which provides pre-computed functional annotations for all the prokaryotic genomes in the GTDB database. Users can easily share their AnnoView sessions using the session link, or download their data into a CSV file that can be reuploaded to AnnoView. Additionally, within AnnoView, we have implemented a clustering algorithm that enables users to sort gene neighborhoods based on the neighboring gene content of a chosen center gene.

2.2 Methods

Database preparation

The AnnoView database was constructed based on the AnnoTree R95 database which is derived from the Genome Taxonomy Database (R95).[85] Gene and functional annotations are ultimately derived from the gene prediction and annotation pipeline described by Mendler et al.[124] In brief, Prodigal[125] with default parameters was used for gene calling from whole genomic fasta files, resulting in GBK/GFF files and predicted protein-coding sequences in FASTA format for all organisms. Pfam and TIGRFAM annotation was performed on all protein sequences using HMMER[46] as well as the pfam_scan.pl (at `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/`) for Pfam domains specifically. For KEGG annotation, DIAMOND[36] was used to identify homologous sequence clusters within the UNIREF100 database,[45] and existing annotations were transferred. DIAMOND homology searches had the following parameters: E-value cutoff $\leq$ 1e-5, percent identity $\geq$ 30%, and query-to-subject and subject-to-query percent alignments $\geq$ 70%. This annotation pipeline results in functional annotations for genes from 30,238 bacterial genomes and 1,672 archaeal genomes.

For the AnnoView database, functional annotations short names, and descriptions were retrieved from the above KEGG, Pfam and TIGRFAM annotation data. Gene coordinates and orientations were extracted from the Prodigal-generated GFF files. An AnnoView SQL database was constructed to store predicted protein sequences, gene functional annotation data (KEGG, Pfam and TIGRFAM), gene coordinates and strand information (+ or -), CDS length, as well as GTDB taxonomy information. All protein sequences in the Annoview database were then used to build a DIAMOND (2.1.8) database[36] that can be searched against to find similar proteins to the query. Once DIAMOND searches are performed by the user, the backend server will run MySQL queries on all sequence hits to identify proteins within a $\pm$10kb window size. Protein-coding genes that span the edges of this window are excluded. This process is highly efficient due to the presence of pre-computed data and indexes in the MySQL database, allowing for quick execution of MySQL queries (0.03 seconds per genome).

Web server implementation

The AnnoView web server adopts the client-server website design (Figure 2.1). For the frontend client, we uses react.js to build the user interface (`https://reactjs.org/`), and d3.js (`https://d3js.org`) for generating the gene neighborhood visualization. The

Python-based Flask web framework (https://flask.palletsprojects.com) was used for backend server development. MySQL (https://www.mysql.com/) was used for storing pre-computed gene neighborhood data, and mongoDB (https://www.mongodb.com/) was used to save gene neighborhood data resulting from users' queries and users' uploaded genomic files. The pseudocode for the algorithm that clusters gene neighborhoods based on a gene of interest is described in Figure 2.2. The clustering algorithm was developed in collaboration with MSc student, Huagang Tan, in the Doxey Lab.

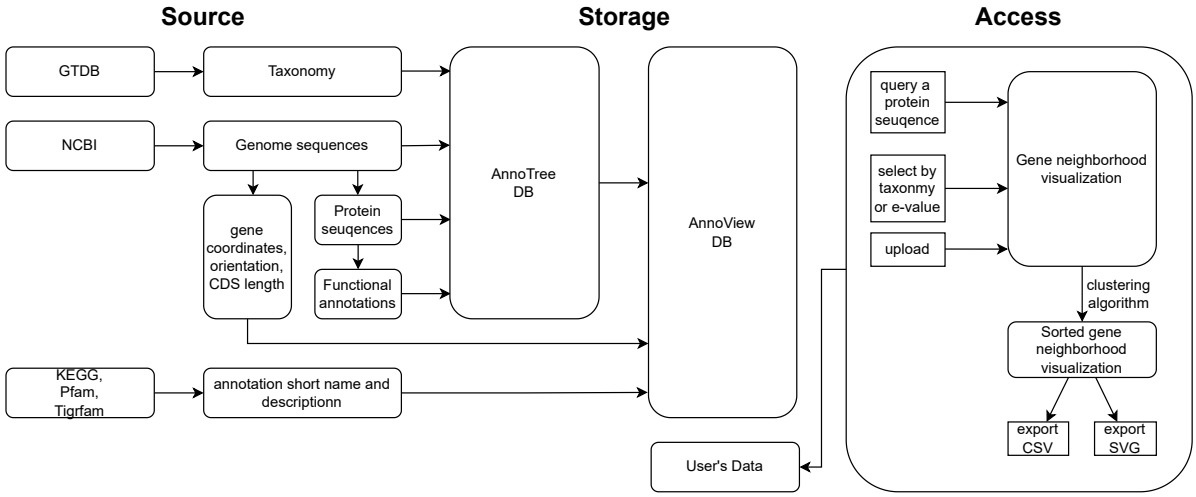


Figure 2.1: The architecture of AnnoView.

Algorithm 1: Smith-Waterman similarity calculation between two genomes

SimCal $(g1, g2)$;
Input : Genome g1, Genome g2
Output: Weighted similarity between g1 and g2
Scoring matrix $H \leftarrow$ Smith-Waterman(g1,g2)
Sim $\leftarrow H_{length(g1),lengh(g2)}$
K $\leftarrow$ index(max(H))
return Sim/K

Algorithm 2: Calculate similarities among all genomes

Similarity (G);
Input : Genomes G
Output: Similarity score matrix H
For every pair $g1, g2 \in G$
    i $\leftarrow$ index of center gene in g1
    j $\leftarrow$ index of center gene in g2
    Leftg1 $\leftarrow$ g1[0:i]
    Leftg2 $\leftarrow$ g2[0:j]
    Rightg1 $\leftarrow$ g1[i+1:]
    Rightg2 $\leftarrow$ g2[i+j:]
    Sim $\leftarrow$ SimCal(Leftg1,leftg2)+ SimCal(Rightg1,Rightg2)
return H

Algorithm 3: Reorder genomes for cluster visualization based on a target center gene.

Center Sort (G);
Input : Genomes G, a list of genomes, each genome is a list of genes
Output: Order O
Similarity score matrix H $\leftarrow$ Similarity(G)
H $\leftarrow$ Min-Max Normalization(H): $1 - \frac{H_{ij} - minH}{max(H) - min(H)}$
H $\leftarrow$ hierarchicalClustring(H)
O $\leftarrow$ FastOptimalLeafOrdering(H)
return O

Figure 2.2: Pseudocode for the clustering algorithm that sorts gene neighborhoods based on a target center gene (gene of interest).

Upload workflow for NCBI data retrieval and gene annotations

In addition to retrieval of gene neighborhood information from the AnnoTree[124]/GTDB[85] database, a bioinformatic workflow was also generated for users who would like to obtain this information directly from the NCBI database. The workflow performs the following steps:

(1) retrieval of gene neighborhood datasets in GBK format from the NCBI; (2) annotation of gene neighborhoods with taxonomic information; (3) KEGG and Pfam functional annotation of gene neighborhoods; (4) sorting of gene neighborhoods by selection of a gene/protein of interest as the "center gene". The detailed workflow and associated code is accessible at the following public github repository: https://github.com/satellite830/AnnoView.

First, the gene neighborhood dataset is retrieved from NCBI using NCBI Entrez Direct,[126] and the downloaded dataset in GBK or GFF format can be uploaded to AnnoView for gene neighborhood visualization. Then, the gene neighborhood data in CSV format is downloaded so that taxonomic information can be added, as well as KEGG and Pfam annotations, or other customized (optional) functional annotations that the user would like to add. The CSV file can then be edited to define the center genes that are used by the AnnoView sorting algorithm. To collect all the protein sequences for functional annotation, the protein sequences of neighboring genes in the CSV file were written into a FASTA file, with duplicated FASTA sequences removed. pfam_scan.pl ((at ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/) and KofamScan[47] are then used to functionally annotate the neighboring protein sequences. Finally, the "Default center" column is added to the data table so that AnnoView can identify the center genes, and sort gene neighborhood based on its neighboring gene content.

2.3 The AnnoView application

AnnoView home page - upload and AnnoTree search

AnnoView is a gene neighborhood browser for prokaryotic genomes and is available at http://annoview.uwaterloo.ca/. There are two modules in AnnoView. The "upload" module allows users to upload genomic neighborhoods in GBK, GFF, and CSV format. An alternative way to prepare these genomic neighborhood files is provided in Methods 2.2. The "AnnoTree search" module requires that users only need to provide a query protein sequence. The query is then searched against our pre-computed AnnoView database (all genomes in AnnoView/GTDB), and then it returns the genomic neighborhoods surrounding identified matches (genes corresponding to protein homologs). Pre-computed functional annotations in KEGG, Pfam, and TIGRFAM are also then available to the user for further exploration and visualization. The protein homology search criteria include E-value, coverage cut-off, and choice of database to search (bacteria/archaea).

After the homology search, the user can further select a subset of the protein hits for genomic neighborhood display by taxonomy or by E-value on an "intermediate" results

page. This is a solution to the potential issue in which an overwhelming number of sequence matches is detected by DIAMOND,[36] and there is a need to further filter results for the user. The "Select by taxonomy" option is intended for the user who is interested in exploring the genomic neighborhoods in a specific lineage, while "Select by E-value" is intended for the user who wants to examine only the top-scoring genomic neighborhoods based on the sequence similarity to the query protein. Importantly, sorting by E-value does not guarantee that the top matches are the most closely related matches phylogenetically, and thus may or may not result in hits that are the closest homologs to the query.

AnnoView gene neighborhood browser

Figure 2.3 shows the interactive browser-component of AnnoView, and allows users to explore multiple gene neighborhoods of interest. A toolbar panel at the top of the browser allows the user to perform different functions (Figure 2.3). These include a zoom-in and zoom-out function, and a download button for downloading the gene neighborhood information in a CSV format. Downloaded CSV data can then be customized by the user to edit existing gene neighborhood information, or add taxonomy information as well as functional annotations (see Methods 2.2). The gene neighborhood plot can also be exported in vector-quality SVG format suitable for publication or offline editing by image-editing software (e.g., Adobe Illustrator or Inkscape).

To quickly navigate to a specific gene, the user can search for the gene name in the search box. The search box is also useful for navigating genes that share a general term, such as "flagellar M-ring protein FliF" and "flagellar motor switch protein FliG". Taxonomy labels allow the user to choose which taxonomy level to display in the visualization panel. Users can change the displayed metadata to color the gene neighborhoods based on different functional annotations.

The visualization panel is a user-interactive interface that allows the exploration of basic gene information, clustering of gene neighborhoods, simple editing of the plot, and data retrievals, such as protein sequences and gene metadata (Figure 2.3). For each gene neighborhood displayed in the visualization panel, its taxonomy information is displayed on the left side of the panel. The taxonomic level can be changed under the taxonomy labels in the toolbar panel. To view the gene information, the user can either hover over a gene, or right-click on a gene for more information such as protein sequence and annotation details. More options are available when right-clicking on a gene for the user to edit the gene neighborhood plot in AnnoView. After choosing a center gene, gene neighborhoods can be clustered using the clustering algorithm implemented in AnnoView, which is based

on similarity of neighboring gene content (see Methods 2.2). Furthermore, users can flip the gene neighborhood orientation, change the color of a gene, and add a gene to the legend. The vertical order of gene neighborhoods displayed can also be changed based on a user's preference (e.g., the user may want to manually re-order neighborhoods based on phylogenetic relationships). A unique token will be generated for each online session and can be used for access within a month. Because of the unique web token, the URL for the user's current session can also be shared with other users over the web to enable data sharing.

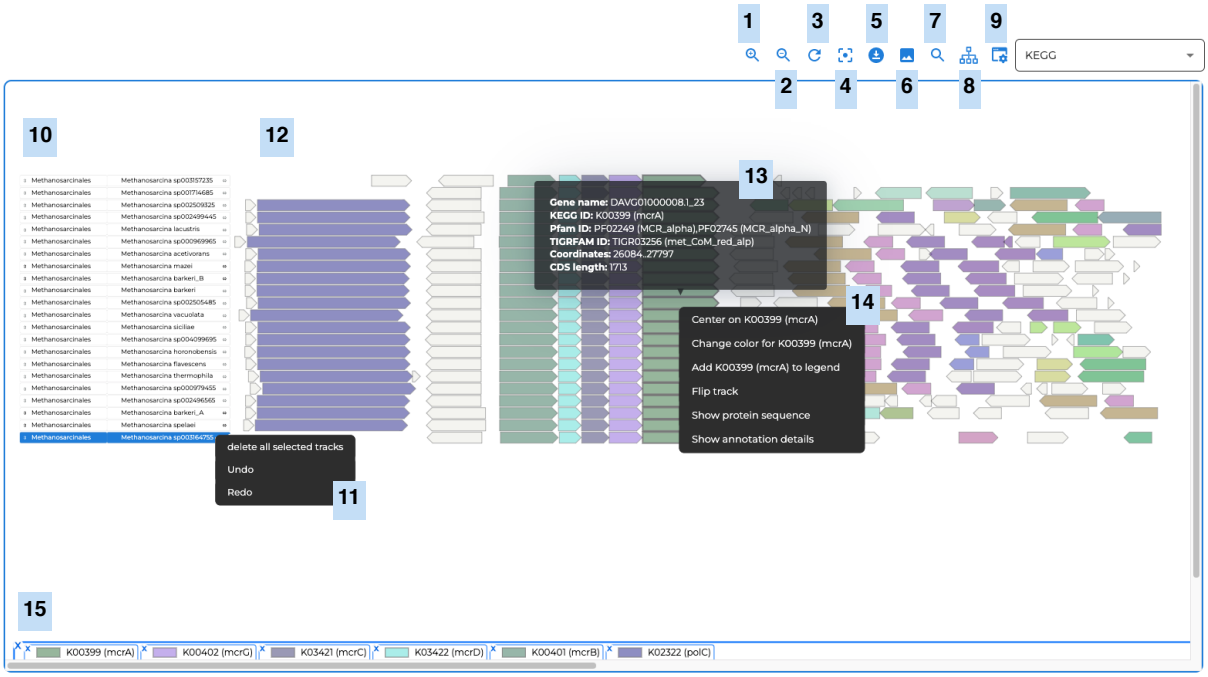


Figure 2.3: Functionalities of AnnoView gene neighborhood browser.(1) Zoom in; (2) Zoom out; (3) Reset the page; (4) Recenter image; (5) Download the gene neighborhood in CSV format; (6) Export image in SVG format; (7) Search box; (8) Display taxonomy labels; (9) Color genes by metadata; the default is colored by KEGG annotations; (10) Taxonomy labels; (11) Delete selected tracks; (12) Gene neighborhood; (13) Gene information when hovering on a gene; (14) More options when right-clicking on a gene.

2.4 Example workflows

AnnoTree-DB search (operon structure prediction - McrA)

It is envisioned that one common use case for AnnoView will be for users to search our pre-computed genome database for matches to their gene/protein of interest. For instance, using AnnoView, users can explore gene neighborhoods surrounding methyl-coenzyme M reductase subunit alpha (McrA) in archaea (Figure 2.4). Methyl-coenzyme M reductase (MCR) is an enzyme that catalyzes the final step of methanogenesis in archaea.[127] Searching the McrA (https://uniprot.org/uniprotkb/P07962/) protein sequence in the AnnoView databases returns 379 homologous sequences. There are two ways of displaying the genomic context of all or part of homologous sequences: either selected hits by taxonomy or by the top N (where N is defined by the user) number of hits ranked by *E*-value. Here, we select McrA homologous sequences in *Methanosarcina* - a genus of archaea that is known for methanogenesis. From the gene neighborhood visualization, it can be observed that the McrA gene is clustered with other genes in its upstream region, and forms a conserved putative operon structure. This operon structure has also been identified in previous studies.[128] The MCR operon in methanogens is known to consist of the essential genes McrA, McrB, and McrG, as well as 2 accessory genes McrC and McrD.

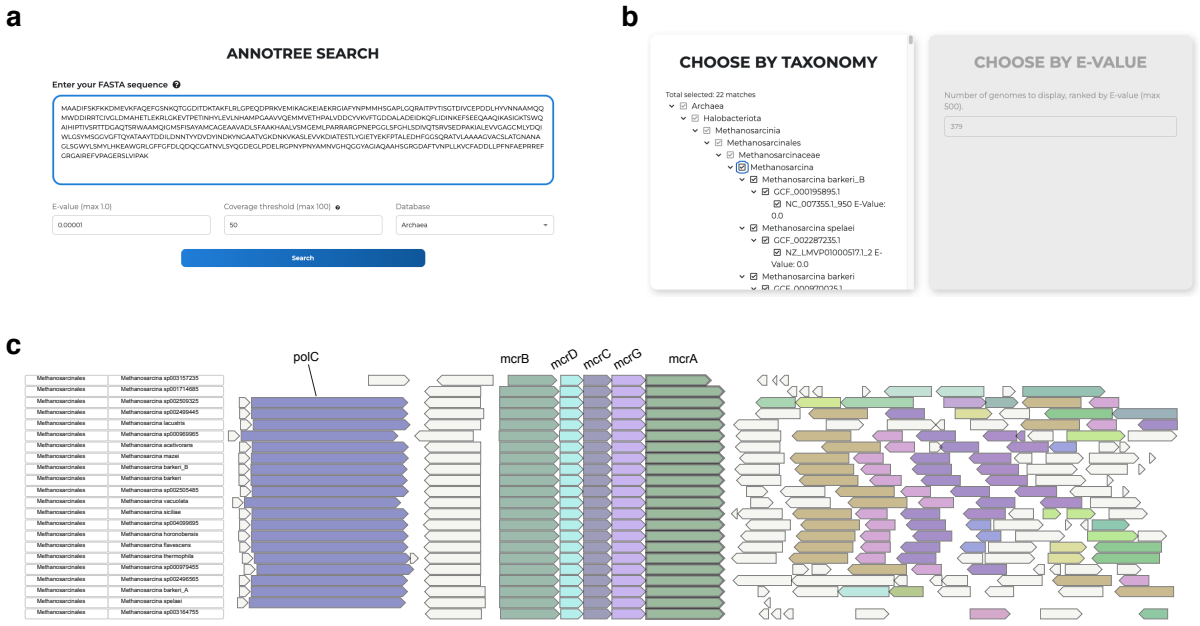


Figure 2.4: Search in the AnnoTree database and generate gene neighborhoods for McrA. (a) AnnoTree search interface. (b) The intermediate page for user to select gene neighborhood by taxonomy or by *E*-value. (c) The genomic context of selected McrA homologs in the *Methanosarcina* genus.

Custom data upload (predict the secretion system for Slr4 and discover a gene insertion event - Slr4)

The other envisioned use case of AnnoView is to upload a gene neighborhood dataset and explore it further using AnnoView's gene neighborhood browser. The gene neighborhood dataset can be generated from public genomic databases such as NCBI or user-generated genomic datasets in GBK or GFF format. To visualize the gene neighborhoods labeled by different functional traits than the default annotations provided by Genbank (e.g., protein product name) such as Pfam or KEGG annotation, a bioinformatic workflow is provided here at https://github.com/satellite830/AnnoView. This workflow allows users to download a CSV table from AnnoView, and then add customized annotations and taxonomy information to the table, and then re-upload the data back to AnnoView for further exploration (See Methods 2.2 for detailed workflow).

In a study by Ali et al.,[129] we used AnnoView to visualize gene neighborhoods surrounding

a newly identified protein called Slr4 from *Pseudoalteromonas tunicata* and its homologs in other species. The gene neighborhood was annotated and visualized based on Pfam annotations. Note that we only colored annotations occurring more than once in this case study. In this study, Slr4 was functionally characterized as a novel surface layer (S-layer) protein that forms a proteinaceous protective layer surrounding cells. The genomic context provides additional supportive evidence of this function, as it revealed nearby genes that have also been implicated previously in S-layer secretion. Specifically, AnnoView visualization reveals that the genes upstream of Slr4 (for all the species shown) include genes involved in the Type II secretion system (Figure 2.5). It was concluded that this upstream operon may encode a dedicated T2S pathway for the novel S-layer protein, which is also a genomic feature of S-layer genes in other species.[130,131]

Interesting deviations from this pattern can also be visually identified using AnnoView, which may represent species-specific genomic rearrangements or insertions. For example, *Cognaticolwellia mytili* has two additional ORFs annotated as PF18765 (Polymerase beta, Nucleotidyltransferase) PF01934 (Ribonuclease HepT-like) within the T2S gene cluster (highlighted in red box), which is likely to be a type II toxin/antitoxin (HEPN/MNT) system that has recently been discovered.[132] It is likely that the toxin/antitoxin system is inserted into the T2S gene cluster due to the high mobile nature of type II toxin/antitoxin system.[133]

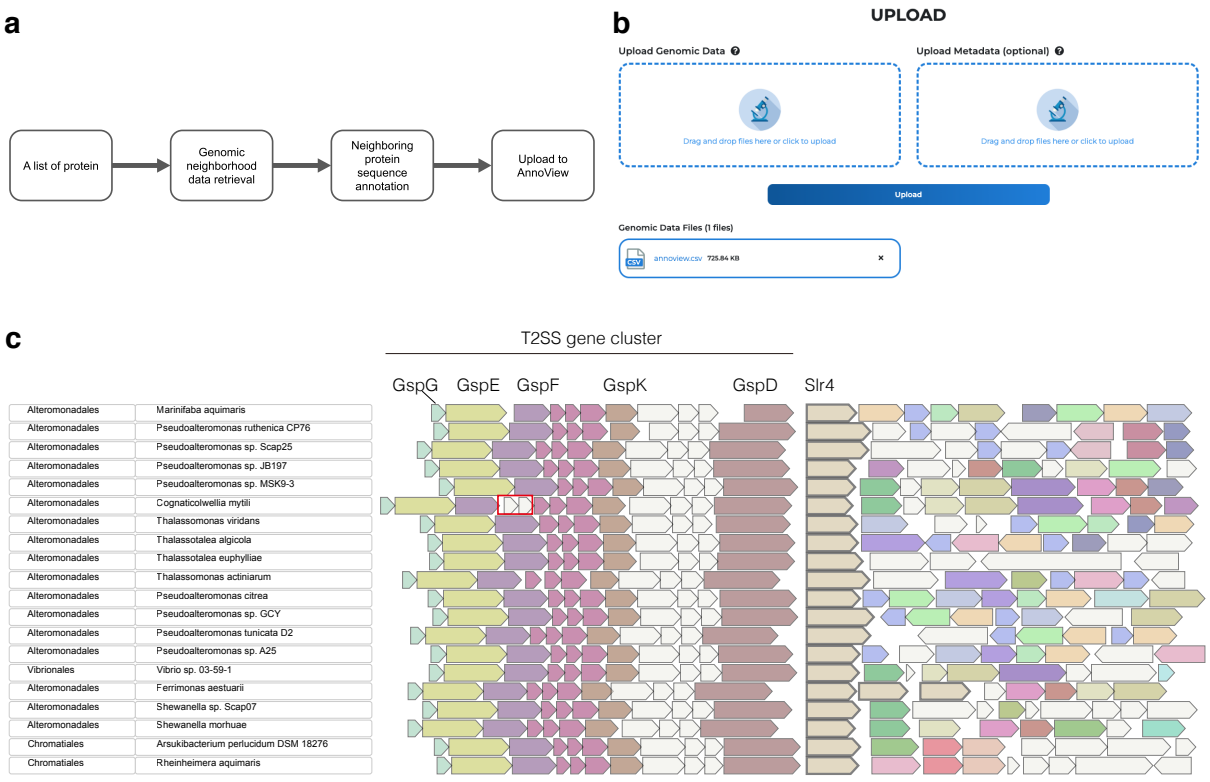


Figure 2.5: Upload gene neighborhood dataset to AnnoView and generate gene neighborhoods for Slr4 and its homologs. (A) Workflow of generating the gene neighborhood dataset. The gene neighborhood dataset in GBK format was downloaded from NCBI, then the neighboring genes were functionally annotated by PFAM. The gene neighborhood dataset along with protein functional annotations was uploaded to AnnoView for visualization. (B) AnnoView upload interface. (C) The gene neighborhoods of Slr4 and its homologs.

Custom data upload (discover genes unique to a specific lineage suggests unique phenotypic characteristics - *Betacoronavirus*)

While primarily intended as a tool to explore gene neighborhoods surrounding genes of interest, AnnoView is also capable of whole genome visualization and comparative analysis for small size genomes, such as viral genomes, plasmids, and even in some cases, smaller bacterial genomes. Figure 2.6 shows AnnoView visualization of several *Betacoronavirus* genomes, which reveals several interesting and biologically relevant differences in genome

structure. As shown in Figure 2.6, the overall genome structure (gene order and composition) of *Betacoronavirus* is highly conserved. Important genes such as ORF1ab, N(nucleocapsid protein), S (spike protein), M (membrane protein), and E(envelope protein) are conserved in all betacoronavirus genomes. However, HE (hemagglutinin esterase) is only present in the *Embecovirus* subgroup. The hemagglutinin esterase is an enzyme that helps attach to the cell surface, but can also destroy sialic acid receptor present on the cell surface. This mechanism allows the virus to escape from non-taraget cells or molecules that has sialic acid present, so that virus is more likely to infect the target cells rather than non-target objects.[134,135] HE is identified in influenza C virus, toroviruses, and coronaviruses subgenus *Embecovirus*, and is likely to be originated from influenza C/D through HGT.[136] The proposed recombination event likely occurred in ancestral members of the *Embecovirus* subgenus before the divergence of other *Betacoronavirus* subgroups, resulting in the unique presence of the HE protein in the *Embecovirus* subgroup.

Moreover, AnnoView visualization reveals that the rabbit coronavirus genome has an insertion of 4 ORFs between the ORF1ab and the downstream HE gene, whereas other genomes from the *Embecovirus* subgroup have one ORF (NS2a) or none between ORF1ab and HE. This insertion of multiple ORFs has been shown to be resulted from the ORF (NS2a) broking into 4 small ORFs in rabbit coronavirus during evolution.[137]

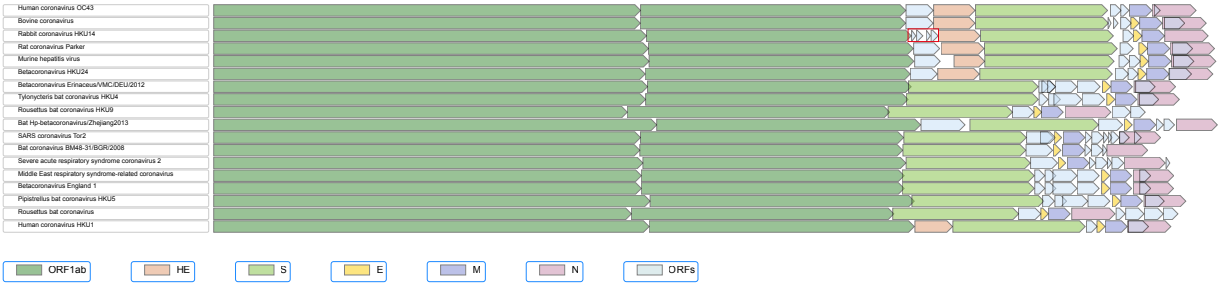


Figure 2.6: Genome structure of *Betacoronavirus*. Genome tracks are sorted by the genome phylogeny from NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/)

2.5 Discussion

In conclusion, AnnoView is a gene neighborhood visualization tool that enables fast exploration of prokaryotic genomes, as well as user-uploaded genomic datasets. The AnnoView database is integrated with the AnnoTree[124]/GTDB[85] database, which provides a comprehensive and consistent taxonomic classification for prokaryotic gene neighborhoods.

The clustering algorithm implemented in AnnoView allows quick identification of gene order conservation patterns from a large-scale gene neighborhood analysis. Our tool also has the potential to be used as a small genome visualization tool. In the future, we hope to add more features to Annoview, such as incorporating InterPro functional annotations into AnnoView, updating AnnoView on a regular basis with the most up-to-date GTDB data, allowing users to view the gene neighborhood with user-defined window size, build a colorblind-friendly palette, and display other genomic features such as tRNAs, rRNAs, CRISPR loci, and pseudogenes. We envision AnnoView being a useful tool for multiple gene neighborhoods exploration, visualization and comparison in prokaryotic genomes.

Chapter 3

Comparative analysis of adenylate *IPT* genes

Material in this chapter has been published as part of Wei et al. (2023).[138] The published manuscript is available here:

> Wei, X., Moreno-Hagelsieb, G., Glick, B. R., & Doxey, A. C. Comparative analysis of adenylate isopentenyl transferase genes in plant growth-promoting bacteria and plant pathogenic bacteria. *Heliyon*, 9(3):e13955, 2023.[138] https://doi.org/10.1016/j.heliyon.2023.e13955

In this chapter, I use comparative genomics approaches including the AnnoView platform to compare adenylate IPT genes and surrounding gene neighborhoods between plant pathogenic and plant growth-promoting bacteria (PGPB). As it is unclear what genomic features contribute to different levels of cytokinin between pathogenic and PGPR, I hypothesized that an analysis of genomic context may help to provide answers to this question.

3.1 Introduction

Cytokinins (CKs) are a group of plant hormones used in agriculture to promote plant growth.[139] These phytohormones have a multiplicity of effects on plant growth and development, including cell differentiation, apical dominance, shoot growth, and delaying of leaf senescence. Since the first CK – kinetin – was discovered in the 1950s, additional ones have

been identified, including iP (isopentenyladenine), and the *trans* and *cis* forms of zeatin.[140] While CKs are mostly studied in plants, they are also commonly found in bacteria, fungi, and insects. These organisms, which include both plant pathogenic bacteria and plant growth-promoting bacteria (PGPB) that mainly reside in soil, air and water, interact with plants and produce CKs that can affect plant growth and development.

Large numbers of soil microorganisms are attracted by plant root exudates, including bacteria, archaea, and fungi that typically inhabit the plant rhizosphere.[116] Different rhizosphere bacteria form interactions with plants that can be beneficial, harmful, or neutral for the plant. The beneficial bacteria (PGPB) can directly promote plant growth by facilitating resource uptake or modulating the plant's level of various phytohormones. The host plant, in turn, through root exudation, provides PGPB with essential nutrients such as organic acids and sugars.[117] In addition, some PGPB can indirectly promote plant growth by limiting the damage to plants caused by various phytopathogens that may be present in the soil.[117,141] On the other hand, phytopathogens in the soil employ various mechanisms to facilitate the colonization of plant cells and cause disease. Some can produce cell wall degrading enzymes that break down the plant cell wall and cause soft rot disease on the plant. Others secrete virulence factors such as exopolysaccharides in the xylem vessels causing disease by blocking nutrient transportation. Bacteria such as *Agrobacterium tumefaciens* and *Rhodococcus* spp. can employ a type IV secretion system that transfers T-DNA into plant cells, resulting in the overproduction of certain phytohormones.[142]

A number of PGPB as well as plant pathogenic bacteria can produce CKs by expressing the isopentenyl transferase (IPT) enzyme. IPT is a key rate-limiting enzyme that controls CK production.[143] In the Pfam domain database, two related protein families, "IPT" (PF01745) and "IPPT" (PF01715), are described as functional domains of IPT proteins. The IPT domain is an isopentenyl transferase/dimethylallyl transferase protein family that synthesizes isopentenyladensosine 5'-monophosphate, and the IPPT domain is a protein tRNA δ (2)-isopentenylpyrophosphate transferase protein family. The evolutionary studies conducted by Nishii et al. (2018) classified the *IPT* genes phylogenetically, where *IPT*s diversified into IPT domain and IPPT domain genes in bacteria, and class I, class II tRNA-*IPT*s, and adenosine-phosphate *IPT*s in plants.[115] In bacteria, proteins with IPT domains are also described as adenylate IPTs, and those with IPPT domain are described as tRNA-IPTs. However, unlike plant tRNA-IPTs, which produce cis-zeatin (cZ) type CK, bacterial tRNA-IPTs can produce various forms of CK, including isopentenyladenine (iP), trans-zeatin (tZ), and *cis*-zeatin (cZ).[144,145] Bacterial tRNA-IPT is ubiquitous in almost all bacterial and plant genomes, although it is not present in Archaea and *Mycoplasma* bacteria. Compared to the tRNA-IPT, bacterial adenylate IPT is only found in a few bacterial lineages, including the phyla *Actinobacteria*, *Cyanobacteria*, and *Proteobacteria*,

as well as in a limited number of eukaryotic species including the slime mold, *Dictyostelium discoideum*.[115]

Most of the bacterial CK genes characterized to date are plant-associated, mainly produced by tRNA-IPTs, with a small portion produced by adenylate IPTs. tRNA-IPTs have been identified in PGPB as well as in plant pathogenic bacteria. *Agrobacterium tumefaciens* was the first characterized bacterial species that produces CK. It has a Ti-plasmid that contains an *IPT* gene encoding CK biosynthesis. After colonization in the endosphere, *A. tumefaciens* transfers its T-DNA into the host cell DNA and alters the plant cell's gene expression, resulting in significant increases in CK and auxin levels and subsequent crown gall formation.[118] Plant pathogenic bacteria have two modes of CK production, with some producing only one type of CK during plant infection, while others produce a diverse set of CK species.[146] For example, the bacterium *Rhodococcus fascians* produces a mixture of CKs that facilitate plant gall formation including iP, tZ, and cZ;[147] the *A. tumefaciens* Ti plasmid has two genes, *tzs* and *tmr*, that synthesize tZ and iP-type CKs, respectively.[118] The bacterial-CK concentration in turn affects bacterial pathogenicity. During the early stages of infection, the lack of tZ synthesis reduced the transformation frequency of several plants by *A. tumefaciens*.[148] Many pathogenic species also increase CK levels at the infection site, which creates an active sink and results in the remobilization of plant nutrients from other areas to the sink, thus avoiding local plant cell death and helping pathogens survive for a longer time.[149]

CKs are also produced by plant growth-promoting bacteria (PGPB), as part of a variety of mechanisms these bacteria use in their facilitation of plant growth. In addition to CK production, such mechanisms also include the acquisition of resources such as fixed nitrogen, iron and phosphorous and the modulation of plant phytohormone levels including ethylene, auxins, and gibberellins.[116,117] PGPB-produced CKs can enhance plant growth by stimulating plant cell division and help to alleviate environmental stress.[150] CK-producing PGPB such as *Bacillus subtilis* inoculated onto *Platycladus orientalis* (a tree in the cypress family) allow the plant to survive in a drought environment, showing a higher shoot dry weight compared to the same plant grown in the absence of the bacterium;[151] *Sinorhizobium* strains engineered to synthesize higher levels of CKs have been found to enhance the drought tolerance of alfalfa;[152] *Methylobacterium* strains produce high levels of CKs that enhance plant productivity and stress resistance.[145] However, these PGPB only produce tRNA-IPTs, and *Nostoc* sp. PCC 7120 is the only PGPB that has been experimentally shown to synthesize adenylate IPTs to produce CK.[153]

An interesting question concerns how CK production can, in some cases, cause harm to the plant host, while promoting beneficial plant growth in other instances. One potential explanation is that the overproduction of CK by plant pathogenic bacteria over-stimulates

plant growth facilitating tumor formation, whereas PGPB secrete a lower amount of active CK that benefits but does not overstimulate the plant. CK homeostasis can be affected by several enzymes in bacteria (Figure 3.1), which may also result in differences in CK levels between PGPB and pathogens. Isopentenyl transferase (IPT) catalyzes CK production by transferring the isoprenoid moiety from dimethylallyl pyrophosphate (DMAPP) or tRNA. The plant *LOG* (LONELY GUY) gene (also present in some bacteria) plays a role in the direct activation of CK by catalyzing a phosphoribohydrolase reaction, which transfers CK from the nucleotide form into the active free-base form.[154] Cytokinin dehydrogenase (CKX) is so far the only enzyme known to catalyze the irreversible degradation of CK into adenine and side chain.[143] Free base CKs can conjugate with other molecules such as riboside and glucose that results in the formation of different CK structures, which lowers their affinity to the plant CK receptor compared to tZ, the most active CK form.[155]

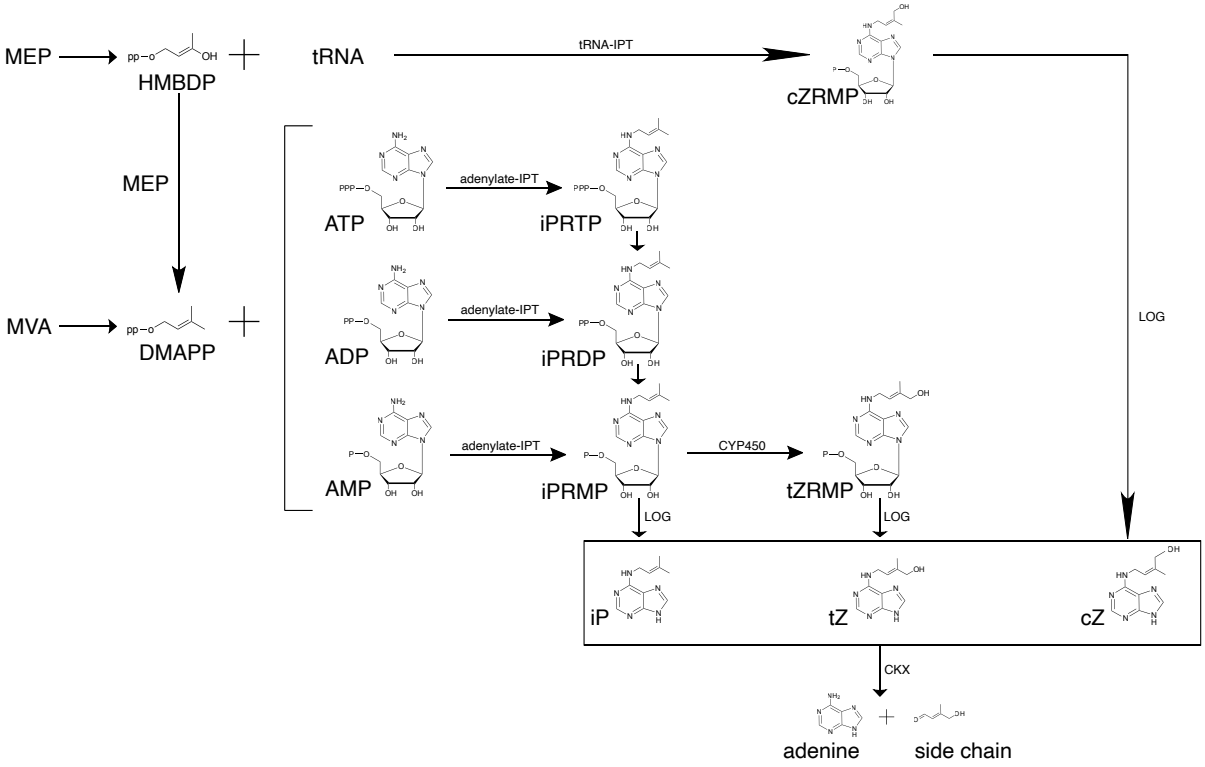


Figure 3.1: General pathway for CK biosynthesis. DMAPP generated from the mevalonate (MVA) or methylerythritol phosphate (MEP) pathway can attach to the adenosine phosphate by adenylate IPT. HMBDP produced as an intermediate in the MEP pathway can bind to tRNA by tRNA-IPT. The LOG enzyme activates CK by releasing the free form of CK from CK nucleotides, such as the transformation from iPRMP to iP. iPRMP can be further hydroxylated into tZRMP by cytochrome P450 monooxygenase (CYP450). tZRMP and cZRMP can be activated into *trans*-zeatin (tZ) and *cis*-zeatin (cZ) by the LOG enzyme. Functional CK including iP, tZ and cZ can be degraded by the CKX enzyme into a side chain and adenine or its derivatives.

To date, adenylate IPT in bacteria is less studied than tRNA-IPT, primarily due to the limited presence of adenylate IPT in bacteria.[115,156] Most of the bacteria that encode adenylate IPT are plant pathogenic.[118,147,157] Thus, the inhibitory effect of adenylate IPT from plant pathogenic bacteria gets more attention than the beneficial effect of adenylate IPT from PGPB. Furthermore, the possible reason for PGPB and plant pathogenic bacteria producing CK that affect plant phenotypes in contrasting ways is less investigated. Therefore,

in the present study, we aimed to computationally investigate the potential differences in adenylate IPT function and regulation, between PGPB and plant pathogenic bacteria. We hypothesize that there are potential differences in IPT regulation and activity between PGPB and plant pathogenic bacteria, and plant pathogenic bacteria may have a higher IPT expression compared to PGPB, leading to an excessive amount of CK production that can overstimulate plant growth through plant-bacteria interaction.

Here, using the AnnoTree[124] and AnnoView platforms for comparative microbial genomics, we have performed a comprehensive bioinformatic survey and analysis of adenylate *IPT* genes across different bacterial species, to gain insights into their evolution, genomic and functional differences. We analyzed 98 *IPT* genes identified from 27 PGPB, 49 pathogenic species, and 22 unclassified species, and compared them based on their genomic neighborhoods, phylogenetic history, copy number, codon usage, and promoter sequences. Our analysis classified bacterial adenylate *IPT* genes into PGPB-associated and pathogen-associated groups phylogenetically, and revealed key differences in *IPT* gene regulation, copy number and genomic context that reflect differences in how these enzymes may be used to either benefit or cause disease to plant hosts.

3.2 Methods

IPT sequence dataset construction

To collect isopentenyl transferase protein sequences in the bacterial genomes, we used AnnoTree, which is a web server that enables protein family exploration across bacteria and archaea.[124] AnnoTree was used to search for protein sequences matching (*E*-value <1e-5) the IPT Pfam domain (PF01745) in the Genome Taxonomy Database (GTDB).[85] AnnoTree results were downloaded as a table that included the protein sequences that matched the IPT domain, gene id, and genome id (Supplementary Table 1). The GTDB species name of each genome was obtained by querying the AnnoTree SQL database with the genome id. To further validate the IPT predictions made by AnnoTree, pfamscan.pl from the EMBL web server was used to analyze all candidate IPT sequences and verify that "IPT" was the top-scoring domain out of all overlapping matches to the Pfam database, with an *E*-value cut-off of 1e-20.[158,159] This step retained 98 IPT sequences and removed 199 matches to IPT belonging to other domain families (e.g., IPPT) (Supplementary Table 2).

Phylogenetic analysis of the IPT protein family

To construct a phylogenetic tree of the IPT protein family, all 98 IPT protein sequences were first aligned to produce a multiple sequence alignment (MSA) using MUSCLE v3.8.31.[160] The MSA was manually trimmed in Aliview to remove highly divergent regions at the beginning and end of the alignment.[161] The trimmed MSA was analyzed using PhyML to infer a maximum likelihood tree with 100 bootstrap replicates, under the LG + G + I + F substitution model.[162]

IPT domain-containing protein sequences from eukaryotes, and IPPT domain-containing sequences from bacteria and eukaryotes, were used as outgroups and aligned with the bacterial IPT protein sequences. The IPPT domain-containing sequences from bacteria and eukaryotes were picked from a previous study by Nishii et al.[115] Candidate IPT domain-containing proteins in eukaryotes were obtained by searching eukaryota for the PF01745 domain in the Pfam database.[30] The presence of IPT domains in candidate eukaryotic IPT proteins was confirmed by searching the PF01745 HMM profile against the protein sequences using HMMER/hmmsearch (version 3.3).[163] Only proteins with *E*-values lower than 1e-10 were considered.

The IPT proteins and IPPT proteins in bacteria and eukaryotes were aligned with MUSCLE v3.8.31.[160] The MSA was also manually edited in Aliview to remove highly divergent regions that could produce phylogenetic artifacts. The phylogenetic tree of bacterial and eukaryotic IPT and IPPT proteins was constructed using PhyML with 100 bootstrap replicates, under the LG + G + I + F substitution model.[162] The output Newick tree files were visualized using ggtree.[164]

Metadata collection

Bacterial species containing predicted IPT proteins were examined in the literature for evidence as to whether these bacteria cause plant disease or promote plant growth. For some of the *IPT*-containing bacterial species, their relationship with the host plants has not been clearly delineated. For example, there is no evidence that *Chondromyces crocatus* is either a PGPB or plant pathogenic bacteria. However, this bacterium seems to produce natural antibacterial compounds.[165] Therefore, we classified it as a PGPB. Some *Rhodococcus* spp. have not been determined to cause plant disease, but through genome sequence analysis, they were found to contain a virulence plasmid. Hence, we classified them as plant pathogenic bacteria.[166] Where no evidence was found, bacteria were classified as unknown. We collected the information and created a metadata table consisting of the isolation source, and associated literature source (Supplementary Table 3).

Promoter retrieval and motif discovery

To find the best promoter candidate rather than simply extract the upstream sequences of the *IPT* genes, a bash script was written to find the closest upstream non-coding sequences of each *IPT* gene over a certain length threshold. To avoid cases where the *IPT* genes were found in an operon, and genes located upstream that are very close or even overlap with the *IPT* genes, the length threshold was set to 100 bp. CD-hit v4.6 removed the redundant sequences with a sequence identity threshold of 99.9%.[167] Promoter sequences exceeding a minimum length of 8 bp were analyzed with the MEME server to scan putative binding site motifs. We tested whether the lengths of the promoters among different IPT groups have significant differences by Wilcoxon rank-sum test. Box plots were created in R using the ggpubr package.[168]

The AME (analysis of motif enrichment) and XSTREME (motif discovery and enrichment analysis) programs in the MEME server were used to perform motif enrichment analysis of the currently known motifs and novel motifs in the regulatory regions of *IPT* genes, respectively.[169,170] The AME program detects the enrichment of known binding motifs in the combined prokaryotes database provided by the MEME server. On the other hand, the XSTREME program detects novel motif enrichment by first using the MEME and STREME (sensitive, thorough, rapid, enriched motif elicitation) programs for *de novo* motif discovery, and then uses the SEA (simple enrichment analysis) program to assess the enrichment of the motifs found in the last step. The *E*-value threshold was set to 0.05 for all the motif analyses.

The predicted promoter regions upstream of *IPT* genes were also scanned for matches to several motifs of interest representing possible plant transcription factor binding sites. These include transcription factor binding motifs for WRKY40 and ARF5, WRKY-boxes (W-boxes) and auxin response elements (AuxREs), which have been previously associated with the regulation of IPT in *Agrobacterium tumefaciens*.[171]

Codon usage analysis

To infer IPT protein abundance from codon usage, we used the effective number of codons (ENC) to quantify the codon usage bias of the *IPT* genes, and the codon adaptation index (CAI) as a proxy of expression levels of the *IPT* genes. The ENC value ranges from 20 (extreme bias) to 61 (no bias), with 20 indicating only one synonymous codon was used for each amino acid, and 61 indicating all synonymous codons were used.[172] The ENC value for each *IPT* gene was calculated using an online server CAIcal.[173] CAI is a widely

used measure for predicting gene expression levels.[174] It refers to a given coding sequence's synonymous codon usage similarity to a reference set of genes, which usually consists of highly expressed genes, such as those coding for ribosomal proteins. This index is calculated by the geometric mean of the 'relative adaptiveness' of codons, with methionine, tryptophan, and stop codons excluded. The 'relative adaptiveness' of a codon is measured by the ratio of the observed codon frequency to the maximum frequency of the codon that codes for the same amino acid in the highly expressed genes. CAI values range from 0 to 1, with 1 indicating the gene contains the most preferred codons in the highly expressed genes. The ribosomal protein genes were extracted from the genome assembly files using Anvio.[175] This program first annotates the raw genome sequences using prodigal, then runs HMMs of bacterial ribosomal protein genes in the annotated genes.[125] A python package, CAI, was used to calculate the CAI value for each *IPT* gene using the ribosomal proteins in the corresponding genomes as the reference gene set.[176] To determine whether *IPT* genes are high or low expression genes, for each *IPT* gene, we calculated the Z-score of its CAI to determine how far the *IPT* gene's CAI deviates from the mean CAI in the corresponding genome. The genomic average CAI was calculated from the CAI of all the coding sequences over 300 bp in the genome. We then tested whether different IPT clades or groups have significant differences in ENC and Z-score of CAI by Wilcoxon rank-sum test. Box plots were plotted in R using the ggpubr package.[168]

Analysis of *IPT* gene neighborhoods

The bacterial genome files in the GFF format and Pfam predictions of open reading frames (orfs) were retrieved from AnnoTree (release 95).[124] The coordinates and Pfam annotations for genes 10 kb upstream and downstream of *IPT* genes were extracted using a bash script. A table representing the presence and absence of important *IPT*-related genes (e.g. *LOG* & *CKX*) in each bacterial genome was then generated and imported into R. Since the LOG enzyme is often annotated as a possible lysine decarboxylase (PF03641) in the Pfam database, neighboring genes assigned by the Pfam domain "PF03641" were regarded as *LOG* genes.[177] However, proteins annotated with the PF03641 domain do not ascertain that they are true LOG enzymes. Therefore, we only searched for possible lysine decarboxylase domain in the genomic neighborhood, because when the target is located very close to the *IPT* genes, there is a better possibility that the target gene encodes a LOG enzyme. The *CKX* genes were identified by the functional domain of CK dehydrogenase (PF09265). In addition, to determine whether the *IPT*-containing genome can produce CK dehydrogenase, the presence of the *CKX* genes in the whole genome rather than the genomic neighborhood of the IPT was analyzed. The presence and absence data and the phylogenetic tree of IPT

proteins were visualized using ggtree.[164]

Evaluating pathogenicity of *IPT* genomic neighborhoods by PathFam scores

The pathogenicity of *IPT* genomic neighborhoods was determined with the help of PathFams, which is a database that quantified the abundance of domain families (Pfam v. 32.0) in pathogen versus non-pathogen genomes. For each Pfam domain, PathFams calculated a pathogen-association score as the fold change of the domain frequency in pathogen proteomes compared to non-pathogen proteomes.[178] Here, we used the pathogen association to determine the pathogenicity score of all proteins encoded within *IPT* genomic neighborhoods. A domain with the strongest pathogen association was selected to represent the pathogenicity of a multi-domain protein. Similarly, a domain with the strongest pathogen association in the IPT genomic neighborhood, excluding the IPT domain, was selected to represent the pathogenicity of the genomic neighborhood and used for pathogenicity comparison between different neighborhoods.

3.3 Results

Identification of IPT domain-containing proteins in bacterial genomes

To collect bacterial IPT domain-containing proteins, we used the PF01745 (IPT) Pfam protein domain to query the AnnoTree database with the default parameters. This search returned a total of 298 protein sequences. Further analysis using pfamscan.pl revealed that most of these sequences were more closely related to the IPPT domain. After the removal of these sequences, 98 verified IPT-domain proteins remained, which are associated with 90 distinct bacterial genomes (Figure 3.2a, see Supplementary Table 2). There are 30,238 bacterial genomes in the AnnoTree database, and the low number of *IPT*-containing bacterial genomes found (0.3%) indicates that IPT proteins are relatively uncommon and sparsely distributed across the bacterial tree of life, consistent with previous analyses.[115] Most of these organisms are from specific lineages within the *Actinobacteria* and *Proteobacteria* phylum, and only a few are from the *Myxococcota* and *Cyanobacteria* phylum (Figure 3.2b–e). Their distribution is also phylogenetically scattered, indicative of multiple independent gains of *IPT* genes in different lineages of the bacterial tree.

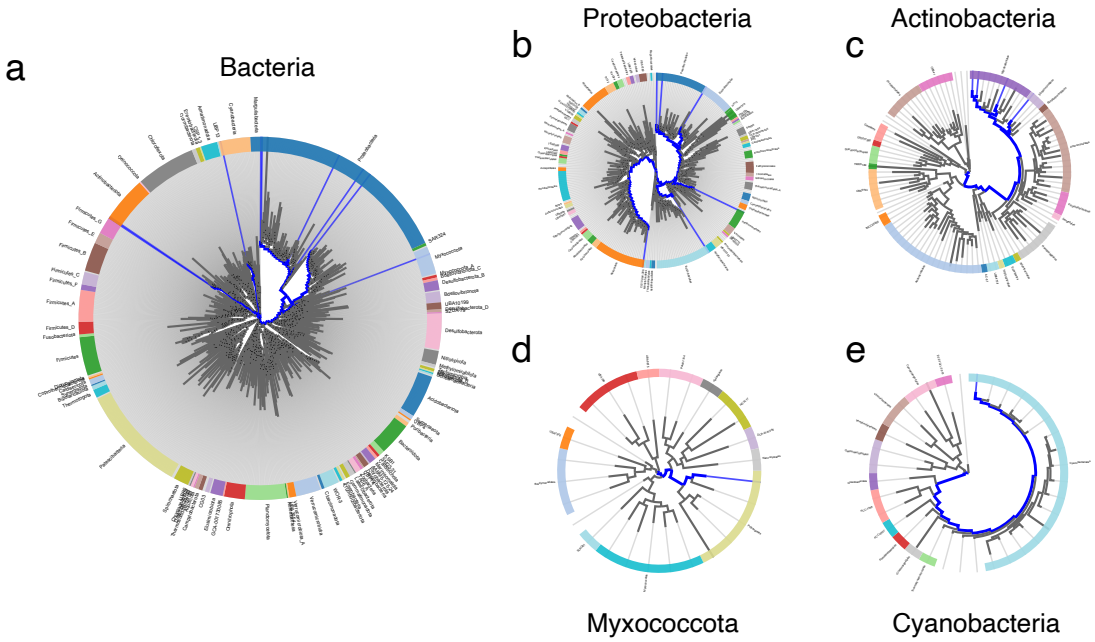


Figure 3.2: The phylogenetic distribution of IPT proteins in AnnoTree. (a) The phylogenetic distribution of the 98 IPT proteins in the bacterial tree of life. (b-e) The phylogenetic distribution of the 98 IPT proteins in the subset of *Proteobacteria*, *Actinobacteria*, *Myxococcota* and *Cyanobacteria*.

Next, we checked the literature to evaluate and classify these bacteria into different phenotypic categories, including plant pathogenic bacteria and PGPB. Those lacking known phenotypes were designated "unknown" (Supplementary Table 3). It is important to note that the genomes detected by AnnoTree are based on the GTDB.[85] The corresponding NCBI species names are provided in Supplementary Table 4. Of the 90 *IPT*-containing bacterial species, 49 are plant pathogens and 27 are PGPB. Most of the plant pathogenic bacterial species occurred within the *Proteobacteria*, and most of the PGPB are within the *Actinobacteria* (Supplementary Table 2).

Phylogenetic analysis of bacterial IPT proteins and comparison between PGPB versus plant pathogenic species

To examine the evolutionary history of bacterial IPT-domain proteins, we constructed a phylogenetic tree of the 98 IPT protein sequences (Figure 3.3). The maximum-likelihood

(ML) phylogeny shows that the IPT protein family can be divided into two main clades, "A" and "B". Interestingly, these two IPT clades showed an association with organism lifestyle, i.e., species in clade A are predominantly plant pathogens from the phylum *Proteobacteria*, while in clade B, except for those species that failed to be classified into either group, most species were found to be plant growth-promoting species from the phylum *Actinobacteria* (Figure 3.3, Supplementary Table 2).

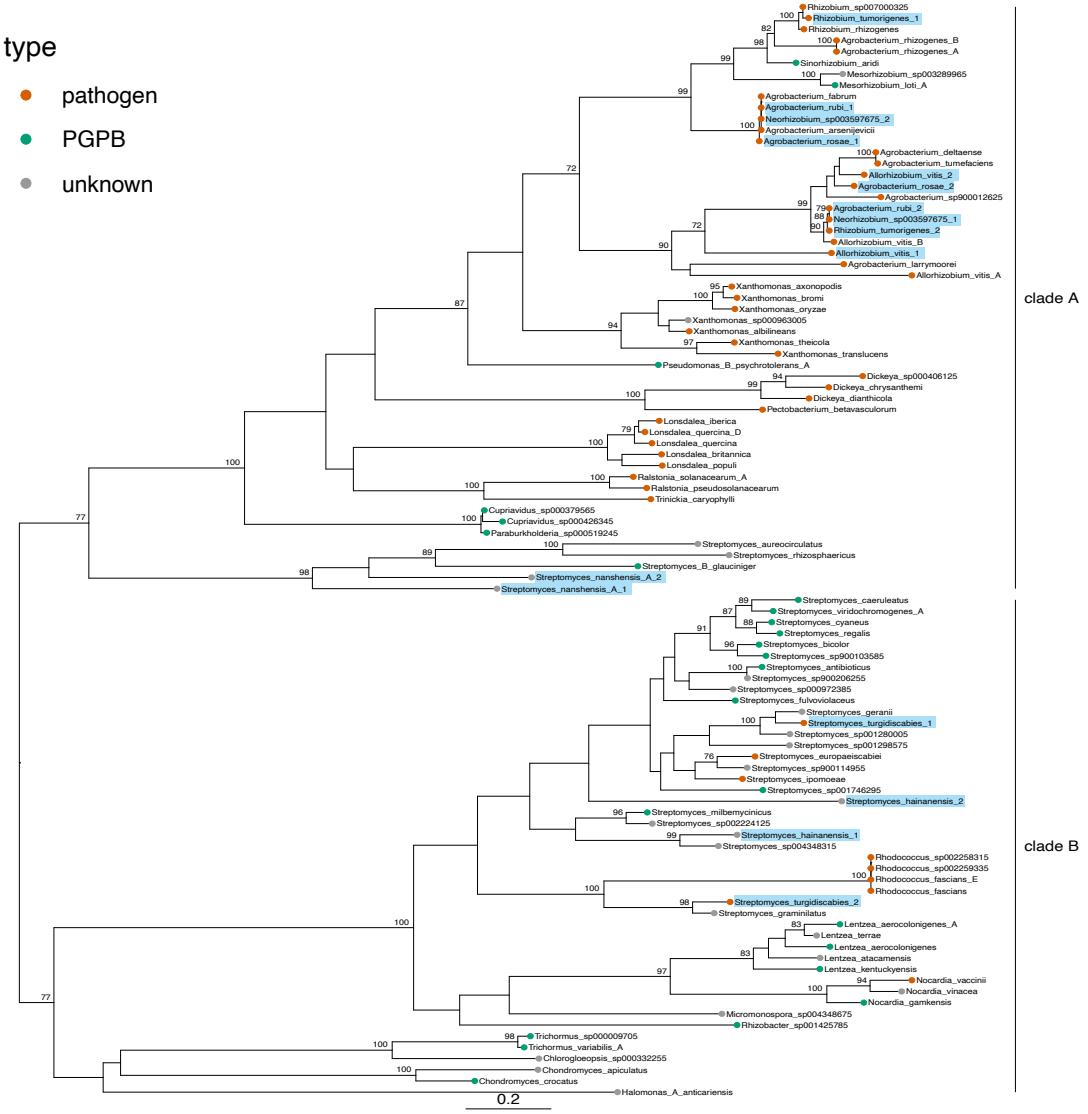


Figure 3.3: Phylogenetic tree of bacterial adenylate IPT proteins. Phylogenetic tree of 98 IPT domain proteins in bacteria inferred using PhyML, with bootstrap values over 70 indicated. The tree was divided into two clades, labelled as clade A and clade B. *IPT*-containing bacteria in clade A shows a trend of being mostly pathogenic to plants, while in clade B, they were mostly harmless or beneficial to plant growth. Species with two *IPT* copies are highlighted in blue.

The IPT protein phylogeny generally follows the species tree, indicative of a vertical evolution of *IPT* genes. However, there are several phylogenetic incongruences that may be due to a non-vertical evolution (e.g., lateral transfer events) of select *IPT* genes. A small clade including *Sinorhizobium aridi*, *Mesorhizobium* spp. and *Neorhizobium* sp., clustered with the *Agrobacterium* clade with strong bootstrap (99/100) support, but was expected to be placed outside the *Agrobacterium* clade according to the species phylogeny. Also, IPT sequences from several *Streptomyces* spp. were included in clade A, rather than clustering with the other *Streptomyces* in clade B. Another incongruence was the phylogenetic placement of IPT sequences from *Rhodococcus* spp. within the *Streptomyces* clade rather than outside as expected. Finally, IPT from *Halomonas anticariensis* was assigned to clade B, although based on taxonomy, it was expected to cluster within the other *Proteobacteria*l sequences in clade A. It is noteworthy that in the IPT phylogeny, there are a few clades with zero branch lengths, because their IPT protein sequences were identical. For example, in clade A, we found a subclade with *Agrobacterium* spp. and *Neorhizobium* sp. including identical proteins. In clade B, different species within the *Rhodococcus* clade also were found to encode identical IPT proteins.

We also included eukaryotic IPT proteins for comparison (Figure 3.4). Eukaryotic IPT sequences were obtained from the Pfam database and verified by pfamscan.pl as IPT-domain containing proteins. After filtering, we identified two "bacterial-like" eukaryotic IPT proteins from *Dictyostelium discoideum* and *Dictyostelium purpureum*. According to the phylogenetic tree of the IPT and IPPT proteins from bacteria and eukaryotes, the IPPT proteins outgrouped from the IPT proteins, and eukaryotic IPT proteins formed a sister lineage that outgrouped from the bacterial IPT proteins. This indicates that the IPT proteins evolved as a distinct protein lineage separate from IPPT family proteins. Also, the extremely rare presence of eukaryotic adenylate IPT domain proteins suggests that they are likely to have been passed from the *IPT*-containing bacteria to the slime molds, as opposed to being transferred from eukaryotes to bacteria. However, the latter scenario cannot be completely discounted.

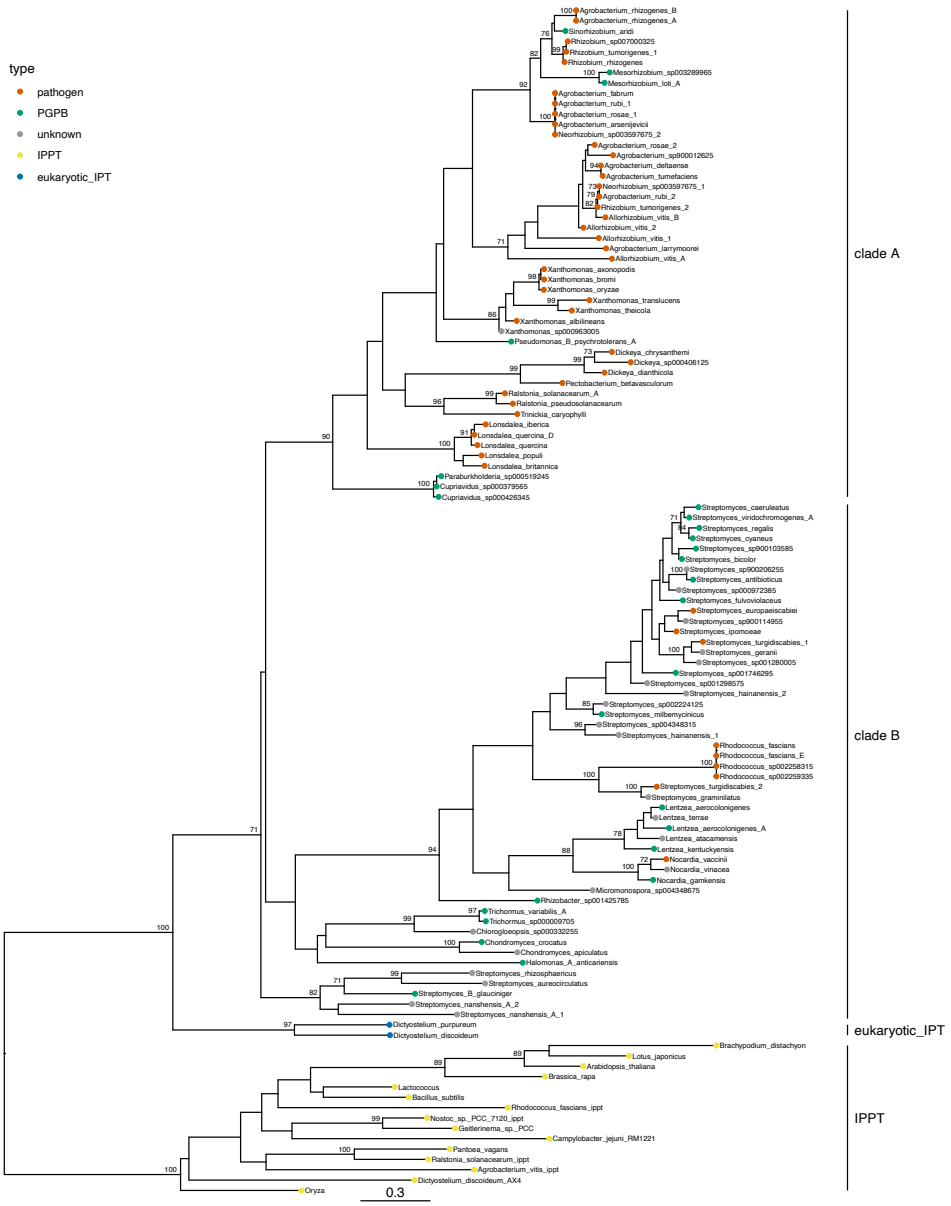


Figure 3.4: Phylogenetic tree of IPTs from bacteria and eukaryotes, and IPPTs from bacteria and eukaryotes. The bacterial and eukaryotic IPPTs are outgrouped from the IPTs. The eukaryotic IPTs from two slime molds are the closest group to the main bacterial IPT clade.

Copy number of *IPT* varies between species

Since 98 IPT proteins were identified in 90 distinct bacterial genomes, this suggests the occurrence of *IPT* gene duplicates, which have been observed in previous studies.[118,179] We therefore examined the number of *IPT* copies in each genome. As seen in Figure 3.3, eight bacterial species were found to possess duplicate *IPT* copies. Interestingly, *IPT duplicates* were predominantly associated with plant pathogens. That is, six pathogenic species were found to possess duplicate *IPT* gene copies: namely, *Rhizobium tumorigenes*, *Neorhizobium* sp., *Agrobacterium rubi*, *Agrobacterium rosae*, *Agrobacterium vitis*, and *Streptomyces turgidiscabies*. Two additional *Streptomyces* species in the "unknown" category also possessed *IPT duplicates* (*Streptomyces nanshensis* and *Streptomyces hainanensis*), whereas duplicate IPTs were not found in any PGPB species. We speculate that the increased occurrence of *IPT duplicates* in pathogens may contribute to increased levels of CK synthesis which is known to result in harmful impacts to the plant host (see Discussion 3.4).

Codon usage analysis of *IPT* genes in bacteria

Next, we sought to determine whether *IPT* genes vary in their codon usage patterns among different groups, as we hypothesized that there may exist different selective pressures acting on *IPT* genes from plant pathogens versus PGPB organisms. In addition to calculating the effective number of codons (ENC), we also calculated the codon adaptation index (CAI) of each *IPT* gene and compared these values to the mean CAI values measured for other genes in the same genome, resulting in a CAI Z-score.[180,181] The Z-score transformation minimizes artifacts resulting from comparisons of CAI scores between genomes of very different codon compositions (*Proteobacteria* versus *Actinobacteria*). We then mapped these CAI Z-scores and the value of ENC onto the phylogenetic tree.

As shown in Figure 3.5a, almost all *IPT* genes have negative CAI Z-score values, indicating lower CAI values than other genes in the same genome, and potentially reduced expression levels compared to other genes. Interestingly, *IPT* genes from clade A (average Z-score of -1.59) deviates from the average to a greater extent than clade B (average Z-score of -1.25) ($p = 0.004$), and *IPT* genes from pathogens (average Z-score of -1.74) deviate to a greater extent from those from PGPB (average Z-score of -1.24) and "unknown" (average Z-score of -0.97) (see Figure 3.5b and 3.5c). We also detected significant differences in ENC values between the two clades and between pathogens and PGPB (Figure 3.5d and 3.5e).

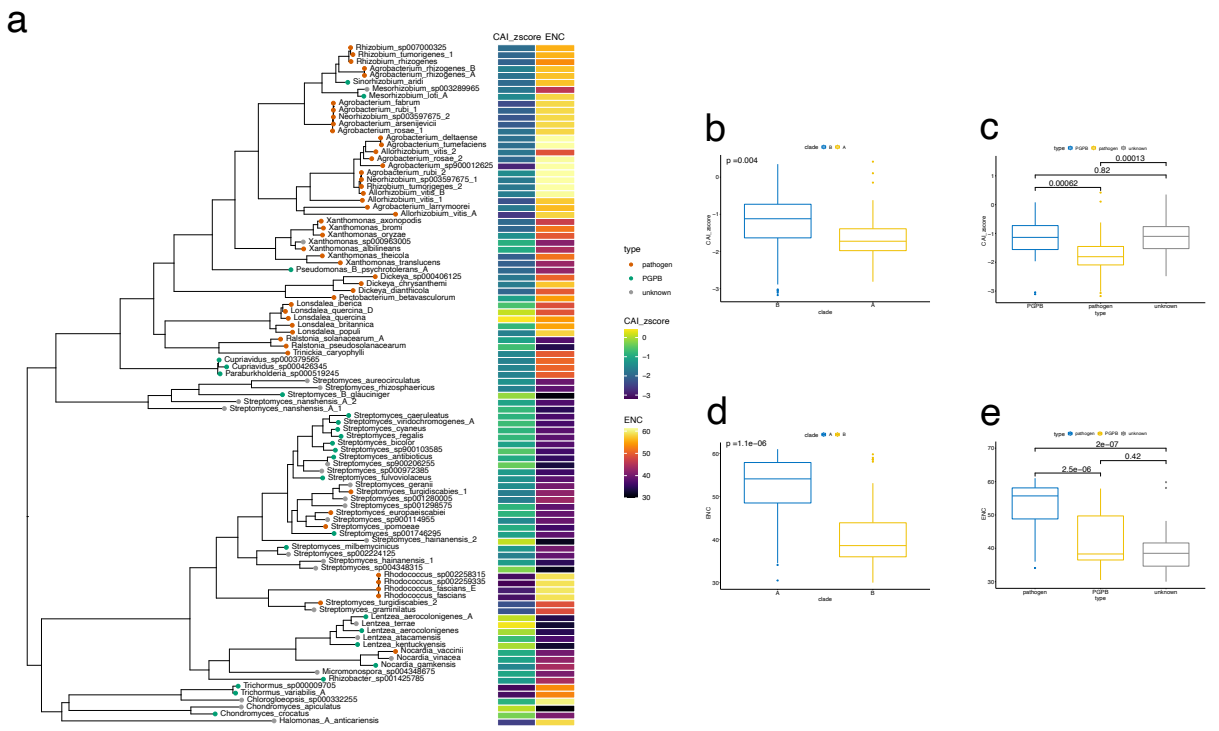


Figure 3.5: Comparison of codon usage patterns between *IPT* genes from the pathogen-associated and PGPB-associated clades. (a) Heatmap showing the CAI deviation scores and ENC values of *IPT* genes in bacterial genomes along with the phylogenetic tree. (b) Boxplot of the *IPT*'s CAI deviation scores in clade A and B. (c) Boxplot of the *IPT*'s CAI deviation scores in PGPB, pathogenic bacteria and unknown groups. (d) Boxplot of *IPT*'s ENC in clade A and B. (e) Boxplot of *IPT*'s ENC in PGPB, pathogenic bacteria and unknown groups.

Ultimately, these results suggest that pathogen-associated *IPT* genes in clade A display atypical codon usage patterns compared to other genes, whereas PGPB-associated *IPT* genes in clade B have codon usage preferences that are more similar to those of other genes in their corresponding genomes.

Genomic neighborhoods surrounding *IPT* genes

To explore the genomic neighborhoods surrounding bacterial *IPT* genes, we used AnnoView to extracted the neighboring genes 10 kb upstream and downstream regions of all predicted

IPT genes and examined their functional annotations (Pfam). This resulted in a total of 408 annotated and 331 unannotated genes. We then analyzed the frequencies of Pfam annotations occurring in these genes (see Figure 3.6a for the top 10 and Supplementary Table 5). Among the top 10 most frequent annotations, we observed several protein families that are related to CK biosynthetic or degradation pathways (Figure 3.1). These include a protein domain annotated as "possible lysine decarboxylase" (PF03641) that encodes the LOG enzyme for CK activation, and the protein domain "cytokinin dehydrogenase 1, FAD and cytokinin binding" (PF09265) that encodes CKX enzymes for CK degradation. We also identified the domain family PF02027 (RolB/RolC glucosidase family), which includes the RolC protein that can release CK from glucoside conjugates.[182] We focused further on the co-occurrence of *LOG* and it CKX and *rolC* genes in the *IPT*-containing bacterial species, paying additional attention to the genes that have the LOG and CKX domains in the same genomic neighborhood as IPT.

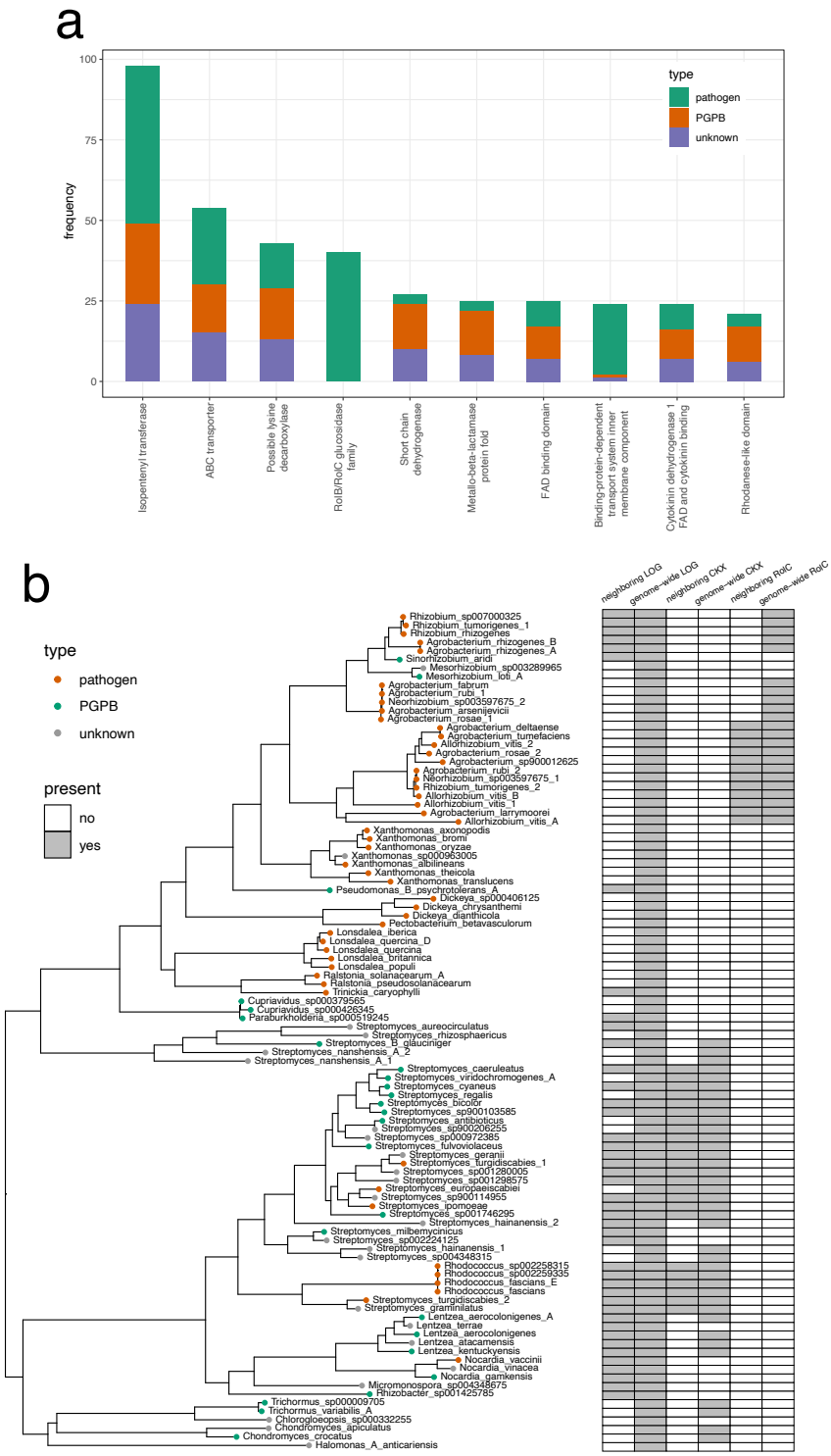

a
type
pathogen
PGPB
unknown
frequency
100
75
50
25
0
Isopentenyl transferase
ABC transporter
Possible lysine decarboxylase
RolB/RolC glucosidase family
Short chain dehydrogenase
Metallo-beta-lactamase protein fold
FAD binding domain
Binding-protein-dependent transport system inner membrane component
Cytokinin dehydrogenase 1 FAD and cytokinin binding
Rhodanese-like domain
b
type
pathogen
PGPB
unknown
present
no
yes
neighboring LOG
genome-wide LOG
neighboring CKX
genome-wide CKX
neighboring RolC
genome-wide RolC
Rhizobium_sp007000325
Rhizobium_tumorigenes_1
Rhizobium_rhizogenes
Agrobacterium_rhizogenes_B
Agrobacterium_rhizogenes_A
Sinorhizobium_aridi
Mesorhizobium_sp003289965
Mesorhizobium_loti_A
Agrobacterium_fabrum
Agrobacterium_rubi_1
Neorhizobium_sp003597675_2
Agrobacterium_arsenijevicii
Agrobacterium_rosae_1
Agrobacterium_deltaense
Agrobacterium_tumefaciens
Allorhizobium_vitis_2
Agrobacterium_rosae_2
Agrobacterium_sp900012625
Agrobacterium_rubi_2
Neorhizobium_sp003597675_1
Rhizobium_tumorigenes_2
Allorhizobium_vitis_B
Allorhizobium_vitis_1
Agrobacterium_larrymoorei
Allorhizobium_vitis_A
Xanthomonas_axonopodis
Xanthomonas_bromi
Xanthomonas_oryzae
Xanthomonas_sp000963005
Xanthomonas_albilineans
Xanthomonas_theicola
Xanthomonas_translucens
Pseudomonas_B_psychrotolerans_A
Dickeya_sp000406125
Dickeya_chrysanthemi
Dickeya_dianthicola
Pectobacterium_betavasculorum
Lonsdalea_iberica
Lonsdalea_quercina_D
Lonsdalea_quercina
Lonsdalea_britannica
Lonsdalea_populi
Ralstonia_solanacearum_A
Ralstonia_pseudosolanacearum
Trinickia_caryophylli
Cupriavidus_sp000379565
Cupriavidus_sp000426345
Paraburkholderia_sp000519245
Streptomyces_aureocirculatus
Streptomyces_rhizosphaericus
Streptomyces_B_glauciniger
Streptomyces_nanshensis_A_2
Streptomyces_nanshensis_A_1
Streptomyces_caeruleatus
Streptomyces_viridochromogenes_A
Streptomyces_cyaneus
Streptomyces_regalis
Streptomyces_bicolor
Streptomyces_sp900103585
Streptomyces_antibioticus
Streptomyces_sp900206255
Streptomyces_sp000972385
Streptomyces_fulvoviolaceus
Streptomyces_geranii
Streptomyces_turgidiscabies_1
Streptomyces_sp001280005
Streptomyces_sp001298575
Streptomyces_europaeiscabiei
Streptomyces_sp900114955
Streptomyces_ipomoeae
Streptomyces_sp001746295
Streptomyces_hainanensis_2
Streptomyces_milbemycinicus
Streptomyces_sp002224125
Streptomyces_hainanensis_1
Streptomyces_sp004348315
Rhodococcus_sp002258315
Rhodococcus_sp002259335
Rhodococcus_fascians_E
Rhodococcus_fascians
Streptomyces_turgidiscabies_2
Streptomyces_griseiniliatus
Lentzea_aerocolonigenes_A
Lentzea_terrae
Lentzea_aerocolonigenes
Lentzea_atacamensis
Lentzea_kentuckyensis
Nocardia_vaccinii
Nocardia_vinacea
Nocardia_gamkensis
Micromonospora_sp004348675
Rhizobacter_sp001425785
Trichormus_sp000009705
Trichormus_variabilis_A
Chlorogloeopsis_sp000332255
Chondromyces_apiculatus
Chondromyces_crocatus
Halomonas_A_anticariensis

Figure 3.6: Genomic context analysis of bacterial *IPT* genes visualized using AnnoView. (a) The histogram of the top 10 annotated Pfam families in the *IPT* genomic neighborhood, colored by PGPB, pathogen and unknown groups. (b) The co-occurrence pattern of the two important genes related to CK levels in the close genomic neighborhoods and in the genome. The left side is the phylogenetic tree of the IPTs. Neighboring genes containing a possible lysine decarboxylase domain (PF03641), a cytokinin dehydrogenase domain (PF09265) and RolB/RolC glucosidase family (PF02027) were regarded as homologs of the *LOG*, *CKX*, and *rolC* sequences, respectively.

The presence/absence of predicted *LOG*, *CKX*, and *rolC* genes was visualized alongside the IPT protein phylogenetic tree (Figure 3.6b). We found that 42 and 24 of the *IPT* genomic neighborhoods contained the *LOG* and *CKX* genes, respectively (Figure 3.6b). Interestingly, we detected the *LOG* gene in only 10 members (19%) of clade A, and in a larger number (32 members, 71%) of clade B. When comparing pathogens and PGPB, *LOG* was present in 14 pathogen genomes (29%) versus 18 PGPB genomes (67%), suggesting a stronger association with PGPB. Similarly, in the IPT genomic neighborhood, the *CKX* gene was found exclusively in *Streptomyces* spp. and *Rhodococcus* spp. in the PGPB-associated clade B (24 members, 45%). When searching for *CKX* genes genome-wide including regions outside of the *IPT* neighborhood, *CKX* was identified in several additional genomes in clade B (11 total members) and three *Streptomyces* in clade A. *CKX* genes were found in 16 PGPB genomes (59%) versus 8 pathogen genomes (16%). Co-occurrence of all three genes (*IPT*, *LOG*, and *CKX*) in the neighborhood occurred exclusively in 19 (36%) members of clade B. Co-occurrence with the *rolC* gene was exclusive to a subclade of A including *Agrobacterium* species (Figure 3.6b).

Overall, these results revealed an increased tendency for *IPT* genes in PGPB organisms to co-occur with other genes within the CK biosynthesis pathway including *LOG* and *CKX*. This finding is consistent with the hypothesis that bacteria with either LOG or *CKX* genes in their genomic neighborhoods may be able to transform the synthesized CK into active forms, or degrade CK when there is an excessive amount. However, failure to identify genes in the same genomic neighborhood does not mean a lack of LOG enzyme to activate CK. At the same time, bacteria containing both genes in the genomic neighborhood are more likely to be capable of dealing with excessive levels of CK than those without *CKX* genes. By examining the genomic neighborhood of *IPT*, in clade A with most of the plant pathogenic bacteria, we rarely found *CKX* genes in the genomic neighborhoods except for 3 *Streptomyces* from clade A, which suggests possible CK accumulation in clade A species.

Pathogen-associated *IPT*s tend to occur in predicted virulence loci

Although pathogen-associated *IPT* neighborhoods show relatively decreased co-occurrence with other CK-related genes, several *IPT* genes from plant pathogens are known to occur in virulence loci. We therefore reasoned that this may be an additional distinguishing characteristic between pathogen-associated and PGPB-associated *IPT* gene neighborhoods.

To investigate this further, we used our previously developed PathFams resource, which has scored each protein family in the Pfam database based on its statistical enrichment in genomes of pathogens (including phytopathogens) versus non-pathogens. As shown in Figure 3.7, we then visualized each *IPT* neighborhood in the AnnoView browser by coloring genes according to their pathogen-association (PathFams score), where scores >1 indicate pathogen overrepresentation and scores <1 indicate underrepresentation (Figure 3.7b-c). As expected, *IPT* neighborhoods from plant pathogens showed an increased tendency to possess numerous genes with high PathFams scores, indicative of virulence loci (Figure 3.8).

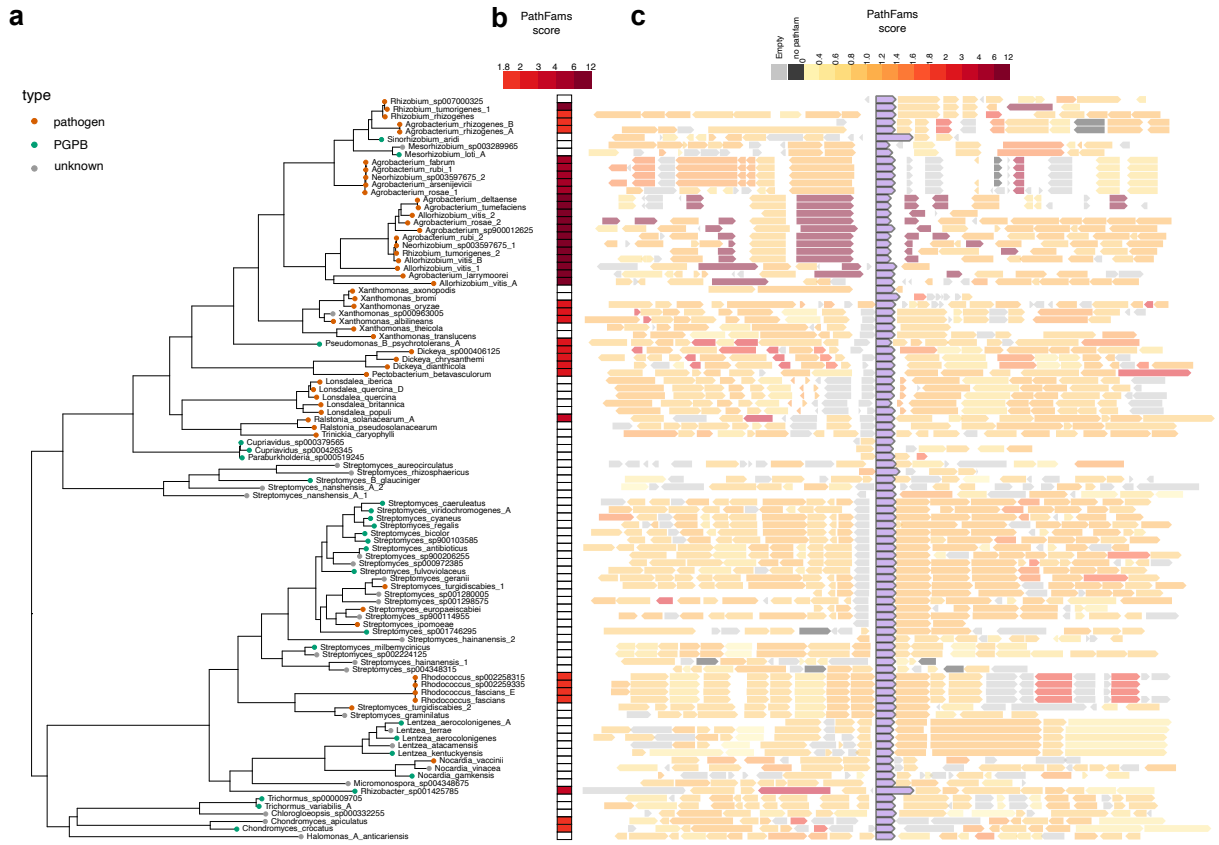


Figure 3.7: Pathogen-associated *IPT* genes tend to reside within predicted virulence loci. (a) Bacterial adenylate IPT phylogenetic tree. (b) Max PathFams score of a gene within the *IPT*-containing genomic region. (c) Genomic neighborhoods surrounding *IPT* genes (pink). Genes are colored according to their PathFams score (fold-change in pathogenic versus non-pathogenic species). Gene neighborhoods were visualized using AnnoView v1.0.

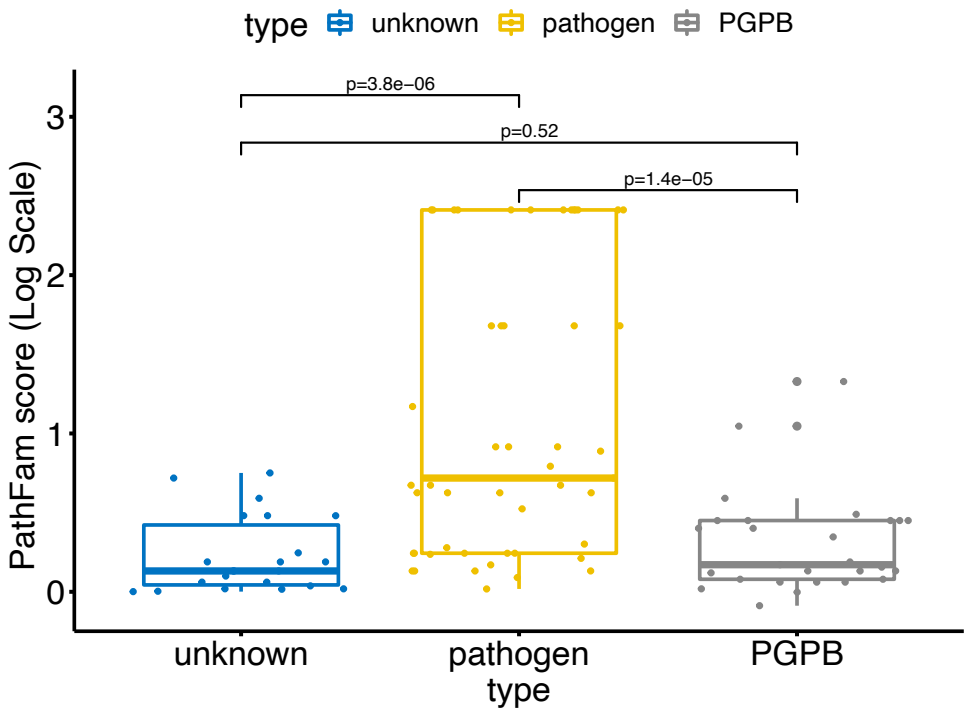


Figure 3.8: Boxplot of the PathFams scores (log scale) of genomic neighborhoods surrounding *IPT* genes in PGPB, pathogenic bacteria and unknown groups

To quantify this pattern, we predicted virulence loci as those containing one or more genes with a PathFams score >1.8, equivalent to a greater than 1.8-fold enrichment in pathogen genomes. (Figure 3.7b). Strikingly, thirty-two (65%) of the pathogen-associated *IPT* neighborhoods satisfied this criterion, whereas only 3 PGPB (11%) and 3 "unknown" (14%) satisfied this criterion. This strongly supported our hypothesis that *IPT* genes are encoded within unique virulence loci in pathogens and cluster with other pathogenicity-associated genes, a trait that distinguishes them from PGPB-associated *IPT*s.

To study *IPT* genomic neighborhoods in greater depth, we chose several representative *IPT* genomic neighborhoods and examined the annotations of neighboring genes (Figure 3.9). *Agrobacterium rubi* has two *IPT*s on different regions of the genome; one is located on the T-DNA region and surrounded by *iaaH* and *iaaM* genes that are involved in auxin biosynthesis; the other *IPT* is located on the virulence region and surrounded by other virulence genes, such as *virA*, *virB* and *virK*. The *iaaH* and *iaaM* genes were also detected in the *IPT* neighborhoods of some other genomes, such as *Agrobacterium rhizogenes*

and *Dickeya chrysanthemi*. The genes that may produce the LOG enzyme were found in the neighborhood of several strains, such as *Agrobacterium rhizogenes*, *Pseudomonas psychrotolerans*, *Lentzea terrae*, *Rhodococcus fascians*, and in one of the *IPT* neighborhoods in *Streptomyces turgidiscabies*. Some other neighboring genes with high PathFams scores were detected in the pathogenic strains. For example, in *Ralstonia solanacearum*, the NEL protein encoding gene that was usually found at the C terminus of bacterial virulence factors was located upstream the IPT coding gene.[183] In *Xanthomonas albilineans*, a lipoprotein coding gene, upstream *IPT*, was involved in the translocation of virulence factors.[184]

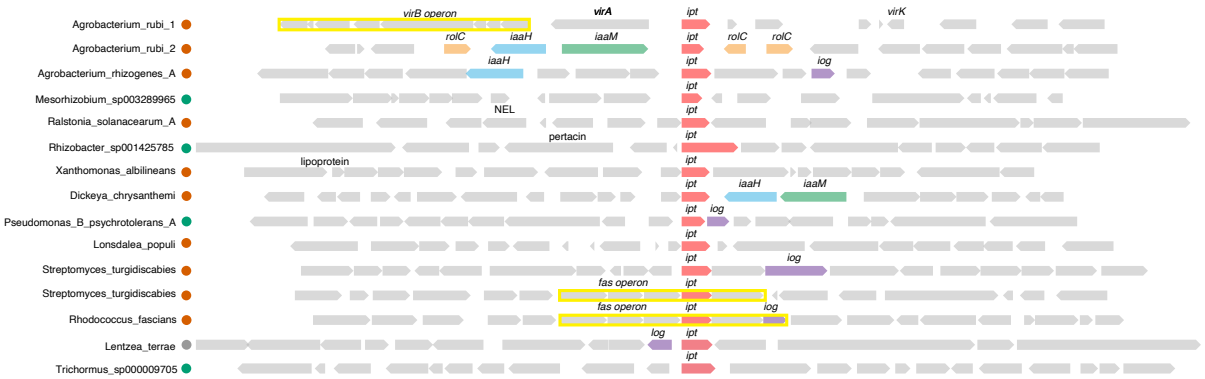


Figure 3.9: Selected *IPT* genomic neighborhoods in PGPB, pathogen and unknown groups

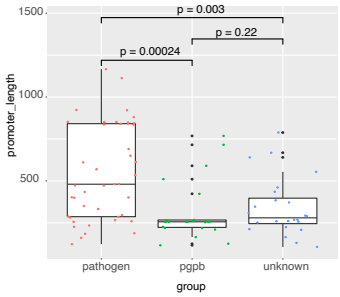
Pathogen-associated *IPT*s have extended promoters with unique regulatory motifs

The upstream non-coding regions surrounding *IPT* genes were examined to assess whether there may be promoter regulatory differences between pathogen and PGPB-associated *IPT* genes. We defined promoter regions as the closest upstream non-coding regions exceeding a length threshold of 100 nt. This procedure identified known *IPT* promoters more accurately than simply retrieving the upstream N nucleotides, which resulted in unwanted overlaps with coding regions of nearby genes.

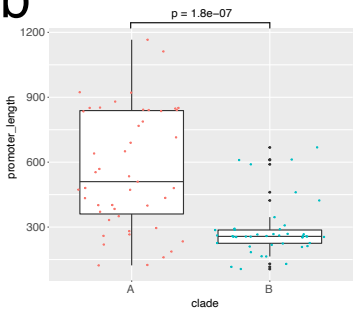
The first pattern that we noticed was a significant difference in promoter length between pathogen-associated *IPT* genes and PGPB-associated *IPT* genes. That is, in pathogens, the closest upstream non-coding regions to *IPT* genes are significantly longer (average length: 573.42 bp), compared to PGPB (average length: 325.88 bp) and "unknown" (average length: 321.36 bp) (Figure 3.10a-b). The identification of extended candidate promoters

upstream of pathogen-associated *IPT* genes suggests that these genes may be under more complex spatiotemporal regulation by transcription factors, requiring specific up-regulation by host-associated environmental cues.

a



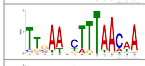
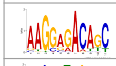
b



c

Known Motif	Name	E-value	TP (%)	Pathogen	PGPB
	Fnr	2.63e-3	14 (11%)	11	0
	XylR	5.99e-3	5 (5.6%)	5	0
	cpxR	1.19e-2	11 (12.2%)	8	0

d

De novo Motif	Rank	E-value	Sits	Pathogen	PGPB
	1	1.80e-010	38	21	9
	2	7.40e-028	40	20	11
	3	7.80e-032	26	17	5
	3	7.80e-015	36	19	12
	5	1.50e-005	22	17	2

e

Motif	Pathogen-associated	PGPB-associated
W-box (TGAC)	43/50	30/40
AuxRE (TGTCNC)	9/50	15/40
AuxRE (TCTCTN)	14/50	12/40
AuxRE (TGTCNC and TCTCTN)	23/50	27/40

Figure 3.10: Analysis of *IPT* promoter sequences. (a) Boxplot of the promoter lengths in the pathogen, PGPB and unknown groups. (b) Boxplot of the promoter lengths in clade A and B. (c) Known motifs discovered in the promoter regions. (d) Novel motifs discovered in the promoter regions. (e) Number of *IPT* promoter sequences with WRKY-boxes (W-boxes) and auxin response elements (AuxREs) in the pathogen-associated clade and PGPB-associated clade.

To identify potential differences in transcription factor motifs, we performed motif discovery and enrichment analyses of the *IPT* upstream promoter regions. First, we used the AME tool from the MEME server to scan for known bacterial motifs enriched in the upstream sequences. The AME analyses revealed that 3 transcription factor binding motifs were significantly enriched in some promoter sequences using an *E*-value threshold of 0.05, which were found to be enriched in the pathogen groups but not in the PGPB group (Figure 3.10c).

Next, we used the XSTREME tool to perform *de novo* motif discovery, which can potentially identify novel motifs highly enriched in the promoter sequences. The enrichment of these novel motifs in the promoter region suggests the potential of these motifs to regulate the expression of IPT proteins in some of the *IPT*-containing bacteria (Figure 3.10d).

We also examined the occurrence of plant W-box and AuxRE regulatory binding motifs, known to be involved in *IPT* regulation when *Agrobacterium* enters plant cells[171] (Figure 3.10e). The W-box motif was found in 43 members of the pathogen-associated clade and 30 members of the PGPB-associated clade, indicating that most *IPT* genes in both groups could be regulated through a W-box motif. AuxREs occurred in 23 members of the pathogen-associated clade and 27 members of the PGPB-associated clade. Thus, fewer bacterial *IPT* genes might be regulated by AuxREs.

Ultimately, these analyses suggest that promoters of *IPT* genes in pathogens may be more sophisticated and possess a unique repertoire of TFBSs to mediate more complex patterns of gene regulation. In the case of *Agrobacterium*, *IPT* loci are transferred to the host plant and therefore utilize host transcription factors for the regulation of expression.[171] In other species, however, further work is required to decipher the TFs involved *IPT* in regulation, and the identified motifs may provide a starting point.

3.4 Discussion

In this work, we performed a comprehensive bioinformatic identification and analysis of adenylate type *IPT* genes in bacteria, and compared these genes and their gene neighborhoods between plant bacterial pathogens and plant growth-promoting bacteria, to understand how differences in *IPT* function may relate to the use of CKs in causing either benefit or harm to plant hosts.

As a rate-limiting enzyme of CK biosynthesis, IPT enzymes, which include adenylate and tRNA-IPT families, are widely distributed across all kingdoms of life.[115] The analysis reported here showed that adenylate type *IPT* genes were rare across bacteria, occurring only in a few bacterial lineages. Most, if not all, bacterial species containing adenylate *IPT* genes were plant-associated and could largely be classified as either plant pathogens or plant growth-promoting bacteria. To date, most of the known adenylate IPTs have been identified in plant pathogenic bacteria, and only the PGPB organism, *Nostoc* sp. PCC 7120, has been demonstrated to produce functional CK in experimental studies.[153] Our work has expanded the number of known adenylated IPTs in PGPB, especially in the *Streptomyces* genus. Note, however, that our survey of adenylate *IPT* genes focused only on the available genomes within the AnnoTree/GTDB database.

The identification of *IPT* genes in plant pathogens and PGPB raises the interesting question as to the mechanism by which the adenylate *IPT* gene may be involved in two seemingly opposing activities: its use by pathogens in causing harm to the plant host by overproduction of CK, and its use by PGPB to provide the right level of CK to benefit the plant host. We examined *IPT* genes from several perspectives including phylogenetics, codon usage, copy number, gene co-occurrence, gene neighborhood, and promoter analysis. This multifaceted analysis revealed at least six trends that, together, provided some insights into how genomic differences in *IPT* genes and their regulation may impact how they contribute to plant pathogenesis versus growth promotion.

The protein phylogeny revealed some separation, although not perfect, between pathogen-associated and PGPB-associated *IPT* genes. That is, clade A was predominantly composed of *Proteobacteria*l pathogens, while clade B was predominantly composed of PGPB and organisms of unknown biological activity. It is possible that this separation merely reflects the evolutionary history of species divergence. However, this phylogenetic separation into clade A and B may also reflect biochemical differences in IPT function and, potentially, the activities of the CKs produced by each organism.

Another distinguishing feature between pathogen and PGPB-associated *IPT* genes was *IPT* gene duplication. That is, *IPT* gene duplicates were observed more frequently in

pathogens than in PGPB. Gene copies in the Agrobacterium and Rhizobium clade, for instance, have resulted from duplication of *IPT*-containing segments (See Figure 3.9). *IPT* gene duplication frequently arises in eukaryotes, especially plants.[185] In plants, copies of *IPT* help express the encoded protein to variable extents,[185] and some of the proteins encoded by these genes may acquire new functions. Thus, *IPT* gene duplication may be viewed as fine-tuning the IPT protein levels in the plant cells.[115] Although *IPT* gene duplication has been well studied in eukaryotes, it is less described in bacteria. Most non-plant genomes retain fewer *IPT* sequences, ranging from 0 to 3 copies compared to the 1–11 copies of this gene in plant genomes.[185] One of the most well-known examples of *IPT* gene duplication in bacteria is the *ipt* (*tmr*) and *tzs* genes identified in the Ti plasmid of *A. tumefaciens*. Both genes possess an IPT domain and encode isopentenyl transferase but are located on different places of the plasmid; ipt resides in the T-DNA region and will incorporate into the plant genome after *A. tumefaciens* infection, whereas *tzs* resides in the non-transferrable part of the plasmid, and only exists in the nopaline type strains of *A. tumefaciens*.[186,187] However, the *IPT* duplication event in *A. tumefaciens* was missed in our phylogenetic tree (Figure 3.3). One possible reason is that in the AnnoTree database, the bacterial genomes are representative genomes in each species cluster. Therefore, it is possible that the chosen representative genome for *A. tumefaciens* is a non-nopaline type strain, which explains the absence of *IPT duplicates* in *A. tumefaciens* in our dataset. Nevertheless, our study revealed additional adenylate *IPT* gene duplicates in bacterial strains other than *Agrobacterium*, including *Neorhizobium*, *Rhizobium*, and *Streptomyces*. On the other hand, we observed that the pathogen-associated clade had a higher occurrence of gene duplicates, suggesting that the organisms in this clade are more likely to produce a higher level of IPT enzymes for CK biosynthesis, therefore have increased pathogenicity and virulence.

Consistent with the idea of differences in IPT activity levels, pathogen-associated and PGPB-associated *IPT* genes also differed in their codon usage patterns, promoter structure and regulatory motifs. This may result in differences in *IPT* gene expression and regulation that govern different levels of CK activity in pathogens versus PGPB. For instance, CAI analysis revealed that pathogen-associated *IPT* genes deviate much more strongly from genome-wide average patterns than *IPT* genes from PGPB. Furthermore, pathogen-associated *IPT* genes were associated with extended upstream promoters and an increased frequency of certain regulatory motifs, which may hint at a more complex mode of regulation for these genes. Previous studies found that bacterial *IPT* gene expression can be regulated in the host eukaryotic cells. For example, a vector that contains the *IPT* gene under the control of the senescence-associated receptor kinase (SARK) gene promoter, was transferred to the plant cells and found to significantly delay leaf senescence in the presence of lowered levels of water by increasing the IPT production.[188] Eukaryotic-host-derived

proteins, including WRKY and ARF transcription factors, were also found to activate transferred bacterial IPT production by binding to the WRKY-boxes (W-boxes) and auxin response elements (AuxREs), respectively,[171] consistent with the frequent occurrences of both regulatory motifs in the bacterial *IPT* promoters. In addition, our analysis also discovered three known regulatory motifs enriched in the pathogen group but not in the PGPB group, which may function as a way of regulating *IPT* gene expression in pathogenic bacteria. Several novel motifs were also found to be enriched in both pathogens and PGPB. However, further studies are required to examine the regulatory function of the detected regulatory motifs experimentally.

Perhaps the strongest feature distinguishing *IPT* genes in pathogens versus PGPB was not related to the *IPT* gene sequence itself, but, rather, its surrounding genomic context. Our analysis of the IPT genomic neighborhood using Pfam and our recently developed PathFams resource showed that, unlike *IPT* genes in PGPB, pathogen-associated *IPT* genes tended to cluster in the bacterial genome with other pathogen-associated genes and reside within candidate virulence loci. Thus, pathogen-associated *IPT* genes may function similarly to those in PGPB at a biochemical level, but function as part of more general virulence pathways through co-expression with other pathogen-associated genes to promote bacterial virulence.

In contrast, PGPB-associated *IPT* genes clustered with non-pathogen-associated genes, and showed a tendency to co-occur with other enzymes in the CK biosynthesis pathway, suggesting added layers of control of CK levels. For example, the *CKX* genes essential for CK degradation were almost missing in the pathogen-associated clade, which may cause CK accumulation and promote bacterial virulence. Although the *LOG* gene necessary for CK activation was not present in the genomic neighborhood of many *IPT*-containing genomes, studies have shown its distribution in major lineages of bacteria. Hence it is possible that these genes were located elsewhere in the genome and may still be able to activate CK.[189] However, since the LOG enzyme was recently found to be misannotated as possible lysine decarboxylase domain due to the high similarity between the two domains,[177] crystal structure analysis will be needed to further confirm the functionality of CK activation of the LOG proteins identified in the genomic neighborhoods.[190,191] The CKX enzyme is well studied in plants but is less represented in bacteria. Only 2.5% of the estimated total bacterial genomes that have been sequenced possess the CKX homologs. Additionally, the function of the CKX proteins in bacteria remains largely undetermined.[156] CKX in different species differs in substrate specificity to various forms of CK, resulting in the dysfunction of degrading various CK forms.[192] Therefore, further experiments such as structural analysis would help determine the functionality of *CKX* genes and its affinity to different forms of CK.

In conclusion, our study has largely expanded an understanding of the roles of adenylate *IPT* genes in bacteria. These genes could be further phylogenetically classified into PGPB-associated *IPT* genes and pathogen-associated *IPT* genes, and were predicted to be related with low or high CK levels produced in bacteria. Pathogens were more likely to contain multiple *IPT* genes, due to gene duplication and horizontal gene transfer, with preferred clustering with other pathogen-associated genes to improve bacterial virulence. Pathogenic organisms lacked the ability to easily degrade CK due to the absence of *CKX* genes, leading to CK accumulation. Furthermore, the promoter analysis revealed potential regulation differences between the two groups, with pathogen-associated *IPT* genes having longer promoter regions, suggesting a more complex regulatory mechanism. Therefore, we predict that the bacteria in the pathogen-associated clade are prone to produce a higher level of IPT, leading to a higher amount of CK that promotes virulence and susceptibility to host plants.

The *IPT*-containing PGPB analyzed in this study are candidates for plant growth-promoting strains applied to crop plants as a biocontrol strategy, with validation of CK production, as well as effectiveness and safety testing required before large-scale application in agriculture. In addition, the CK species generated by plant pathogenic bacteria and PGPB need to be characterized experimentally to examine whether the CK structure differs, since plant CK receptors may have high or low affinity to different CK substrates. Such differences, combined with CK quantity, could influence whether CK results in a harmful or beneficial effect to the plant host. It would also be interesting to quantify the CK levels and compare them between the plant pathogenic bacteria and PGPB. Finally, we predict that the CKX enzymes in PGPB may avoid an over accumulation of CK through regulated CK degradation. However, the CKX enzyme may itself have different affinities to various types of CK. For example, the CKX enzyme was reported to strongly prefer iP-type CK in *Rhodococcus fascians*.[193] Therefore, it is essential to study the substrate preferences of the CKX enzymes as it may also affect the CK homeostasis in bacteria.

Chapter 4

Novel Clostridial neurotoxins identified outside *Clostridium*

Material in this section has resulted in two preprints (Wei et al. (2022),[194] Wei et al. (2023)[195]) available at *biorxiv*. The preprints are available here:

> X. Wei, T. Wentz, B. Lobb, M. Mansfield, W. Zhen, H. Tan, Z. Wu, S. Pellett, M. Dong, and A. C. Doxey. Identification of divergent botulinum neurotoxin homologs in *Paeniclostridium ghonii*. *bioRxiv*, 2022.[194] https://doi.org/10.1101/2022.08.17.504336
>
> X. Wei, B. Lobb, K. Wang, M. Dong, and A. C. Doxey. Identification of a botulinum neurotoxin-like gene cluster in *Bacillus toyonensis*. *bioRxiv*, 2023.[195] https://doi.org/10.1101/2023.07.21.550100

In this chapter, I present the identification of two putative clostridial neurotoxins outside the genus *Clostridium*. These have been designated as PG toxins from *Paeniclostridium ghonii* and BTNT from *Bacillus toyonensis*. I analyze the sequence, structure, phylogenetic relationships of the two putative novel toxins in comparison with other BoNTs. Then, I use the AnnoView platform to analyze the gene neighborhoods of these toxins to gain insights into their potential functions and predict their target hosts. It is hypothesized that the novel toxin genes and their neighboring genes might share functional similarities. By annotating the functions of these neighboring genes, I hope to derive valuable insights about the target hosts of the putative toxins.

4.1 Introduction

Clostridial neurotoxins (CNT) which include botulinum neurotoxin (BoNT) and tetanus neurotoxin (TeNT) are notable as the most potent toxins known to humankind, and they are the molecular agents responsible for botulism and tetanus neuroparalytic diseases. BoNTs target motor nerve terminals and enter neurons via receptor-mediated endocytosis, and then transfer their LC protease domains into the cytosol where they cleave host SNARE (soluble N-ethylmaleimide-sensitive factor attachment protein receptor) proteins including VAMP1/2/3, syntaxin 1, and SNAP25, resulting in blockage of neurotransmitter release.[196–198] *tent* shares the same mode of action with BoNTs, but it undergoes retrograde transport along the axon after entering motor neurons, reaching the spinal cord, getting released from the motor neuron soma, and then reentering inhibitory interneurons, eventually resulting in a distinct phenotype of spastic paralysis. Both forms of paralysis cause death due to respiratory failure.

BoNTs are typical A-B bacterial toxins composed of two polypeptide chains: the light chain (LC), which is the protease domain responsible for SNARE proteolysis, and the heavy chain (HC), which includes the translocation domain (Hn) that transports the LC across the endosomal membrane into the cytosol, and the receptor-binding domain (Hc) that binds to a number of different host cell receptors including polysialogangliosides, synaptotagmin I/II and glycosylated SV2, depending on BoNT serotype.[198] The LC and HC are encoded on a single gene and initially synthesized as a single polypeptide. The linker region between LC and HC needs to be proteolytically cleaved, by either bacterial proteases or host proteases, which is essential for the activity of the toxin. After the linker is cleaved, the LC and HC remain connected via a single interchain disulfide bond. This disulfide bond would be reduced once the LC reaches the cytosol, thus releasing the LC into the cytosol.

BoNTs were classified into seven serotypes (A-G), which include many subtypes.[199,200] Serotypes A, B, E, and F are associated with human botulism, whereas subtypes C, D and hybrid CD are primarily associated with botulism outbreaks in wildlife. The host specificity of BoNT/G is less clear. BoNTs are produced not only by four groups of *C. botulinum* (Group I strains are human-associated and Group III are animal-associated), but also by strains of *C. noyvi*, *C. baratii*, *C. sporogenes*, and *C. butyricum*. BoNTs are encoded chromosomally, on plasmids, or on phages depending on the subtype and strain.[201] Related non-toxigenic strains are also common.[202]

With the growing availability of genomes in public databases, researchers have identified a growing number of BoNT-related toxins bioinformatically (reviewed here[203–206]). In 2015, Mansfield et al. reported the first identification of a BoNT homolog outside of *Clostridium*,

which occurred in the genome of the organism, *Weissella oryzae*.[120] This toxin, labeled "BoNT/Wo" was reported to cleave VAMP within a novel site.[207] Although BoNT/Wo conserves all BoNT domains, phylogenetically it sits on a divergent branch and outgroups the entire BoNT phylogenetic tree. This suggests that the lineage including BoNT/A-G may have evolved as a sublineage of a broader BoNT-like toxin family. Following the identification of BoNT/Wo, Zhang et al. identified BoNT/X in a *C. botulinum* strain, another distinct member of the BoNT family which formed its own lineage and serotype.[208] BoNT/X was demonstrated to cleave multiple VAMPs including several non-canonical substrates that cannot be cleaved by any other BoNTs. A year later, another BoNT-like toxin was discovered within a strain of commensal organism, *Enterococcus faecium*.[122,123] This BoNT, labeled BoNT/En, was shown to cleave both SNAP25 and VAMP2 at unique sites, and reside on a common conjugative plasmid, suggesting a mode of circulation within *Enterococcus* strains. Interestingly, BoNT/En showed several similarities (phylogenetic and gene neighborhood structure) to BoNT/X, clustering with it in the BoNT phylogenetic tree. In 2019, a new family of toxins were also identified in the genome of *Chryseobacterium piperi*, which resembled BoNTs in terms of their domain structure.[209] These toxins have not been shown to cleave SNAREs and to date have unknown substrate specificity. However, their sequence and structural similarities to BoNTs suggest that they are more distant evolutionary relatives, even more so than the divergent BoNT homolog, BoNT/Wo.

Finally, the most recent addition to the BoNT family tree has come from studies of the insecticidal organism, *Paraclostridium bifermentans*. Contreras et al. demonstrated that this organism produces a BoNT-like neurotoxin, labeled PMP1, that has insecticidal activity toward anopheline mosquitoes by cleavage of the mosquito SNARE protein, syntaxin 1.[121] PMP1 clustered within the divergent lineage of BoNTs including BoNT/X and BoNT/En, despite possessing only 34-36% identity to these toxins. In addition, PMP1 shares several genomic features with BoNT/X and BoNT/En, namely, the presence of neighboring *ORFX* genes, which were suggested to play a role in facilitation of PMP1 absorption in the insect gut, since their presence along with PMP1 increased its toxicity.[121] The identification of PMP1 as an insecticidal BoNT-like toxin aligns with earlier work which highlighted the numerous links between BoNTs and insecticidal pathogens.[204] Particularly, these links include the presence of homologs *ORFX* and *HA* genes in insecticidal gene clusters, the detection of partial homologs of BoNTs encoded within the genomes of entomopathogenic fungi, and the identification of BoNT-like gene fragments insect gut metagenomes.[206]

In this work, we identified a novel BoNT homolog gene cluster from *Paeniclostridium ghonii*, as well as a novel putative BoNT-like toxin from a particular strain of *Bacillus toyonensis* ("isolate CH177"), whose genome was recently added to the NCBI database on Apr 12, 2023. The novel gene cluster in *Paeniclostridium ghonii* is conserved in two

distinct strains and also contains a neighboring gene encoding a putative insecticidal Cry toxin.[210] We analyzed these new BoNT homologs in terms of their sequence, structure, phylogenetic relationships to other BoNTs, and gene neighborhood. Our work adds to the growing diversity of BoNT-like toxins.

4.2 Methods

Toxin identification

To identify potential new clostridial neurotoxins in non-clostridium species, the BLAST webserver (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to perform a search for clostridial neurotoxin homologs in the NCBI nr protein database, restricting our search to organisms outside of the *Clostridium* genus. We used BoNT/A1 (accession # WP_021136346.1) as a query with default parameters.

Phylogenetic analysis

The BoNT-like toxins in *P. ghonii* and *B. toyonensis* , PMP1 (WP_150887772.1) as well as representative BoNT sequences for each subtype (downloaded from https://bontbase.org) were used to create multiple protein sequence alignments. Multiple alignments were then constructed using MAFFT v7.407 with the L-INS-i algorithm.[211] Jalview was used for alignment visualization. IQ-TREE version 2.0.3 was used to generate phylogenetic trees for BoNT-like toxins in *P. ghonii* and in *B. toyonensis*, respectively.[212] The alignment for BoNT-like toxins in *P. ghonii* was split into the heavy and light chain regions corresponding to domain boundary definitions from PDB 3BTA. Conserved alignment blocks were then manually selected and used to produce two separate phylogenetic trees with the IQ-TREE algorithm.[212] We then midpoint rooted the trees, collapsed some clades that include toxins from the same subtype, and removed bootstrap values lower than 50%.

Structural modeling of PG-toxin LC and HC sequences

The PG-toxin LC and HC sequences both exhibit sequence similarities with BoNTs. We are interested if they also demonstrate structural similarities to the protease domain, as well as the translocation domain and receptor-binding domain, respectively. Thus, the PG-toxin LC and HC sequences were submitted to the SwissModel resource for automated

protein homology modeling.[213] Structural models were generated using a variety of template structures of existing BoNT related proteins. For PGT1-LC, PDB ID 6bvd.1.A (Botulinum Neurotoxin Serotype HA Light Chain) was used as the top identified template, which had 28.09% identity. For PGT2-LC, PDB ID 5bqn.1.A (Botulinum neurotoxin type D) was used as the top identified template, which had 26.36% identity. For PGT1-HC, 6g5g.1.B (Botulinum neurotoxin type B) was used as the top identified template, which had 25.32% identity. For PGT2-HC, 7by5.1.A (tetanus toxin) was used as the top identified template, which had 26.28% identity. Structural models were colored by QMEANDisCo Local quality estimate scores.

Phylogenetic analysis of *P. ghonii* and *B. toyonensis* strain CH177

For *P. ghonii* strains, we chose the closely related species of *Paeniclostridium ghonii* to build the species phylogeny, including all species in the *Paeniclostridium* genus, as well as *Paraclostridium bifermentans*, which was shown to contain PMP1 toxin, and *Paraclostridium benzoelyticum* that was reported as the species with the highest ANI in the NCBI database. The assembly files in .fna format of all the genomes were download from the NCBI (https://www.ncbi.nlm.nih.gov/assembly) and JGI database.[63] Then, the fastANI program was used to compute pairwise ANI and generate a distance matrix with identity values for the list of the assemblies.[214] The R package phangorn was used to build a phylogenetic tree using the neighbor-joining method based on the ANI distance matrix.[215] The ANIs for each of closely related genome to both *P. ghonii* genomes were averaged and indicated in Figure 4.4.

For *B. toyonensis* strain CH177, a dataset of 310 *B. toyonensis* genomes was downloaded from the NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=155322). A whole genome SNP-based phylogeny was created using Parsnp v1.2[216] with default parameters and the *B. toyonensis* P18 strain (accession # GCF_016605985.1) selected as a reference genome. The predicted proteome (.fasta file) was submitted to the VFanalyzer tool in the VFDB database[217] with the *Bacillus* genus chosen as a reference taxon. Default parameters were used.

Genomic neighborhood analysis

The genomic neighborhoods for ±20kb surrounding the neurotoxin genes were obtained from the NCBI and IMG database in .gbk format. For PG-toxin genes, we downloaded the complete scaffold to collect as much sequence information as possible. The .gbk files

were uploaded to an in-house tool called AnnoView (https://annoview.uwaterloo.ca) for gene neighborhood visualization (manuscript in preparation). Genes in *B. toyonensis* were annotated based on BLAST searches against the NCBI-nr protein database, while all other gene annotations were derived from NCBI and IMG annotations.

Experimental characterization assays

The experimental characterization assays were performed by our collaborators Dr. Pyung-Gang Lee and Dr. Min Dong from Boston Children's hospital, Harvard.

DNA constructs. The DNAs encoding LC/PG1 (residues 1-392), HC/PG1 (residues 1-838), LC/PG2 (residues 1-395), HC/PG2 (residues 1-846), *Drosophila melanogaster* (fly) SNAP25 (residues 1-212, GenBank no. NP_001036641.1) *Anopheles stephensi* (mosquito) SNAP25 (residues 1-212, GenBank no. XP_035899569.1), *Plutella xylostella* (moth) SNAP25 (residues 1-213, GenBank no. KAG7297990.1), *Ixodes scapularis* (tick) SNAP25 (residues 1-214, GenBank no. XP_040073329.1), *Dendroctonus ponderosae* (beetle) SNAP25 (residues 1-211, GenBank no. XP_019769279.1), *Penaeus vannamei* (shrimp) SNAP25 (residues 1-211, GenBank no. ROT68655.1), *Drosophila melanogaster* SNAP24 (residues 1-212, GenBank no. NP_524298.1), *Drosophila melanogaster* SNAP29 (residues 1-284, GenBank no. NP_523831.1) were synthesized by Twist bioscience (South San Francisco, CA).

Protein purification. All the recombinant proteins were expressed and purified from *E. coli* BL21 (DE3). In general, the recombinant *E. coli* was cultured in autoinduction medium (Formedium) at 20 °C overnight. Bacterial pellets were suspended in lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl) and disrupted by sonication. Lysates were collected after centrifugation at 30,000 g from 30 min at 4 °C. His-tagged proteins were purified using Ni-NTA resin (Qiagen), while GST-fused proteins were purified using GSTtrap column (Cytiva). To obtain the complex toxin, Ni-NTA resin and subsequent StrepTrap XT column (Cytiva) were used to purify the complex toxins, which consisted of Strep-II tagged LC and His-tagged HC.

Cleavage assay. The reaction buffer consisted of 50 mM HEPES pH7.4, 5 mM NaCl, 0.1% (v/v) Tween-20, 2 mM DTT and 10 µM of ZnOAc. 10 µM of each SNAP25 were incubated with 2 µM of each LCs for 90 min at 37 °C. The cleaved samples were run on the SDS-PAGE gel, followed by Coomassie blue staining and destaining.

Co-transfection assay. 0.25 µg/well of pcDNA3.1 plasmid encoding HA-tagged SNAP25s (or fly SNAP24) and 0.25 µg/well LC/PG1, PG2, A, E, En were coexpressed in 293T cells (cultured in 24-well) via transient transfection using Polyjet transfection reagents (SignaGen, MD) following the manufacturer's instructions. After 24 h culture, cell lysates were harvested

in RIPA buffer (50 mM Tris, 1% NP40, 150 mM NaCl, 0.5% sodium deoxycholate, 0.1% SDS) and subjected to immunoblot analysis. To detect cleavage of SNAP25s, mouse anti-HA antibody (Biolegend, 16B12) and goat anti-mouse HRP-conjugated IgG (Biorad, Star207P) were used as primary and secondary antibodies, respectively.

Identification of cleavage sites using LC-MS/MS. Samples containing cleavage products were quenched by adding trifluoroacetic acid (TFA) and incubating at 95 °C for 10 min. As a negative control, equivalent samples were prepared as pre-TFA treated and heat-inactivated, before cleavage. The cleaved small peptides were extracted from each samples and subsequently analyzed with LC-MS/MS to determine the molecular weight. Samples were analyzed at Taplin Biological Mass Spectrometry Core Facility at Harvard Medical School. The molecular weight of each ion was aligned to predicted b and y ions of fly SNAP25 or mosquito SNAP25 to identify the cleavage sites of LC/PG1 and LC/PG2.

Fly injection. The Oregon-R flies enclosed for 3 to 7 days were selected for injection. To prepare the toxin injection solution, the toxins were diluted in PBS with 0.25% gelatin at pH 6.3. In each experimental group, 20 to 30 flies were anesthetized using CO2. Subsequently, 100 nL of the toxin solution was injected into the thorax of each fly using a Nanoject III (Drummond).[218] Following injection, the flies were placed in vials with fly medium. For survival testing, the injected flies were monitored for 2 days. For imaging purposes, the flies were transferred to a 60 mm dish after a 4 hr post-injection period.

4.3 Results

Discovery and characterization of PG toxins

BLAST-based identification of BoNT-related genes in *P. ghonii*

On Jul 5, 2022, we performed a BLAST-based search for clostridial neurotoxin homologs in the NCBI database, restricting our search to organisms outside of the *Clostridium* genus. Our search, using BoNT/A1 (accession # WP_021136346.1) as a query identified two previously unidentified matches in the genome of organism, *Paeniclostridium ghonii* (strain NCTR 3900). The BoNT/A1 light chain (LC) region encoding the metalloprotease domain matched a protein labeled as "M91 family zinc metallopeptidase" (WP_250673407.1) with an *E*-value of 1e-21 and 34% amino acid identity. The BoNT/A1 heavy chain (HC) region matched a second protein labeled as "hypothetical protein" (WP_250673408.1) with an *E*-value of 1e-54 and 28% amino acid identity. We designated these proteins "PG-toxin1-LC" and "PG-toxin1-HC" or the short forms, "PGT1-LC" and "PGT1-HC".

Next, we examined the JGI genome database for additional strains of *P. ghonii*, and found one additional genome (unique from NCTR 3900), labeled as strain "DSM 15049". A similar toxin gene cluster was identified in the genome of this strain, including orthologs of both proteins identified in NCTR 3900. Interestingly, these orthologs possess 39% identity for the LC protease gene and 61% identity for the HC gene, suggesting possible rapid divergence of the toxin sequences since its acquisition by an ancestral *P. ghonii* strain similar to those observed for BoNTs in Group I and II *C. botulinum*. We designated these proteins "PG-toxin2-LC" and "PG-toxin2-HC" or the short forms, "PGT2-LC" and "PGT2-HC".

A table of pairwise percentage identities to a representative subset of BoNT-like proteins is shown in Figure 4.1a for the LC and Figure 4.1b for the HC. Overall, combining their LC and HC sequences, PG-toxin1 and Pg-toxin2 share 54% sequence identity. The PG-toxin LC and HC sequences consistently show the highest percentage identities to PMP1, BoNT/En and BoNT/X compared to other BoNT-like sequences, suggesting a possible phylogenetic affiliation with this lineage.

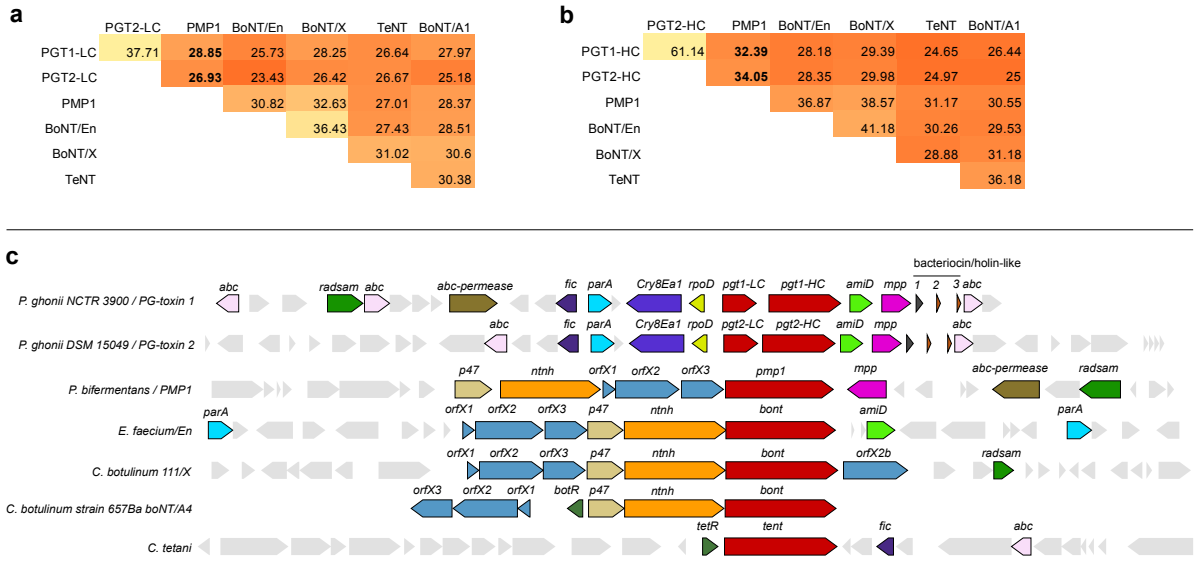


Figure 4.1: Identification of BoNT-like genes in genomes of two *Paeniclostridium ghonii* strains. Pairwise amino acid sequence identities between PG-toxin-LC (a) and PG-toxin-HC (b) and a subset of BoNT proteins. (c) Genomic neighborhoods surrounding PG-toxin genes and other BoNT genes.

Gene neighborhood analysis

Next, we investigated the genomic neighborhoods surrounding both PG-toxin genes (Figure 4.1c). In both *P. ghonii* genomes, the PG-toxin-LC gene encoding the predicted zinc metalloprotease is immediately upstream of the PG-toxin-HC gene encoding the translocase and binding domain. To verify that this is not due to sequencing errors, we analyzed the region further and determined that the intervening region is non-coding with numerous stop-codons in all reading frames.

To examine the source of the fragment (chromosomal versus plasmid), we analyzed the gene content up and downstream of the PG-toxin gene cluster. Within the scaffold from DSM 15049 containing the PG-toxin2 genes, which was more complete (64,948 bp) than the toxin-encoding contig from NCTR 3900 (29,173 bp), we identified genes encoding plasmid replication initiation proteins. BLAST analysis of these proteins confirmed a close relationship to orthologous genes from *P. bifermentans*, suggesting a link to the toxin-encoding megaplasmids of *Paraclostridium*.[121] To investigate the source of the contigs in their entirety, contig level assemblies were passed through rfPlasmid (v0.0.16)[219] using a *Clostridium* model and default settings. The contigs carrying the BoNT-like gene sequences for *Paeniclostridium ghonii* DSM15049 and NCTR 3900 were predicted as plasmids with 73.28% and 74.66% percent of the vote, in excess of the 40-60% threshold where incorrect classifications typically fall (rfPlasmid[219]). Similarly, another near-*Clostridium* species with a known BoNT-like sequence on a reported circularized plasmid, *Paraclostridium bifermentans* parabai, was predicted as a plasmid with 79.16% of the vote. We therefore speculate that the toxin loci may reside on a plasmid.

The PG-toxin1 locus from strain NCTR 3900 is similar but non-identical from that of the PG-toxin2 locus from strain DSM 15049. A notable unique feature of the PG-toxin locus is the presence of a nearby gene encoding an insecticidal Cry delta endotoxin (accession # WP_250673405.1). A cry toxin prediction was confirmed by the Conserved Domain Database, with matches to all three domains (N, M, and C-terminal region) with significant *E*-values all less than 1e-15. BLAST analysis of this protein also confirmed its relationship to a Cry8Ea1 family insecticidal delta endotoxin from *Bacillus thuringiensis* (accession # WP_172452114.1) with 40% amino acid identity. The closest sequence of known structure is B. thuringiensis cry3Aa (PDB ID 4QX0_A) at 32% identity. In both *P. ghonii* genomes, this Cry toxin gene is encoded in the opposite direction to the PGT-LC genes, which suggests that it may be regulated differently.

In addition, The PG-toxin locus also encodes several genes shared with those found in other BoNT gene clusters. These include *amiD* (n-acetylmuramoyl-l-alanine *amiD*ase gene) and *parA* genes which are also present in the BoNT/En locus of E. faecium; and *abc* (ABC

transporter ATP-binding protein) and *fic* (filamentation induced by cAMP protein) genes that are also found in the *tent* locus of *C. tetani* . Three nearby genes are also shared with the PMP1 locus of *P. bifermentans*: an *mpp* gene encoding a putative metallophosphatase family protein, an *abc-permease* (ABC transporter ATP-binding protein/permease) and a *radsam* gene encoding a radical SAM domain containing protein. Interestingly, the *orfx*, *ha* and *ntnh* genes present in other BoNT gene clusters are absent. Downstream of the *mpp* gene in both *P. ghonii* genomes is a cluster of genes encoding a BhlA/UviB family holin-like peptide and two genes encoding aureocin-like type II bacteriocins (Figure 4.1c). Other toxins are known to be exported via holin-dependent mechanisms.[220]

Protein sequence and domain analysis

Next, we compared the PG-toxin sequences to other BoNTs to examine the presence/absence patterns of key motifs and structural features important for neurotoxin activity. While typical BoNTs are proteolytically cleaved into two distinct chains (LC and HC) that remain connected via a disulfide bond (Figure 4.2a), a unique feature of the *P. ghonii* BoNT-like toxins is the split of the LC and HC into two distinct coding sequences (Figure 4.2b). The correspondence of the PG-toxin-LC and neighboring PG-toxin-HC sequences with the LC and HC of BoNTs can also be seen at a structural level, as revealed by the homology models shown in Figure 4.2b.

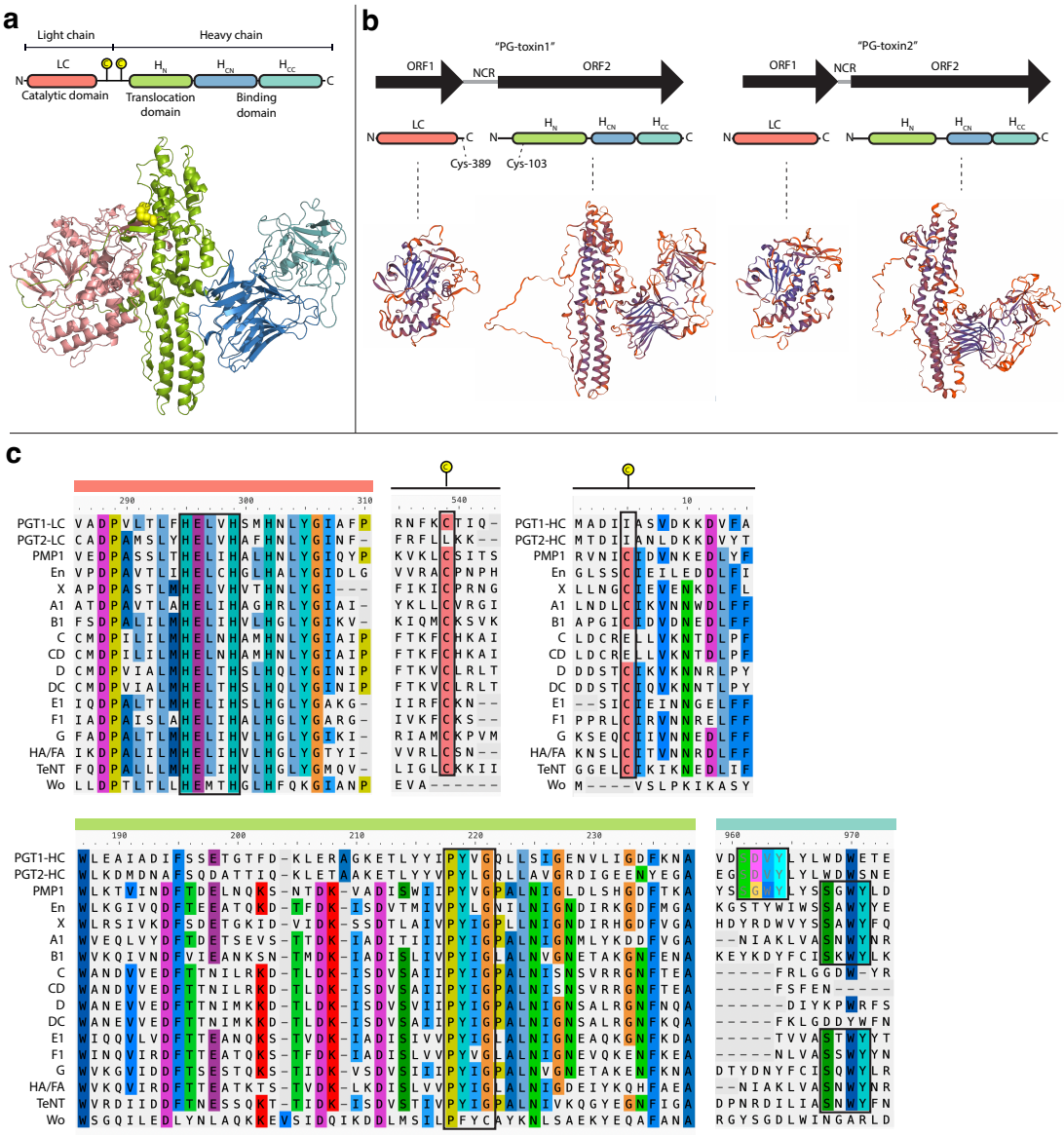

a
Light chain
Heavy chain
LC
H_N
H_CN
H_CC
N
C
Catalytic domain
Translocation domain
Binding domain
b
"PG-toxin1"
"PG-toxin2"
ORF1
NCR
ORF2
Cys-389
Cys-103
c
PGT1-LC
PGT2-LC
PMP1
En
X
A1
B1
C
CD
D
DC
E1
F1
G
HA/FA
TeNT
Wo
PGT1-HC
PGT2-HC

Figure 4.2: Structural and motif analysis of PG-toxins in relation to BoNT. (a) Domain structure of BoNT/A1 (above) depicting the light and heavy chain and corresponding crystal structure (below) from PDB 3BTA. (b) Protein domain architectures of PG-toxin open reading frames (ORFs). *P. ghonii* strain NCTR 3900 and DSM 15049 both encode two ORFs separated by a non-coding region (NCR) which correspond to the LC and HC regions of BoNT, which are collectively referred to as PG-toxin1 and PG-toxin2, respectively. Structural models of these regions confirm their correspondence to the BoNT LC and HC. (c) Selected amino acid alignments depicting key functional residues of known importance in BoNT. The catalytic HExxH motif required for zinc endopeptidase activity is conserved in PGT1-LC and PGT2-LC, while all but one of the canonical Cys residues involved in the LC-HC disulfide bond are missing. The PxxG motif within the translocation domain is also conserved, and the canonical SxWY ganglioside binding motif is absent, although a second upstream SxxY motif is present in an equivalent site to a second SGWY motif previously identified in PMP1.

The gene split between the LC and HC raises the question of how these domains may physically interact if they are assumed to function equivalently to BoNTs. Interestingly, PG-toxin1-LC contains a Cys residue at its C-terminal region (Cys-389), whereas this Cys residues appear to be absent in PG-toxin2-LC. The closest second Cys residue in PG-toxin-1-HC occurs at residue Cys-103 but this is within the translocase domain. Thus, it appears unlikely that the PG-toxin-LC and HC domains interact via the same disulfide bond as in BoNTs. Structural models of the PG-toxin chains (Figure 4.2b), however, may provide some insights into a possible interaction between the LC and HC domains. That is, the "belt" region of the translocase domain in the HC domains, which is known to wrap around the LC in other BoNTs,[221,222] is present in both PG-toxin1-HC and Pg-toxin2-HC. We therefore speculate that this belt may provide a mechanism for forming a complex between the light and heavy chains, although this hypothesis remains to be evaluated experimentally.

Both LC protease sequences (PG-toxin1-LC and PG-toxin2-LC) contain the conserved HExxH motif required for zinc-endopeptidase activity, and more specifically an HELVH motif identical to that in BoNT/X (Figure 4.2c). Overall, PG-toxin1-LC and PG-toxin2-LC regions possess a maximum of 28.9% and 26.9% identity to their closest BoNT-like homolog (PMP1), respectively (Figure 4.1a).

A region of sequence conservation for the translocation domain is also shown in Figure 4.2c. The PxxG motif, which has previously been identified as a strongly conserved motif across BoNTs (including divergent homologs)[209] is conserved in the PG-toxin translocation

domains. This PxxG motif is also conserved in translocation domains of diphtheria toxins (DT) and DT-like homologs,[223] as well as large clostridial toxins (LCTs).[224] Overall, the putative translocation domains from PG-toxin1-HC and PG-toxin2-HC share 48.2% identity, and possess a maximum of 30% to other BoNT-related proteins (BoNT/En).

Lastly, we examined the C-terminal binding domain (Hcc) of both PG-toxin HC sequences. In the Hcc domains of BoNT/A, B, E, F, G, X, En, PMP1, and TeNT, a conserved SxWY gangloside binding motif (which we designated "motif 1") has been identified that is known to bind ganglioside cell receptors.[225] Importantly, in the mosquitocidal PMP1, a second SxWY motif ("motif 2") is present immediately upstream of the canonical SxWY motif, as well as a third SWYG motif ("motif 3") which was found to be critical for toxicity.[121] Interestingly, both PG-toxin-HC sequences lack the canonical SxWY motif ("motif 1") but possess an SDVY motif in place of motif 2, which we hypothesize may play a role in receptor binding (Figure 4.2c).

Phylogenetic analysis

To explore their evolutionary relationships to other BoNT proteins, we separately aligned the PG-toxin LC and HC protein sequences to corresponding regions from a representative set of available BoNTs retrieved from `https://bontbase.org` (also including PMP1). We then built maximum-likelihood phylogenies for the LC (Figure 4.3a) and HC (Figure 4.3b).

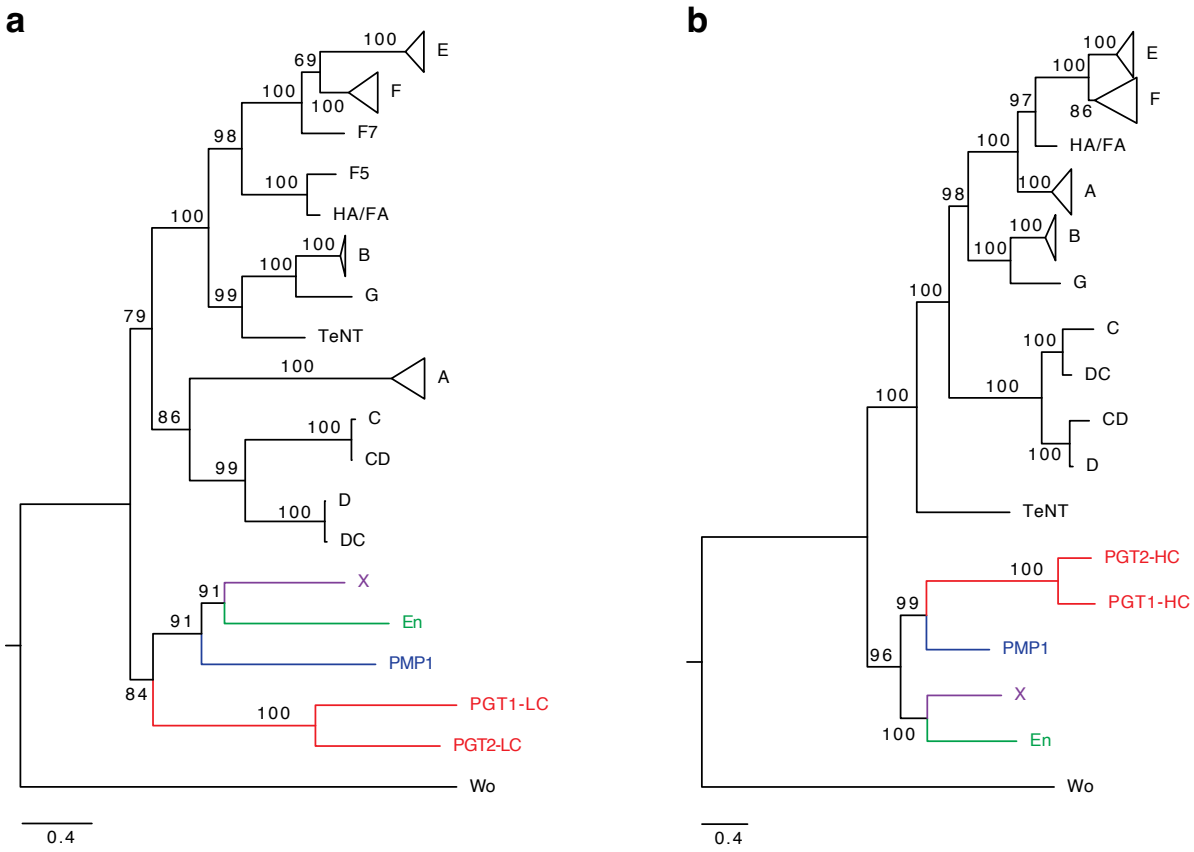


Figure 4.3: Phylogeny of the BoNT family including PG-toxin. (a) Phylogeny of the LC encoding the protease domain. (b) Phylogeny of the HC encoding the translocation domain and binding domain. In both trees, PG-toxin-LC and PG-toxin-HC cluster with the X/En/PMP1 lineage. Bootstrap values are indicated for all major clades.

The phylogenetic placement of the PG-toxin-LC and PG-toxin-HC sequences was similar. In both cases, they clustered with the X/En/PMP1 lineage (Figure 4.3). The PG-toxin-LC clustered as a sister lineage to X/En/PMP1 with moderate (84%) clade support, whereas the PG-toxin-HC clustered specifically with PMP1 with very strong (99%) clade support. As expected, the PG-toxin genes from both *P. ghonii* strains clustered as neighbors indicating a common origin in an ancestral *P. ghonii* strain.

Summary and phylogenetic analysis of the organism, *Paeniclostridium ghonii*

Paeniclostridium ghonii was first identified in the 1930s and was originally named *Clostridium ghoni* by Prévot. It was later reclassified as a member of the genus *Paeniclostridium*.[226] The organism has been found in soil, marine sediments, rumen, soft-tissue infections in humans and human faeces, and in alfalfa silage.[227] The species is also notable as it was identified as a member of the microbiome in the colon of the Iceman.[228] One of the most closely related species is *Paeniclostridium sordellii* (*Clostridium sordellii*), a Gram-positive pathogen that causes severe oedemic, myonecrotic or enterotoxic infections in humans, and farm animals.[229]

To examine the phylogenomic similarities of *P. ghonii* to other available sequenced organisms, we retrieved genomes of all available *Paeniclostridium* species, as well as *Eubacterium tenue* and members of the genus, *Paraclostridium*, given their previously identified relationships to *P. ghonii* based on 16S analysis.[226] Based on genome-wide average nucleotide identity (ANI) clustering, the two strains of *P. ghonii* cluster as a sister lineage to a clade of *Paraclostridium* which includes *Paraclostridium bifermentans* and *P. benzoelyticum* (Figure 4.4). *P. ghonii* displays ANI values of 87.4-87.6% to these *Paraclostridium* species and lower ANI values to *Eubacterium tenue* (83.4%) and *Paeniclostridium sordelli* (82.7%). The relatively close relationship between *P. ghonii* and *P. bifermentans* is noteworthy given the presence of BoNT-like genes associated with these species, and suggests the existence of a broader lineage of potential *Paraclostridium* and *Paeniclostridium* species that encode BoNT-like toxins.

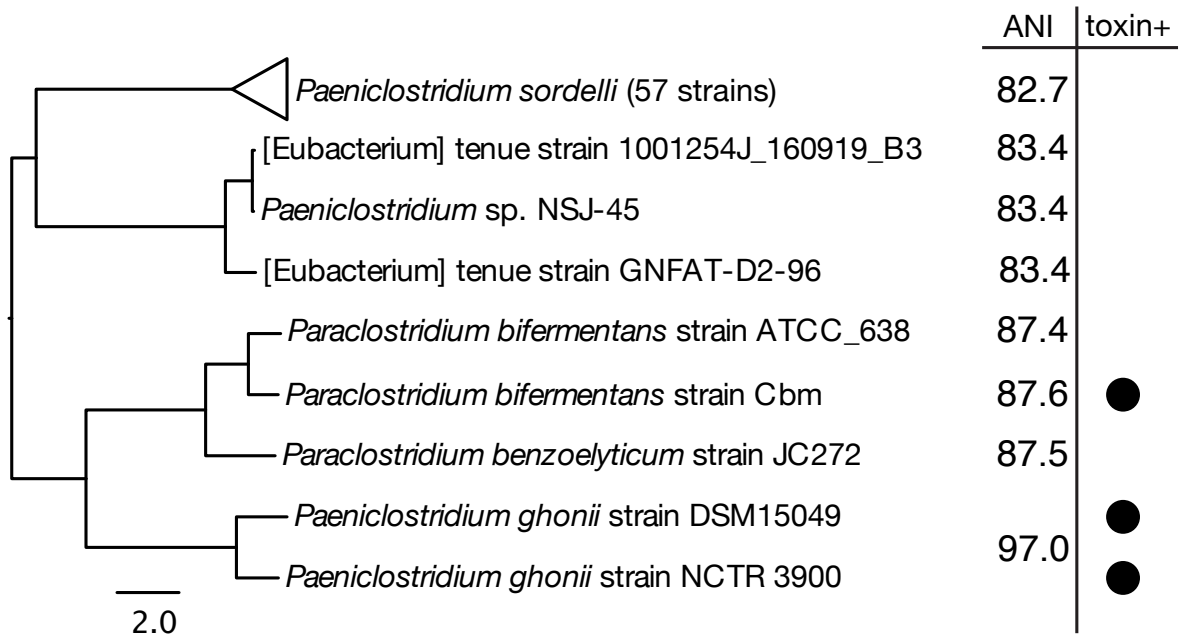


Figure 4.4: Phylogenetic clustering of *P. ghonii* genomes with the top 64 closest genomes in the NCBI database based on average nucleotide identity. Average nucleotide identity (ANI) values are shown on the right of the tree, and are averaged for the two *P. ghonii* strains.

Discovery of BTNT toxins

BLAST-based identification of BoNT-related genes in *B. toyonensis*

On July 6, 2023, we performed a BLASTp search against the NCBI nr database with default parameters using BoNT/A1 (NCBI accession # WP_011948511) as a query. A previously unreported homolog annotated as "hypothetical protein" (accession # HDR7951433) from *Bacillus toyonensis* was detected with an *E*-value of 3×10^{-53}, and amino acid sequence identity of 26.87%. A second protein (HDR7951432), also from *B. toyonensis* , was detected with an *E*-value of 4×10^{-33} and 24.6% amino acid sequence identity. Inspection of both sequences and subsequent BLAST analysis revealed that the first protein (HDR7951433) is a homolog of BoNT, while the second (HDR7951432) is a homolog of the NTNH protein, whose gene is typically encoded upstream of bont in *bont* gene clusters. Consistent with

this, a zinc-metalloprotease HExxH motif unique to BoNTs was found in HDR7951433 but was absent in HDR7951432. We therefore designated HDR7951433 as BTNT for putative *Bacillus toyonensis neurotoxin* and HDR7951432 as *bt-ntnh* .

Next, we aligned the putative *B. toyonensis* neurotoxin (BTNT) with other members of the BoNT family (Figure 4.5). It is most similar to PMP1 from *Paraclostridium bifermentans*, but is still highly divergent at only $\sim$ 29% identity. This degree of sequence divergence is expected for different serotypes within the BoNT family as shown in Figure 4.5.

	PMP1	BoNT/A1	BoNT/X	BoNT/En	TeNT
BTNT	28.99	24.15	26.97	25.75	26.22
PMP1		30.27	36.07	34.78	29.35
BoNT/A1			30.51	28.92	34.23
BoNT/X				38.62	29.13
BoNT/En					29.05

Figure 4.5: Pairwise sequence similarity of BoNT toxins including the newly identified BoNT-like toxin from *B. toyonensis* (HDR7951433.1).

Gene neighborhood analysis

To examine the gene neighborhood surrounding the newly identified *B. toyonensis* BoNT-like toxin gene, we retrieved the associated contig (DAOJGQ010000093.1) and gene annotations from the NCBI Assembly (GCA_029709895.1), and visualized it using AnnoView (http://annoview.uwaterloo.ca). We also manually analyzed and annotated the nearby genes using BLAST. *btnt* has a gene neighborhood structure that is characteristic of other *bont* genes with the *bt-ntnh* gene located immediately upstream as expected, as well as three *P47/ORFX*-like genes (Figure 4.6). BLAST analysis revealed that the P47/RFX containing proteins putatively match the proteins ORFX3, P47, and ORFX2 in *P. bifermentans*, *E. faecium*, and *C. botulinum*. This general gene arrangement is conserved among other members of the X/En/PMP1 clade of BoNT-like toxins, and also some other BoNT genes (e.g., BoNT/A4). It is noticeable that p47 is located within the *orfX* locus in the genomic neighborhood of BNTN, whereas it is usually located upstream or downstream of the *orfX* locus in the genomic neighborhood of other BoNT toxins. As the available contig is short, the identity of genes further upstream and downstream are currently unknown.

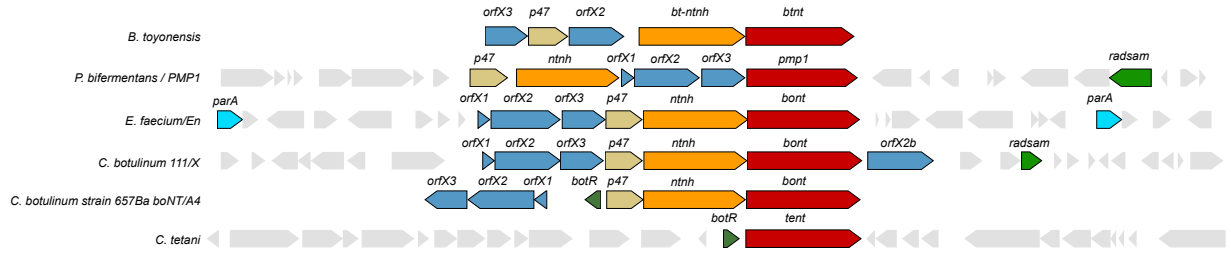


Figure 4.6: Visualization of a newly identified BoNT-like gene cluster in *B. toyonensis* containing homologs of *bont*, *ntnh*, and P47-domain containing proteins. Also shown for comparison are *bont* gene clusters from other species.

Sequence analysis

Next, we aligned BTNT with all representatives of all other BoNT subtypes, and examined the alignment for presence/absence of key functional residues/motifs (Figure 4.7). Although BTNT contains the conserved HExxH motif as described earlier, it appears to lack the double Cys residues within the disulfide linker region at the canonical positions conserved in other BoNTs. This conserved disulfide bond connects the heavy and light chain, and is essential for the toxicity of neurotoxins.[230] However, other Cys residues may perform this role, such as the Cys-590 located eight residues further downstream. BTNT also possesses the conserved PxxG motif identified in the translocation domain of BoNTs and more distantly related BoNT homologs.[209] It lacks the SxWY ganglioside cell receptor-binding motif[231] but possesses a similar SSEY motif at this site, which could indicate an ability to bind a similar receptor.

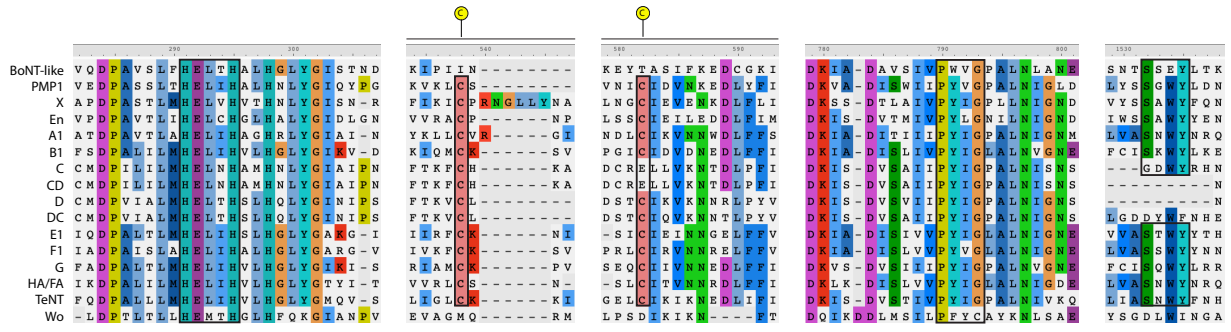


Figure 4.7: Alignment of *B. toyonensis* BoNT-like toxin with other subtypes and presence/absence of key functional motifs.

Phylogenetic analysis

To investigate the evolutionary relationships between BTNT and other BoNTs, we constructed a maximum-likelihood tree using IQ-TREE (Figure 4.8). BTNT clustered with PG toxins and the X/En/PMP1 lineage as the outermost branch. However, the bootstrap support (57%) was not strong. Next, we produced a second phylogenetic tree by including both BTNT and the newly identified PG toxins (Figure 4.9). In this tree, again the BTNT and PG toxins clustered together within the X/En/PMP1 lineage, and the clade support was even higher at (93%). BTNT's closer evolutionary relationship to this BoNT lineage is consistent with its gene neighborhood structure (Figure 4.6).

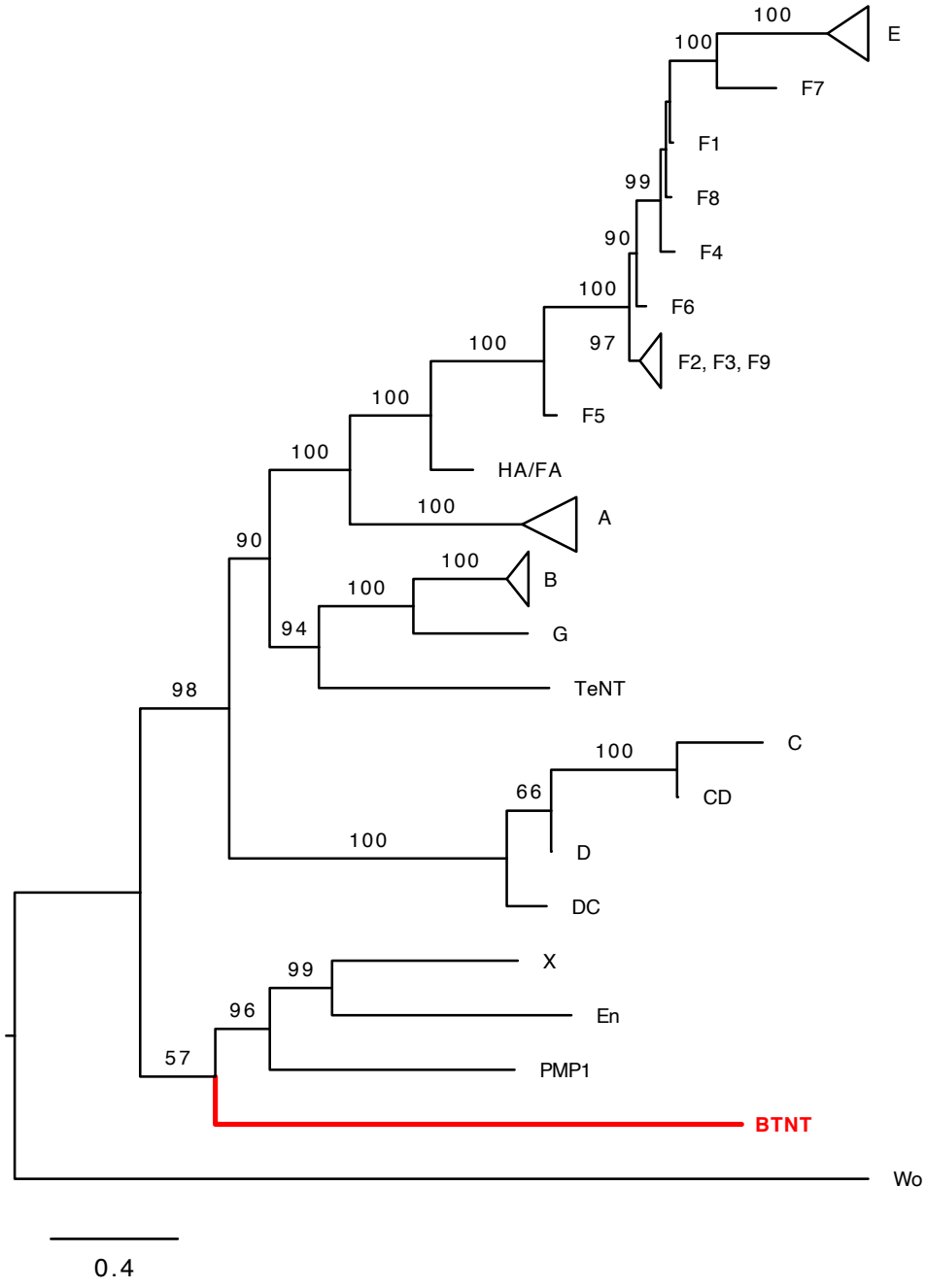


Figure 4.8: Maximum-likelihood phylogenetic tree of BoNTs and related toxins including the newly identified *B. toyonensis* BoNT-like toxin. Bootstrap values >50 are indicated above the nodes.

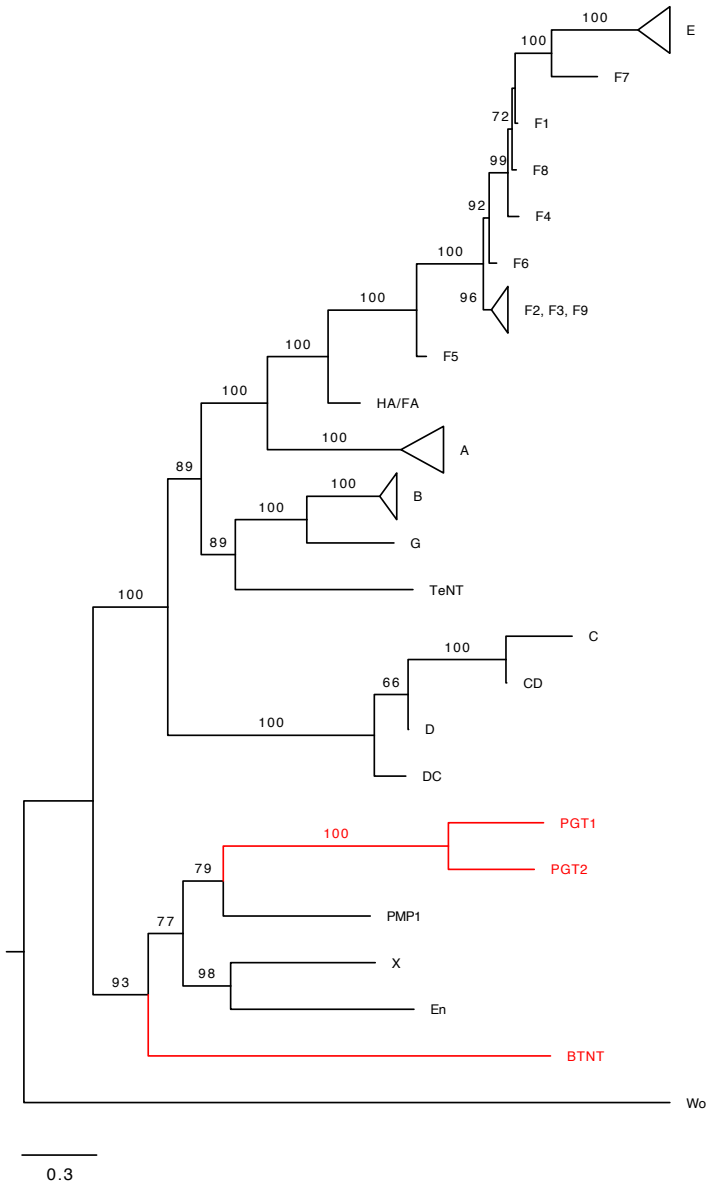


Figure 4.9: Maximum-likelihood phylogenetic tree of BoNTs and related toxins including the newly identified BoNT-like toxins in *B. toyonensis* and *P. ghonii*. Bootstrap values >50 are indicated above the nodes.

Genomic analysis *B. toyonensis* isolate CH177

Based on available information in the NCBI, the BTNT-containing *B. toyonensis* strain was recently added to the NCBI database under the accession # GCA_029709895.1 and the isolate name "CH177". The assembly has 296 contigs, and the draft WGS was generated using the SKESA v2.2. de novo assembler.[232]

B. toyonensis isolate CH177 was found to be phylogenetically and genomically similar to other *B. toyonensis* strains (data not shown). All currently assembled *B. toyonensis* strains lack detectable homologs of BoNT genes, suggesting that the BTNT gene cluster is a recent acquisition into *B. toyonensis* isolate CH177. Further sequencing and assembly will be required to establish whether it is chromosomally encoded or whether it resides on a plasmid that has been uniquely acquired by this strain.

Next, we examined the CH177 assembly for additional evidence of pathogenicity-related genes using VFanalyzer within the VFDB database.[217] VFanalyzer detected hemolytic enterotoxin genes (*hblA*, *hblC*, and *hblD*), and a non-hemolytic enterotoxin (*nheC*), all of which have been detected in other related *Bacillus* species. Notably, in a commonly used probiotic *B. toyonensis* strain, these toxin genes are not detectably expressed or are expressed at much lower levels than *B. cereus*.[233] Also detected was the PlcR-PapR quorum sensing system, a hemolysin III family protein, haemolysin *xhlA* gene, and sphingomyelinase. The *islA* gene was also detected, which is a virulence factor in *Bacillus cereus* and has been shown to be expressed in the insect hemocoel and under iron-depleted conditions.[234,235] Ultimately, these predictions suggest a pathogenicity-related function for CH177, but the host is still unknown.

4.4 Experimental characterization of PG toxins

Although genomic context analysis predicted an insect host for PG toxin in particular, ultimately, additional experiments are required to validate this prediction. Therefore, through a collaboration with the Min Dong Lab (Boston Children's hospital, Harvard), the predicted neurotoxin activity against insect hosts was further examined. It is anticipated that the complete methodological details will be available in a forthcoming publication by the collaborators.

Survival analyses reveal that adult *Drosophila* subjected to PG1 and PG2 injections exhibit considerably reduced survival rates in comparison to the other treatment groups and the control. Notably, LC or HC of the toxins alone appears to have a negligible

effect on *Drosophila* survival (Figure 4.10a-b). The survival rate decreases more sharply as the dose increases, with PG2 displaying higher potency than PG1. The survival rate for *Drosophila* treated with 1 amol of PG1 drops to 0% after around 20hrs of injection, whereas PG2-treated *Drosophila* maintains a survival rate of just below 50% at the end of the experiment (Figure 4.10c-d). Further EC 50 (half maximal effective concentration) analysis highlights differences in potency between the two toxins. The effective dose causing 50% paralysis for PG1 is 108 fmol, while for PG2, it is 401 amol (Figure 4.10e-f). The recombined PG221 and PG112 treatment seems to have a survival curve intermediate between PG1 and PG2 (Figure 4.10g). In general, PG2 is more toxic than PG1 according to the dose assay. Finally, as shown in Figure 4.10h, after shaking the test tube, *Drosophila* treated with PG1 LC moved to the top of the tube, displaying vigorous escape behavior. While for PG1 complex group, almost all *Drosophila* exhibiting paralysis and stayed at the bottom of the test tube. This further suggests that the protease domain alone does not have significant toxicity.

Ultimately, these assays demonstrate that PG1 and PG2 toxins induce paralysis and death, which confirms the bioinformatics prediction that these proteins are neurotoxins capable of targeting insect hosts.

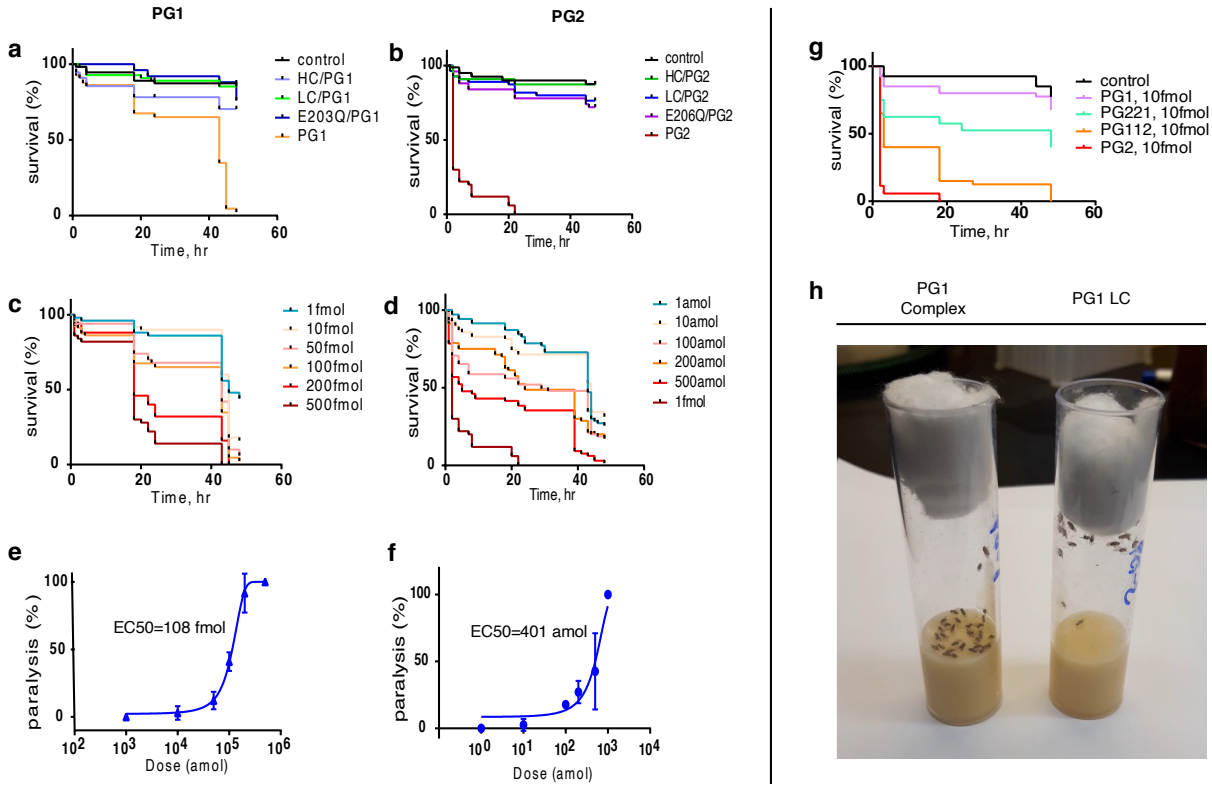


Figure 4.10: PG1 and PG2 injection in adult *Drosophila* induce paralysis and death. (a) Survival curves of adult *Drosophila* injected with 100 fmol of PG1, HC/PG1, LC/PG1, inactivate mutant form E203Q/PG1 and PBS injection as negtive control. (b) Survival curves of adult *Drosophila* injected with 1 fmol of PG2, HC/PG2, LC/PG2, inactivate mutant form E206Q/PG2 and PBS injection as negative control. (c) Survival curves of adult *Drosophila* injected with different dosages of PG1. (d) Survival curves of adult *Drosophila* injected with different dosages of PG2. (e) Percentage of paralysis rate in adult *Drosophila* 24 hrs after injection with PG1. (f) Percentage of paralysis rate in adult *Drosophila* 24 hrs after injection with PG2. (g) Survival curves of adult *Drosophila* injected with 10 fmol of PG1, PG2 and recombined PG221, PG112, and PBS injection as negtive control. N=25-30 for each injection time, and data is analyzed by GraphPad Prism. (h) PG1 complex and PG1 LC paralyzed flies. The data presented is a combination of three repeated experiments. The experiments were performed by the collaborators Dr. Pyung-Gang Lee and Dr. Min Dong.

Next, the biochemical activities of these toxins were examined, and particularly, their

ability to cleave canonical BoNT substrates. PG toxin protease assays show that both LC/PG1 and LC/PG2 are able to cleave f-SNAP25 but not h-SNAP25, suggesting that the PG1 and PG2 toxins may not be toxic to humans (Figure 4.11a). The cleavage sites for PG toxins are situated at the C-terminus of SNAP25, distinct from the cleavage sites of BoNT/A and BoNT/E on h-SNAP25. Notably, PG1 and PG2 have unique cleavage sites (Figure 4.11b-c). Then, the specificity of LC/PG1 and LC/PG2 across various insect species was evaluated. Both LC/PG1 and PG2 demonstrate the ability to cleave fly, moth, beetle and mosquito SNAP25, while LC/PG2 additionally cleaves shrimp SNAP25 and fly-SNAP24, providing evidence that LC/PG2 may target a broader host range (Figure 4.11d). Finally, the purification results for both LC/PG and HC/PG toxins indicate their ability to form complex molecular structures (Figure 4.11e).

Ultimately, these results indicate that the protease domain (LC) of PG1 and PG2 exhibit proteolytic activity against canonical substrates from insects and a range of other invertebrates, consistent with our earlier bioinformatic predictions.

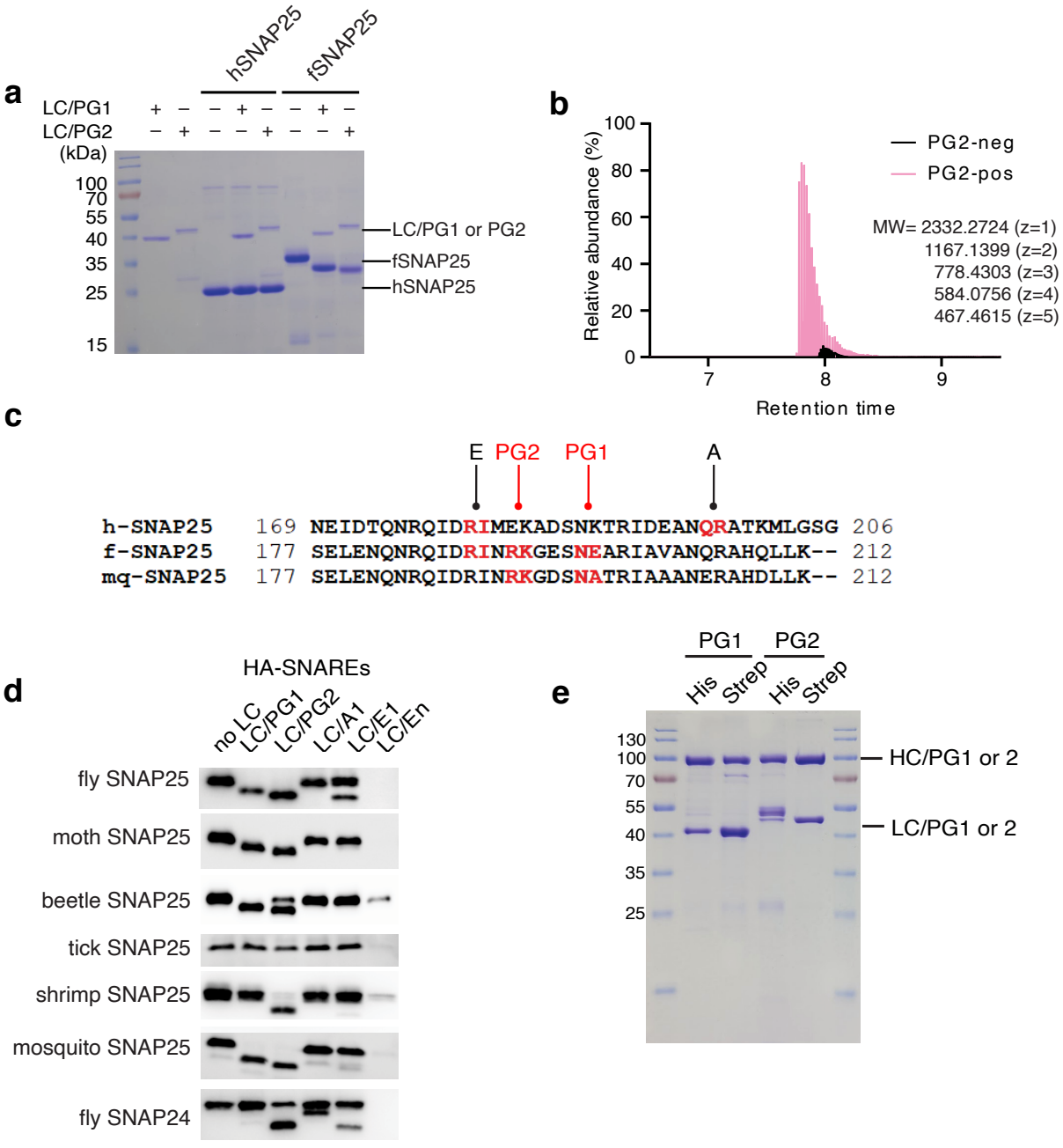

a
hSNAP25
fSNAP25
LC/PG1
LC/PG2
(kDa)
100
70
55
40
35
25
15
LC/PG1 or PG2
fSNAP25
hSNAP25
b
Relative abundance (%)
Retention time
PG2-neg
PG2-pos
MW= 2332.2724 (z=1)
1167.1399 (z=2)
778.4303 (z=3)
584.0756 (z=4)
467.4615 (z=5)
c
E
PG2
PG1
A
h-SNAP25 169 NEIDTQNRQIDRIMEKADSNKTRIDEANQRATKMLGSG 206
f-SNAP25 177 SELENQNRQIDRINRKGESNEARIAVANQRAHQLLK-- 212
mq-SNAP25 177 SELENQNRQIDRINRKGDSNATRIAAANERAHDLLK-- 212
d
HA-SNAREs
no LC
LC/PG1
LC/PG2
LC/A1
LC/E1
LC/En
fly SNAP25
moth SNAP25
beetle SNAP25
tick SNAP25
shrimp SNAP25
mosquito SNAP25
fly SNAP24
e
PG1
PG2
His
Strep
His
Strep
130
100
70
55
40
35
25
HC/PG1 or 2
LC/PG1 or 2

Figure 4.11: Characterization of PG1 and PG2 toxins. (a) Recombinant f-SNAP25 or h-SNAP25 were incubated with LC/PG1 or PG2, respectively. Both LC/PG1 and PG2 cleaved f-SNAP25, but they did not cleave h-SNAP25. (b) f-SNAP25 was incubated with LC/PG1 and LC/PG2, respectively. As a negative control, equivalent samples, which were pre-TFA treated and heat-inactivated, were prepared. The cleaved small peptides were extracted from each samples and subsequently analyzed with LC-MS/MS to determine the molecular weight of cleaved peptides, finally to identify cleavage sites of LC/PG1 and PG2. (c) Sequence alignment of h-SNAP25 and f-SNAP25 with the cleavage sites for BoNT/A, E, PG1 and PG2. All the cleavage sties are located in the C-terminus of SNAP25. (d) HA-tagged SNAP25s and LC/PG1, PG2, A, E, En were coexpressed in 293T cells via transient transfection. After 24 h culture, cell lysates were harvested and subjected to immunoblot analysis. LC/PG1 cleaved fly, moth, beetle and mosquito SNAP25, while LC/PG2 additionally cleaved shrimp SNAP25 and fly-SNAP24. (e) Strep-II tagged LC/PG1 or 2 and His-tagged HC/PG1 or 2 were coexpressed and purified using Ni-NTA column. Both PG1 and PG2 were copurified as LC and HC, suggesting the complex can be maintained through purification condition. The following Strep-II tag purification enhanced both purity and complex ratio (LC/HC) compared to His-tag purification. The experiments were performed by the collaborators Dr. Pyung-Gang Lee and Dr. Min Dong.

4.5 Discussion

Here, we report two novel putative BoNT-like toxins from non-*Clostridium* lineages, PG toxins from two *P. ghonii* strains and BTNT from *B. toyonensis* isolate CH177. These homologs exhibit low to moderate sequence similarities with known BoNTs. Protein alignments reveal that both PG toxins and BTNT retain residues identical or similar to the functional motifs vital for toxicity, suggesting that these putative toxins may function as neurotoxins. Phylogenetic analyses place PG toxins and BTNT alongside the X/En/PMP1 cluster, which further suggests these toxins form a distinct lineage with unique specificities compared to BoNTs in *Clostridium* species.

In July 2022, we first discovered PG toxins in two *P. ghonii* strains. We designated them PG-toxin 1 (PGT1) and PG-toxin 2 (PGT2), which share 54% protein sequence identity. Unlike any other known BoNT homologs, PGT1 and PGT2 are composed of two separate subunits encoded on two neighboring genes: one encoding the protease domain (light chain, LC) with a conserved HExxH motif, and the second encoding the heavy-chain (HC) containing the putative translocation domain and receptor-binding domain. Phylogenetic

analysis of both the LC and HC reveal that it is a divergent member of the lineage of BoNT that also includes BoNT/X, BoNT/En and the insecticidal PMP1. The gene clusters harboring PGT1 and PGT2 also include a putative insecticidal delta-endotoxin, Cry8Ea1, as well as putative endolysin and bacteriocin genes that may facilitate lytic toxin secretion, suggesting a possibility that this gene cluster might serve an insecticidal purpose. Further experimental validation by our collaborators provides evidence that both PG toxins are insecticidal and are able to target a wide range of insects including fly, moth, beetle and mosquito, with PG2 displaying higher potency than PG1.

Next, we identified another BoNT homolog in B. toyonensis isolate CH177, distinct from earlier strains. This putative toxin shares 26-29% identity with its closest BoNT relatives, suggesting that it is likely a novel BoNT-like toxin. It possesses key functional motifs (e.g., HExxH) indicative of toxin protease activity, contains the four characteristic BoNT domains, and is located in a BoNT-like genomic neighborhood containing the upstream non-toxic non-hemagglutinin (NTNH) gene as well as several P47-related genes. Phylogenetically, the toxin clusters as a divergent member of the recently discovered lineage of BoNT-like toxins that includes BoNT/X, BoNT/En, and the insecticidal PMP1. Genomic analysis of the *B. toyonensis* isolate CH177 reveals additional virulence factors and toxin genes indicative of potential pathogenicity targeting an unknown host species.

The power of genomic methodologies in toxin identification and characterization is evident. Gene neighborhood analysis, in particular, aids in novel toxin classification and host prediction.[236] In this study, we identified two novel putative BoNT toxins outside the *Clostridium* genus and successfully used gene neighborhood data for host prediction. Although genomic context approaches provide considerable information for toxin characterization, one limitation is that the amount of contextual information relies on the contig length housing the gene of interest. While we acquired extensive neighboring genes for PG toxins, the BTNT's short contig provided limited data. In this case, additional genomic analyses might be helpful for host prediction, such as examining biological pathways or functions that are enriched in the bacteria, or analyzing the associated metadata including the isolation source to obtain clues about the bacteria target host.

In summary, our findings propose PG toxins and BTNT as novel BoNT variants in non-*Clostridium* species. The identification of these novel BoNT toxins, along with the other BoNTs identified in non-*Clostridium* species, broadens our understanding of the evolution and distribution of botulinum toxins, which is important to understand the evolutionary emergence of pathogens and their specificity for different hosts. Furthermore, it is also expected that the identification of new specificities in these new putative CNTs can be harnessed in existing botulinum toxin therapeutics for human drugs. More importantly, The discovery of PG toxins as potential insecticidal neurotoxins also opens new avenues

for agricultural pest control, by creating an exciting new and highly effective formula of biological insecticide that targets insects but is safe for humans at the same time.

Chapter 5

Comparative genomics of a novel *Clostridium* sp. X identified from archaeological samples

Material in this chapter has been published as part of Hodgins et al. (2023).[237] The published manuscript is available here:

> H. P. Hodgins, P. Chen, B. Lobb, X. Wei, B. J. M. Tremblay, M. J. Mansfield, V. C. Y. Lee, P.-G. Lee, J. Coffin, A. T. Duggan, A. E. Dolphin, G. Renaud, M. Dong, and A. C. Doxey. Ancient *Clostridium* DNA and variants of tetanus neurotoxins associated with human archaeological remains. *Nature Communications*, 14(1):5475, 2023.[237] https://doi.org/10.1038/s41467-023-41174-0

In this chapter, I extend the MSc work of Harold Hodgins, who conducted a large-scale data mining study of millions of sequencing datasets that led to the discovery of lineage Y and clade X strains of *Clostridium*.[237,238] Building upon this work, I further analyze the gene content and gene neighborhoods (powered by AnnoView) of modern *C. tetani* genomes as well as novel *Clostridium* species using comparative genomics approaches, and place emphasis on the gene neighborhoods unique to clade X strains and those of key virulence factors. I hypothesize that the unique genomic structures in clade X strains provide additional support to the idea that *Clostridium* sp. X/Y is a unique *Clostridium* species from *C. tetani*.

5.1 Introduction

Clostridium tetani, the causative agent of the neuroparalytic disease tetanus, is an important bacterial pathogen of humans and animals. The *C. tetani* disease life cycle is as follows: after its spores enter wounds, it germinates, diffuses through oxygen-depleted and necrotic tissue, and produces a highly potent neurotoxin (tetanus neurotoxin, TeNT) that paralyzes hosts.[119,239] This ultimately leads to spastic paralysis instead of the flaccid paralysis that is caused by *C. botulinum* derived botulinum toxin in botulism. This may present as a local or a systemic effect, which can result in death due to paralysis of respiratory muscles and subsequent respiratory failure.

As a wound-associated infectious disease, tetanus is estimated to have plagued *Homo sapiens* throughout history. Accounts of tetanus-like diagnostic features ("lockjaw") date back as far as the ancient Greeks through written descriptions by Hippocrates (c. 380 BCE)[240,241] and the ancient Egyptians as seen in reports from the Edwin Smith papyrus (c. 1600 BCE).[242]

The first isolate of *C. tetani* originated in 1889[243] and an early isolate of *C. tetani* (the 1920 Harvard E88 strain) is still used as a reference today. The genome sequence of the E88 reference strain revealed that has a single chromosome of 2.8 Mb in length and a plasmid of 74 kb in length.[244] All known strains of *C. tetani* maintain this genomic organization, but different strains exhibit variation in their plasmid sizes.[245–247] The plasmid is critical to pathogenicity as it contains the key virulence genes including *tent* which encodes the tetanus neurotoxin and *colT* which encodes a collagenase enzyme involved in tissue degradation. Based on phylogenetic analysis of modern *C. tetani* genomes, *C. tetani* strains cluster into two clades.[246] Different strains exhibit low genetic variation with average nucleotide identities of 96-99%. Interestingly, the *tent* gene is also extremely conserved with 98% to 100% nucleotide identity across strains. Modern *C. tetani* genomes therefore offer a limited perspective on the true diversity of *C. tetani* and its evolutionary history as a human pathogen. Moreover, no other *C. tetani* related species have yet been identified that contain a similar plasmid or *tent* gene, which raises questions regarding the origin of *C. tetani* and its neurotoxin.

In a parallel analysis by Harold Hodgins (Doxey Lab), a large-scale data mining study of millions of sequencing datasets was done to identify datasets containing putative *C. tetani* DNA.[237,238] This work led to the unexpected discovery of *C. tetani* in 38 ancient DNA samples associated with human archaeological remains including teeth, bone, and tissue samples. Using taxonomic binning, *C. tetani* contigs from each sample were pooled together to form draft genomes or "bins". Contigs were then aligned with existing modern *C. tetani* genomes to build a phylogenetic tree as shown in Figure 5.1a below.

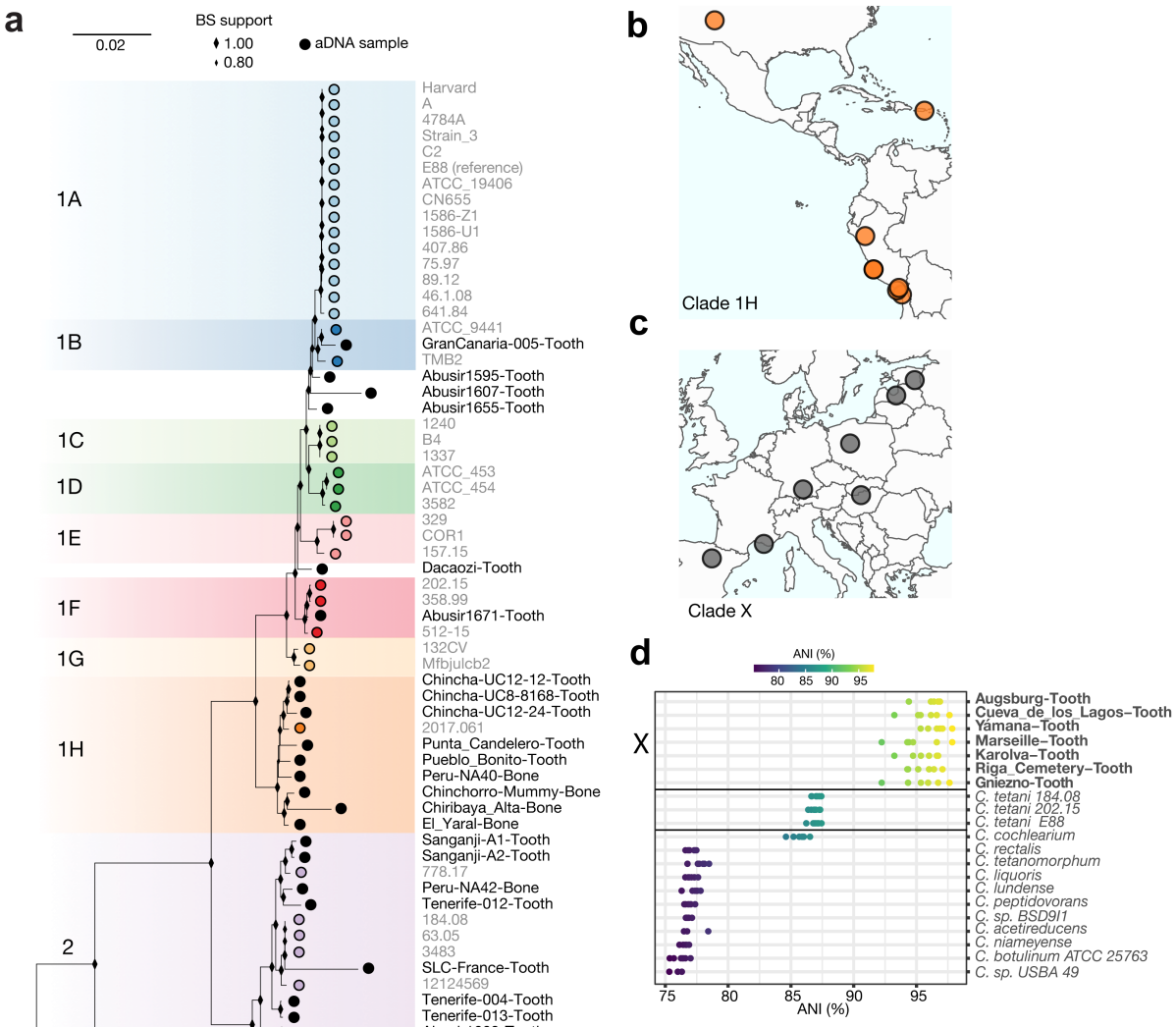

a
BS support
0.02
1.00
0.80
aDNA sample
1A
1B
1C
1D
1E
1F
1G
1H
2
Y
X
Harvard
A
4784A
Strain_3
C2
E88 (reference)
ATCC_19406
CN655
1586-Z1
1586-U1
407.86
75.97
89.12
46.1.08
641.84
ATCC_9441
GranCanaria-005-Tooth
TMB2
Abusir1595-Tooth
Abusir1607-Tooth
Abusir1655-Tooth
1240
B4
1337
ATCC_453
ATCC_454
3582
329
COR1
157.15
Dacaozi-Tooth
202.15
358.99
Abusir1671-Tooth
512-15
132CV
Mfbjulcb2
Chincha-UC12-12-Tooth
Chincha-UC8-8168-Tooth
Chincha-UC12-24-Tooth
2017.061
Punta_Candelero-Tooth
Pueblo_Bonito-Tooth
Peru-NA40-Bone
Chinchorro-Mummy-Bone
Chiribaya_Alta-Bone
El_Yaral-Bone
Sanganji-A1-Tooth
Sanganji-A2-Tooth
778.17
Peru-NA42-Bone
Tenerife-012-Tooth
184.08
63.05
3483
SLC-France-Tooth
12124569
Tenerife-004-Tooth
Tenerife-013-Tooth
Abusir1668-Tooth
Tenerife-011-Tooth
GranCanaria-008-Tooth
Vác-Mummy-Tissue
Riga_Cemetery-Tooth
Yámana-Tooth
Marseille-Tooth
Augsburg-Tooth
Karolva-Tooth
Cueva_de_los_Lagos-Tooth
Gniezno-Tooth
C. cochlearium
b
Clade 1H
c
Clade X
d
ANI (%)
80 85 90 95
X
Augsburg-Tooth
Cueva_de_los_Lagos-Tooth
Yámana-Tooth
Marseille-Tooth
Karolva-Tooth
Riga_Cemetery-Tooth
Gniezno-Tooth
C. tetani 184.08
C. tetani 202.15
C. tetani E88
C. cochlearium
C. rectalis
C. tetanomorphum
C. liquoris
C. lundense
C. peptidovorans
C. sp. BSD9I1
C. acetireducens
C. niameyense
C. botulinum ATCC 25763
C. sp. USBA 49
75 80 85 90 95
ANI (%)

Figure 5.1: Previous work identified two putative *C. tetani* related species (*Clostridium* sp. X and Y) from ancient DNA associated with human archaeological remains. (a) Phylogenetic tree of *C. tetani* related genome bins from ancient samples and modern *C. tetani* genomes. Novel branches are labeled "X" and "Y", which are phylogenetically distinct. (b-c) Geographical location of clade 1H and clade X associated archaeological samples. (d) Average nucleotide identities (ANIs) between clade X bins and a range of *Clostridium* species.

As shown in the phylogenetic tree, the *C. tetani* related genome "bins" from ancient samples cluster quite broadly within and outside of the *C. tetani* tree. Some of the bins clustered within clade 1 with the largest expansion seen in clade 1H. Interestingly, these bins tend to come from ancient DNA samples derived from the Americas including several from ancient South American archaeological samples (Figure 5.1b). In addition, nine genome bins clustered with clade 2 *C. tetani* strains.

However, potentially the most intriguing finding from this phylogenetic analysis is the identification of several genome bins that cluster outside of the entire *C. tetani* tree. One of these clustered as a single lineage ("Y") while eight bins clustered as a monophyletic group (clade "X"). The lineage Y strain was assembled from an aDNA sample (GranCanaria-008-Tooth) taken from an archeological specimen dated to 936CE. The clade X associated samples show a non-random geographic distribution as they occur predominantly in Europe (Figure 5.1c). But are these clade X strains a new lineage of *C. tetani*, or something else entirely?

Interestingly, based on ANI analysis, the clade X strains appear to be distinct from *C. tetani* strains possessing only 86-87% sequence identity, which is below the threshold used to group strains within the same species. Clade X strains also exhibit a similarly low ANI to *C. cochlearium*, which is the closest known species to *C. tetani*. Ultimately, this finding suggests that these newly identified genome bins from archaeological samples represent a potentially new species of *Clostridium*. BLAST searches of the contigs from these bins confirmed that they are distinct from any organisms that are currently represented in the NCBI database as well.

Although the MSc work of Hodgins et al.[238] led to the discovery of lineage Y and clade X strains of *Clostridium*, little to nothing is known about these organisms, their genomic content, and their functional and virulence potential (e.g., whether clade X strains may be neurotoxigenic clostridia). The goal of this chapter is therefore to perform a deeper comparative genomic analysis of these putatively novel *Clostridium* spp. The research questions of this study are as follows: (1) Do *Clostridium* sp. X/Y possess a tetanus

neurotoxin? (2) Are *Clostridium* sp. X/Y pathogens? (3) Are there unique genes and gene neighborhood structures that differentiate *Clostridium sp.* X and Y from *C. tetani*? To answer these questions, I analyzed the gene neighborhoods (powered by AnnoView) of modern *C. tetani* genomes as well as novel *Clostridium* species. Based on this analysis, I propose that *Clostridium* sp. X/Y has low or no neurotoxicity. Two *Clostridium* sp. X genomes possess a fragmented *tent* gene due to potential gene pseudogenization. *Clostridium* sp. X/Y, their plasmids retain plasmid marker genes and critical virulence factors including *repA*, *colT* and *tent* despite genomic rearrangement activities being observed, suggesting that these two putative novel strains could also be pathogenic but with unique biological characteristics.

5.2 Methods

Annotation and comparative genomics of acBins with modern *C. tetani* strains

I annotated all contigs from the 38 acBins strains using prodigal v2.6.2.[125] Anonymous annotation mode was used as it is intended for metagenomes and low-quality draft genomes. Sequences encoding rRNA, tRNA and tmRNAs were also annotated as described above. These annotations resulted in predicted proteomes for all 38 acBins, which were then compared with the existing proteomes from the 41 modern *C. tetani* genomes. To cluster identified protein coding sequences into groups of orthologous sequence clusters, Orthofinder v2.5.4 was used[40] with default settings. Orthofinder was applied to modern *Clostridium tetani* proteomes and prodigal-predicted acBin proteomes, resulting in a set of 4,334 total orthogroups. Any orthogroup that was made up of only partial coding sequences (as predicted by prodigal) was removed to limit orthogroups resulting from gene fragments or pseudogenes. Orthogroups without any coding sequences found on *C. tetani* BLAST-validated contigs were also removed as a filter for any potential bin contamination. Any contigs with a *C. tetani* top hit screened via BLASTN search were considered *C. tetani* to be BLAST-validated. We then performed a gene set comparison between acBins and modern strains by comparing the presence/absence of all identified orthogroups. This resulted in the identification of common orthogroups shared between modern and acBin strains as well as orthogroups unique to acBins, which were investigated further through gene neighborhood analysis. The orthology inference was done in collaboration with Dr. Briallen Lobb, from Doxey lab. In particular, I focused on orthogroups unique to *Clostridium* sp. X and investigated their gene neighborhoods and their conservation with modern reference

C. tetani genomes. For gene neighborhood visualization and alignment, I used AnnoView (https://annoview.uwaterloo.ca). The contigs with the *tent* gene in Augsburg-Tooth and in strain 12124569 were compared using BLASTN, and mapped regions were visualized using Artemis ACT.[56]

Measurement and visualization of genome coverage and alignments

Genome coverage for the recovered acBins was further assessed and visualized in two ways. First, Bowtie2 v2.4.2[248] was used to map reads from individual runs to the E88 chromosome (accession NC_004557.1) and plasmid (accession NC_004565.1). Bowtie2 was run with the following parameters (–local -D 20 -R 3 -N 1 -L 20 -i S,1,0.50 -bSF4). Using samtools v1.12, the resultant BAM files were then sorted, indexed, and merged based on their BioSample ID. Total (average # of reads per base) and percent (number of bases with 1 or more reads divided by total number of bases) coverage was calculated for the entire chromosome and plasmid as well as the *tent*, *colT*, and *repA* regions. Coverage was visualized using Python v3.8.5 and matplotlib v3.3.2. Coverage plots (Figure 5.7a) were created by loading BAM files into R v4.1.0 with the Rsamtools library v2.8.0 and plotted as area plots using ggplot2 v3.3.5. Coverage of the plasmid sequences was calculated by averaging the number of reads per base in 100bp bins.

As a second approach, we also visualized the multiple genome alignment generated by snippy (described below) using ggplot2 v3.3.5 after loading it into R v4.1.0 using Biostrings 2.62.0. This work was done in collaboration with Benjamin Tremblay.

5.3 Results

Comparison of clade X gene content to *C. tetani*

I first carried out a comprehensive comparison of genome content and structure between the acBins and modern *C. tetani* strains. I clustered protein-coding sequences from all modern genomes and acBins into a set of 3,729 orthologous groups, and compared their presence/absence across all strains (see Methods 5.2). Based on this analysis, I observed considerable overlap in gene content between the acBins versus the modern reference genomes, with the greatest overlap observed between acBins from *C. tetani* lineages (1 and 2) and the smallest overlap observed for *Clostridium* sp. X (Figure 5.2). Since there is only one species in *Clostridium* sp. Y, it is hard to estimate the confidence of *Clostridium* sp. Y

specific genes being conserved in this species. Therefore, only *Clostridium* sp. X specific genes were considered for this analysis.

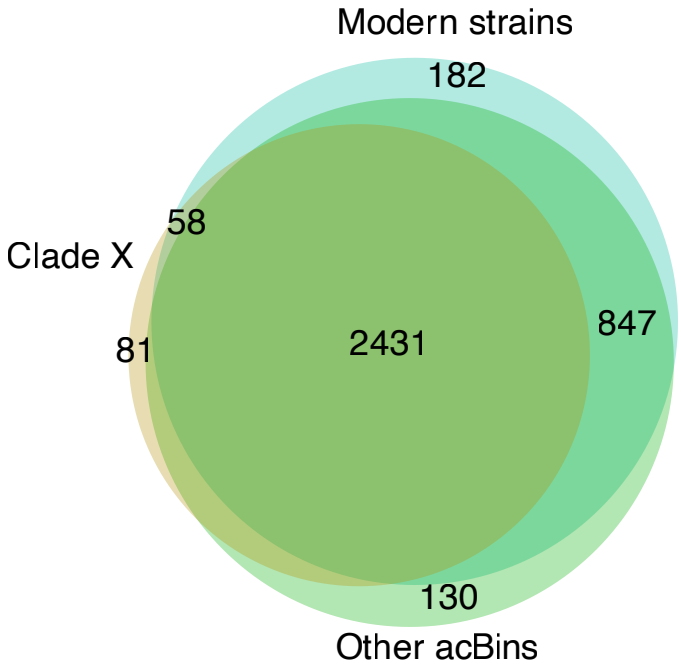


Figure 5.2: Venn diagram indicating overlap of predicted orthogroups between acBins, clade X, and modern *C. tetani* genomes.

Consistent with the idea that clade X represents a distinct species from *C. tetani*, I identified fourteen genes that are unique to four or more clade X members and absent from all other *C. tetani* genomes. As shown in Figure 5.3, most clade X specific genes were detected in two clade X genomes. Only a small subset of genes were detected in half or more than half of clade X genomes (Table 5.1). The potential functions of selected strongly conserved genes in clade X are mentioned in Figure 5.4. It is also important to note that since these bins are assembled from aDNA shotgun datasets (i.e., essentially metagenomes), they may "lack" genes due to absence of data. Thus, levels of gene detection do not necessarily reflect their true levels of conservation.

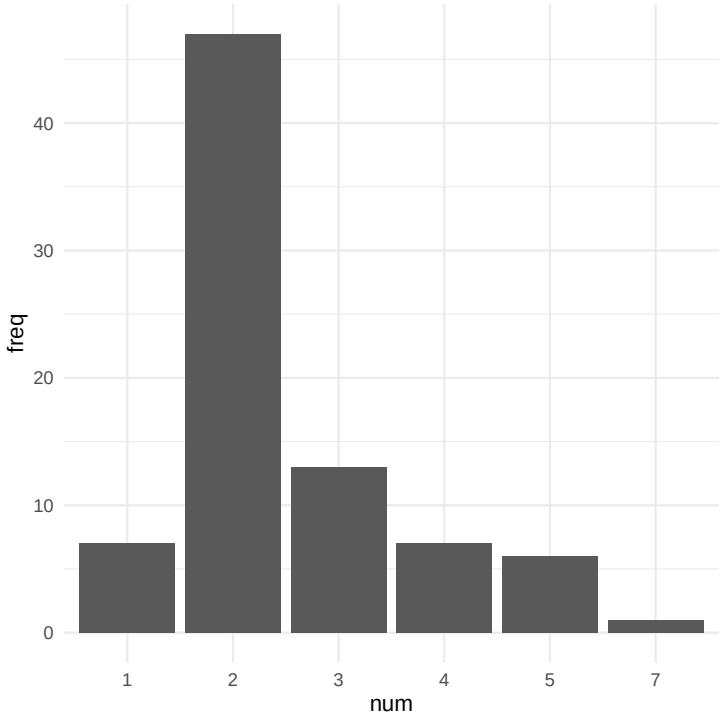


Figure 5.3: The number of clade X specific genes shared by one or more Clade X genomes.

Table 5.1: Unique orthologous groups that occur in more than half of the clade X genomes. Frequency indicates the number of Clade X genomes that contain the corresponding orthologous group.

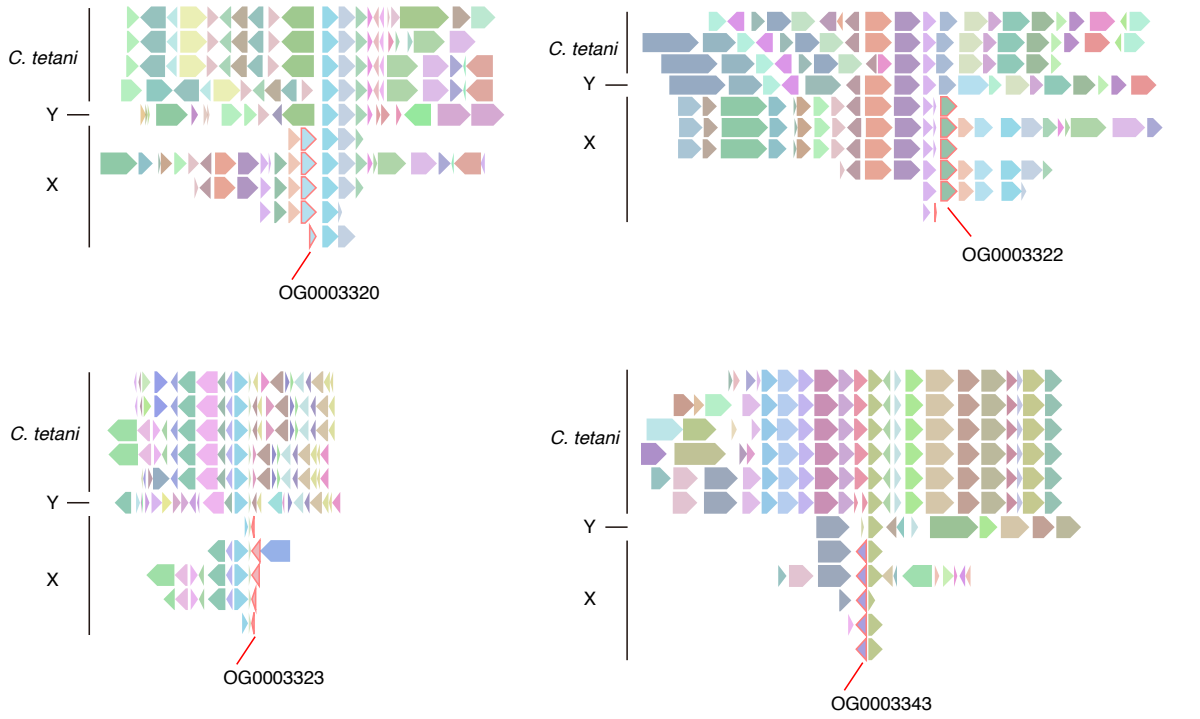
Frequency	**Orthologous group**	**Closest BLAST match**
7	OG0003136	sodium-dependent transporter
5	OG0003208	radical SAM protein
5	OG0003320	MYG1 family protein
5	OG0003322	EFR1 family ferrodoxin
5	OG0003323	SWIM zinc finger family protein
5	OG0003325	YjfB family protein
5	OG0003343	no detected homologs
4	OG0003340	unnamed protein product
4	OG0003513	methylglyoxal synthase
4	OG0003514	hypothetical protein
4	OG0003519	hydrogenase expression/formation protein HypE
4	OG0003520	hypothetical protein
4	OG0003555	hypothetical protein

Analysis of gene neighborhoods surrounding unique genes found in clade X and Y

Given the high similarity in gene content between *Clostridium sp.* X and *C. tetani*, it is therefore important to also compare the genomic organization. I then examined the gene neighborhoods surrounding the cladeX/Y-specific genes (labeled by orthogroup). It was hypothesized that these unique genes may be evolutionary gains of genes in clade X/Y relative to *C. tetani*, or that *C. tetani* may have lost these genes during evolution. By visualizing the gene neighborhoods of these unique genes in clade X/Y, I expected to observe conservation in neighborhood content/structure between *C. tetani* and clade X/Y, but with an absence of the X/Y-unique gene in *C. tetani*.

In order to compare the gene neighborhoods between clade X/Y and *C. tetani* species, I chose four genes that occur in more than half of the clade X genomes, as they are most likely to be genuine gain or loss of genes rather than binning, sequencing or prediction errors (See Table 5.1 for additional unique genes that occur frequently in clade X genomes).

The genomic context of four of these genes (labeled by orthogroup) is shown in Figure 5.4. As hypothesized, although these genes are unique to clade X, at least one side of their surrounding genomic neighborhood structures are conserved in other *C. tetani* genomes, implying that genome rearrangements may have resulted in these genes being either gained in *Clostridium* sp. X or lost in *C. tetani*.



Closest BLAST matches:

OG0003320 - MYG1 family protein in *Lutibacter* sp. B2 (MBF8984673.1), 79.8 % identity
OG0003322 - EFR1 family ferrodoxin in *Desulforamulus reducens* (WP_011878543.1), 79.8 % identity
OG0003323 - SWIM zinc finger family protein in *Anaerophilus nitritogenes* (WP_129596215.1), 83.6 %ID
OG0003343 - No detected homologs

Figure 5.4: Unique gene neighborhood structures in clade X and Y acBins compared to modern reference *C. tetani* strains. Orthology-based clustering of all protein sequences from our dataset identified fourteen orthogroups (genes) unique to four or more clade X members and absent among all *C. tetani* genomes. The genomic context of four of these genes (labeled by their orthogroup ID) is shown above. The highlighted genes in red are unique to clade X. The genes surrounding these X-specific genes are conserved in other *C. tetani* genomes.

Genomic context comparison of plasmid associated genes

Next, I focused on the plasmid as it contains the critical virulence factors (*colT*, and *TeNT*) of *C. tetani* and provides *C. tetani* with its unique biological characteristics. The plasmid marker genes *repA* and *colT* are also of interest. It is hypothesized that if species from clade X/Y are divergent enough from modern *C. tetani*, the gene neighborhood of the plasmid marker genes and virulence factors (*repA*, *colT* and *tent*) is likely to be different in clade X/Y than in *C. tetani*, therefore suggesting genomic rearrangement activities may have happened in *Clostridium* sp. X/Y or *C. tetani*.

To examine differences in plasmid gene content and structure directly, I then compared the gene neighborhoods surrounding the plasmid marker genes *repA* (Figure 5.5) and *colT* (Figure 5.6). In several acBins from *C. tetani* lineages 1 or 2, the gene neighborhoods surrounding these genes are similar to that in modern strains. However, particularly in *Clostridium* sp. X and Y, I identified unique gene clusters distinct from those in modern strains. For example, in two *Clostridium* sp. X genomes and the *Clostridium* sp. Y genome, I identified a conserved toxin/antitoxin pair and a phage integrase flanking the *repA* gene (Figure 5.5). I also observed a unique gene arrangement surrounding *colT* that is conserved in two clade X genomes (Figure 5.6). Additional differences were identified in a few lineage 2 acBins; for example, Tenerife-004-Tooth contains unique genes neighboring *repA*, and the Tenerife-013-Tooth acBin uniquely encodes the *repA* gene adjacent to its *tent* and *tetR* gene (Figure 5.5).

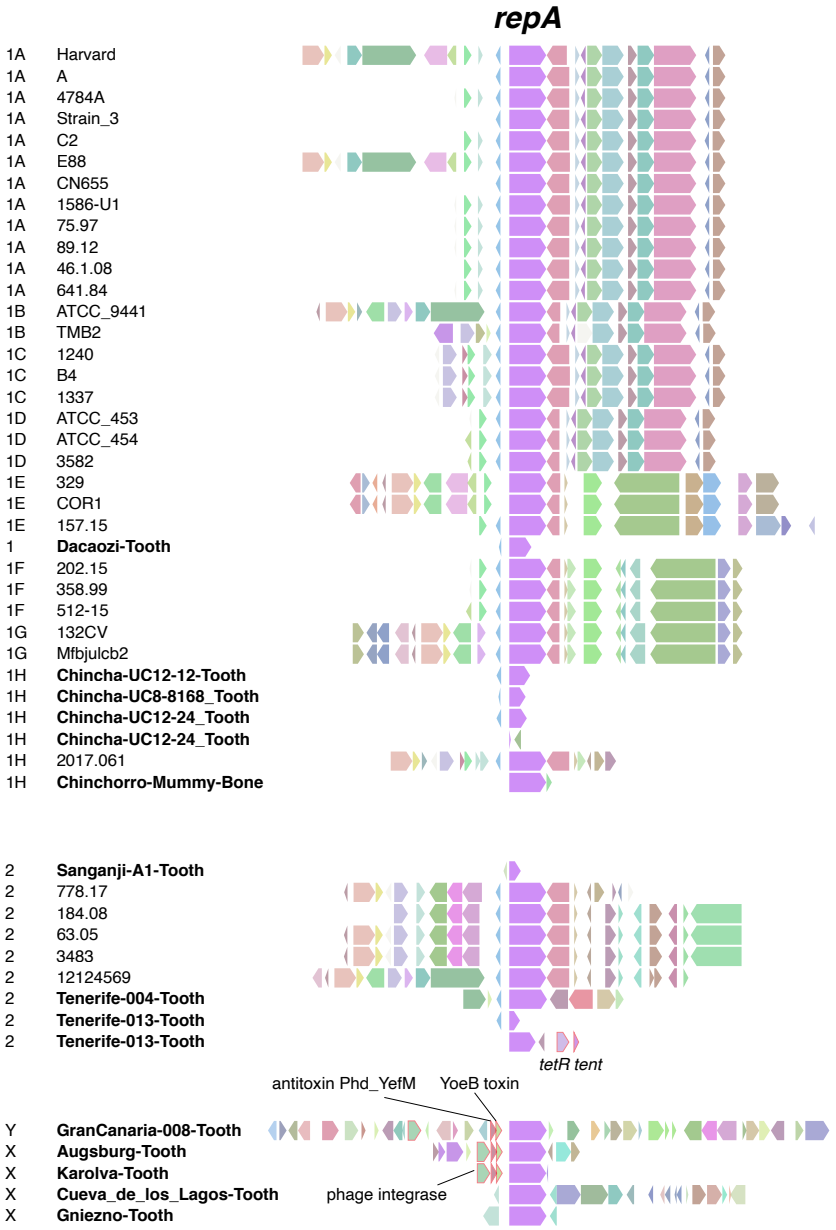


Figure 5.5: Gene neighborhoods surrounding the *repA* genes in select acBins and reference *C. tetani* genomes. Only assembled gene clusters with one or more genes surrounding the center gene were selected for display.

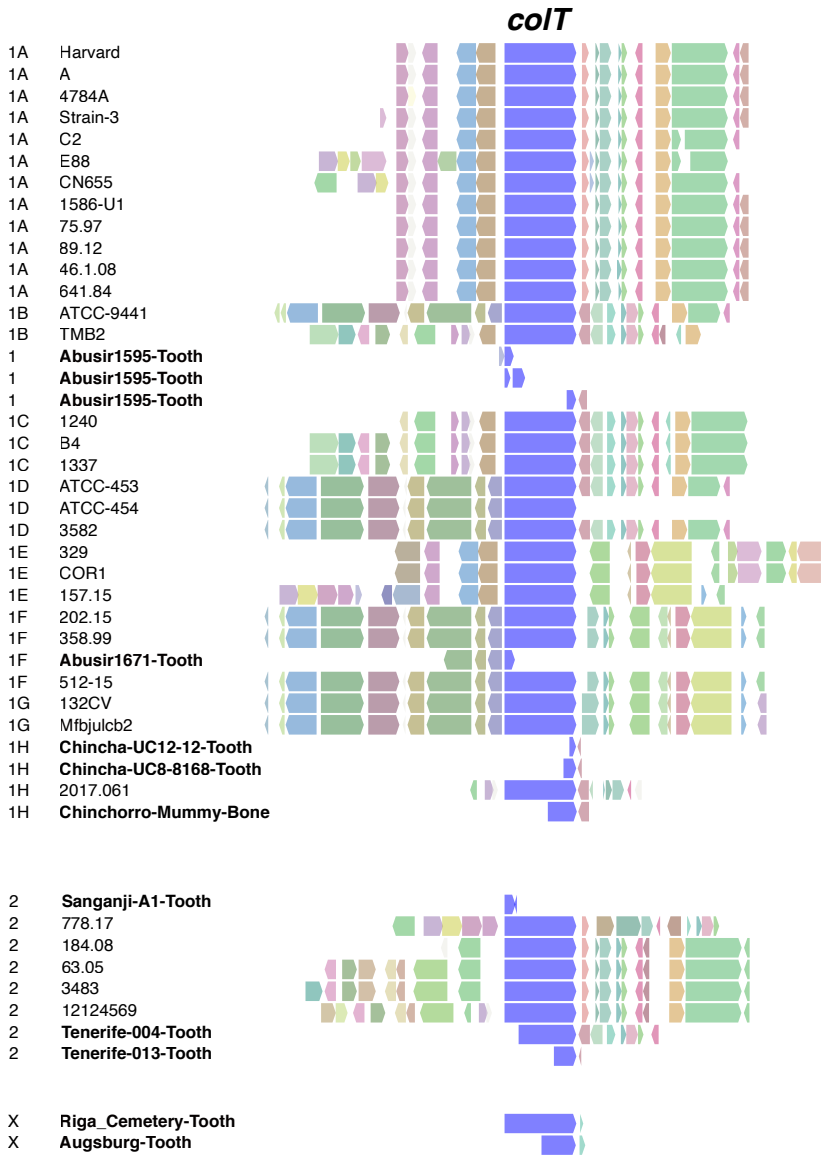


Figure 5.6: Gene neighborhoods surrounding the *colT* genes in select acBins and reference *C. tetani* genomes.

Identification of a putative *tent* pseudogene fragment resulting from a deletion event

I then performed a detailed comparison of the plasmid-encoded neurotoxin gene, *tent*, and its gene neighborhood (where possible) across the strains. As shown in Figure 5.7, based on mapped read coverage to these regions, the *tent* gene was detected at a relatively high depth of coverage in acBins from *C. tetani* lineages 1 and 2 as expected. The *tent* gene neighborhood structure from lineage 1 or 2 acBin strains is also similar or identical to that in modern strains, with the exception of Tenerife-013-Tooth (as it encodes the *repA* gene nearby) (Figure 5.8).

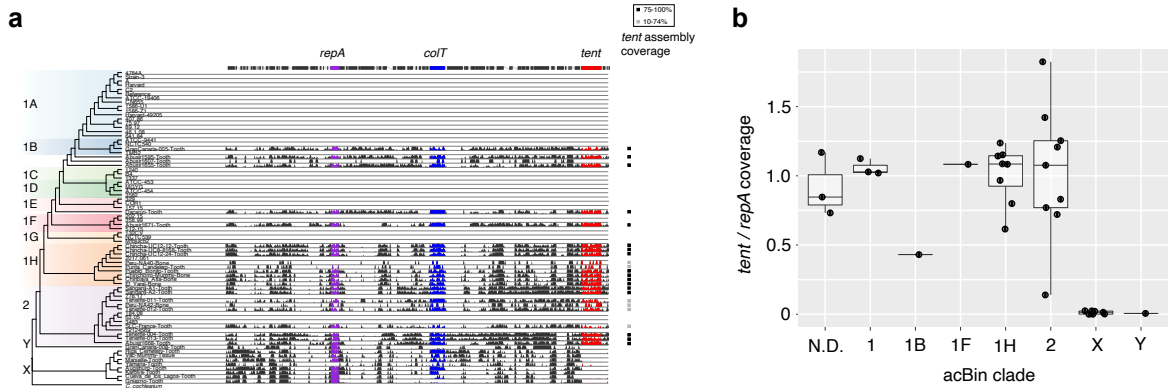


Figure 5.7: Mapped read coverage to the reference *C. tetani* E88 plasmid and *tent* genes. (a) Visualization of mapped read coverage to the *C. tetani* E88 plasmid. The genes (*tent*, *colT*, and *repA*) are coloured red, blue and purple, respectively. The plot reveals a relative lack of coverage in clade X acBins across the *tent* gene. Coverage values were capped to the 80th percentile to prevent high coverage regions from obscuring other regions. Genes were plotted as black bars using RefSeq annotations. On the right of the plot, the boxes indicate the acBins for which *tent* sequences were assembled with high coverage (black) or partially (gray) using SNP profiling as described in the Methods. (b) The coverage was calculated as the average depth of coverage based on mapped reads to each gene. The boxplots depict the lower quartile, median, and upper quartile of the data, with whiskers extending to 1.5 times the interquartile range (IQR) above the third quartile or below the first quartile. This figure was constructed in collaboration with Benjamin Tremblay, who wrote R code for plotting.

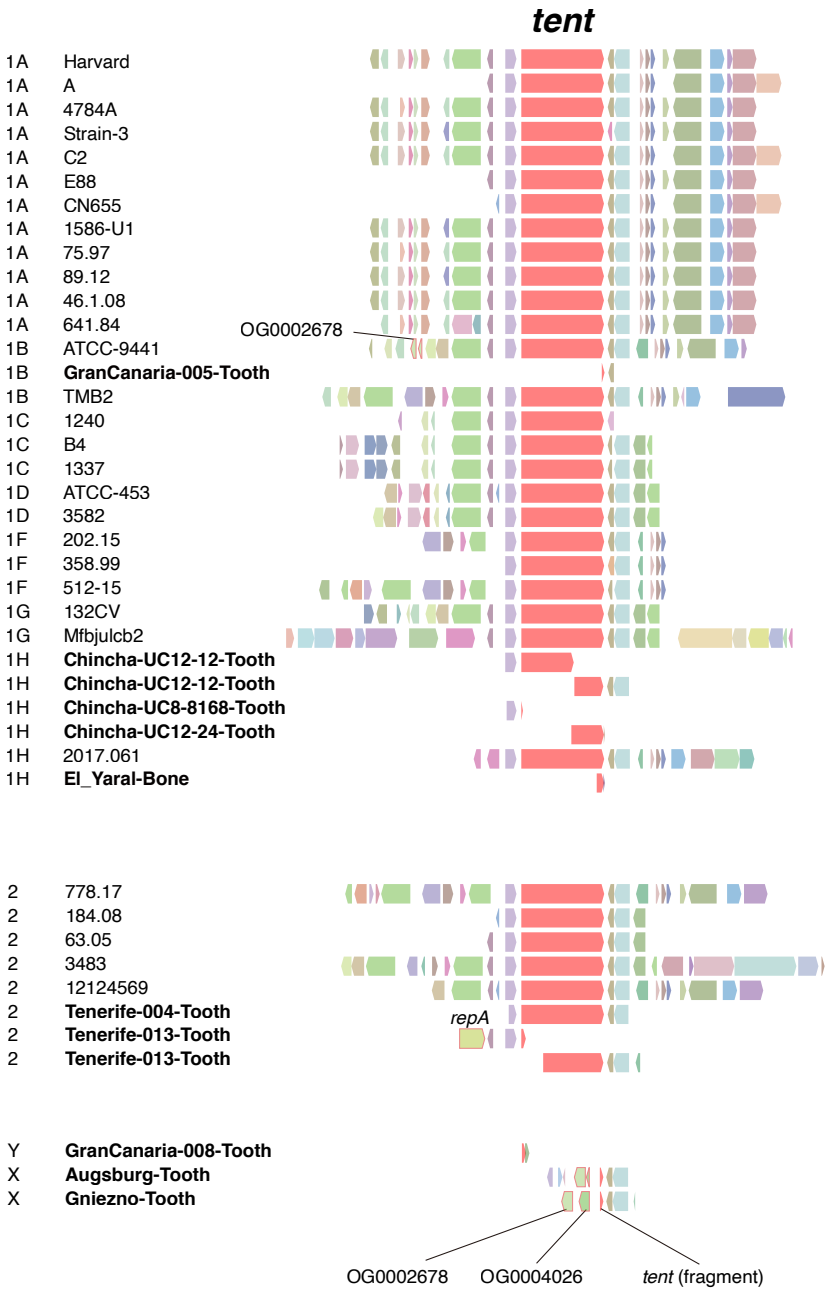


Figure 5.8: Gene neighborhoods surrounding the *tent* genes in select acBins and reference *C. tetani* genomes.

However, in the acBins from lineage X and Y, the *tent* gene was either missing or was fragmented, suggesting a possible gene loss or pseudogenization event (Figure 5.7B). This pattern can be seen clearly in read coverage plots (Figure 5.7A) and when normalizing *tent* depth of coverage to that of the plasmid-marker gene, *repA* (Figure 5.7B). The *tent* locus in the two *Clostridium* sp. X genomes for which assembly data is available over this region, seems to have undergone a deletion event resulting in the deletion of over 90% of the *tent* sequence as well as 3-4 neighboring genes (Figure 5.8 and Figure 5.9). This analysis further supports the idea that the *tent* fragment may be a non-functional pseudogene in these clade X strains. Other hypotheses of the formation of fragmented *tent* in *Clostridium* sp. X were mentioned in discussion.

Ultimately, our comparative genomic analysis of gene content and neighborhood structure demonstrates that the plasmids in several of our ancient samples (particularly those of *Clostridium* sp. X) are distinct from modern *C. tetani* plasmids, while the plasmids of acBins from lineages 1 and 2 are similar to those of existing *C. tetani* strains. The critical virulence factor, *tent*, was either missing or fragmented in clade X/Y acBins, while *colT* was identified in clade X/Y acBins, suggesting that *Clostridium* sp. X/Y might still be pathogenic but with a weakened toxicity compared to modern *C. tetani*. Further gene neighborhood analysis of the clade X specific genes shows that neighboring gene content and orders are conserved in modern *C. tetani*, implying that the unique genes were lost in *C. tetani* or gained in novel *Clostridium* species. Together, these analysis suggest that clade X and branch Y represent new *Clostridium* species that are closely related to *C. tetani* but reduced in their neurotoxicity and virulence capabilities.

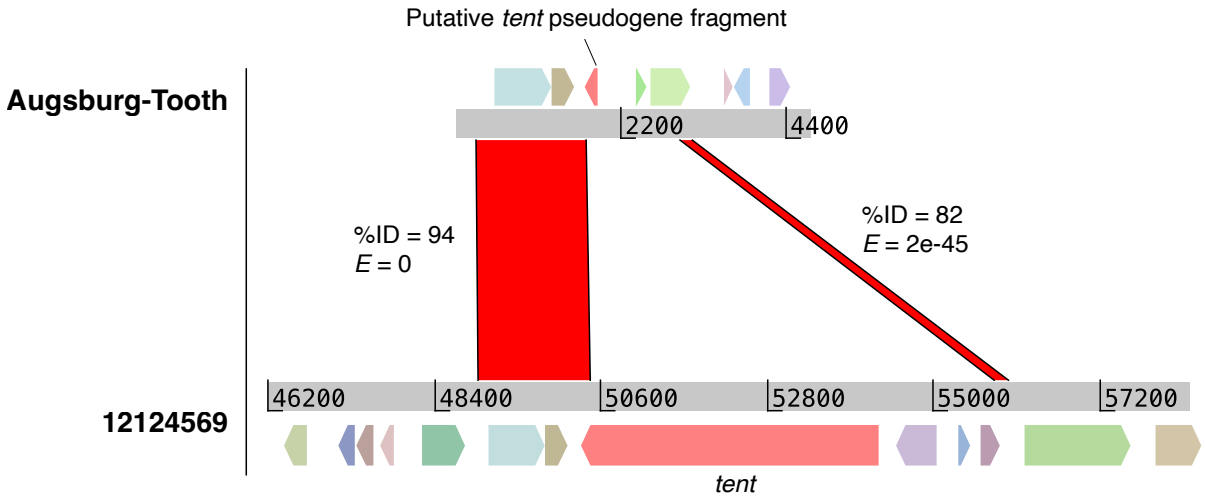


Figure 5.9: Identification of a putative deletion event resulting in a possible *tent* pseudogene in clade X acBins. Two clade X acBins possess an assembled contig containing a *tent* fragment as shown in Figure 5.8. A BLASTN comparison of this contig from Augsburg-Tooth against a representative *tent* locus from strain 12124569 reveals two regions of homology that flank the *tent* gene in strain 12124569. This indicates a possible rearrangement or deletion that has resulted in the removal of the *tent* gene and some adjacent sequence in Augsburg-Tooth.

5.4 Discussion

In this study, I analyzed gene content and genomic organization of putative, new *Clostridium* species found in ancient human samples. Genomic context analysis of the plasmid marker genes and virulence factors shows that gene neighborhoods in *C. tetani* and *Clostridium* sp. X/Y are distinct. A fragmented *tent* gene was identified in contigs from *Clostridium* sp. X. This may represent the first identification of a tetanus toxin-like gene (even if it is just a pseudogene) in a species outside of *C. tetani*, and it further supports the close relationship between *Clostridium* sp. X/Y and *C. tetani*. In addition, *Clostridium* sp. X/Y shares most genes with *C. tetani*, but with unique genes and genomic organizations identified, which provides another piece of evidence supporting the idea that *Clostridium* sp. X/Y is a unique *Clostridium* species from *C. tetani*. This reinforces other evidence including phylogenetic analysis and marker genes phylogenetics.[238] However, as this identification relies on metagenomic predictions, there is still some uncertainty, since the possibility

of assembly errors or strain contamination by *C. tetani* cannot be completely ruled out that may have produced artifactual results. Further whole-genome sequencing or genomic discovery and analysis of clade X species in other environments may be needed to reinforce these findings.

To explain the fragmented *tent* in clade X genomes, three hypotheses are proposed: (1) There was a gene loss in *tent* sequence due to pseudogenization (Figure 5.10A). The pseudogenization event is caused by the accumulation of mutation events including premature stop codons in the *tent* gene, the gene decays, and becomes different from the ancestral gene. However, the gene loss of *tent* leads to another question: why would this important toxin gene be lost in clade X? One idea is that it is possible that an ancestral clade X species carrying a plasmid with *tent* may have experienced a drastic environmental change, and that there was no selective advantage to maintain the toxin gene that requires highly metabolic cost. This is a plausible scenario because clade X strains were identified in *human* archaeological samples, but it is possible that the ancestral toxin (and clade X species) has a different host specificity or that its toxin loses its fitness advantage in later stages of tissue colonization. It could also be that there is a fitness advantage for non-toxigenic strains to develop throughout the course of tissue decomposition, since strains with lower metabolic cost may have a growth advantage. (2) Modern *C. tetani* acquired the *tent* gene via HGT. Since clade X retained a fragmented *tent* gene, it is less likely that the DNA happen to insert near the stop codon of the fragmented *tent* and form a new extended *tent* gene (Figure 5.10B). (3) The contigs may be contamination from non-toxigenic *C. tetani* genomes that haven't been sequenced before. While previously reported non-toxigenic *C. tetani* strains either have the whole plasmid missing, or their *tent* gene was deleted, thus rendering the strains non-toxigenic.[247] There have been no reports of a fragmented *tent* gene.

Although there appears to be some partial conservation of a *C. tetani* like plasmid in *Clostridium* sp. X and Y with genes such as *repA* and *colT* consistently detected, the *tent* gene was either not detected or was fragmented indicative of a pseudogenization event. These findings support the delineation of the species boundary between these groups, and highlight *tent* as a key gene separating *C. tetani* from closely related species. However, even if they are only pseudogenes, the detection of partial *tent* fragments in the X acBins raises the intriguing possibility that some members of *Clostridium* sp. X may have at one point encoded a full-length *tent* gene, and that *tent*-carrying plasmids may have been circulating in the past outside of *C. tetani* in other closely related *Clostridium* species. This is in line with recent genomic studies that have demonstrated the potential for species outside of the *Clostridium* genus to carry neurotoxin genes.[120,122,204] Future efforts to sequence the microbiomes associated with archaeological samples, as well as environmental *Clostridium*

isolates, may shed further light on these questions.

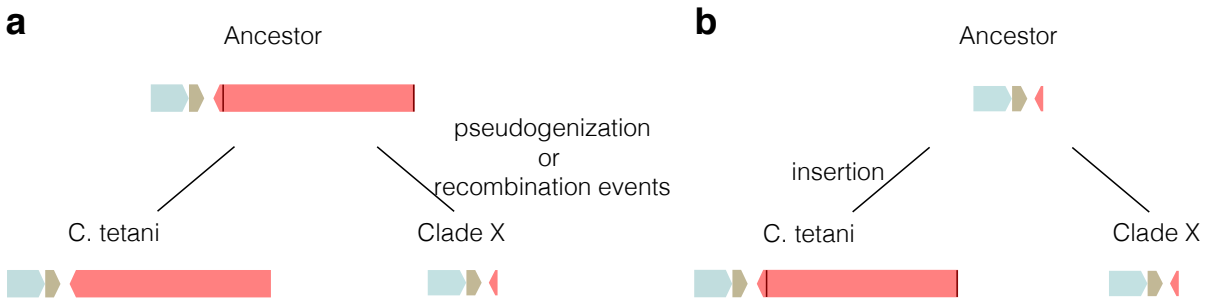


Figure 5.10: Evolutionary models of *tent* gene in *Clostridium* sp. X and *C. tetani*. (a) The *tent* gene have undergone a gene loss event due to pseudogenization or recombination events. (b) The *tent* gene have undergone a gene gain event due to an insertion of DNA fragment via HGT.

These analysis shows that in *Clostridium* sp. X/Y, the tetanus toxin genes were either found to be fragmented or missing, suggesting *tent* is partially deleted compared to normal tetanus toxin genes according to the genomic context analysis. Genomic context analysis of unique orthologous groups revealed additional genomic neighborhood differences between *Clostridium* sp. X/Y and *C. tetani*, reinforcing our earlier analysis indicating that clade X and branch Y represent a new *Clostridium* sp. that is closely related to but distinct from *C. tetani*.

Chapter 6

Discussion

> Although context methods usually provide only rather general predictions, they represent a new and important development in genomics that explicitly takes advantage of the rapidly growing collection of sequenced genomes.
>
> *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics.*
> Koonin, E., & Galperin, M. Y.[249]

Gene neighborhood approaches have been widely used in genomics studies in the past decades. With the advancement in sequencing techniques, gene neighborhood information can be easily accessed from genomic data. However, exploring gene neighborhoods for custom genes and species of interest can be challenging, highlighting a need for a tool that enables gene neighborhood data retrieval, visualization, comparison, and customization. The wealth of neighboring gene information offers valuable insights into functional linkages between characterized and uncharacterized genes, as well as insights into the evolutionary history of gene clusters. Investigating gene neighborhoods in pathogen genomes helps investigate the gain or loss of virulence genes, explore the genetic background of virulence factors, infer potential evolutionary events related to pathogenicity traits, characterize genes and gene clusters involved in pathogenesis, and even predict potential host targets based on gene cluster functionalities. In this chapter, I review the contributions presented in this thesis, discuss current limitations in gene neighborhood analysis, and conclude with a perspective for this field.

6.1 Summary of findings

6.1.1 Identification of *IPT* gene clusters

Analyzing gene neighborhoods yields valuable insight into gene functional associations, facilitating the characterization of genes and gene clusters. As seen in Chapter 3, several different types of clusters of *IPT* genes were identified across different bacterial lineages by genomic context analysis. Cytokinin is generally considered beneficial to plant growth. However, some bacteria can overstimulate plant growth and can sometimes be detrimental due to CK overexpression. The distribution of adenylate *IPT* genes is highly limited to a few bacterial lineages, which includes both PGPB and pathogens. In Chapter 3, I investigated the CK biosynthesis pathway in further detail by examining *IPT* gene clusters across different lineages, along with other comparative studies including phylogenetic analysis, promoter analysis, and codon usage analysis. The protein phylogeny showed that this protein family is divided into two different clades. One clade is abundant with plant-pathogenic bacteria, and the other clade contains bacteria with mixed types of bacteria-plant interactions, including plant growth promoting bacteria, plant pathogens, and several species with unknown lifestyles. This clade was largely considered to be PGPB-associated. It is important to mention a key limitation of this study: *IPT*-containing bacteria were binary classified as beneficial or pathogenic. This classification overlooks the complexity of bacterial behavior and interactions, particularly in cases where bacteria do not fit neatly into either category, such as pathobionts, or other opportunistic or context-dependent pathogens.

Pathogen-associated *IPT*s were found to be frequently coupled with predicted virulence loci. Thirty-two (65%) of the *IPT* gene neighborhoods from the pathogen-associated clade were found to contain pathogen-associated genes, while only 3 PGPB and 3 unknown (14%) IPT gene clusters satisfy this criterion. The pathogen-associated *IPT* gene clusters were then studied in greater depth. *Agrobacterium rubi* was found to contain two *IPT* genes in different regions of the genome; one is located on the T-DNA region and surrounded by *iaaH* and *iaaM* genes that are involved in auxin biosynthesis; the other IPT gene is located in a virulence region and surrounded by other virulence genes, such as *virA*, *virB* and *virK*. The *iaaH* and *iaaM* genes were also detected in the *IPT* neighborhoods of some other genomes, such as *Agrobacterium rhizogenes* and *Dickeya chrysanthemi*. Some other neighboring genes with high PathFams scores were detected in the pathogenic strains. For example, in *Ralstonia solanacearum*, the NEL protein encoding gene that is usually found at the C-terminus of bacterial virulence factors is located upstream the IPT coding gene.[183] In *Xanthomonas albilineans*, a lipoprotein coding gene, upstream *IPT*, was involved in the translocation of virulence factors. It should be noted that the observed pattern of

IPT genes clustering with predicted virulence genes might also be influenced by a general tendency of pathogens to contain an increased frequency of virulence genes. One method to potentially address this issue is to compare the observed pathogen-association scores for *IPT* neighborhoods in pathogens with those derived from other regions of the genome. This analysis could help determine if the proximity of *IPT* to virulence genes is a distinct feature or simply a result of its presence in pathogenic genomes.

On the other hand, PGPB-associated *IPT*s are in close proximity with genes related to cytokinin regulation. The product of these genes includes possible lysine decarboxylase domain (PF03641) that encodes the LOG enzyme for CK activation, "cytokinin dehydrogenase 1, FAD and cytokinin binding" domain (PF09265) that encodes CKX enzymes for CK degradation, RolB/RolC glucosidase family domain (PF02027) that encodes RolB/RolC protein that can release CK from glucoside conjugates. Forty-two and twenty-four out of 98 of the *IPT* gene clusters contain the *LOG* and *CKX* genes, respectively. The *LOG* gene was found in only 10 members (19%) of pathogen associated clade, but was found in a larger number (32 members, 71%) of PGPB associated clade. The *CKX* gene was found exclusively in *Streptomyces* spp. and *Rhodococcus* spp. in the PGPB-associated clade (24 members, 45%). The gene co-occurrence data shows that *CKX* was identified in several additional genomes in clade B (11 total members) and three *Streptomyces* in clade A. Co-occurrence of all three genes (*IPT*, *LOG*, and *CKX*) in the neighborhood occurred exclusively in 19 (36%) members of PGPB associated clade. Co-occurrence with the *rolB/rolC* gene was exclusive to a subclade of pathogen associated clade including *Agrobacterium* species. It is suggested that these CK regulation related genes may play a role in transform the synthesized CK into active forms, and then degrade CK when there is an excessive amount.

A putative novel operon structure is also identified in the *IPT* gene neighborhood from the non-pathogenic clade. A gene cluster that contains *IPT*, *CKX* and sometimes *LOG* genes is conserved in multiple *Streptomyces* strains. This known operon structure is very similar to part of the *fas* operon structure, which is found in *S. tugidiscabies* and several other *Rhodococcus* strains.[250] The *fas* operon is known to regulate the fascination of the bacteria, and helps invade the plant cells.

6.1.2 Identification of two novel clostridial neurotoxins

An additional application of genomic context analysis is to provide novel functional insights for unknown genes/proteins by applying guilt-by-association (GBA) approaches. As seen in Chapter 4, the putative BoNT-like toxin genes in two *Paeniclostridium ghonii* strains

(PG toxins) were predicted to target insect hosts. The BoNT family of toxins is the most potent toxin known to humankind, and traditionally is associated with *Clostridium* lineages. In recent years, several new BoNT-like toxins or homologs have been identified within or outside the *Clostridium* species. In Chapter 4, two putative *bont*-like gene clusters were identified in *Paeniclostridium ghonii*. This BoNT-like gene cluster contains two genes that shares sequence similarity with BoNT. The neighboring genes were annotated and compared to investigate the function and target host of putative novel neurotoxins.

In both *P. ghonii* genomes, the PG-toxin-LC gene encoding the predicted zinc metalloprotease is immediately upstream of the PG-toxin-HC gene encoding the translocase and binding domain. The PG-toxin1 locus from strain NCTR 3900 is similar but non-identical from that of the PG-toxin2 locus from strain DSM 15049. A notable unique feature of the PG-toxin locus is the presence of a nearby gene encoding an insecticidal Cry delta endotoxin (accession # WP_250673405.1). In both *P. ghonii* genomes, this Cry toxin gene is encoded in the opposite direction to the PGT-LC genes, which suggests that it may be regulated differently.

In addition, The PG-toxin locus also encodes several genes shared with those found in other BoNT gene clusters. These include *amiD* (n-acetylmuramoyl-l-alanine amidase gene) and *parA* genes which are also present in the BoNT/En locus of *E. faecium*; and *abc* (ABC transporter ATP-binding protein) and *fic* (filamentation induced by cAMP protein) genes that are also found in the *tent* locus of *C. tetani*. Three nearby genes are also shared with the PMP1 locus of *P. bifermentans*: an *mpp* gene encoding a putative metallophosphatase family protein, an *abc-permease* (ABC transporter ATP-binding protein/permease) and a *radsam* gene encoding a radical SAM domain containing protein. Interestingly, the *orfx*, *ha* and *ntnh* genes present in other BoNT gene clusters are absent. Downstream of the *mpp* gene in both *P. ghonii* genomes is a cluster of genes encoding a BhlA/UviB family holin-like peptide and two genes encoding aureocin-like type II bacteriocins (Figure 4.1c). Other toxins are known to be exported via holin-dependent mechanisms.[220] The bacteriocins and putative endolysins that may facilitate lytic toxin secretionin, together with insecticidal Cry8Ea1, suggesting a possibility that this gene cluster might serve an insecticidal purpose.

In general, PG toxins and their neighboring genes form a gene cluster that is divergent from the BoNT gene clusters in terms of neighboring gene composition, which suggests that the PG toxins may be functionally divergent genes from other BoNT subtypes. This result is consistent with the protein phylogeny. As expected, the PG-toxin genes from both *P. ghonii* strains clustered as neighbors indicating a common origin in an ancestral *P. ghonii* strain. The phylogenetic placement of the PG-toxin-LC and PG-toxin-HC sequences was similar. Both LC and HC of the PG toxins cluster with X/En/PMP1 lineage, form a distant subclade from other BoNT subtypes.

The PG toxin loci may reside on a plasmid. Within the scaffold from DSM 15049 containing the PG-toxin2 genes, we identified genes encoding plasmid replication initiation proteins. BLAST analysis of these proteins confirmed a close relationship to orthologous genes from *P. bifermentans*, suggesting a link to the toxin-encoding megaplasmids of *Paraclostridium*.[121] The contigs carrying the BoNT-like gene sequences for *P. ghonii* DSM15049 and NCTR 3900 were predicted as plasmids with 73.28% and 74.66% percent of the vote, in excess of the 40-60% threshold where incorrect classifications typically fall.[219]

BTNT from *Bacillus toyonensis* has a gene neighborhood structure that is characteristic of other BoNT genes with the *bt-ntnh* gene located immediately upstream as expected, as well as three P47/ORFX-like genes (Figure 4.6). BLAST analysis revealed that the P47/ORFX containing proteins putatively match the proteins ORFX3, P47, and ORFX2 in *P. bifermentans*, *E. faecium*, and *C. botulinum*. This general gene arrangement is conserved among other members of the X/En/PMP1 clade of BoNT-like toxins, and also some other BoNT genes (e.g., BoNT/A4). It is noticeable that *p47* is located within the *orfX* locus in the genomic neighborhood of BNTN, whereas it is usually located upstream or downstream of the *orfX* locus in the genomic neighborhood of other BoNT toxins.

6.1.3 Identification of genomic traits unique to *Clostridium* sp. X

Gene neighborhood comparison can also be applied to infer genomic rearrangement events that might have occurred in specific strains. As seen in Chapter 5, two putative new *Clostridium* species (*Clostridium* sp. X and Y) that are closely related to *C. tetani* from ancient DNA samples were identified by Hodgins et al.[237] In Chapter 5, the proteomes of *Clostridium* sp. X and Y and reference genomes are annotated by orthology inference methods. Then *Clostridium* sp. X and Y were compared against to modern *Clostridium* reference genomes by comparative genomic analysis including gene content and genomic context comparison.

There is a considerable overlap in gene content between the acBins versus the modern reference genomes, with the greatest overlap observed between acBins from *C. tetani* lineages (1 and 2) and the smallest overlap observed for *Clostridium* sp. X. In terms of differences, fourteen genes present in four or more clade X members and absent from all other *C. tetani* genomes. Although these genes are unique to clade X, their surrounding genes are conserved in other *C. tetani* genomes, implying that genome rearrangements may have resulted in these genes being either gained in *Clostridium* sp. X or lost in *C. tetani*.

The genomic context surrounding the plasmid marker genes *repA* and *colT* are different

between *Clotridium* sp. X and Y and modern *Clostridium* genomes. In *Clostridium* sp. X and Y, unique gene clusters distinct from those in modern strains. For example, in two *Clostridium* sp. X genomes and the *Clostridium* sp. Y genome, I identified a conserved toxin/antitoxin pair and a phage integrase flanking the *repA* gene. I also observed a unique gene arrangement surrounding *colT* that is conserved in two clade X genomes.

The plasmid-encoded neurotoxin gene, *tent*, is either missing or was fragmented, suggesting a possible gene loss or pseudogenization event. This pattern can be seen clearly in read coverage plots and when normalizing *tent* depth of coverage to that of the plasmid-marker gene, *repA*, provided by the collaborators. The *tent* locus in the two *Clostridium* sp. X genomes for which assembly data is available over this region, appears to have undergone a deletion event resulting in the deletion of over 90% of the *tent* sequence as well as 3 neighboring genes. This analysis further supports the idea that the *tent* fragment may be a non-functional pseudogene in these clade X strains.

6.2 Limitations

Genomic data quality

Careful data selection is required when preparing gene neighborhood datasets for comparative analysis. Highly divergent species may have significant differences in genome content and structure, making them less comparable due to limited or even no shared genetic features. Conversely, closely related species may exhibit no detectable differences in gene order, since they haven't had sufficient time to accumulate significant evolutionary changes. Thus, when comparing the genomic context of a protein/gene family, it is imperative to choose species with an appropriate evolutionary distance. In addition, one should be cautious about the potential skewing effect posed by closely related species. There are some strategies to mitigate the overrepresentation of similar gene sequences. For example, one could select a representative gene neighborhood for each taxa, or apply a sequence similarity cutoff to reduce sequence redundancies before extracting gene neighborhood, or weight the functional association score by sequence similarity. Gene cluster grouping tools might also be beneficial for generating representative gene neighborhood and facilitating comparisons across distant clades.[84]

The selection of species also depends on the research question and the sequencing data available. Closely related species provide insight into more recent evolutionary changes, whereas more distantly related species offer a broader perspective on the evolutionary history

of the whole protein family. To ensure high-quality genomic data, many studies opt for reference genomes and sequences from specialized, hand-curated databases. These databases leverages experts insights into specific research areas, such as the *Pseudomonas* genome database,[251] and *bont* database (at https://bontbase.org) for clostridial neurotoxins. In the absence of such specialized databases, it is common practice to opt for reference genomes from widely recognized public databases or genomic data that were often used by the research community. For example, reference genomic data might be sourced from Refseq, or from GTDB, where genomes are selected based on their representativeness of species clusters, as determined by GTDB taxonomic classification system.[85] Reference genome databases not only minimize genome redundancies in the database that are mainly caused by human attention and the popularity of the organism, but also covers as many organisms as possible while still being representative of current prokaryotic genomic data. However, one pitfall of using such reference genomes is the potential loss of intra-species variation. In Chapter 3, reference genomes from GTDB were selected to examine the evolutionary history of the adenlyate IPT gene family. Although nopaline type strain of *A. tumefaciens* is known for having two *ipt* copies, it was not included in GTDB, leading to its absence in the gene copies analysis.[186,187]

The reference genomic databases also ensure that high-quality genome assemblies are used for genomic context analysis. High quality assemblies contain fewer and longer scaffolds compared to low quality assemblies, facilitating a comprehensive understanding of neighboring genes for almost any genes of interests. Many genomic context tools, such as microbesonline,[64] Gecoviz,[69] fast.genomics,[76] and including AnnoView, leverage reference genomes to guarantee consistent annotations and minimal contamination in high-quality genomes. The AnnoView precomputed database includes representative species from GTDB species clusters, which are required to pass an integrated quality score. These quality check measures includes criteria such as checkm score, whether the genome is from a type species or strain, the number of contigs, etc.[85] Similarly, proGenomes is another prokaryotic reference genome database that is often used for comparative studies.[252] It also implements several metrics for genomic data quality control and serves as a data source for the comparative genomic context tool Gecoviz.

Functional dark matter

One of the major weakness of genomic neighborhood analysis is the prevalence of functional dark matter (FDM) in prokaryotic genomes. The amount of FDM greatly influences the information content derived from genomic context analysis. FDM includes genes without annotation, hypothetical proteins, and those encoding proteins that match ambiguous

protein signatures, such as the domain of unknown function (DUF). FDM makes up a significant portion of prokaryotic proteomics.[253] For example, GTDB r86 proteomics contains 21% FDM that cannot be annotated by protein or domain-based homology search approaches.[253] Up to 2020, the NCBI protein database has greater than 30% FDM in archaea and nearly 20% FDM in bacteria.[254] A great proportion of FDM is also seen in Chapter 3, where only 55.21% of coding sequences can be annotated by Pfam domains, leaving the rest of genes unannotated. Nonetheless, many genes with undefined functions demonstrate conservation across various species, suggesting their potential role in cellular functionality. Therefore, despite their ambiguous nature, FDM can still be informative to comparative functional analyses when annotated using orthology inference. This method postulates that proteins within the same orthologous group typically perform similar functions, and can be seen as the same gene. Consequently, unknown proteins can be categorized into respective orthologous groups, assisting in the identification of conserved protein families and determining protein functional associations without prior functional understanding. As observed in Chapter 5, the proteomes of *C. tetani* genomes benefited from orthology inference. However, this strategy does not guarantee that every protein sequence will find an orthologous sequence in the database, leaving some proteins without functional annotations or assigned orthologous groups. The proportion of such proteins in the prokaryotic protenomes is not a trivial amount. In a reference prokaryotic database (proGenomes2), it is reported that 18.8% of the proteome does not match any orthologous groups.[69]

The proportion of FDM in metagenomes derived from environmental samples is even higher since metagenome-assembled genomes (MAGs) are generally harder to annotate. Considering the rapid advancement and accumulation of genomic and metagenomic data, we can anticipate a corresponding rise in the identification of FDM. In particular, some efforts have been made in recent years for FDM annotation, incorporating methods ranging from data mining to the application of deep learning techniques.[254–256] It is expected that these new technologies help boost our understanding and annotation of FDM. Interestingly, genomic context approaches emerges as a valuable strategy to assign functions to FDM based on protein association within genomic loci. As FDM is better annotated through these various approaches, the granularity and depth of data from neighboring gene studies are expected to see significant improvements.

GBA is Error-prone

As one of the GBA approaches, genomic context approach inherently carries a higher risk of errors due to its reliance on indirect connections and associations rather than direct causation.

One example of the false negatives can be seen in gene clusters that exhibit diverse functions. For example, FAD synthase is in close proximity to the protein translation-related genes across different bacterial lineages, and RNA modification/degradation genes cluster with the central metabolic enzyme enolase in euryarchaeota and a subset of proteobacteria.[257,258] In Chapter 4, the PG toxin clusters are in close proximity to *amiD* (N-acetylmuramoyl-L-alanine amidase), despite that *amiD* does not have functional similarity to neurotoxins or any toxin-related properties. In some cases, proteins with diverse functions falling under the same general functional theme are clustered together in the genome, e.g., carbohydrate metabolism-related genes and carbohydrate transport genes, DNA topological change-related genes and cellular response to DNA damage stimulus-related genes.[257] Employing GBA in such cases may offer insights into functional categories, but further analysis is necessary to infer strong and more detailed functional association between the target gene and the neighboring genes.

Another benefit of combining genomic context analysis with multiple approaches is to mitigate the false negatives and draw more accurate conclusions for function prediction. Genomic context can be integrated with other approaches that can reveal functional association evidence, such as phylogenetic co-occurrence profiles,[259] gene co-expression profiles,[260] and protein interactions.[17,260] In Chapter 3, genomic context combined with phylogenetic co-occurrence to infer the functional association between *IPT* and other cytokinin regulation-related genes. Moreover, multiple lines of evidence derived from integrated results help provide a broader functional perspective, aiding in the recognition of protein multifunctionality.[261]

Gene-centric analysis

Another major drawback of gene neighborhood analysis is that it is not able to reveal most of the genetic differences that exist within genes, such as key substitutions, insertions, and deletions. Therefore, apart from analysis that focus on gene functional association, gene neighborhood analysis is occasionally coupled with gene centric analysis. Although orthologs or proteins from the same family are generally considered to share the same functionality, a few key mutations can significantly alter protein structure and function.[262] Gene centric analysis such as protein alignment or domain analysis can provide more information about a protein's biochemical role. In Chapter 4, the protein alignment analysis of neurotoxins provides further insights on the functional differences between PG toxins and other neurotoxins by examining residues that were known to partake in important functions.

Nevertheless, certain intra-genetic variations can be detected by genomic context analysis. For example, gene deletion events leading to significant length reduction can be identified by visualizing the genomic context. As seen in Chapter 5, genomic context analysis of *tent* gene reveals a deletion event resulting in the deletion of over 90% of the *tent* sequence in two *Clostridium* sp. X strains. Another instance is that when neighboring genes are annotated by protein functional domains such as Pfam. In such cases, genomic context analysis may identify genetic variations results from domain gain and loss events. Still, it might overlook variations result from domain shuffling and segment duplication/deletion, as these alterations would not change the domain composition, but may change the protein function significantly.

Annotation heterogeneity

Annotation heterogeneity refers to the scenario where the same gene from different species receives different annotations.[263] Such inconsistency could obscure functional conservation from genomic context analysis. Annotation heterogeneity often appears in public databases such as NCBI Genbank, where different annotation methods and pipelines have been applied to different genomes and species, depending on the choices and preference of the uploader. A potential solution to this issue is to apply a consistent annotation method to all genomes. However, even in databases adopting a homogeneous annotation strategy, such annotation heterogeneity can emerge. This is typically the results of using arbitrary bioinformatics thresholds to determine whether a gene/protein sequence qualifies for a specific functional annotation or category. For example, a gene might just barely pass the threshold in one genome but almost approach the threshold in another genome and hence gets no annotation. Such discrepancies, often termed "homology detection failures", are frequently observed in research and can mistakenly classify some highly divergent orthologs as lineage-specific genes.[263–265] To improve the accuracy of gene annotation in comparative studies, incorporating remote homology detection into annotation pipelines can be beneficial, such as HHpred and HHblits, which utilize profile hidden Markov models and structure comparisons.[266] Recent data mining initiatives within the Alphafold protein structure database are set to further boost remote homology detection capabilities.[267,268] While beyond the abilities of current technologies to detect remote homologies, it is generally safe to assume that genes sharing the same genomic context are orthologs. Nevertheless, one should be cautious of occasional false negatives where non-homologous genes exhibit synteny, such as the emergence of novel genes within the synteny region, translocation of non-homologous genes therein, or gene losses followed by neighboring gene duplications.[264]

6.3 Final remarks

Gene neighborhood analysis has emerged as a powerful tool, allowing researchers to uncover functional associations and evolutionary history of genes based on the visualization and annotation of genomic loci. In general, the implications of genomic neighborhood analysis extend far beyond basic genomic research. In microbial genomics, it aids in predicting potential operon structures. In evolutionary biology, understanding the genomic context offers insights into the evolutionary trajectories of genomes, revealing conserved blocks and adaptive modifications. In synthetic biology, the identification of biosynthetic gene clusters helps to elucidate metabolic pathways and provide insights into metabolic engineering.

With advances in genomic sequencing and functional genomics, genomic context analysis is anticipated to draw more accurate and insightful results. However, integration of multiple gene-centric approaches remains essential to offer a more comprehensive perspective on the functional and evolutionary aspects of a gene/protein family. In addition, the overprediction of functional association based on gene proximity should raise awareness, and combining with other approaches that offer supporting evidence will yield a better prediction. Moreover, improved methods for annotating and categorizing gene families such as orthology inference and protein family assignment provide a holistic view of the functional conservation among diverse species. Finally, the development of tools for fast genomic context visualization and analysis needs to be adjusted and updated frequently based on different research community to better assist a broad range of researchers.

Letter of Copyright Permission

The published articles included in this thesis are made with the permission of all publishers. At the time of writing (October 2023), the policies of the publishers Elsevier and Nature do not require additional permissions for the reproduction of articles in theses. The relevant policies can be found at the following links:

1. Elsevier

   www.elsevier.com/about/policies/copyright

2. Nature

   www.nature.com/nature-portfolio/reprints-and-permissions/permissions-requests

References

[1] João C. Setubal, Jens Stoye, and Peter F. Stadler. *Comparative Genomics: Methods and Protocols*, volume 1704 of *Methods in Molecular Biology*. Springer, 2018.

[2] Xuhua Xia. *What is Comparative Genomics?* SpringerBriefs in Genetics. Springer, 2013.

[3] L. Aravind. Guilt by association: Contextual information in genome analysis. *Genome Research*, 10(8):1074–1077, 2000.

[4] Shany Doron, Sarah Melamed, Gal Ofir, Azita Leavitt, Anna Lopatina, Mai Keren, Gil Amitai, and Rotem Sorek. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, 359(6379):eaar4120, 2018.

[5] Sergey A. Shmakov, Kira S. Makarova, Yuri I. Wolf, Konstantin V. Severinov, and Eugene V. Koonin. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 115(23):E5307–E5316, 2018.

[6] Derrick E. Fouts, Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Research*, 40(22):e172, 2012.

[7] Adi Millman, Sarah Melamed, Gil Amitai, and Rotem Sorek. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nature Microbiology*, 5(12):1608–1615, 2020.

[8] Pascal Schläpfer, Peifen Zhang, Chuan Wang, Taehyong Kim, Michael Banf, Lee Chae, Kate Dreher, Arvind K. Chavali, Ricardo Nilo-Poyanco, Thomas Bernard, Daniel Kahn, and Seung Y. Rhee. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, 173(4):2041–2059, 2017.

[9] Rémi Zallot, Nils Oberg, and John A. Gerlt. The EFI web resource for genomic enzymology tools: Leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry*, 58(41):4169–4182, 2019.

[10] Zhenhua Liu, Hernando G. Suarez Duran, Yosapol Harnvanichvech, Michael J. Stephenson, M. Eric Schranz, David Nelson, Marnix H. Medema, and Anne Osbourn. Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the *Brassicaceae*. *New Phytologist*, 227(4):1109–1123, 2020.

[11] Marnix H Medema, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene De Bruijn, Yit Heng Chooi, Jan Claesen, R Cameron Coates, et al. Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology*, 11(9):625–631, 2015.

[12] Julien Guglielmini and Laurence Van Melderen. Bacterial toxin-antitoxin systems. *Mobile Genetic Elements*, 1(4):283–306, 2011.

[13] Jan O. Korbel, Lars J. Jensen, Christian von Mering, and Peer Bork. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology*, 22(7):911–917, 2004.

[14] Thomas Dandekar, Berend Snel, Martijn Huynen, and Peer Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–328, 1998.

[15] Ross Overbeek, Michael Fonstein, Mark D'Souza, Gordon D. Pusch, and Natalia Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, 96(6):2896–2901, 1999.

[16] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4285–4288, 1999.

[17] Matteo Pellegrini. Using phylogenetic profiles to predict functional relationships. *Bacterial molecular networks: methods and protocols*, pages 167–177, 2012.

[18] David Ochoa and Florencio Pazos. Practical aspects of protein co-evolution. *Frontiers in Cell and Developmental Biology*, 2, 2014.

[19] Tamir Tuller, Yifat Felder, and Martin Kupiec. Discovering local patterns of co-evolution: computational aspects and biological examples. *BMC Bioinformatics*, 11(1):43, 2010.

[20] Enrique Morett, Jan O. Korbel, Emmanuvel Rajan, Gloria Saab-Rincon, Leticia Olvera, Maricela Olvera, Steffen Schmidt, Berend Snel, and Peer Bork. Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nature Biotechnology*, 21(7):790–795, 2003.

[21] Anton J. Enright, Ioannis Iliopoulos, Nikos C. Kyrpides, and Christos A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, 1999.

[22] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[23] Itai Yanai, Adnan Derti, and Charles DeLisi. Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. *Proceedings of the National Academy of Sciences*, 98(14):7940–7945, 2001.

[24] Mohamed Samir El-Robh and Stephen J W Busby. The *Escherichia coli* cAMP receptor protein bound at a single target can activate transcription initiation at divergent promoters: a systematic study that exploits new promoter probe plasmids. *Biochemical Journal*, 368(Pt 3):835–843, 2002.

[25] Kyu Y. Rhee, Michael Opel, Elaine Ito, She-pin Hung, Stuart M. Arfin, and G. Wesley Hatfield. Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the *ilvYC* operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 96(25):14294–14299, 1999.

[26] Dina Svetlitsky, Tal Dagan, and Michal Ziv-Ukelson. Discovery of multi-operon colinear syntenic blocks in microbial genomes. *Bioinformatics*, 36(Supplement_1):i21–i29, 2020.

[27] Qingyu Chen, Justin Zobel, and Karin Verspoor. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database: The Journal of Biological Databases and Curation*, 2017:baw163, 2017.

[28] Michael Y Galperin, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V Koonin. COG database update: focus on microbial

diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, 49(D1):D274–D281, 2021.

[29] Ana Hernández-Plaza, Damian Szklarczyk, Jorge Botas, Carlos P Cantalapiedra, Joaquín Giner-Lamia, Daniel R Mende, Rebecca Kirsch, Thomas Rattei, Ivica Letunic, Lars J Jensen, Peer Bork, Christian von Mering, and Jaime Huerta-Cepas. eggNOG 6.0: enabling comparative genomics across 12535 organisms. *Nucleic Acids Research*, 51(D1):D389–D394, 2023.

[30] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021.

[31] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[32] D. H. Haft, B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research*, 29(1):41–43, 2001.

[33] David M. Kristensen, Yuri I. Wolf, Arcady R. Mushegian, and Eugene V. Koonin. Computational methods for gene orthology inference. *Briefings in Bioinformatics*, 12(5):379–391, 2011.

[34] Adrian M Altenhoff, Natasha M Glover, and Christophe Dessimoz. Inferring orthology and paralogy. *Evolutionary genomics: statistical and computational methods*, pages 149–175, 2019.

[35] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[36] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, 2021.

[37] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

[38] Julie E. Hernández-Salmerón and Gabriel Moreno-Hagelsieb. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. *BMC Genomics*, 21(1):741, 2020.

[39] Benjamin Linard, Ingo Ebersberger, Shawn E McGlynn, Natasha Glover, Tomohiro Mochizuki, Mateus Patricio, Odile Lecompte, Yannis Nevers, Paul D Thomas, Toni Gabaldón, Erik Sonnhammer, Christophe Dessimoz, Ikuo Uchiyama, and QFO Consortium. Ten years of collaborative progress in the quest for orthologs. *Molecular Biology and Evolution*, 38(8):3033–3045, 2021.

[40] David M. Emms and Steven Kelly. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238, 2019.

[41] Natasha Glover, Christophe Dessimoz, Ingo Ebersberger, Sofia K Forslund, Toni Gabaldón, Jaime Huerta-Cepas, Maria-Jesus Martin, Matthieu Muffato, Mateus Patricio, Cécile Pereira, et al. Advances and applications in the quest for orthologs. *Molecular Biology and Evolution*, 36(10):2157–2164, 2019.

[42] 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*, 25(17):3389–3402, 1997.

[43] Wenjun Li, Kathleen R O'Neill, Daniel H Haft, Michael DiCuccio, Vyacheslav Chetvernin, Azat Badretdin, George Coulouris, Farideh Chitsaz, Myra K Derbyshire, A Scott Durkin, Noreen R Gonzales, Marc Gwadz, Christopher J Lanczycki, James S Song, Narmada Thanki, Jiyao Wang, Roxanne A Yamashita, Mingzhang Yang, Chanjuan Zheng, Aron Marchler-Bauer, and Françoise Thibaud-Nissen. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Research*, 49(D1):D1020–D1028, 2020.

[44] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

[45] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, and Cathy H. Wu. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[46] Robert D. Finn, Jody Clements, and Sean R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–W37, 2011.

[47] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, 2020.

[48] Minoru Kanehisa, Yoko Sato, and Kanae Morishima. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology*, 428(4):726–731, 2016.

[49] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, 34(8):2115–2122, 2017.

[50] Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunić, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, and Alex Bateman. InterPro in 2022. *Nucleic Acids Research*, 51(D1):D418–D427, 2023.

[51] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.

[52] Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Noreen R. Gonzales, Marc Gwadz, Shennan Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Roxanne A. Yamashita, Mingzhang Yang, Dachuan Zhang, Chanjuan Zheng, Christopher J. Lanczycki, and Aron Marchler-Bauer. The conserved domain database in 2023. *Nucleic Acids Research*, 51(D1):D384–D388, 2023.

[53] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[54] João C. Sequeira, Miguel Rocha, M. Madalena Alves, and Andreia F. Salvador. UPIMAPI, reCOGnizer and KEGGCharter: Bioinformatics tools for functional annotation and visualization of (meta)-omics datasets. *Computational and Structural Biotechnology Journal*, 20:1798–1810, 2022.

[55] Sitao Wu, Zhengwei Zhu, Liming Fu, Beifang Niu, and Weizhong Li. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics*, 12(1):444, 2011.

[56] Tim J. Carver, Kim M. Rutherford, Matthew Berriman, Marie-Adele Rajandream, Barclay G. Barrell, and Julian Parkhill. ACT: the Artemis comparison tool. *Bioinformatics*, 21(16):3422–3423, 2005.

[57] Aaron C.E. Darling, Bob Mau, Frederick R. Blattner, and Nicole T. Perna. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–1403, 2004.

[58] Colin Diesh, Garrett J. Stevens, Peter Xie, Teresa De Jesus Martinez, Elliot A. Hershberg, Angel Leung, Emma Guo, Shihab Dider, Junjun Zhang, Caroline Bridge, Gregory Hogue, Andrew Duncan, Matthew Morgan, Tia Flores, Benjamin N. Bimber, Robin Haw, Scott Cain, Robert M. Buels, Lincoln D. Stein, and Ian H. Holmes. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biology*, 24(1):74, 2023.

[59] Venkat Bandi and Carl Gutwin. Interactive exploration of genomic conservation. *Proceedings of Graphics Interface*, 2020.

[60] Haibao Tang, Matthew D. Bomhoff, Evan Briones, Liangsheng Zhang, James C. Schnable, and Eric Lyons. SynFind: Compiling syntenic regions across any set of genomes on demand. *Genome Biology and Evolution*, 7(12):3286–3298, 2015.

[61] Daniel Veltri, Martha Malapi Wight, and Jo Anne Crouch. SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species. *Nucleic Acids Research*, 44(Web Server issue):W41–W45, 2016.

[62] Yupeng Wang, Haibao Tang, Jeremy D. DeBarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-ho Lee, Huizhe Jin, Barry Marler, Hui Guo, Jessica C. Kissinger, and Andrew H. Paterson. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40(7):e49, 2012.

[63] I-Min A Chen, Ken Chu, Krishnaveni Palaniappan, Anna Ratner, Jinghua Huang, Marcel Huntemann, Patrick Hajek, Stephan Ritter, Neha Varghese, Rekha Seshadri, Simon Roux, Tanja Woyke, Emiley A Eloe-Fadrosh, Natalia N Ivanova, and Nikos C Kyrpides. The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Research*, 49(D1):D751–D763, 2021.

[64] Paramvir S. Dehal, Marcin P. Joachimiak, Morgan N. Price, John T. Bates, Jason K. Baumohl, Dylan Chivian, Greg D. Friedland, Katherine H. Huang, Keith Keller,

Pavel S. Novichkov, Inna L. Dubchak, Eric J. Alm, and Adam P. Arkin. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Research*, 38(Database issue):D396–400, 2010.

[65] Marius Alfred Dieckmann, Sebastian Beyvers, Rudel Christian Nkouamedjo-Fankep, Patrick Harald Georg Hanel, Lukas Jelonek, Jochen Blom, and Alexander Goesmann. EDGAR3.0: comparative genomics and phylogenomics on a scalable infrastructure. *Nucleic Acids Research*, 49(W1):W185–W192, 2021.

[66] Eric Lyons and Michael Freeling. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, 53(4):661–673, 2008.

[67] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.

[68] David Vallenet, Alexandra Calteau, Mathieu Dubois, Paul Amours, Adelme Bazin, Mylène Beuvin, Laura Burlot, Xavier Bussell, Stéphanie Fouteau, Guillaume Gautreau, Aurélie Lajus, Jordan Langlois, Rémi Planel, David Roche, Johan Rollin, Zoe Rouy, Valentin Sabatet, and Claudine Médigue. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research*, 48(D1):D579–D589, 2020.

[69] Jorge Botas, Álvaro Rodríguez del Río, Joaquín Giner-Lamia, and Jaime Huerta-Cepas. GeCoViz: genomic context visualisation of prokaryotic genes from a functional and evolutionary perspective. *Nucleic Acids Research*, 50(W1):W352–W357, 2022.

[70] Alan M Cleary and Andrew D Farmer. Genome Context Viewer (GCV) version 2: enhanced visual exploration of multiple annotated genomes. *Nucleic Acids Research*, 51(W1):W225–W231, 2023.

[71] Pierre Simon Garcia, Frédéric Jauffrit, Christophe Grangeasse, and Céline Brochier-Armanet. GeneSpy, a user-friendly and flexible genomic context visualizer. *Bioinformatics*, 35(2):329–331, 2019.

[72] Vadim M. Gumerov and Igor B. Zhulin. TREND: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucleic Acids Research*, 48(W1):W72–W76, 2020.

[73] Katherine J Harrison, Valérie de Crécy-Lagard, and Rémi Zallot. Gene Graphics: a genomic neighborhood data visualization web application. *Bioinformatics*, 34(8):1406–1408, 2018.

[74] Xiangyang Li, Fang Chen, and Yunpeng Chen. Gcluster: a simple-to-use tool for visualizing and comparing genome contexts for numerous genomes. *Bioinformatics*, 36(12):3871–3873, 2020.

[75] Joana Pereira. GCsnap: Interactive snapshots for the comparison of protein-coding genomic contexts. *Journal of Molecular Biology*, 433(11):166943, 2021.

[76] Morgan N. Price and Adam P. Arkin. A fast comparative genome browser for diverse bacteria and archaea. *bioRxiv*, 2023.

[77] Chayan Kumar Saha, Rodrigo Sanches Pires, Harald Brolin, Maxence Delannoy, and Gemma Catherine Atkinson. FlaGs and webFlaGs: discovering novel biology through the analysis of gene neighbourhood conservation. *Bioinformatics*, 37(9):1312–1314, 2021.

[78] Jacques Oberto. SyntTax: a web server linking synteny to prokaryotic taxonomy. *BMC bioinformatics*, 14:4, 2013.

[79] Marina Marcet-Houben, Ismael Collado-Cala, Diego Fuentes-Palacios, Alicia D. Gómez, Manuel Molina, Andrés Garisoain-Zafra, Uciel Chorostecki, and Toni Gabaldón. EvolClustDB: Exploring eukaryotic gene clusters with evolutionarily conserved genomic neighbourhoods. *Journal of Molecular Biology*, 435(14):168013, 2023.

[80] Cameron L M Gilchrist and Yit-Heng Chooi. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics*, 37(16):2473–2475, 2021.

[81] Kai Blin, Marnix H. Medema, Renzo Kottmann, Sang Yup Lee, and Tilmann Weber. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research*, 45(D1):D555–D559, 2017.

[82] Marnix H. Medema, Eriko Takano, and Rainer Breitling. Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Molecular Biology and Evolution*, 30(5):1218–1223, 2013.

[83] Mark Stam, Jordan Langlois, Céline Chevalier, Guillaume Reboul, Karine Bastard, Claudine Médigue, and David Vallenet. NetSyn: genomic context exploration of protein families. *bioRxiv*, 2023.

[84] Alexander G McFarland, Nolan W Kennedy, Carolyn E Mills, Danielle Tullman-Ercek, Curtis Huttenhower, and Erica M Hartmann. Density-based binning of gene clusters to infer function or evolutionary history using GeneGrouper. *Bioinformatics*, 38(3):612–620, 2022.

[85] Donovan H. Parks, Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996–1004, 2018.

[86] Anne E. Osbourn and Ben Field. Operons. *Cellular and Molecular Life Sciences*, 66(23):3755–3775, 2009.

[87] Gabriel Moreno-Hagelsieb. The power of operon rearrangements for predicting functional associations. *Computational and Structural Biotechnology Journal*, 13:402–406, 2015.

[88] Eugene V. Koonin. Evolution of genome architecture. *The international journal of biochemistry and cell biology*, 41(2):298–306, 2009.

[89] Romilio T. Espejo and Nicolás Plaza. Multiple ribosomal RNA operons in bacteria; their concerted evolution and potential consequences on the rate of evolution of their 16S rRNA. *Frontiers in Microbiology*, 9, 2018.

[90] Syed Shujaat Ali Zaidi and Xuegong Zhang. Computational operon prediction in whole-genomes and metagenomes. *Briefings in Functional Genomics*, 16(4):181–193, 2017.

[91] Blanca Taboada, Karel Estrada, Ricardo Ciria, and Enrique Merino. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics*, 34(23):4118–4120, 2018.

[92] Xizeng Mao, Qin Ma, Chuan Zhou, Xin Chen, Hanyuan Zhang, Jincai Yang, Fenglou Mao, Wei Lai, and Ying Xu. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Research*, 42(D1):D654–D659, 2014.

[93] G. Moreno-Hagelsieb and J. Collado-Vides. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, 18(Suppl 1):S329–S336, 2002.

[94] Rutger W. W. Brouwer, Oscar P. Kuipers, and Sacha A. F. T. van Hijum. The relative value of operon predictions. *Briefings in Bioinformatics*, 9(5):367–375, 2008.

[95] Li-Yeh Chuang, Hsueh-Wei Chang, Jui-Hung Tsai, and Cheng-Hong Yang. Features for computational operon prediction in prokaryotes. *Briefings in Functional Genomics*, 11(4):291–299, 2012.

[96] Dina Svetlitsky, Tal Dagan, Vered Chalifa-Caspi, and Michal Ziv-Ukelson. CSBFinder: discovery of colinear syntenic blocks across thousands of prokaryotic genomes. *Bioinformatics*, 35(10):1634–1643, 2019.

[97] Sascha Winter, Katharina Jahn, Stefanie Wehner, Leon Kuchenbecker, Manja Marz, Jens Stoye, and Sebastian Böcker. Finding approximate gene clusters with Gecko 3. *Nucleic Acids Research*, 44(20):9600–9610, 2016.

[98] Gal Horesh, Alexander Harms, Cinzia Fino, Leopold Parts, Kenn Gerdes, Eva Heinz, and Nicholas Robert Thomson. SLING: a tool to search for linked genes in bacterial datasets. *Nucleic Acids Research*, 46(21):e128, 2018.

[99] Sanjida H. Rangwala, Anatoliy Kuznetsov, Victor Ananiev, Andrea Asztalos, Evgeny Borodin, Vladislav Evgeniev, Victor Joukov, Vadim Lotov, Ravinder Pannu, Dmitry Rudnev, Andrew Shkeda, Eric M. Weitz, and Valerie A. Schneider. Accessing NCBI data using the NCBI sequence viewer and genome data viewer (GDV). *Genome Research*, 31(1):159–169, 2021.

[100] Jairo Navarro Gonzalez, Ann S. Zweig, Matthew L. Speir, Daniel Schmelter, Kate R. Rosenbloom, Brian J. Raney, Conner C. Powell, Luis R. Nassar, Nathan D. Maulding, Christopher M. Lee, Brian T. Lee, Angie S. Hinrichs, Alastair C. Fyfe, Jason D. Fernandes, Mark Diekhans, Hiram Clawson, Jonathan Casper, Anna Benet-Pagès, Galt P. Barber, David Haussler, Robert M. Kuhn, Maximilian Haeussler, and W. James Kent. The UCSC genome browser database: 2021 update. *Nucleic Acids Research*, 49(D1):D1046–D1057, 2021.

[101] Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, et al. Ensembl 2022. *Nucleic Acids Research*, 50(D1):D988–D995, 2022.

[102] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

[103] Emile Gluck-Thaler, Aude Cerutti, Alvaro L Perez-Quintero, Jules Butchacas, Verónica Roman-Reyna, Vishnu Narayanan Madhavan, Deepak Shantharaj, Marcus V Merfa, Céline Pesce, Alain Jauneau, et al. Repeated gain and loss of a single gene modulates the evolution of vascular plant pathogen lifestyles. *Science Advances*, 6(46):eabc4516, 2020.

[104] Anthony T. Maurelli. Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiology Letters*, 267(1):1–8, 2007.

[105] You Che, Yu Yang, Xiaoqing Xu, Karel Břinda, Martin F. Polz, William P. Hanage, and Tong Zhang. Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proceedings of the National Academy of Sciences*, 118(6):e2008731118, 2021.

[106] Allison K. Guitor, Efrah I. Yousuf, Amogelang R. Raphenya, Eileen K. Hutton, Katherine M. Morrison, Andrew G. McArthur, Gerard D. Wright, and Jennifer C. Stearns. Capturing the antibiotic resistome of preterm infants reveals new benefits of probiotic supplementation. *Microbiome*, 10(1):136, 2022.

[107] Patrick Munk, Christian Brinch, Frederik Duus Møller, Thomas N. Petersen, Rene S. Hendriksen, Anne Mette Seyfarth, Jette S. Kjeldgaard, Christina Aaby Svendsen, Bram van Bunnik, Fanny Berglund, D. G. Joakim Larsson, Marion Koopmans, Mark Woolhouse, and Frank M. Aarestrup. Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nature Communications*, 13(1):7251, 2022.

[108] Mislav Acman, Ruobing Wang, Lucy van Dorp, Liam P. Shaw, Qi Wang, Nina Luhmann, Yuyao Yin, Shijun Sun, Hongbin Chen, Hui Wang, and Francois Balloux. Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene blaNDM. *Nature Communications*, 13(1):1131, 2022.

[109] Giarlã Cunha da Silva, Osiel Silva Gonçalves, Jéssica Nogueira Rosa, Kiara Campos França, Janine Thérèse Bossé, Mateus Ferreira Santana, Paul Richard Langford, and Denise Mara Soares Bazzolli. Mobile genetic elements drive antimicrobial resistance gene spread in *Pasteurellaceae* species. *Frontiers in Microbiology*, 12, 2022.

[110] Fatima Aysha Hussain, Javier Dubert, Joseph Elsherbini, Mikayla Murphy, David VanInsberghe, Philip Arevalo, Kathryn Kauffman, Bruno Kotska Rodino-Janeiro, Hannah Gavin, Annika Gomez, Anna Lopatina, Frédérique Le Roux, and Martin F. Polz. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science*, 374(6566):488–492, 2021.

[111] Markus H K Johansson, Valeria Bortolaia, Supathep Tansirichaiya, Frank M Aarestrup, Adam P Roberts, and Thomas N Petersen. Detection of mobile genetic elements associated with antibiotic resistance in *Salmonella enterica* using a newly developed web tool: MobileElementFinder. *Journal of Antimicrobial Chemotherapy*, 76(1):101–109, 2021.

[112] Marvin Djukic, Elzbieta Brzuszkiewicz, Anne Fünfhaus, Jörn Voss, Kathleen Gollnow, Lena Poppinga, Heiko Liesegang, Eva Garcia-Gonzalez, Elke Genersch, and Rolf Daniel. How to kill the honey bee larva: Genomic potential and virulence mechanisms of *Paenibacillus larvae*. *PLOS ONE*, 9(3):e90914, 2014.

[113] Nir Dover, Jason R. Barash, Karen K. Hill, Karen W. Davenport, Hazuki Teshima, Gary Xie, and Stephen S. Arnon. *Clostridium botulinum* strain Af84 contains three neurotoxin gene clusters: *Bont/A2*, *bont/F4* and *bont/F5*. *PLOS ONE*, 8(4):e61205, 2013.

[114] Holly R. Ramage, Lynn E. Connolly, and Jeffery S. Cox. Comprehensive functional analysis of *Mycobacterium tuberculosis* toxin-antitoxin systems: Implications for pathogenesis, stress responses, and evolution. *PLOS Genetics*, 5(12):e1000767, 2009.

[115] Kanae Nishii, Frank Wright, Yun-Yu Chen, and Michael Möller. Tangled history of a multigene family: The evolution of *ISOPENTENYLTRANSFERASE* genes. *PLOS ONE*, 13(8):e0201198, 2018.

[116] Bernard R. Glick. *Beneficial Plant-Bacterial Interactions*. Springer International Publishing, 2 edition, 2020.

[117] Bernard R. Glick. Plant growth-promoting bacteria: Mechanisms and applications. *Scientifica*, 2012:e963401, 2012.

[118] Nanae Ueda, Mikiko Kojima, Katsunori Suzuki, and Hitoshi Sakakibara. *Agrobacterium tumefaciens* tumor morphology root plastid localization and preferential usage of hydroxylated prenyl donor is important for efficient gall formation. *Plant Physiology*, 159(3):1064–1072, 2012.

[119] Michel R. Popoff. Tetanus in animals. *Journal of Veterinary Diagnostic Investigation*, 32(2):184–191, 2020.

[120] Michael J. Mansfield, Jeremy B. Adams, and Andrew C. Doxey. Botulinum neurotoxin homologs in non-*Clostridium* species. *FEBS Letters*, 589(3):342–348, 2015.

[121] Estefania Contreras, Geoffrey Masuyer, Nadia Qureshi, Swati Chawla, Harpal S. Dhillon, Han Lim Lee, Jianwu Chen, Pål Stenmark, and Sarjeet S. Gill. A neurotoxin that specifically targets *Anopheles* mosquitoes. *Nature Communications*, 10(1):2869, 2019.

[122] Sicai Zhang, Francois Lebreton, Michael J. Mansfield, Shin-Ichiro Miyashita, Jie Zhang, Julia A. Schwartzman, Liang Tao, Geoffrey Masuyer, Markel Martínez-Carranza, Pål Stenmark, Michael S. Gilmore, Andrew C. Doxey, and Min Dong. Identification of a botulinum neurotoxin-like toxin in a commensal strain of *Enterococcus faecium*. *Cell Host and Microbe*, 23(2):169–176.e6, 2018.

[123] Jason Brunt, Andrew T. Carter, Sandra C. Stringer, and Michael W. Peck. Identification of a novel botulinum neurotoxin gene cluster in *Enterococcus*. *FEBS Letters*, 592(3):310–317, 2018.

[124] Kerrin Mendler, Han Chen, Donovan H Parks, Briallen Lobb, Laura A Hug, and Andrew C Doxey. AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Research*, 47(9):4442–4448, 2019.

[125] Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 2010.

[126] Jonathan Kans. *Entrez Direct: E-utilities on the Unix Command Line*. National Center for Biotechnology Information (US), 2020.

[127] U. Ermler, W. Grabarse, S. Shima, M. Goubeaud, and R. K. Thauer. Crystal structure of methyl-coenzyme M reductase: the key enzyme of biological methane formation. *Science*, 278(5342):1457–1462, 1997.

[128] Aleksei Gendron and Kylie D. Allen. Overview of diverse methyl/alkyl-coenzyme M reductases and considerations for their potential heterologous expression. *Frontiers in Microbiology*, 13, 2022.

[129] Sura Ali, Benjamin Jenkins, Jiujun Cheng, Briallen Lobb, Xin Wei, Suhelen Egan, Trevor C. Charles, Brendan J. McConkey, John Austin, and Andrew C. Doxey. Slr4, a newly identified S-layer protein from marine *Gammaproteobacteria*, is a major biofilm matrix component. *Molecular Microbiology*, 114(6):979–990, 2020.

[130] Hein J. Boot and Peter H. Pouwels. Expression, secretion and antigenic variation of bacterial S-layer proteins. *Molecular Microbiology*, 21(6):1117–1123, 1996.

[131] J. M. Tomás. The main *Aeromonas* pathogenic factors. *International Scholarly Research Notices*, 2012:e256261, 2012.

[132] Jianyun Yao, Xiangkai Zhen, Kaihao Tang, Tianlang Liu, Xiaolong Xu, Zhe Chen, Yunxue Guo, Xiaoxiao Liu, Thomas K Wood, Songying Ouyang, and Xiaoxue Wang. Novel polyadenylylation-dependent neutralization mechanism of the HEPN/MNT toxin/antitoxin system. *Nucleic Acids Research*, 48(19):11054–11067, 2020.

[133] Nathan Fraikin, Frédéric Goormaghtigh, and Laurence Van Melderen. Type II toxin-antitoxin systems: Evolution and revolutions. *Journal of Bacteriology*, 202(7):e00763–19, 2020.

[134] Raoul J. de Groot. Structure, function and evolution of the hemagglutinin-esterase proteins of corona- and toroviruses. *Glycoconjugate Journal*, 23(1):59–72, 2006.

[135] Qinghong Zeng, Martijn A. Langereis, Arno L. W. van Vliet, Eric G. Huizinga, and Raoul J. de Groot. Structure of coronavirus hemagglutinin-esterase offers insight into corona and influenza virus evolution. *Proceedings of the National Academy of Sciences*, 105(26):9065–9069, 2008.

[136] Patrick C. Y. Woo, Yi Huang, Susanna K. P. Lau, and Kwok-Yung Yuen. Coronavirus genomics and bioinformatics analysis. *Viruses*, 2(8):1804–1820, 2010.

[137] Susanna K. P. Lau, Patrick C. Y. Woo, Cyril C. Y. Yip, Rachel Y. Y. Fan, Yi Huang, Ming Wang, Rongtong Guo, Carol S. F. Lam, Alan K. L. Tsang, Kenneth K. Y. Lai, Kwok-Hung Chan, Xiao-Yan Che, Bo-Jian Zheng, and Kwok-Yung Yuen. Isolation and characterization of a novel *Betacoronavirus* subgroup A coronavirus, rabbit coronavirus HKU14, from domestic rabbits. *Journal of Virology*, 86(10):5481–5496, 2012.

[138] Xin Wei, Gabriel Moreno-Hagelsieb, Bernard R. Glick, and Andrew C. Doxey. Comparative analysis of adenylate isopentenyl transferase genes in plant growth-promoting bacteria and plant pathogenic bacteria. *Heliyon*, 9(3):e13955, 2023.

[139] Radoslav Koprna, Nuria De Diego, Lucie Dundálková, and Lukáš Spíchal. Use of cytokinins as agrochemicals. *Bioorganic and Medicinal Chemistry*, 24(3):484–492, 2016.

[140] Joseph J. Kieber and G. Eric Schaller. Cytokinins. *The Arabidopsis Book / American Society of Plant Biologists*, 12:e0168, 2014.

[141] Prajna Priyadarshini Das, Kshitij RB Singh, Gunjan Nagpure, Aadil Mansoori, Ravindra Pratap Singh, Irfan Ahmad Ghazi, Anirudh Kumar, and Jay Singh. Plant-soil-microbes: A tripartite interaction for nutrient acquisition and better plant growth for sustainable agricultural practices. *Environmental Research*, 214:113821, 2022.

[142] John Mansfield, Stephane Genin, Shimpei Magori, Vitaly Citovsky, Malinee Sriariyanum, Pamela Ronald, Max Dow, Valérie Verdier, Steven V. Beer, Marcos A. Machado, Ian Toth, George Salmond, and Gary D. Foster. Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology*, 13(6):614–629, 2012.

[143] Ivo Frébort, Marta Kowalska, Tomáš Hluska, Jitka Frébortová, and Petr Galuszka. Evolution of cytokinin biosynthesis and degradation. *Journal of Experimental Botany*, 62(8):2431–2452, 2011.

[144] Kaori Miyawaki, Petr Tarkowski, Miho Matsumoto-Kitano, Tomohiko Kato, Shusei Sato, Danuse Tarkowska, Satoshi Tabata, Göran Sandberg, and Tatsuo Kakimoto. Roles of *Arabidopsis* ATP/ADP isopentenyltransferases and tRNA isopentenyltransferases in cytokinin biosynthesis. *Proceedings of the National Academy of Sciences*, 103(44):16598–16603, 2006.

[145] Daniel Palberg, Anna Kisiała, Gabriel Lemes Jorge, and R. J. Neil Emery. A survey of *Methylobacterium* species and strains reveals widespread production and varying profiles of cytokinin phytohormones. *BMC Microbiology*, 22(1):49, 2022.

[146] David Giron and Gaëlle Glevarec. Cytokinin-induced phenotypes in plant-insect interactions: Learning from the bacterial world. *Journal of Chemical Ecology*, 40(7):826–835, 2014.

[147] Ine Pertry, Kateřina Václavíková, Stephen Depuydt, Petr Galuszka, Lukáš Spíchal, Wim Temmerman, Elisabeth Stes, Thomas Schmülling, Tatsuo Kakimoto, Marc C. E. Van Montagu, Miroslav Strnad, Marcelle Holsters, Petr Tarkowski, and Danny Vereecke. Identification of *Rhodococcus fascians* cytokinins and their modus operandi

to reshape the plant. *Proceedings of the National Academy of Sciences*, 106(3):929–934, 2009.

[148] Hau-Hsuan Hwang, Ming-Hsuan Wang, Ying-Ling Lee, Yun-Long Tsai, Yi-Ho Li, Fong-Jhih Yang, Yu-Chen Liao, Shao-Kai Lin, and Erh-Min Lai. *Agrobacterium*-produced and exogenous cytokinin-modulated *Agrobacterium*-mediated plant transformation. *Molecular Plant Pathology*, 11(5):677–690, 2010.

[149] Muhammad Naseem, Mirko Wölfling, and Thomas Dandekar. Cytokinins for immunity beyond growth, galls and green islands. *Trends in Plant Science*, 19(8):481–484, 2014.

[150] Saqib Saleem Akhtar, Mengistu F. Mekureyaw, Chandana Pandey, and Thomas Roitsch. Role of cytokinins for interactions of plants with microbial pathogens and pest insects. *Frontiers in Plant Science*, 10:1777, 2020.

[151] Fangchun Liu, Shangjun Xing, Hailin Ma, Zhenyu Du, and Bingyao Ma. Cytokinin-producing, plant growth-promoting rhizobacteria that confer resistance to drought stress in *Platycladus* orientalis container seedlings. *Applied Microbiology and Biotechnology*, 97(20):9155–9164, 2013.

[152] Ji Xu, Xiao-Lin Li, and Li Luo. Effects of engineered *Sinorhizobium meliloti* on cytokinin synthesis and tolerance of alfalfa to extreme drought stress. *Applied and Environmental Microbiology*, 78(22):8056–8061, 2012.

[153] Jitka Frébortová, Marta Greplová, Michael F. Seidl, Alexander Heyl, and Ivo Frébort. Biochemical characterization of putative adenylate dimethylallyltransferase and cytokinin dehydrogenase from *Nostoc* sp. PCC 7120. *PLOS ONE*, 10(9):e0138468, 2015.

[154] Takashi Kurakawa, Nanae Ueda, Masahiko Maekawa, Kaoru Kobayashi, Mikiko Kojima, Yasuo Nagato, Hitoshi Sakakibara, and Junko Kyozuka. Direct control of shoot meristem activity by a cytokinin-activating enzyme. *Nature*, 445(7128):652–655, 2007.

[155] G. A. Romanov. How do cytokinins affect the cell? *Russian Journal of Plant Physiology*, 56(2):268–290, 2009.

[156] Jitka Frébortová and Ivo Frébort. Biochemical and structural aspects of cytokinin biosynthesis and degradation in bacteria. *Microorganisms*, 9(6):1314, 2021.

[157] R. de Pedro-Jové, M. Puigvert, P. Sebastià, A. P. Macho, J. S. Monteiro, N. S. Coll, J. C. Setúbal, and M. Valls. Dynamic expression of *Ralstonia solanacearum* virulence factors and metabolism-controlling genes during plant infection. *BMC Genomics*, 22(1):170, 2021.

[158] Fábio Madeira, Young mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn, and Rodrigo Lopez. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, 47(W1):W636–W641, 2019.

[159] Jaina Mistry, Alex Bateman, and Robert D. Finn. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics*, 8(1):298, 2007.

[160] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[161] Anders Larsson. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278, 2014.

[162] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, 2010.

[163] S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[164] Guangchuang Yu, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2017.

[165] Nestor Zaburannyi, Boyke Bunk, Josef Maier, Jörg Overmann, and Rolf Müller. Genome analysis of the fruiting body-forming Myxobacterium *Chondromyces crocatus* reveals high potential for natural product biosynthesis. *Applied and Environmental Microbiology*, 82(6):1945–1957, 2016.

[166] Elizabeth A Savory, Skylar L Fuller, Alexandra J Weisberg, William J Thomas, Michael I Gordon, Danielle M Stevens, Allison L Creason, Michael S Belcher, Maryna Serdani, Michele S Wiseman, Niklaus J Grünwald, Melodie L Putnam, and Jeff H Chang. Evolutionary transitions between beneficial and phytopathogenic *Rhodococcus* challenge disease management. *eLife*, 6:e30925, 2017.

[167] Weizhong Li, Lukasz Jaroszewski, and Adam Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, 2001.

[168] Alboukadel Kassambara. ggpubr: 'ggplot2' based publication ready plots, 2022.

[169] Charles E. Grant and Timothy L. Bailey. XSTREME: Comprehensive motif analysis of biological sequence datasets. *BioRxiv*, page 2021.09.02.458722, 2021.

[170] Robert C. McLeay and Timothy L. Bailey. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, 11(1):165, 2010.

[171] Yi Zhang, Chil-Woo Lee, Nora Wehner, Fabian Imdahl, Veselova Svetlana, Christoph Weiste, Wolfgang Dröge-Laser, and Rosalia Deeken. Regulation of oncogene expression in T-DNA-transformed host plant cells. *PLOS Pathogens*, 11(1):e1004620, 2015.

[172] Frank Wright. The 'effective number of codons' used in a gene. *Gene*, 87(1):23–29, 1990.

[173] Pere Puigbò, Ignacio G. Bravo, and Santiago Garcia-Vallve. CAIcal: A combined set of tools to assess codon usage adaptation. *Biology Direct*, 3(1):38, 2008.

[174] P. M. Sharp and W. H. Li. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3):1281, 1987.

[175] A Murat Eren, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E Miller, Matthew S Schechter, Isaac Fink, Jessica N Pan, Mahmoud Yousef, Emily C Fogarty, et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology*, 6(1):3–6, 2021.

[176] Benjamin D. Lee. Python implementation of codon adaptation index. *Journal of Open Source Software*, 3(30):905, 2018.

[177] Muhammad Naseem, Elena Bencurova, and Thomas Dandekar. The cytokinin-activating LOG-family proteins are not lysine decarboxylases. *Trends in Biochemical Sciences*, 43(4):232–236, 2018.

[178] Briallen Lobb, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, and Andrew C. Doxey. PathFams: statistical detection of pathogen-associated protein domains. *BMC Genomics*, 22(1):663, 2021.

[179] G K Powell and R O Morris. Nucleotide sequence and expression of a *Pseudomonas savastanoi* cytokinin biosynthetic gene: homology with *Agrobacterium tumefaciens tmr* and *tzs* loci. *Nucleic Acids Research*, 14(6):2555–2565, 1986.

[180] Maddalena Dilucca, Sergio Forcelloni, Alexandros G. Georgakilas, Andrea Giansanti, and Athanasia Pavlopoulou. Codon usage and phenotypic divergences of SARS-CoV-2 genes. *Viruses*, 12(5):498, 2020.

[181] Marc del Grande and Gabriel Moreno-Hagelsieb. The loose evolutionary relationships between transcription factors and other gene products across prokaryotes. *BMC Research Notes*, 7:928, 2014.

[182] J J Estruch, D Chriqui, K Grossmann, J Schell, and A Spena. The plant oncogene rolC is responsible for the release of cytokinins from glucoside conjugates. *The EMBO Journal*, 10(10):2889–2895, 1991.

[183] Cindy M. Quezada, Stuart W. Hicks, Jorge E. Galán, and C. Erec Stebbins. A family of *Salmonella* virulence factors functions as a distinct class of autoregulated E3 ubiquitin ligases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(12):4864–4869, 2009.

[184] A. Kovacs-Simon, R. W. Titball, and S. L. Michell. Lipoproteins of bacterial pathogens. *Infection and Immunity*, 79(2):548–561, 2011.

[185] Xiaojing Wang, Shanshan Lin, Decai Liu, Lijun Gan, Richard McAvoy, Jing Ding, and Yi Li. Evolution and roles of cytokinin genes in angiosperms 1: Do ancient IPTs play housekeeping while non-ancient IPTs play regulatory roles? *Horticulture Research*, 7(1):1–15, 2020.

[186] Wolfgang Heinemeyer, Ilka Buchmann, Dave W. Tonge, John D. Windass, Juliane Alt-Moerbe, Elmar W. Weiler, Thomas Botz, and Joachim Schröder. Two *Agrobacterium tumefaciens* genes for cytokinin biosynthesis: Ti plasmid-coded isopentenyltransferases adapted for function in prokaryotic or eukaryotic cells. *Molecular and General Genetics MGG*, 210(1):156–164, 1987.

[187] Zhao-Fen Han, David M. Hunter, Susan Sibbald, Ji-Shu Zhang, and Lining Tian. Biological activity of the *tzs* gene of nopaline *Agrobacterium tumefaciens* GV3101 in plant regeneration and genetic transformation. *Molecular Plant-Microbe Interactions*, 26(11):1359–1365, 2013.

[188] Rosa M. Rivero, Mikiko Kojima, Amira Gepstein, Hitoshi Sakakibara, Ron Mittler, Shimon Gepstein, and Eduardo Blumwald. Delayed leaf senescence induces extreme drought tolerance in a flowering plant. *Proceedings of the National Academy of Sciences*, 104(49):19631–19636, 2007.

[189] Muhammad Naseem, Edita Sarukhanyan, and Thomas Dandekar. LONELY-GUY knocks every door: Crosskingdom microbial pathogenesis. *Trends in Plant Science*, 20(12):781–783, 2015.

[190] Lenka Dzurová, Federico Forneris, Simone Savino, Petr Galuszka, Josef Vrabka, and Ivo Frébort. The three-dimensional structure of "Lonely Guy" from *Claviceps purpurea* provides insights into the phosphoribohydrolase function of Rossmann fold-containing lysine decarboxylase-like proteins. *Proteins: Structure, Function, and Bioinformatics*, 83(8):1539–1546, 2015.

[191] Hogyun Seo and Kyung-Jin Kim. Structural basis for a novel type of cytokinin-activating protein. *Scientific Reports*, 7(1):45985, 2017.

[192] David Kopečný, Radka Končitíková, Hana Popelka, Pierre Briozzo, Armelle Vigouroux, Martina Kopečná, David Zalabák, Marek Šebela, Jana Skopalová, Ivo Frébort, and Solange Moréra. Kinetic and structural investigation of the cytokinin oxidase/dehydrogenase active site. *The FEBS Journal*, 283(2):361–377, 2016.

[193] Ine Pertry, Katerina Václavíková, Markéta Gemrotová, Lukás Spíchal, Petr Galuszka, Stephen Depuydt, Wim Temmerman, Elisabeth Stes, Annick De Keyser, Michael Riefler, Stefania Biondi, Ondrej Novák, Thomas Schmülling, Miroslav Strnad, Petr Tarkowski, Marcelle Holsters, and Danny Vereecke. *Rhodococcus fascians* impacts plant development through the dynamic fas-mediated production of a cytokinin mix. *Molecular plant-microbe interactions: MPMI*, 23(9):1164–1174, 2010.

[194] Xin Wei, Travis Wentz, Briallen Lobb, Michael Mansfield, William Zhen, Huagang Tan, Zijing Wu, Sabine Pellett, Min Dong, and Andrew C. Doxey. Identification of divergent botulinum neurotoxin homologs in *Paeniclostridium ghonii*. *bioRxiv*, 2022.

[195] Xin Wei, Briallen Lobb, Kang Wang, Min Dong, and Andrew C. Doxey. Identification of a botulinum neurotoxin-like gene cluster in *Bacillus toyonensis*. *bioRxiv*, 2023.

[196] Cesare Montecucco and Glampietro Schiavo. Mechanism of action of tetanus and botulinum neurotoxins. *Molecular Microbiology*, 13(1):1–8, 1994.

[197] F. Clementi, R. Fesce, J. Meldolesi, F. Valtorta, Rossella Pellizzari, Ornella Rossetto, Giampietro Schiavo, and Cesare Montecucco. Tetanus and botulinum neurotoxins: mechanism of action and therapeutic uses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 354(1381):259–268, 1999.

[198] Marco Pirazzini, Ornella Rossetto, Roberto Eleopra, and Cesare Montecucco. Botulinum neurotoxins: Biology, pharmacology, and toxicology. *Pharmacological Reviews*, 69(2):200–235, 2017.

[199] Karen K Hill and Theresa J Smith. Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes. *Botulinum neurotoxins*, pages 1–20, 2013.

[200] Michael W. Peck, Theresa J. Smith, Fabrizio Anniballi, John W. Austin, Luca Bano, Marite Bradshaw, Paula Cuervo, Luisa W. Cheng, Yagmur Derman, Brigitte G. Dorner, Audrey Fisher, Karen K. Hill, Suzanne R. Kalb, Hannu Korkeala, Miia Lindström, Florigio Lista, Carolina Lúquez, Christelle Mazuet, Marco Pirazzini, Michel R. Popoff, Ornella Rossetto, Andreas Rummel, Dorothea Sesardic, Bal Ram Singh, and Sandra C. Stringer. Historical perspectives and guidelines for botulinum neurotoxin subtype nomenclature. *Toxins*, 9(1):38, 2017.

[201] Travis G. Wentz, Benjamin J. M. Tremblay, Marite Bradshaw, Andrew C. Doxey, Shashi K. Sharma, John-Demian Sauer, and Sabine Pellett. Endogenous CRISPR-Cas systems in group I *Clostridium botulinum* and *Clostridium sporogenes* do not directly target the botulinum neurotoxin gene cluster. *Frontiers in Microbiology*, 12, 2022.

[202] Jason Brunt, Arnoud H. M. van Vliet, Andrew T. Carter, Sandra C. Stringer, Corinne Amar, Kathie A. Grant, Gauri Godbole, and Michael W. Peck. Diversity of the genomes and neurotoxins of strains of *Clostridium botulinum* group I and *Clostridium sporogenes* associated with foodborne, infant and wound botulism. *Toxins*, 12(9):586, 2020.

[203] Domenico Azarnia Tehran and Marco Pirazzini. Novel botulinum neurotoxins: Exploring underneath the iceberg tip. *Toxins*, 10(5):190, 2018.

[204] Michael J Mansfield and Andrew C Doxey. Genomic insights into the evolution and ecology of botulinum neurotoxins. *Pathogens and Disease*, 76(4):fty040, 2018.

[205] Andrew C. Doxey, Michael J. Mansfield, and Briallen Lobb. Exploring the evolution of virulence factors through bioinformatic data mining. *mSystems*, 4(3):10.1128/msystems.00162–19, 2019.

[206] Andrew C. Doxey, Michael J. Mansfield, and Cesare Montecucco. Discovery of novel bacterial toxins by genomics and computational biology. *Toxicon*, 147:2–12, 2018.

[207] Irene Zornetta, Domenico Azarnia Tehran, Giorgio Arrigoni, Fabrizio Anniballi, Luca Bano, Oneda Leka, Giuseppe Zanotti, Thomas Binz, and Cesare Montecucco. The first non Clostridial botulinum-like toxin cleaves VAMP within the juxtamembrane domain. *Scientific Reports*, 6(1):30257, 2016.

[208] Sicai Zhang, Geoffrey Masuyer, Jie Zhang, Yi Shen, Daniel Lundin, Linda Henriksson, Shin-Ichiro Miyashita, Markel Martínez-Carranza, Min Dong, and Pål Stenmark. Identification and characterization of a novel botulinum neurotoxin. *Nature Communications*, 8(1):14130, 2017.

[209] Michael James Mansfield, Travis Gwynn Wentz, Sicai Zhang, Elliot Jeon Lee, Min Dong, Shashi Kant Sharma, and Andrew Charles Doxey. Bioinformatic discovery of a toxin family in *Chryseobacterium piperi* with sequence similarity to botulinum neurotoxins. *Scientific Reports*, 9(1):1634, 2019.

[210] Liliana Pardo-López, Mario Soberón, and Alejandra Bravo. *Bacillus thuringiensis* insecticidal three-domain Cry toxins: mode of action, insect resistance and consequences for crop protection. *FEMS Microbiology Reviews*, 37(1):3–22, 2013.

[211] Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[212] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2015.

[213] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303, 2018.

[214] Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):5114, 2018.

[215] Klaus Peter Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

[216] Todd J. Treangen, Brian D. Ondov, Sergey Koren, and Adam M. Phillippy. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(11):524, 2014.

[217] Bo Liu, Dandan Zheng, Qi Jin, Lihong Chen, and Jian Yang. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Research*, 47(D1):D687–D692, 2019.

[218] Sarah H Merkling and Ronald P Van Rij. Analysis of resistance and tolerance to virus infection in *Drosophila*. *Nature Protocols*, 10(7):1084–1097, 2015.

[219] Linda van der Graaf-van Bloois, Jaap A. Wagenaar, and Aldert L. Zomer. RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. *Microbial Genomics*, 7(11):000683, 2021.

[220] Revathi Govind and Bruno Dupuy. Secretion of *Clostridium difficile* toxins A and B requires the holin-like protein TcdE. *PLOS Pathogens*, 8(6):e1002727, 2012.

[221] D. Borden Lacy, William Tepp, Alona C. Cohen, Bibhuti R. DasGupta, and Raymond C. Stevens. Crystal structure of botulinum neurotoxin type A and implications for toxicity. *Nature Structural Biology*, 5(10):898–902, 1998.

[222] Subramanyam Swaminathan and Subramaniam Eswaramoorthy. Structural analysis of the catalytic and binding sites of *Clostridium botulinum* neurotoxin B. *Nature Structural Biology*, 7(8):693–699, 2000.

[223] Michael J. Mansfield, Seiji N. Sugiman-Marangos, Roman A. Melnyk, and Andrew C. Doxey. Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus. *FEBS Letters*, 592(16):2693–2705, 2018.

[224] Kathleen E. Orrell, Michael J. Mansfield, Andrew C. Doxey, and Roman A. Melnyk. The *C. difficile* toxin B membrane translocation machinery is an evolutionarily conserved protein delivery apparatus. *Nature Communications*, 11(1):432, 2020.

[225] Paul Emsley, Constantina Fotinou, Isobel Black, Neil F. Fairweather, Ian G. Charles, Colin Watts, Eric Hewitt, and Neil W. Isaacs. The structures of the HC fragment of tetanus toxin with carbohydrate subunit complexes provide insight into ganglioside binding. *Journal of Biological Chemistry*, 275(12):8889–8894, 2000.

[226] T. S. Sasi Jyothsna, L. Tushar, Ch. Sasikala, and Ch. V. Ramana. *Paraclostridium benzoelyticum* gen. nov., sp. nov., isolated from marine sediment and reclassification of *Clostridium bifermentans* as *Paraclostridium bifermentans* comb. nov. Proposal of a new genus *Paeniclostridium* gen. nov. to accommodate *Clostridium sordellii* and *Clostridium ghonii*. *International Journal of Systematic and Evolutionary Microbiology*, 66(3):1268–1274, 2016.

[227] Mingli Zheng, Dongze Niu, Sasa Zuo, Peichun Mao, Lin Meng, and Chuncheng Xu. The effect of cultivar, wilting and storage period on fermentation and the clostridial community of alfalfa silage. *Italian Journal of Animal Science*, 17(2):336–346, 2018.

[228] Raul J. Cano, Friedrich Tiefenbrunner, Massimo Ubaldi, Clarissa Del Cueto, Stefania Luciani, Tobe Cox, Paula Orkand, Karl H. Künzel, and Franco Rollo. Sequence analysis of bacterial DNA in the colon and stomach of the Tyrolean Iceman. *American Journal of Physical Anthropology*, 112(3):297–309, 2000.

[229] Callum J. Vidor, Dieter Bulach, Milena Awad, and Dena Lyras. *Paeniclostridium sordellii* and *Clostridioides difficile* encode similar and clinically relevant tetracycline resistance loci in diverse genomic locations. *BMC Microbiology*, 19(1):53, 2019.

[230] Marco Pirazzini, Domenico Azarnia Tehran, Giulia Zanetti, Aram Megighian, Michele Scorzeto, Silvia Fillo, Clifford C. Shone, Thomas Binz, Ornella Rossetto, Florigio Lista, and Cesare Montecucco. Thioredoxin and its reductase are present on synaptic vesicles, and their inhibition prevents the paralysis induced by botulinum neurotoxins. *Cell Reports*, 8(6):1870–1878, 2014.

[231] Pål Stenmark, Min Dong, Jérôme Dupuy, Edwin R. Chapman, and Raymond C. Stevens. Crystal structure of the botulinum neurotoxin type G binding domain: Insight into cell surface binding. *Journal of Molecular Biology*, 397(5):1287–1297, 2010.

[232] Alexandre Souvorov, Richa Agarwala, and David J. Lipman. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biology*, 19(1):153, 2018.

[233] Amir Abdulmawjood, Jens Herrmann, Susanne Riede, Guillermo Jimenez, Andre Becker, and Gerhard Breves. Evaluation of enterotoxin gene expression and enterotoxin production capacity of the probiotic strain *Bacillus toyonensis* BCT-7112T. *PLOS ONE*, 14(4):e0214536, 2019.

[234] Sinda Fedhila, Nadine Daou, Didier Lereclus, and Christina Nielsen-LeRoux. Identification of *Bacillus cereus* internalin and other candidate virulence genes specifically induced during oral infection in insects. *Molecular Microbiology*, 62(2):339–355, 2006.

[235] Nadine Daou, Christophe Buisson, Michel Gohar, Jasmina Vidic, Hélène Bierne, Mireille Kallassy, Didier Lereclus, and Christina Nielsen-LeRoux. IlsA, a unique surface protein of *Bacillus cereus* required for iron acquisition from heme, hemoglobin and ferritin. *PLOS Pathogens*, 5(11):e1000675, 2009.

[236] Yucheng Zhang, Min Gao, Stacy D. Singer, Zhangjun Fei, Hua Wang, and Xiping Wang. Genome-wide identification and analysis of the TIFY gene family in grape. *PLOS ONE*, 7(9):e44465, 2012.

[237] Harold P. Hodgins, Pengsheng Chen, Briallen Lobb, Xin Wei, Benjamin J. M. Tremblay, Michael J. Mansfield, Victoria C. Y. Lee, Pyung-Gang Lee, Jeffrey Coffin, Ana T. Duggan, Alexis E. Dolphin, Gabriel Renaud, Min Dong, and Andrew C. Doxey. Ancient *Clostridium DNA* and variants of tetanus neurotoxins associated with human archaeological remains. *Nature Communications*, 14(1):5475, 2023.

[238] Harold Hodgins. Petabase-scale data mining identifies novel clostridial species and neurotoxins associated with ancient human dna. Master's thesis, UWSpace, 2022.

[239] Aram Megighian, Marco Pirazzini, Federico Fabris, Ornella Rossetto, and Cesare Montecucco. Tetanus and tetanus neurotoxin: From peripheral uptake to central nervous tissue targets. *Journal of Neurochemistry*, 158(6):1244–1253, 2021.

[240] Steven H. Miles. Hippocrates and informed consent. *Lancet*, 374(9698):1322–1323, 2009.

[241] Georgios Pappas, Ismene J. Kiriaze, and Matthew E. Falagas. Insights into infectious disease in the era of Hippocrates. *International journal of infectious diseases*, 12(4):347–350, 2008.

[242] Gonzalo Moreno Sanchez and Alwyn Louise Burridge. Decision making in head injury management in the Edwin Smith Papyrus. *Neurosurgical Focus*, 23(1):E5, 2007.

[243] S. Kitasato. Ueber den Tetanusbacillus. *Zeitschrift für Hygiene*, 7(1):225–234, 1889.

[244] Holger Brüggemann, Sebastian Bäumer, Wolfgang Florian Fricke, Arnim Wiezer, Heiko Liesegang, Iwona Decker, Christina Herzberg, Rosa Martínez-Arias, Rainer Merkl, Anke Henne, and Gerhard Gottschalk. The genome sequence of *Clostridium*

tetani, the causative agent of tetanus disease. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):1316–1321, 2003.

[245] Holger Brüggemann, Elzbieta Brzuszkiewicz, Diana Chapeton-Montes, Lucile Plourde, Denis Speck, and Michel R. Popoff. Genomics of *Clostridium tetani*. *Research in Microbiology*, 166(4):326–331, 2015.

[246] Diana Chapeton-Montes, Lucile Plourde, Christiane Bouchier, Laurence Ma, Laure Diancourt, Alexis Criscuolo, Michel Robert Popoff, and Holger Brüggemann. The population structure of *Clostridium tetani* deduced from its pan-genome. *Scientific Reports*, 9(1):11220, 2019.

[247] Jonathan E. Cohen, Rong Wang, Rong-Fong Shen, Wells W. Wu, and James E. Keller. Comparative pathogenomics of *Clostridium tetani*. *PLOS ONE*, 12(8):e0182909, 2017.

[248] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[249] Eugene Koonin and Michael Y. Galperin. *Sequence — Evolution — Function: Computational Approaches in Comparative Genomics*. Springer Science and Business Media, 2002.

[250] Madhumita V. Joshi and Rosemary Loria. *Streptomyces turgidiscabies* possesses a functional cytokinin biosynthetic pathway and produces leafy galls. *Molecular Plant-Microbe Interactions*, 20(7):751–758, 2007.

[251] Geoffrey L. Winsor, Emma J. Griffiths, Raymond Lo, Bhavjinder K. Dhillon, Julie A. Shay, and Fiona S. L. Brinkman. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the Pseudomonas genome database. *Nucleic Acids Research*, 44(D1):D646–653, 2016.

[252] Daniel R Mende, Ivica Letunic, Oleksandr M Maistrenko, Thomas S B Schmidt, Alessio Milanese, Lucas Paoli, Ana Hernández-Plaza, Askarbek N Orakov, Sofia K Forslund, Shinichi Sunagawa, Georg Zeller, Jaime Huerta-Cepas, Luis Pedro Coelho, and Peer Bork. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Research*, 48(D1):D621–D625, 2020.

[253] Briallen Lobb, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, and Andrew C. Doxey. An assessment of genome annotation coverage across the bacterial tree of life. *Microbial Genomics*, 6(3):e000341, 2020.

[254] Pedro Escudeiro, Christopher S. Henry, and Ricardo P. M. Dias. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Current Research in Microbial Sciences*, 3:100159, 2022.

[255] A. Hoarfrost, A. Aptekmann, G. Farfañuk, and Y. Bromberg. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nature Communications*, 13(1):2606, 2022.

[256] Yuguo Zha, Hui Chong, Pengshuo Yang, and Kang Ning. Microbial dark matter: from discovery to applications. *Genomics, Proteomics and Bioinformatics*, 20(5):867–881, 2022.

[257] Matej Mihelčić, Tomislav Šmuc, and Fran Supek. Patterns of diverse gene functions in genomic neighborhoods predict gene function and phenotype. *Scientific Reports*, 9(1):19537, 2019.

[258] Igor B. Rogozin, Kira S. Makarova, Janos Murvai, Eva Czabarka, Yuri I. Wolf, Roman L. Tatusov, Laszlo A. Szekely, and Eugene V. Koonin. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Research*, 30(10):2212–2223, 2002.

[259] Benjamin J.-M. Tremblay, Briallen Lobb, and Andrew C. Doxey. PhyloCorrelate: inferring bacterial gene-gene functional associations through large-scale phylogenetic profiling. *Bioinformatics*, 37(1):17–22, 2021.

[260] Peter Kharchenko, Lifeng Chen, Yoav Freund, Dennis Vitkup, and George M. Church. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, 7(1):177, 2006.

[261] Adriana Espinosa-Cantú, Erika Cruz-Bonilla, Lianet Noda-Garcia, and Alexander DeLuna. Multiple forms of multifunctional proteins in health and disease. *Frontiers in Cell and Developmental Biology*, 8, 2020.

[262] Mohammad A. Siddiq, Georg Ka Hochberg, and Joseph W. Thornton. Evolution of protein specificity: insights from ancestral protein reconstruction. *Current Opinion in Structural Biology*, 47:113–122, 2017.

[263] Caroline M. Weisman, Andrew W. Murray, and Sean R. Eddy. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current biology : CB*, 32(12):2632–2639.e2, 2022.

[264] Nikolaos Vakirlis, Anne-Ruxandra Carvunis, and Aoife McLysaght. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife*, 9:e53500, 2020.

[265] Caroline M. Weisman, Andrew W. Murray, and Sean R. Eddy. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biology*, 18(11):e3000862, 2020.

[266] Felix Gabler, Seung-Zin Nam, Sebastian Till, Milot Mirdita, Martin Steinegger, Johannes Söding, Andrei N. Lupas, and Vikram Alva. Protein sequence analysis using the MPI bioinformatics toolkit. *Current Protocols in Bioinformatics*, 72(1):e108, 2020.

[267] Janani Durairaj, Andrew M. Waterhouse, Toomas Mets, Tetiana Brodiazhenko, Minhal Abdullah, Gabriel Studer, Gerardo Tauriello, Mehmet Akdel, Antonina Andreeva, Alex Bateman, Tanel Tenson, Vasili Hauryliuk, Torsten Schwede, and Joana Pereira. Uncovering new families and folds in the natural protein universe. *Nature*, pages 1–3, 2023.

[268] Liisa Holm, Aleksi Laiho, Petri Törönen, and Marco Salgado. DALI shines a light on remote homologs: One hundred discoveries. *Protein Science: A Publication of the Protein Society*, 32(1):e4519, 2023.

[269] Nemanja Kuzmanović, Joanna Puławska, Anelka Prokić, Milan Ivanović, Nevena Zlatković, Jeffrey B. Jones, and Aleksa Obradović. *Agrobacterium arsenijevicii* sp. nov., isolated from crown gall tumors on raspberry and cherry plum. *Systematic and Applied Microbiology*, 38(6):373–378, 2015.

[270] M. Dequivre, B. Diel, C. Villard, O. Sismeiro, M. Durot, J. Y. Coppée, X. Nesme, L. Vial, and F. Hommais. Small RNA deep-sequencing analyses reveal a new regulator of virulence in *Agrobacterium fabrum* C58. *Molecular Plant-Microbe Interactions*, 28(5):580–589, 2015.

[271] H. Bouzar and J. B. Jones. *Agrobacterium larrymoorei* sp. nov., a pathogen isolated from aerial tumours of *Ficus benjamina*. *International Journal of Systematic and Evolutionary Microbiology*, 51(Pt 3):1023–1026, 2001.

[272] Mily Ron, Kaisa Kajala, Germain Pauluzzi, Dongxue Wang, Mauricio A. Reynoso, Kristina Zumstein, Jasmine Garcha, Sonja Winte, Helen Masson, Soichi Inagaki, Fernán Federici, Neelima Sinha, Roger B. Deal, Julia Bailey-Serres, and Siobhan M.

Brady. Hairy root transformation using *Agrobacterium rhizogenes* as a tool for exploring cell type-specific gene expression and function using tomato as a model. *Plant Physiology*, 166(2):455–469, 2014.

[273] Nemanja Kuzmanović, Joanna Puławska, Kornelia Smalla, and Xavier Nesme. *Agrobacterium rosae* sp. nov., isolated from galls on different agricultural crops. *Systematic and Applied Microbiology*, 41(3):191–197, 2018.

[274] A. M. Alippi, A. C. Lopez, and P. A. Balatti. First report of *Agrobacterium rubi* and *A. rhizogenes* causing crown and root gall and hairy root on blueberry in argentina. *Plant Disease*, 94(8):1064, 2010.

[275] Jochen Gohlke and Rosalia Deeken. Plant responses to *Agrobacterium tumefaciens* and crown gall development. *Frontiers in Plant Science*, 5:155, 2014.

[276] I. H. Tolba and M. F. Zaki. Characterization of *Agrobacterium vitis* isolates obtained from galled grapevine plants in Egypt. *Annals of Agricultural Sciences*, 56(2):113–119, 2011.

[277] Sofie E. De Meyer, Elena Fabiano, Rui Tian, Peter Van Berkum, Rekha Seshadri, TBK Reddy, Victor Markowitz, Natalia Ivanova, Amrita Pati, Tanja Woyke, John Howieson, Nikos Kyrpides, and Wayne Reeve. High-quality permanent draft genome sequence of the *Parapiptadenia rigida*-nodulating *Burkholderia* sp. strain UYPR1.413. *Standards in Genomic Sciences*, 10(1):31, 2015.

[278] L. Shimkets and C. R. Woese. A phylogenetic analysis of the myxobacteria: basis for their classification. *Proceedings of the National Academy of Sciences*, 89(20):9459–9463, 1992.

[279] Sofie E. De Meyer, Matthew Parker, Peter Van Berkum, Rui Tian, Rekha Seshadri, T.B.K. Reddy, Victor Markowitz, Natalia Ivanova, Amrita Pati, Tanja Woyke, Nikos Kyrpides, John Howieson, and Wayne Reeve. High-quality permanent draft genome sequence of the *Mimosa asperata*-nodulating *Cupriavidus* sp. strain AMP6. *Standards in Genomic Sciences*, 10:80, 2015.

[280] Amornpan Klanchui, Supapon Cheevadhanarak, Peerada Prommeenate, and Asawin Meechai. Exploring components of the CO2-concentrating mechanism in alkaliphilic *Cyanobacteria* through genome-based analysis. *Computational and Structural Biotechnology Journal*, 15:340–350, 2017.

[281] A. Palacio-Bielsa, M. A. Cambra, and M. M. López. Characterisation of potato isolates of *Dickeya chrysanthemi* in Spain by a microtitre system for biovar determination. *Annals of Applied Biology*, 148(2):157–164, 2006.

[282] Yannick Raoul des Essarts, Jérémy Cigna, Angélique Quêtu-Laurent, Aline Caron, Euphrasie Munier, Amélie Beury-Cirou, Valérie Hélias, and Denis Faure. Biocontrol of the potato blackleg and soft rot diseases caused by *Dickeya dianthicola*. *Applied and Environmental Microbiology*, 82(1):268–278, 2016.

[283] Nicole Hugouvieux-Cotte-Pattat, Guy Condemine, Erwan Gueguen, and Vladimir E. Shevchik. *Dickeya* plant pathogens. In *eLS*, pages 1–10. American Cancer Society, 2020.

[284] Antoine Le Quéré, Nisha Tak, Hukam Singh Gehlot, Celine Lavire, Thibault Meyer, David Chapulliot, Sonam Rathi, Ilham Sakrouhi, Guadalupe Rocha, Marine Rohmer, Dany Severac, Abdelkarim Filali-Maltouf, and Jose-Antonio Munive. Genomic characterization of *Ensifer aridi*, a proposed new species of nitrogen-fixing rhizobium recovered from Asian, African and American deserts. *BMC Genomics*, 18(1):85, 2017.

[285] Inmaculada Sampedro, Daniel Pérez-Mendoza, Laura Toral, Esther Palacios, César Arriagada, and Inmaculada Llamas. Effects of halophyte root exudates and their components on chemotaxis, biofilm formation and colonization of the halophilic bacterium *Halomonas Anticariensis* FP35T. *Microorganisms*, 8(4):575, 2020.

[286] Hiroyasu Onaka, Shin-ichi Taniguchi, Yasuhiro Igarashi, and Tamotsu Furumai. Characterization of the biosynthetic gene cluster of rebeccamycin from *Lechevalieria aerocolonigenes* ATCC 39243. *Bioscience, Biotechnology, and Biochemistry*, 67(1):127–138, 2003.

[287] Chinyere K. Okoro, Alan T. Bull, Ankur Mutreja, Xiaoying Rong, Ying Huang, and Michael Goodfellow. *Lechevalieria atacamensis* sp. nov., *Lechevalieria deserti* sp. nov. and *Lechevalieria roselyniae* sp. nov., isolated from hyperarid soils. *International Journal of Systematic and Evolutionary Microbiology*, 60(Pt 2):296–300, 2010.

[288] Ekaterina Gavrish, Clarissa S. Sit, Shugeng Cao, Olga Kandror, Amy Spoering, Aaron Peoples, Losee Ling, Ashley Fetterman, Dallas Hughes, Anthony Bissell, Heather Torrey, Tatos Akopian, Andreas Mueller, Slava Epstein, Alfred Goldberg, Jon Clardy, and Kim Lewis. Lassomycin, a ribosomally synthesized cyclic peptide, kills *Mycobacterium tuberculosis* by targeting the ATP-dependent protease ClpC1P1P2. *Chemistry and Biology*, 21(4):509–518, 2014.

[289] Dongmei Li, Hao Jiang, Liyuan Han, Yuanyuan Li, Junwei Zhao, Shanwen Jiang, Xiangjing Wang, and Wensheng Xiang. *Lentzea terrae* sp. nov., isolated from soil and an emended description of *Lentzea* soli. *International Journal of Systematic and Evolutionary Microbiology*, 68(11):3528–3533, 2018.

[290] Yong Li, Han Xue, Li-min Guo, András Koltay, Ana Palacio-Bielsa, Jupu Chang, Shoujiang Xie, and Xuqi Yang. Elevation of three subspecies of *Lonsdalea quercina* to species level: *Lonsdalea britannica* sp. nov., *Lonsdalea iberica* sp. nov. and *Lonsdalea populi* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 67(11):4680–4684, 2017.

[291] Tímea Tóth, Tamás Lakatos, and András Koltay. *Lonsdalea quercina* subsp. *populi* subsp. nov., isolated from bark canker of poplar trees. *International Journal of Systematic and Evolutionary Microbiology*, 63(Pt_6):2309–2313, 2013.

[292] Takakazu Kaneko, Yasukazu Nakamura, Shusei Sato, Erika Asamizu, Tomohiko Kato, Shigemi Sasamoto, Akiko Watanabe, Kumi Idesawa, Atsuko Ishikawa, Kumiko Kawashima, et al. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA research*, 7(6):331–338, 2000.

[293] Luisa Caroline Ferraz Helene, Rebeca Fuzinatto Dall'Agnol, Jakeline Renata Marçon Delamuta, and Mariangela Hungria. *Mesorhizobium atlanticum* sp. nov., a new nitrogen-fixing species from soils of the Brazilian Atlantic Forest biome. *International Journal of Systematic and Evolutionary Microbiology*, 69(6):1800–1806, 2019.

[294] Mindia Haryono, Yi-Ming Tsai, Chien-Ting Lin, Fan-Chen Huang, Yan-Chen Ye, Wen-Ling Deng, Hau-Hsuan Hwang, and Chih-Horng Kuo. Presence of an *Agrobacterium*-type tumor-inducing plasmid in *Neorhizobium* sp. NCHU2750 and the link to phytopathogenicity. *Genome Biology and Evolution*, 10(12):3188–3195, 2018.

[295] Marilize le Roes and Paul R. Meyers. *Nocardia gamkensis* sp. nov. *Antonie van Leeuwenhoek*, 90(3):291–298, 2006.

[296] J. B. Demaree and N. R. Smith. *Nocardia vaccinii* n.sp. causing galls on blueberry plants. *Phytopathology*, 42:249–52, 1952.

[297] Mari T. Iwasawa, Yaei Togawa, Noriaki Kamada, Naotomo Kambe, Hiroyuki Matsue, Katsukiyo Yazawa, Takashi Yaguchi, and Yuzuru Mikami. Lymphocutaneous type of nocardiosis caused by *Nocardia vinacea* in a patient with polymyositis. *Mycopathologia*, 172(1):47–53, 2011.

[298] Manuel Brenes-Álvarez, Jan Mitschke, Elvira Olmedo-Verd, Jens Georg, Wolfgang R. Hess, Agustín Vioque, and Alicia M. Muro-Pastor. Elements of the heterocyst-specific transcriptome unravelled by co-expression analysis in *Nostoc* sp. PCC 7120. *Environmental Microbiology*, 21(7):2544–2558, 2019.

[299] Louis Gardan, Cécile Gouy, Richard Christen, and Régine Samson. Elevation of three subspecies of *Pectobacterium carotovorum* to species level: *Pectobacterium atrosepticum* sp. nov., *Pectobacterium betavasculorum* sp. nov. and *Pectobacterium wasabiae* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 53(Pt 2):381–391, 2003.

[300] Ruifang Liu, Yu Zhang, Ping Chen, Haiyan Lin, Guoyou Ye, Zhongwei Wang, Changwei Ge, Bo Zhu, and Deyong Ren. Genomic and phenotypic analyses of *Pseudomonas psychrotolerans* PRS08-11306 reveal a turnerbactin biosynthesis gene cluster that contributes to nitrogen fixation. *Journal of Biotechnology*, 253:10–13, 2017.

[301] Stéphane Genin and Timothy P. Denny. Pathogenomics of the *Ralstonia solanacearum* species complex. *Annual Review of Phytopathology*, 50(1):67–89, 2012.

[302] Zhirong Guo, Xi Cheng, Xinjie Hui, Xingsheng Shu, Aaron P. White, Yueming Hu, and Yejun Wang. Prediction of new bacterial type III secreted effectors with a recursive hidden markov model profile-alignment strategy. *Current Bioinformatics*, 13(3):280–289, 2018.

[303] Nemanja Kuzmanović, Patric Behrens, Elke Idczak, Stefan Wagner, Monika Götz, Cathrin Spröer, Boyke Bunk, Jörg Overmann, and Kornelia Smalla. A novel group of *Rhizobium tumorigenes* -like *Agrobacteria* associated with crown gall disease of rhododendron and blueberry. *Phytopathology*, 109(11):1840–1848, 2019.

[304] Koen Goethals, Danny Vereecke, Mondher Jaziri, Marc Van Montagu, and Marcelle Holsters. Leafy gall formation by *Rhodococcus fascians*. *Annual Review of Phytopathology*, 39(1):27–52, 2001.

[305] Jucimara Anunciação de Jesus Sousa and Fabio Lopes Olivares. Plant growth promotion by streptomycetes: ecophysiology, mechanisms and applications. *Chemical and Biological Technologies in Agriculture*, 3(1):24, 2016.

[306] Kaiqin Li, Yihui Guo, Junzhen Wang, Zhiyong Wang, Jiarong Zhao, and Jian Gao. *Streptomyces aquilus* sp. nov., a novel actinomycete isolated from a Chinese medicinal plant. *International Journal of Systematic and Evolutionary Microbiology*, 70(3):1912–1917, 2020.

[307] Anja Greule, Marija Marolt, Denise Deubel, Iris Peintner, Songya Zhang, Claudia Jessen-Trefzer, Christian De Ford, Sabrina Burschel, Shu-Ming Li, Thorsten Friedrich, Irmgard Merfort, Steffen Lüdeke, Philippe Bisel, Michael Müller, Thomas Paululat, and Andreas Bechthold. Wide distribution of foxicin biosynthetic gene clusters in *Streptomyces* strains – an unusual secondary metabolite with various properties. *Frontiers in Microbiology*, 8:221, 2017.

[308] Yacine Goudjal, Miyada Zamoum, Nasserdine Sabaou, Florence Mathieu, and Abdelghani Zitouni. Potential of endophytic *Streptomyces* spp. for biocontrol of *Fusarium* root rot disease and growth promotion of tomato seedlings. *Biocontrol Science and Technology*, 26(12):1691–1705, 2016.

[309] M. W. Dees, A. Sletten, and A. Hermansen. Isolation and characterization of *Streptomyces* species from potato common scab lesions in Norway. *Plant Pathology*, 62(1):217–225, 2013.

[310] G. Bubici. *Streptomyces* spp. as biocontrol agents against *Fusarium* species. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, 13(050), 2018.

[311] Xiaoguang Li, Xinting Lai, Longzhan Gan, Xiufeng Long, Yanyan Hou, Yuqing Zhang, and Yongqiang Tian. *Streptomyces geranii* sp. nov., a novel endophytic actinobacterium isolated from root of *Geranium carolinianum* L. *International Journal of Systematic and Evolutionary Microbiology*, 68(8):2562–2567, 2018.

[312] Hassan Mohamed Awad, Hesham Ali El-Enshasy, Siti Zulaiha Hanapi, Eman Ramadan Hamed, and Bustanur Rosidi. A new chitinase-producer strain *Streptomyces glauciniger* WICC-A03: isolation and identification as a biocontrol agent for plants phytopathogenic fungi. *Natural Product Research*, 28(24):2273–2277, 2014.

[313] Hyo-Jin Lee and Kyung-Sook Whang. *Streptomyces graminilatus* sp. nov., isolated from bamboo litter. *International Journal of Systematic and Evolutionary Microbiology*, 64(Pt 2):528–532, 2014.

[314] Yi Jiang, Shu-Kun Tang, Jutta Wiese, Li-Hua Xu, Johannes F. Imhoff, and Cheng-Lin Jiang. *Streptomyces hainanensis* sp. nov., a novel member of the genus *Streptomyces*. *International Journal of Systematic and Evolutionary Microbiology*, 57(Pt 11):2694–2698, 2007.

[315] Dongli Guan, Brenda L. Grau, Christopher A. Clark, Carol M. Taylor, Rosemary Loria, and Gregg S. Pettis. Evidence that thaxtomin C is a pathogenicity determinant

of *Streptomyces ipomoeae*, the causative agent of *Streptomyces* soil rot disease of sweet potato. *Molecular Plant-Microbe Interactions*, 25(3):393–401, 2011.

[316] S. Wang, F. Lu, Z. Yang, Z. Li, and Y. Tian. Combining ribosomal engineering with heterologous expression of a regulatory gene to improve milbemycin production in *Streptomyces milbemycinicus* A2079. *Applied Biochemistry and Microbiology*, 57(3):303–310, 2021.

[317] Zhikai Guo, Guohui Pan, Zhengren Xu, Dong Yang, Hindra, Xiangcheng Zhu, Yong Huang, Li-Xing Zhao, Yi Jiang, Yanwen Duan, and Ben Shen. New isofuranonaphthoquinones and isoindolequinones from *Streptomyces* sp. CB01883. *The Journal of Antibiotics*, 70(4):414–422, 2017.

[318] Christopher M. M. Franco, Eric M. Adetutu, Hoang X. Le, Ross A. Ballard, Ricardo Araujo, Shanan S. Tobe, Bobby Paul, Sandeep Mallya, and Kapaetu Satyamoorthy. Complete genome sequences of the endophytic *Streptomyces* sp. strains LUP30 and LUP47B, isolated from lucerne plants. *Genome Announcements*, 5(24):e00556–17, 2017.

[319] Bruna Domingues Vieira, Henrique Niero, Rafael de Felício, Luiz Fernando Giolo Alves, Cristina Freitas Bazzano, Renata Sigrist, Luciana Costa Furtado, Gabriela Felix Persinoti, Leticia Veras Costa-Lotufo, and Daniela Barretto Barbosa Trivella. Production of epoxyketone peptide-based proteasome inhibitors by *Streptomyces* sp. BRA-346: regulation and biosynthesis. *Frontiers in Microbiology*, 13, 2022.

[320] Madhumita Joshi, Xin Rong, Simon Moll, Johan Kers, Christopher Franco, and Rosemary Loria. *Streptomyces turgidiscabies* secretes a novel virulence protein, Nec1, which facilitates infection. *Molecular Plant-Microbe Interactions*, 20(6):599–608, 2007.

[321] Hannelore Drautz, Hans Zähner, Ernst Kupfer, and Walter Keller-Schierlein. Stoffwechselprodukte von Mikroorganismen 205. Mitteilung.. Isolierung und Struktur von Streptazolin. *Helvetica Chimica Acta*, 64(6):1752–1765, 1981.

[322] A. S. Gladkikh, O. I. Belykh, I. V. Klimenkov, and I. V. Tikhonova. Nitrogen-fixing cyanobacterium *Trichormus variabilis* of the lake baikal phytoplankton. *Microbiology*, 77(6):726–733, 2008.

[323] W. S. Dowson. On the stem rot or wilt disease of carnation. *Annals of Applied Biology*, 16:261–280, 1929.

[324] Isabelle Pieretti, Stéphane Cociancich, Stéphanie Bolot, Sébastien Carrère, Alexandre Morisset, Philippe Rott, and Monique Royer. Full genome sequence analysis of two isolates reveals a novel *Xanthomonas* species close to the sugarcane pathogen *Xanthomonas albilineans*. *Genes*, 6(3):714–733, 2015.

[325] Germán Dunger, Cecilia G. Garofalo, Natalia Gottig, Betiana S. Garavaglia, María C. Pereda Rosa, Chuck S. Farah, Elena G. Orellano, and Jorgelina Ottado. Analysis of three *Xanthomonas axonopodis* pv. citri effector proteins in pathogenicity and their interactions with host plant proteins. *Molecular Plant Pathology*, 13(8):865–876, 2012.

[326] Lena Hersemann, Daniel Wibberg, Jochen Blom, Franco Widmer, and Roland Kölliker. Draft genome sequence of the *Xanthomonas bromi* type strain LMG 947. *Genome Announcements*, 4(5), 2016.

[327] Jillian M. Lang, Alvaro L. Pérez-Quintero, Ralf Koebnik, Elysa DuCharme, Soungalo Sarra, Hinda Doucoure, Ibrahim Keita, Janet Ziegle, Jonathan M. Jacobs, Ricardo Oliva, Ousmane Koita, Boris Szurek, Valérie Verdier, and Jan E. Leach. A pathovar of *Xanthomonas oryzae* infecting wild grasses provides insight into the evolution of pathogenicity in rice agroecosystems. *Frontiers in Plant Science*, 10, 2019.

[328] Ralf Koebnik, Daiva Burokiene, Claude Bragard, Christine Chang, Marion Fischer-Le Saux, Roland Kölliker, Jillian M. Lang, Jan E. Leach, Emily K. Luna, Perrine Portier, Angeliki Sagia, Janet Ziegle, Stephen P. Cohen, and Jonathan M. Jacobs. The complete genome sequence of *Xanthomonas theicola*, the causal agent of canker on tea plants, reveals novel secretion systems in clade-1 Xanthomonads. *Phytopathology*, 111(4):611–616, 2020.

[329] Suraj Sapkota, Mohamed Mergoum, and Zhaohui Liu. The translucens group of *Xanthomonas translucens*: Complicated and important pathogens causing bacterial leaf streak on cereals. *Molecular Plant Pathology*, 21(3):291–302, 2020.

APPENDICES

Supplementary Tables

Supplementary Table 1

Table 1: AnnoTree results of searching PF01745 in the GTDB database with E-value <1e-5.

Bitscore	E-value	Gene ID	GTDB ID	Pfam ID
28	1.10E-07	MFJD01000001.1_308	GB_GCA_001779095.1	PF01745
24.3	7.50E-07	MFZT01000032.1_1	GB_GCA_001789515.1	PF01745
150.8	1.60E-44	NZ_BDBM01000003.1_180	RS_GCF_001612985.1	PF01745
225.3	2.60E-67	NZ_AUFE01000129.1_6	RS_GCF_000426345.1	PF01745
26.8	2.30E-07	MHCL01000021.1_15	GB_GCA_001816915.1	PF01745
178.8	5.80E-53	NZ_FNHV01000013.1_22	RS_GCF_900103585.1	PF01745
22.8	0.000007	PCXB01000052.1_37	GB_GCA_002787535.1	PF01745
49.1	4.20E-14	HF999519.1_7	GB_GCA_000434015.1	PF01745
27.9	1.30E-07	MNER01000063.1_20	GB_GCA_001917775.1	PF01745
305.7	4.60E-92	NZ_CP031697.1_3319	RS_GCF_004136375.1	PF01745
26.5	6.00E-07	PLHC01000080.1_9	GB_GCA_003138795.1	PF01745
63.8	6.50E-18	NZ_CP026304.1_91	RS_GCF_003054555.1	PF01745
27.7	1.30E-07	PBWA01000046.1_29	GB_GCA_002728855.1	PF01745
33.8	1.50E-09	QBYD01000033.1_1	GB_GCA_003282975.1	PF01745
26.4	2.30E-07	MGKQ01000023.1_6	GB_GCA_001795735.1	PF01745
22.6	0.0000031	MFZQ01000027.1_4	GB_GCA_001789445.1	PF01745
190.8	1.30E-56	NZ_NCSM01000005.1_275	RS_GCF_002224125.1	PF01745
28.8	6.00E-08	LCEX01000007.1_20	GB_GCA_000995415.1	PF01745
58.6	1.30E-16	DINJ01000136.1_11	GB_GCA_002423665.1	PF01745
28.3	8.70E-08	PFSC01000160.1_7	GB_GCA_002788865.1	PF01745

Bitscore	E-value	Gene ID	GTDB ID	Pfam ID
173.7	1.80E-51	NZ_CP012159.1_1556	RS_GCF_001189295.1	PF01745
44.7	1.50E-12	DCSU01000079.1_1	GB_GCA_002325785.1	PF01745
191.2	1.00E-56	NZ_MUNE01000113.1_6	RS_GCF_002154605.1	PF01745
362.9	1.90E-109	NZ_CP030831.1_171	RS_GCF_003597675.1	PF01745
239.9	7.60E-72	NZ_CP030831.1_15	RS_GCF_003597675.1	PF01745
30.6	1.10E-08	QCNX01000019.1_8	GB_GCA_003210015.1	PF01745
355	5.10E-107	NZ_CP039893.1_154	RS_GCF_005221325.1	PF01745
33.2	4.50E-09	DEEW01000136.1_4	GB_GCA_002350465.1	PF01745
171.4	9.50E-51	NZ_PJME01000030.1_62	RS_GCF_002954775.1	PF01745
28.3	5.90E-08	LBVN01000008.1_2	GB_GCA_000993115.1	PF01745
30	1.50E-08	MFFV01000045.1_23	GB_GCA_001777995.1	PF01745
28.1	1.10E-07	LUZH01000204.1_12	GB_GCA_001602885.1	PF01745
239.9	6.40E-72	NZ_BBJU01000024.1_64	RS_GCF_000739935.1	PF01745
362.9	1.60E-109	NZ_BBJU01000024.1_106	RS_GCF_000739935.1	PF01745
179.7	2.90E-53	NZ_LGUR01000002.1_359	RS_GCF_001270495.1	PF01745
53.3	6.40E-15	PLXD01000248.1_3	GB_GCA_003151295.1	PF01745
306.7	2.00E-92	NZ_FLTX01000115.1_8	RS_GCF_900092025.1	PF01745
29.4	2.80E-08	MHCP01000030.1_150	GB_GCA_001816655.1	PF01745
169.1	5.60E-50	NZ_LLZG01000411.1_109	RS_GCF_001509475.1	PF01745
237.6	3.40E-71	NZ_MBEV01000035.1_135	RS_GCF_001758275.1	PF01745
316.3	2.90E-95	NZ_LMVL02000015.1_19	RS_GCF_001541345.2	PF01745
240.3	4.80E-72	NZ_LMVL02000017.1_3	RS_GCF_001541345.2	PF01745
32.8	2.60E-09	MFBZ01000004.1_4	GB_GCA_001776545.1	PF01745
33.9	4.50E-09	NZ_CP013862.1_672	RS_GCF_001307805.1	PF01745
28.5	5.20E-08	MGDU01000020.1_67	GB_GCA_001790815.1	PF01745
172.2	5.60E-51	NZ_LJCV01000022.1_18	RS_GCF_001298575.1	PF01745
27.7	6.80E-08	MNZO01000050.1_11	GB_GCA_001873665.1	PF01745
188.9	3.30E-56	NZ_LJGW01000377.1_72	RS_GCF_001751255.1	PF01745
189.2	2.80E-56	NZ_LJGW01000252.1_26	RS_GCF_001751255.1	PF01745
26.7	1.90E-07	MFCV01000002.1_18	GB_GCA_001777015.1	PF01745
23.5	0.0000025	MFGN01000011.1_48	GB_GCA_001780285.1	PF01745
23.6	0.0000073	PMJE01000152.1_16	GB_GCA_003157315.1	PF01745
37.1	3.10E-10	NZ_BAJX01000002.1_31	RS_GCF_000613925.1	PF01745
24.3	8.60E-07	LCFM01000006.1_43	GB_GCA_000999025.1	PF01745
180.4	1.80E-53	NZ_KQ948422.1_1234	RS_GCF_001514065.1	PF01745
299.1	4.30E-90	NZ_CAPJ01000301.1_7	RS_GCF_000331775.1	PF01745
169.7	3.00E-50	NZ_LGDW01000086.1_5	RS_GCF_001280005.1	PF01745
27.4	1.70E-07	MFJM01000060.1_12	GB_GCA_001779205.1	PF01745
164.5	1.20E-48	NZ_JOAP01000046.1_42	RS_GCF_000720475.1	PF01745
23.1	0.0000026	PEUF01000033.1_18	GB_GCA_002780795.1	PF01745
33.8	2.30E-09	ONXS01000055.1_13	GB_GCA_900321855.1	PF01745
22.9	0.0000061	NC_014657.1_1319	RS_GCF_000166335.1	PF01745
22.9	0.0000063	NC_014392.1_1056	RS_GCF_000145215.1	PF01745

Bitscore	E-value	Gene ID	GTDB ID	Pfam ID
38.3	5.80E-11	PADT01000025.1_1	GB_GCA_002700465.1	PF01745
22.4	0.0000044	LMZU01000017.1_16	GB_GCA_001567245.1	PF01745
31.3	9.30E-09	QOAL01000082.1_1	GB_GCA_003978135.1	PF01745
23.4	0.0000046	QBLP01000075.1_6	GB_GCA_003070825.1	PF01745
26	2.90E-07	MFEO01000030.1_1	GB_GCA_001777585.1	PF01745
28.4	1.10E-07	NZ_NEEY01000020.1_27	RS_GCF_002252085.1	PF01745
24.6	5.20E-07	MNZH01000050.1_2	GB_GCA_001873515.1	PF01745
28.2	6.10E-08	MGAO01000001.1_15	GB_GCA_001788215.1	PF01745
184	1.40E-54	NZ_JXTE01000013.1_17	RS_GCF_000972385.1	PF01745
41.8	3.00E-11	BAXV01000013.1_61	GB_GCA_001310715.1	PF01745
27	7.70E-07	NZ_BAWC01000039.1_38	RS_GCF_000576365.1	PF01745
28.1	4.80E-08	MFRV01000025.1_2	GB_GCA_001784695.1	PF01745
27.1	1.70E-07	DJJA01000003.1_430	GB_GCA_002433895.1	PF01745
30.2	5.80E-08	MKRS01000806.1_4	GB_GCA_001897035.1	PF01745
34.6	2.60E-09	MKRS01000546.1_1	GB_GCA_001897035.1	PF01745
23.2	0.0000048	LYMA02000076.1_33	GB_GCA_001654785.2	PF01745
23.4	0.0000019	MEXE01000052.1_7	GB_GCA_001775555.1	PF01745
27.5	1.00E-07	LCLX01000001.1_205	GB_GCA_001002285.1	PF01745
28.9	3.70E-08	DBDS01000004.1_105	GB_GCA_002293685.1	PF01745
22.4	0.0000097	APCU01000014.1_14	GB_GCA_000347575.1	PF01745
36.8	1.40E-10	PCWD01000009.1_2	GB_GCA_002787255.1	PF01745
27.4	4.90E-07	VBKW01000259.1_3	GB_GCA_005879405.1	PF01745
27.2	1.20E-07	PFAT01000050.1_26	GB_GCA_002773575.1	PF01745
183.7	1.40E-54	NZ_SMKI01000157.1_4	RS_GCF_004348415.1	PF01745
171.4	7.90E-51	NZ_SMKI01000289.1_14	RS_GCF_004348415.1	PF01745
241.1	3.10E-72	NZ_PCDQ01000093.1_36	RS_GCF_003240565.1	PF01745
332.6	3.20E-100	NZ_PCDQ01000085.1_1	RS_GCF_003240565.1	PF01745
23.5	0.0000085	BBZO01000015.1_38	GB_GCA_001304875.1	PF01745
33.8	1.80E-09	URMC01000104.1_1	GB_GCA_900548735.1	PF01745
31.5	6.90E-09	LBTJ01000007.1_24	GB_GCA_000990445.1	PF01745
192.8	1.90E-57	NC_003272.1_223	RS_GCF_000009705.1	PF01745
41.6	1.50E-11	DGYJ01000115.1_70	GB_GCA_002396665.1	PF01745
21.7	0.0000088	MFJA01000054.1_14	GB_GCA_001779055.1	PF01745
224.6	3.00E-67	NZ_CP018758.1_2073	RS_GCF_001913135.1	PF01745
269.5	4.60E-81	NZ_CM001974.1_3701	RS_GCF_000406105.1	PF01745
325.6	4.40E-98	NZ_CP022482.1_1232	RS_GCF_003999715.1	PF01745
31.3	7.20E-09	LBWY01000023.1_5	GB_GCA_000995135.1	PF01745
282.5	5.10E-85	NZ_JQHM01000003.1_363	RS_GCF_000749845.1	PF01745
46.2	3.80E-13	SFPH01000036.1_9	GB_GCA_004552115.1	PF01745
22.3	0.0000045	PMJF01000037.1_6	GB_GCA_003157295.1	PF01745
33.2	1.50E-08	NZ_KV757701.1_22	RS_GCF_001712795.1	PF01745
25.8	4.50E-07	ONQV01000040.1_6	GB_GCA_900320065.1	PF01745
183.4	2.40E-54	NZ_MJAH01000006.1_593	RS_GCF_001746295.1	PF01745

Bitscore	E-value	Gene ID	GTDB ID	Pfam ID
24.9	0.0000029	MHBF01000189.1_4	GB_GCA_001803225.1	PF01745
28.5	5.70E-08	MHDE01000035.1_2	GB_GCA_001803465.1	PF01745
330.2	1.60E-99	NZ_CP014703.1_1080	RS_GCF_001586135.1	PF01745
28.7	2.40E-08	PFNV01000017.1_1	GB_GCA_002792165.1	PF01745
21.8	0.0000067	LLEV01000006.1_201	GB_GCA_001567515.1	PF01745
21.3	0.0000091	LCNZ01000006.1_248	GB_GCA_001002325.1	PF01745
26.8	1.60E-07	DDWQ01000005.1_17	GB_GCA_002345745.1	PF01745
271.7	8.00E-82	NZ_JIBO01000010.1_28	RS_GCF_000688655.1	PF01745
182.1	5.60E-54	NZ_LRTR01000929.1_5	RS_GCF_001550375.1	PF01745
25.6	3.00E-07	MFSF01000032.1_11	GB_GCA_001784815.1	PF01745
22.6	0.00001	NZ_SNWX01000029.1_7	RS_GCF_004362045.1	PF01745
23.5	0.0000042	DFGF01000011.1_3	GB_GCA_002371715.1	PF01745
286.9	3.20E-86	NZ_PNXZ01000006.1_136	RS_GCF_002879875.1	PF01745
196.4	1.50E-58	NC_007413.1_2780	RS_GCF_000204075.1	PF01745
25.5	2.10E-07	PEZH01000042.1_13	GB_GCA_002774085.1	PF01745
25.6	3.10E-07	PFCG01000016.1_19	GB_GCA_002773095.1	PF01745
30.8	2.00E-08	QCGG01000007.1_1	GB_GCA_003280055.1	PF01745
27.7	1.40E-07	MFJU01000022.1_62	GB_GCA_001779315.1	PF01745
55.3	9.40E-16	DENY01000199.1_6	GB_GCA_002359825.1	PF01745
21.3	0.0000076	MGFD01000024.1_6	GB_GCA_001791715.1	PF01745
297.4	1.80E-89	NZ_BCWW01000017.1_35	RS_GCF_001894785.1	PF01745
257.9	2.90E-77	NZ_KB903063.1_1	RS_GCF_000379565.1	PF01745
22.4	0.0000038	MHCC01000008.1_2	GB_GCA_001816475.1	PF01745
315.2	5.30E-95	NZ_JFAQ01000171.1_2	RS_GCF_001304695.1	PF01745
24.1	0.0000022	PLXF01000187.1_5	GB_GCA_003151315.1	PF01745
30.5	1.10E-08	MFFY01000006.1_4	GB_GCA_001778055.1	PF01745
156.9	2.90E-46	NZ_JYJG01000301.1_7	RS_GCF_000955955.1	PF01745
21.9	0.0000043	PEVT01000013.1_14	GB_GCA_002780175.1	PF01745
33.2	1.70E-09	MGFX01000001.1_347	GB_GCA_001792115.1	PF01745
25.4	4.40E-07	MEWB01000001.1_754	GB_GCA_001773215.1	PF01745
27.6	1.10E-07	MHJU01000009.1_39	GB_GCA_001818155.1	PF01745
321.3	5.20E-97	NZ_JZHZ01000007.1_2404	RS_GCF_000963005.1	PF01745
25.9	2.80E-07	QIIF01000082.1_4	GB_GCA_003231075.1	PF01745
23.1	0.0000049	DFDV01000068.1_2	GB_GCA_002369025.1	PF01745
312.3	4.20E-94	NZ_MIGX01000097.1_15	RS_GCF_002940605.1	PF01745
180.4	1.20E-53	NZ_LMDI01000045.1_268	RS_GCF_001425785.1	PF01745
44.2	1.80E-12	URHG01000034.1_27	GB_GCA_900547555.1	PF01745
28	8.00E-08	MFDZ01000010.1_8	GB_GCA_001777455.1	PF01745
240.8	3.60E-72	NZ_LMVK01000049.1_146	RS_GCF_001541315.1	PF01745
26.1	2.60E-07	MFCX01000005.1_24	GB_GCA_001777055.1	PF01745
272.3	5.40E-82	NZ_LUTP01000007.1_26	RS_GCF_002111585.1	PF01745
22.8	0.0000034	MFZK01000003.1_89	GB_GCA_001789395.1	PF01745
39.3	7.90E-11	MVQR01000432.1_1	GB_GCA_002067815.1	PF01745

Bitscore	E-value	Gene ID	GTDB ID	Pfam ID
28.2	7.30E-08	PFSB01000021.1_58	GB_GCA_002788855.1	PF01745
26.8	3.70E-07	DHVG01000027.1_1	GB_GCA_002412545.1	PF01745
332.3	4.10E-100	NZ_SCFK01000035.1_2	RS_GCF_007000325.1	PF01745
173.4	2.90E-51	NZ_MUMC01000754.1_2	RS_GCF_002155885.1	PF01745
24.8	0.0000016	NC_014654.1_1137	RS_GCF_000166415.1	PF01745
333.3	2.40E-100	NZ_BAYX01000012.1_47	RS_GCF_000696095.1	PF01745
24.7	0.000001	MNYD01000113.1_1	GB_GCA_001872925.1	PF01745
197.5	1.00E-58	NZ_FZOF01000021.1_96	RS_GCF_900188405.1	PF01745
22.4	0.0000027	MNWF01000023.1_25	GB_GCA_001872305.1	PF01745
27.2	9.40E-08	NZ_CP025989.1_386	RS_GCF_003072485.1	PF01745
24.8	7.70E-07	DBQG01000016.1_15	GB_GCA_002305125.1	PF01745
174.8	9.10E-52	NZ_KQ948929.1_499	RS_GCF_001514235.1	PF01745
25.5	5.50E-07	MFXK01000022.1_19	GB_GCA_001788835.1	PF01745
323	1.60E-97	NZ_MDCB01000005.1_57	RS_GCF_002939705.1	PF01745
26.9	1.40E-07	PFAR01000046.1_10	GB_GCA_002773655.1	PF01745
21.2	0.0000073	MHLB01000042.1_9	GB_GCA_001818895.1	PF01745
274.5	1.40E-82	NZ_CM001841.1_350	RS_GCF_000365305.1	PF01745
24	0.0000065	NZ_QQAV01000001.1_23	RS_GCF_003350475.1	PF01745
240.6	4.30E-72	NZ_CP026926.1_67	RS_GCF_002943835.1	PF01745
33.9	8.90E-10	LCJU01000006.1_8	GB_GCA_000994815.1	PF01745
26.4	2.80E-07	MGHC01000007.1_26	GB_GCA_001793015.1	PF01745
23.8	0.0000017	LBYD01000004.1_102	GB_GCA_000995965.1	PF01745
179.6	3.20E-53	NZ_CP034463.1_8229	RS_GCF_003955715.1	PF01745
362.9	1.90E-109	NZ_JWIT01000027.1_10	RS_GCF_000949895.1	PF01745
26.7	1.90E-07	MNYS01000003.1_159	GB_GCA_001873215.1	PF01745
36.3	5.10E-10	DCIP01000100.1_5	GB_GCA_002317115.1	PF01745
23.3	0.0000022	MHNR01000009.1_23	GB_GCA_001819875.1	PF01745
23.6	0.0000011	PFEE01000069.1_13	GB_GCA_002793745.1	PF01745
23.4	0.0000055	NZ_FTNC01000014.1_4	RS_GCF_900156285.1	PF01745
29.4	3.80E-08	DCTD01000042.1_16	GB_GCA_002335355.1	PF01745
182.1	6.10E-54	NZ_JOEY01000020.1_101	RS_GCF_000718165.1	PF01745
24.6	0.0000013	MGAG01000026.1_33	GB_GCA_001789645.1	PF01745
33.2	9.50E-09	NZ_AKKF01000096.1_8	RS_GCF_000281955.1	PF01745
28.8	3.50E-08	MFAI01000029.1_6	GB_GCA_001774555.1	PF01745
22.8	0.0000024	MFGA01000018.1_22	GB_GCA_001778095.1	PF01745
22.6	0.0000042	DGGY01000039.1_26	GB_GCA_002392405.1	PF01745
24.9	0.0000017	NZ_NIQC01000002.1_126	RS_GCF_002216005.1	PF01745
170.5	1.30E-50	NZ_SMKN01000120.1_6	RS_GCF_004348675.1	PF01745
22.2	0.000005	PETS01000130.1_9	GB_GCA_002781265.1	PF01745
25.7	4.70E-07	PFMK01000042.1_15	GB_GCA_002785215.1	PF01745
288	8.80E-87	NZ_LUTM01000008.1_74	RS_GCF_002111555.1	PF01745
22.6	0.0000026	DFKW01000005.1_8	GB_GCA_002373935.1	PF01745
28.6	3.30E-08	PFHP01000007.1_6	GB_GCA_002782365.1	PF01745

Bitscore	E-value	Gene ID	GTDB ID	Pfam ID
335.6	3.60E-101	NZ_CP019703.3_158	RS_GCF_002005205.3	PF01745
206.8	1.40E-61	NZ_LIQQ01000089.1_10	RS_GCF_001418565.1	PF01745
27.2	1.30E-07	LBYG01000008.1_45	GB_GCA_000996085.1	PF01745
274.8	9.00E-83	NZ_JIBQ01000010.1_223	RS_GCF_000688695.1	PF01745
151.5	1.20E-44	NZ_QLTT01000008.1_239	RS_GCF_003269295.1	PF01745
28.1	7.20E-08	QIFI01000056.1_1	GB_GCA_003229735.1	PF01745
27.6	9.40E-08	DBQA01000007.1_274	GB_GCA_002305675.1	PF01745
25.2	0.0000013	APKF01000086.1_3	GB_GCA_000353875.1	PF01745
32.2	6.30E-09	SFWW01000093.1_9	GB_GCA_004559845.1	PF01745
30	3.80E-08	DGGQ01000045.1_10	GB_GCA_002392275.1	PF01745
32.4	9.10E-09	SLIE01000430.1_6	GB_GCA_007135795.1	PF01745
205.9	3.10E-61	NZ_AEJB01000361.1_396	RS_GCF_000331005.1	PF01745
169.4	4.40E-50	NZ_AEJB01000023.1_14	RS_GCF_000331005.1	PF01745
35.8	8.40E-10	ONSV01000192.1_5	GB_GCA_900320585.1	PF01745
24.5	0.0000018	NZ_FUXC01000001.1_33	RS_GCF_900167025.1	PF01745
32.5	1.70E-08	NZ_CP013209.1_1	RS_GCF_001447315.1	PF01745
191.2	7.10E-57	NZ_SMKC01000179.1_7	RS_GCF_004348315.1	PF01745
363.2	1.40E-109	NZ_NXEJ01000013.1_168	RS_GCF_002915175.1	PF01745
245.4	1.30E-73	NZ_NXEJ01000013.1_37	RS_GCF_002915175.1	PF01745
28.9	1.70E-07	PMZC01000126.1_1	GB_GCA_003170555.1	PF01745
307.5	1.40E-92	NZ_CP039694.1_185	RS_GCF_005145045.1	PF01745
34	1.30E-09	MFCA01000025.1_137	GB_GCA_001775095.1	PF01745
29.4	2.70E-08	MFSE01000021.1_46	GB_GCA_001784805.1	PF01745
26.1	3.10E-07	LBWP01000004.1_13	GB_GCA_000993685.1	PF01745
23.6	0.0000019	DJCU01000035.1_70	GB_GCA_002430725.1	PF01745
297.4	1.80E-89	NZ_NOYD01000054.1_35	RS_GCF_002258315.1	PF01745
297.5	1.70E-89	NZ_JMEX01000009.1_76	RS_GCF_000760735.1	PF01745
30.6	9.70E-09	QZIU01000035.1_34	GB_GCA_003599625.1	PF01745
260.1	3.30E-78	NZ_CM001975.1_345	RS_GCF_000406125.1	PF01745
25.4	5.30E-07	MFJZ01000012.1_1	GB_GCA_001781025.1	PF01745
27	1.90E-07	BDTH01000029.1_21	GB_GCA_002923195.1	PF01745
33.1	6.00E-09	QHNB01000166.1_8	GB_GCA_003217965.1	PF01745
147.8	1.70E-43	NZ_BAGN01000129.1_63	RS_GCF_000308835.1	PF01745
222.8	1.70E-66	NZ_LPWA01000123.1_30	RS_GCF_001510895.1	PF01745
362.9	1.70E-109	NC_003065.3_163	RS_GCF_000092025.1	PF01745
25	0.0000011	DLLE01000065.1_2	GB_GCA_002477955.1	PF01745
156.4	4.30E-46	NZ_QFWW01000001.1_517	RS_GCF_003265345.1	PF01745
186.7	2.20E-55	NZ_AEJC01000522.1_3	RS_GCF_000317595.1	PF01745
22.3	0.000004	MNWN01000039.1_32	GB_GCA_001872545.1	PF01745
26.6	6.10E-07	PLDZ01000136.1_10	GB_GCA_003136295.1	PF01745
27.5	1.10E-07	PFWU01000049.1_79	GB_GCA_002791295.1	PF01745
28.4	4.70E-08	PFVO01000018.1_41	GB_GCA_002790655.1	PF01745
24.1	0.0000014	MHCJ01000003.1_768	GB_GCA_001816585.1	PF01745

Bitscore	E-value	Gene ID	GTDB ID	Pfam ID
36.4	5.10E-10	NZ_BAJY01000030.1_16	RS_GCF_000613945.1	PF01745
22.8	0.0000035	PFOB01000002.1_8	GB_GCA_002792315.1	PF01745
177.6	1.20E-52	NZ_JOFS01000067.1_23	RS_GCF_000719285.1	PF01745
148.9	7.00E-44	NZ_BDCC01000100.1_7	RS_GCF_001613305.1	PF01745
169.7	3.10E-50	NZ_ASRX01000048.1_50	RS_GCF_000601485.1	PF01745
27	1.30E-07	PCRW01000003.1_13	GB_GCA_002772115.1	PF01745
27.4	8.30E-08	MGLG01000057.1_4	GB_GCA_001795975.1	PF01745
23.9	0.0000017	MFJN01000016.1_11	GB_GCA_001779235.1	PF01745
27.2	8.50E-08	DCTW01000004.1_33	GB_GCA_002329465.1	PF01745
43.2	4.20E-12	ONHS01000051.1_1	GB_GCA_900317715.1	PF01745
335.6	3.80E-101	NZ_SGNY01000016.1_16	RS_GCF_007002745.1	PF01745
206.3	8.80E-62	NZ_LUTN01000009.1_131	RS_GCF_002111595.1	PF01745
24.3	9.10E-07	MGFG01000023.1_28	GB_GCA_001791745.1	PF01745
22.7	0.0000026	MHWW01000009.1_5	GB_GCA_001824275.1	PF01745
23.1	0.000002	DIVO01000011.1_15	GB_GCA_002428285.1	PF01745
28.7	5.50E-08	LBQR01000026.1_4	GB_GCA_000990945.1	PF01745
29.2	2.30E-07	AP019514.1_2491	GB_GCA_007182875.1	PF01745
24.1	0.0000015	PFTH01000089.1_5	GB_GCA_002789465.1	PF01745
28.4	1.90E-07	NZ_BBAB01000019.1_69	RS_GCF_001311035.1	PF01745
179.6	2.80E-53	NZ_FOUV01000016.1_167	RS_GCF_900114955.1	PF01745
31.9	1.50E-08	NATG01000284.1_6	GB_GCA_002085215.1	PF01745
22.5	0.0000027	MZGJ01000006.1_69	GB_GCA_002084955.1	PF01745
177.3	7.20E-53	NZ_KB235926.1_464	RS_GCF_000332255.1	PF01745
30	3.70E-08	SKUN01000002.1_19	GB_GCA_007129945.1	PF01745
184.9	3.90E-55	NZ_KE332392.1_123	RS_GCF_000409775.1	PF01745
24.6	5.10E-07	PGYQ01000001.1_39	GB_GCA_002839745.1	PF01745
352.7	2.40E-106	NC_011982.1_225	RS_GCF_000016285.1	PF01745
24.3	7.90E-07	LCKU01000008.1_28	GB_GCA_001001625.1	PF01745
24	0.0000031	NZ_CAUI01000005.1_574	RS_GCF_000350165.1	PF01745
34.4	2.80E-09	NZ_BBMK01000006.1_65	RS_GCF_000755285.1	PF01745
28.4	2.10E-07	PKCP01000513.1_3	GB_GCA_002862225.1	PF01745
44.8	7.90E-13	DBBY01000047.1_7	GB_GCA_002292855.1	PF01745
297.4	1.80E-89	NZ_NPFX01000024.1_78	RS_GCF_002259335.1	PF01745
29.9	3.40E-08	NC_022582.1_604	RS_GCF_000479315.1	PF01745
35.2	6.70E-10	DBRA01000056.1_17	GB_GCA_002305445.1	PF01745
167.8	1.30E-49	NZ_MUYM01000084.1_43	RS_GCF_002150765.1	PF01745
22.3	0.0000083	NZ_MWOX01000012.1_12	RS_GCF_002026005.1	PF01745
25.7	9.30E-07	ONDP01000002.1_3	GB_GCA_900316645.1	PF01745
32.2	2.80E-08	NZ_RZQL01000345.1_7	RS_GCF_003997935.1	PF01745
24.6	0.0000013	PLYC01000007.1_126	GB_GCA_003151625.1	PF01745
324.6	1.10E-97	NZ_AZNX01000014.1_105	RS_GCF_000510685.1	PF01745
32.8	5.60E-09	NCKC01000041.1_30	GB_GCA_002282715.1	PF01745
24.9	5.90E-07	MGGH01000036.1_51	GB_GCA_001792305.1	PF01745

Bitscore	E-value	Gene ID	GTDB ID	Pfam ID
23.5	0.0000012	DIAR01000023.1_16	GB_GCA_002415845.1	PF01745
25.5	3.80E-07	PFLE01000019.1_49	GB_GCA_002784465.1	PF01745
23.4	0.0000054	NZ_QAXS01000002.1_40	RS_GCF_003053565.1	PF01745
22.3	0.0000035	LCOA01000017.1_7	GB_GCA_001002405.1	PF01745
22.8	0.000003	MGGC01000041.1_15	GB_GCA_001792215.1	PF01745
153.9	2.50E-45	NZ_BBOJ01000028.1_77	RS_GCF_000974445.1	PF01745
29.8	2.40E-08	LBVU01000003.1_39	GB_GCA_000993205.1	PF01745
219.8	1.10E-65	NZ_QMBQ01000009.1_130	RS_GCF_003289965.1	PF01745
181.4	8.50E-54	NZ_RQJW01000016.1_154	RS_GCF_004028635.1	PF01745
26.1	3.00E-07	MHII01000006.1_25	GB_GCA_001817425.1	PF01745
32.5	2.80E-09	MHIE01000015.1_4	GB_GCA_001817345.1	PF01745
28.1	9.60E-08	MFNJ01000029.1_27	GB_GCA_001782015.1	PF01745
32.1	2.10E-08	BBCA01000035.1_12	GB_GCA_001311805.1	PF01745
30	1.60E-08	MGJU01000012.1_7	GB_GCA_001794435.1	PF01745
23.4	0.0000011	PFEM01000034.1_2	GB_GCA_002793635.1	PF01745
26.8	1.50E-07	MFEL01000001.1_37	GB_GCA_001777575.1	PF01745
256.2	1.30E-76	NZ_JAFD01000245.1_2	RS_GCF_000519245.1	PF01745
25.6	5.50E-07	SOIK01000221.1_4	GB_GCA_004377245.1	PF01745
28.3	1.90E-07	QKGJ01000346.1_4	GB_GCA_003251275.1	PF01745

Supplementary Table 2

Table 2: 98 IPT protein sequences after removing 199 matches belonging to other domain families (e.g., IPPT).

Genome accession (GTDB ID)	NCBI species name	Gene ID	Phylum	Clade
RS_GCF_001298575.1	Actinobacteria bacterium OK074	NZ_LJCV01000022.1_18	actinobacteria	B
RS_GCF_000949895.1	Agrobacterium arsenijevicii	NZ_JWIT01000027.1_10	proteobacteria	A
RS_GCF_000092025.1	Agrobacterium fabrum str. C58	NC_003065.3_163	proteobacteria	A
RS_GCF_005145045.1	Agrobacterium larrymoorei	NZ_CP039694.1_185	proteobacteria	A
RS_GCF_002005205.3	Agrobacterium rhizogenes	NZ_CP019703.3_158	proteobacteria	A
RS_GCF_007002745.1	Agrobacterium rhizogenes	NZ_SGNY01000016.1_16	proteobacteria	A
RS_GCF_000696095.1	Agrobacterium rhizogenes NBRC 13257	NZ_BAYX01000012.1_47	proteobacteria	A
RS_GCF_002915175.1	Agrobacterium rosae	NZ_NXEJ01000013.1_168	proteobacteria	A
RS_GCF_002915175.1	Agrobacterium rosae	NZ_NXEJ01000013.1_37	proteobacteria	A
RS_GCF_000739935.1	Agrobacterium rubi TR3 = NBRC 13261	NZ_BBJU01000024.1_64	proteobacteria	A
RS_GCF_000739935.1	Agrobacterium rubi TR3 = NBRC 13261	NZ_BBJU01000024.1_106	proteobacteria	A
RS_GCF_002943835.1	Agrobacterium tumefaciens	NZ_CP026926.1_67	proteobacteria	A
RS_GCF_005221325.1	Agrobacterium tumefaciens	NZ_CP039893.1_154	proteobacteria	A
RS_GCF_001541315.1	Agrobacterium tumefaciens str. B6	NZ_LMVK01000049.1_146	proteobacteria	A
RS_GCF_001541345.2	Agrobacterium vitis	NZ_LMVL02000015.1_19	proteobacteria	A
RS_GCF_001541345.2	Agrobacterium vitis	NZ_LMVL02000017.1_3	proteobacteria	A
RS_GCF_001758275.1	Agrobacterium vitis	NZ_MBEV01000035.1_135	proteobacteria	A
RS_GCF_000016285.1	Agrobacterium vitis S4	NC_011982.1_225	proteobacteria	A
RS_GCF_000519245.1	Burkholderia sp. UYPR1.413	NZ_JAFD01000245.1_2	proteobacteria	A
RS_GCF_000601485.1	Chondromyces apiculatus DSM 436	NZ_ASRX01000048.1_50	myxococcota	B
RS_GCF_001189295.1	Chondromyces crocatus	NZ_CP012159.1_1556	myxococcota	B
RS_GCF_000426345.1	Cupriavidus sp. amp6	NZ_AUFE01000129.1_6	proteobacteria	A

Genome accession (GTDB ID)	NCBI species name	Gene ID	Phylum	Clade
RS_GCF_000379565.1	Cupriavidus sp. UYPR2.512	NZ_KB903063.1_1	proteobacteria	A
RS_GCF_000332255.1	Cyanobacterium PCC 7702	NZ_KB235926.1_464	cyanobacteria	B
RS_GCF_000406105.1	Dickeya chrysanthemi NCPPB 402	NZ_CM001974.1_3701	proteobacteria	A
RS_GCF_000365305.1	Dickeya dianthicola NCPPB 453	NZ_CM001841.1_350	proteobacteria	A
RS_GCF_000406125.1	Dickeya poaceiphila	NZ_CM001975.1_345	proteobacteria	A
RS_GCF_000510685.1	Ensifer aridi	NZ_AZNX01000014.1_105	proteobacteria	A
RS_GCF_000409775.1	Halomonas anticariensis FP35 = DSM 16096	NZ_KE332392.1_123	proteobacteria	B
RS_GCF_000955955.1	Lechevalieria aerocolonigenes	NZ_JYJG01000301.1_7	actinobacteria	B
RS_GCF_000974445.1	Lechevalieria aerocolonigenes	NZ_BBOJ01000028.1_77	actinobacteria	B
RS_GCF_003269295.1	Lechevalieria atacamensis	NZ_QLTT01000008.1_239	actinobacteria	B
RS_GCF_002150765.1	Lentzea kentuckyensis	NZ_MUYM01000084.1_43	actinobacteria	B
RS_GCF_003265345.1	Lentzea terrae	NZ_QFWW01000001.1_517	actinobacteria	B
RS_GCF_002111595.1	Lonsdalea britannica	NZ_LUTN01000009.1_131	proteobacteria	A
RS_GCF_002111585.1	Lonsdalea iberica	NZ_LUTP01000007.1_26	proteobacteria	A
RS_GCF_002111555.1	Lonsdalea populi	NZ_LUTM01000008.1_74	proteobacteria	A
RS_GCF_000688655.1	Lonsdalea quercina subsp. quercina	NZ_JIBO01000010.1_28	proteobacteria	A
RS_GCF_000688695.1	Lonsdalea quercina subsp. quercina	NZ_JIBQ01000010.1_223	proteobacteria	A
RS_GCF_001510895.1	Mesorhizobium loti	NZ_LPWA01000123.1_30	proteobacteria	A
RS_GCF_003289965.1	Mesorhizobium sp. CNPSo 3140	NZ_QMBQ01000009.1_130	proteobacteria	A
RS_GCF_004348675.1	Micromonospora sp. KC606	NZ_SMKN01000120.1_6	actinobacteria	B
RS_GCF_003597675.1	Neorhizobium sp. NCHU2750	NZ_CP030831.1_171	proteobacteria	A
RS_GCF_003597675.1	Neorhizobium sp. NCHU2750	NZ_CP030831.1_15	proteobacteria	A
RS_GCF_001612985.1	Nocardia gamkensis NBRC 108242	NZ_BDBM01000003.1_180	actinobacteria	B
RS_GCF_001613305.1	Nocardia vaccinii NBRC 15922	NZ_BDCC01000100.1_7	actinobacteria	B
RS_GCF_000308835.1	Nocardia vinacea NBRC 16497	NZ_BAGN01000129.1_63	actinobacteria	B

Genome accession (GTDB ID)	NCBI species name	Gene ID	Phylum	Clade
RS_GCF_000009705.1	Nostoc sp. PCC 7120 = FACHB-418	NC_003272.1_223	cyanobacteria	B
RS_GCF_000749845.1	Pectobacterium betavasculorum	NZ_JQHM01000003.1_363	proteobacteria	A
RS_GCF_001913135.1	Pseudomonas psychrotolerans	NZ_CP018758.1_2073	proteobacteria	A
RS_GCF_001586135.1	Ralstonia solanacearum	NZ_CP014703.1_1080	proteobacteria	A
RS_GCF_003999715.1	Ralstonia solanacearum	NZ_CP022482.1_1232	proteobacteria	A
RS_GCF_001425785.1	Rhizobacter sp. Root1221	NZ_LMDI01000045.1_268	proteobacteria	B
RS_GCF_007000325.1	Rhizobium sp. rho-6.2	NZ_SCFK01000035.1_2	proteobacteria	A
RS_GCF_003240565.1	Rhizobium tumorigenes	NZ_PCDQ01000093.1_36	proteobacteria	A
RS_GCF_003240565.1	Rhizobium tumorigenes	NZ_PCDQ01000085.1_1	proteobacteria	A
RS_GCF_000760735.1	Rhodococcus fascians A44A	NZ_JMEX01000009.1_76	actinobacteria	B
RS_GCF_001894785.1	Rhodococcus fascians NBRC 12155 = LMG 3623	NZ_BCWW01000017.1_35	actinobacteria	B
RS_GCF_002258315.1	Rhodococcus sp. 06-462-5	NZ_NOYD01000054.1_35	actinobacteria	B
RS_GCF_002259335.1	Rhodococcus sp. 14-2483-1-1	NZ_NPFX01000024.1_78	actinobacteria	B
RS_GCF_001514065.1	Streptomyces antibioticus	NZ_KQ948422.1_1234	actinobacteria	B
RS_GCF_003955715.1	Streptomyces aquilus	NZ_CP034463.1_8229	actinobacteria	B
RS_GCF_000720475.1	Streptomyces aureocirculatus	NZ_JOAP01000046.1_42	actinobacteria	A
RS_GCF_000719285.1	Streptomyces bicolor	NZ_JOFS01000067.1_23	actinobacteria	B
RS_GCF_001514235.1	Streptomyces caeruleatus	NZ_KQ948929.1_499	actinobacteria	B
RS_GCF_004028635.1	Streptomyces cyaneus	NZ_RQJW01000016.1_154	actinobacteria	B
RS_GCF_001550375.1	Streptomyces europaeiscabiei	NZ_LRTR01000929.1_5	actinobacteria	B
RS_GCF_000718165.1	Streptomyces fulvoviolaceus	NZ_JOEY01000020.1_101	actinobacteria	B
RS_GCF_002954775.1	Streptomyces geranii	NZ_PJME01000030.1_62	actinobacteria	B
RS_GCF_900188405.1	Streptomyces glauciniger	NZ_FZOF01000021.1_96	actinobacteria	A

Genome accession (GTDB ID)	NCBI species name	Gene ID	Phylum	Clade
RS_GCF_001418565.1	Streptomyces graminilatus	NZ_LIQQ01000089.1_10	actinobacteria	B
RS_GCF_004348415.1	Streptomyces hainanensis	NZ_SMKI01000157.1_4	actinobacteria	B
RS_GCF_004348415.1	Streptomyces hainanensis	NZ_SMKI01000289.1_14	actinobacteria	B
RS_GCF_000317595.1	Streptomyces ipomoeae	NZ_AEJC01000522.1_3	actinobacteria	B
RS_GCF_002154605.1	Streptomyces milbemycinicus	NZ_MUNE01000113.1_6	actinobacteria	B
RS_GCF_001751255.1	Streptomyces nanshensis	NZ_LJGW01000377.1_72	actinobacteria	A
RS_GCF_001751255.1	Streptomyces nanshensis	NZ_LJGW01000252.1_26	actinobacteria	A
RS_GCF_001509475.1	Streptomyces regalis	NZ_LLZG01000411.1_109	actinobacteria	B
RS_GCF_002155885.1	Streptomyces rhizosphaericus	NZ_MUMC01000754.1_2	actinobacteria	A
RS_GCF_004348315.1	Streptomyces sp. 8K308	NZ_SMKC01000179.1_7	actinobacteria	B
RS_GCF_900114955.1	Streptomyces sp. cf124	NZ_FOUV01000016.1_167	actinobacteria	B
RS_GCF_900103585.1	Streptomyces sp. cf386	NZ_FNHV01000013.1_22	actinobacteria	B
RS_GCF_001746295.1	Streptomyces sp. LUP47B	NZ_MJAH01000006.1_593	actinobacteria	B
RS_GCF_002224125.1	Streptomyces sp. NBS 14/10	NZ_NCSM01000005.1_275	actinobacteria	B
RS_GCF_001280005.1	Streptomyces sp. NRRL WC-3618	NZ_LGDW01000086.1_5	actinobacteria	B
RS_GCF_000972385.1	Streptomyces sp. WM6386	NZ_JXTE01000013.1_17	actinobacteria	B
RS_GCF_000331005.1	Streptomyces turgidiscabies Car8	NZ_AEJB01000361.1_396	actinobacteria	B
RS_GCF_000331005.1	Streptomyces turgidiscabies Car8	NZ_AEJB01000023.1_14	actinobacteria	B
RS_GCF_001270495.1	Streptomyces viridochromogenes	NZ_LGUR01000002.1_359	actinobacteria	B
RS_GCF_000204075.1	Trichormus variabilis	NC_007413.1_2780	cyanobacteria	B
RS_GCF_002879875.1	Trinickia caryophylli	NZ_PNXZ01000006.1_136	proteobacteria	A
RS_GCF_002939705.1	Xanthomonas albilineans	NZ_MDCB01000005.1_57	proteobacteria	A
RS_GCF_001304695.1	Xanthomonas axonopodis	NZ_JFAQ01000171.1_2	proteobacteria	A
RS_GCF_900092025.1	Xanthomonas bromi	NZ_FLTX01000115.1_8	proteobacteria	A

Genome accession (GTDB ID)	NCBI species name	Gene ID	Phylum	Clade
RS_GCF_004136375.1	Xanthomonas oryzae pv. oryzae	NZ_CP031697.1_3319	proteobacteria	A
RS_GCF_000963005.1	Xanthomonas sp. GPE 39	NZ_JZHZ01000007.1_2404	proteobacteria	A
RS_GCF_002940605.1	Xanthomonas theicola	NZ_MIGX01000097.1_15	proteobacteria	A
RS_GCF_000331775.1	Xanthomonas translucens pv. translucens DSM 18974	NZ_CAPJ01000301.1_7	proteobacteria	A

Supplementary Table 3

Table 3: Metadata table of IPT-containing bacteria, including isolation source, evidence of PGPB or pathogens, and their associated literature source.

NCBI species name	Gene ID	Isolation source	Evidence of pgpb or pathogen	Justification
Actinobacteria bacterium OK074	NZ_LJCV01000022.1_18	endosphere of plant Populus trichocarpa	unknown	-
Agrobacterium arsenijevicii	NZ_JWIT01000027.1_10	-	pathogen	cause crown gall tumors on raspberry and cherry plum[269]
Agrobacterium fabrum str. C58	NC_003065.3_163	-	pathogen	has Ti plasmid that induce tumor formation[270]
Agrobacterium larrymoorei	NZ_CP039694.1_185	-	pathogen	isolated from tumors on Ficus benjamina[271]
Agrobacterium rhizogenes	NZ_CP019703.3_158	-	pathogen	induce hairy roots[272]
Agrobacterium rhizogenes	NZ_SGNY01000016.1_16	-	pathogen	induce hairy roots[272]
Agrobacterium rhizogenes NBRC 13257	NZ_BAYX01000012.1_47	-	pathogen	induce hairy roots[272]
Agrobacterium rosae	NZ_NXEJ01000013.1_168	-	pathogen	isolated from galls[273]
Agrobacterium rosae	NZ_NXEJ01000013.1_37	-	pathogen	isolated from galls[273]
Agrobacterium rubi TR3 = NBRC 13261	NZ_BBJU01000024.1_64	-	pathogen	cause crown and root gall[274]
Agrobacterium rubi TR3 = NBRC 13261	NZ_BBJU01000024.1_106	-	pathogen	cause crown and root gall[274]
Agrobacterium tumefaciens	NZ_CP026926.1_67	-	pathogen	cause crown gall disease[275]
Agrobacterium tumefaciens	NZ_CP039893.1_154	-	pathogen	cause crown gall disease[275]
Agrobacterium tumefaciens str. B6	NZ_LMVK01000049.1_146	apple seedling crown gall	pathogen	cause crown gall disease[275]
Agrobacterium vitis	NZ_LMVL02000015.1_19	-	pathogen	induce galls and necrosis in grapevines[276]
Agrobacterium vitis	NZ_LMVL02000017.1_3	-	pathogen	induce galls and necrosis in grapevines[276]
Agrobacterium vitis	NZ_MBEV01000035.1_135	crown gall tumor	pathogen	induce galls and necrosis in grapevines[276]
Agrobacterium vitis S4	NC_011982.1_225	-	pathogen	induce galls and necrosis in grapevines[276]
Burkholderia sp. UYPR1.413	NZ_JAFD01000245.1_2	-	PGPB	isolated from a root nodule of Parapiptadenia rigida, its a highly efficient nitrogen fixing symbioint with host P.rigida[277]
Chondromyces apiculatus DSM 436	NZ_ASRX01000048.1_50	-	unknown	myxobacteria, form fruiting bodies[278]
Chondromyces crocatus	NZ_CP012159.1_1556	soil	PGPB	produce antibacterial compounds[165]
Cupriavidus sp. amp6	NZ_AUFE01000129.1_6	-	PGPB	isolated from root nodule of Mimosa asperata, develop nitrogen fixing symbiosis with Mimosa species[279]

NCBI species name	Gene ID	Isolation source	Evidence of pgpb or pathogen	Justification
Cupriavidus sp. UYPR2.512	NZ_KB903063.1_1	native forest; Artigas; Uruguay	PGPB	isolated from a root nodule of Parapiptadenia rigida, nodulate and promote the growth of P. rigida, as well as Mimosa pudica plants[277]
Cyanobacterium PCC 7702	NZ_KB235926.1_464	-	unknown	isolated from alkaline siliceous hot springs, nitrogen fixing bacteria[280]
Dickeya chrysanthemi NCPPB 402	NZ_CM001974.1_3701	-	pathogen	cause soft rot and wilting disease[281]
Dickeya dianthicola NCPPB 453	NZ_CM001841.1_350	-	pathogen	cause blackleg infection[282]
Dickeya poaceiphila	NZ_CM001975.1_345	-	pathogen	isolated from diseased sugar cane[283]
Ensifer aridi	NZ_AZNX01000014.1_105	-	PGPB	a nitrogen fixing bacteria, can develop a beneficial symbiotic interaction[284]
Halomonas anticariensis FP35 = DSM 16096	NZ_KE332392.1_123	-	PGPB	enhance growth of Salicornia. Halophilic bacteria, produce EPSs, isolated from saline wetland soil[285]
Lechevalieria aerocolonigenes	NZ_JYJG01000301.1_7	-	PGPB	produce antitumor compournd rebeccamycin[286]
Lechevalieria aerocolonigenes	NZ_BBOJ01000028.1_77	-	PGPB	produce antitumor compournd rebeccamycin[286]
Lechevalieria atacamensis	NZ_QLTT01000008.1_239	-	unknown	isolated from hyperarid soils[287]
Lentzea kentuckyensis	NZ_MUYM01000084.1_43	-	PGPB	isolated from an equine placenta, produce bactericidal compound lassomycin[288]
Lentzea terrae	NZ_QFWW01000001.1_517	-	unknown	isolated from soil[289]
Lonsdalea britannica	NZ_LUTN01000009.1_131	-	pathogen	associated with acute oak decline in Great Britain and bark canker of species of Quercus in Spain[290]
Lonsdalea iberica	NZ_LUTP01000007.1_26	-	pathogen	associated with acute oak decline in Great Britain and bark canker of species of Quercus in Spain[290]
Lonsdalea populi	NZ_LUTM01000008.1_74	-	pathogen	causal agent of bark canker[291]
Lonsdalea quercina subsp. quercina	NZ_JIBO01000010.1_28	forest	pathogen	cause bleeding cankers in the oak bark[290]
Lonsdalea quercina subsp. quercina	NZ_JIBQ01000010.1_223	forest	pathogen	cause bleeding cankers in the oak bark[290]
Mesorhizobium loti	NZ_LPWA01000123.1_30	soil from zinc mining area	PGPB	nitrogen-fixing symbiotic bacterium[292]
Mesorhizobium sp. CNPSo 3140	NZ_QMBQ01000009.1_130	-	PGPB	isolated from nitrogen fixing nodule of Phaseolus vulgaris[293]
Micromonospora sp. KC606	NZ_SMKN01000120.1_6	-	unknown	-
Neorhizobium sp. NCHU2750	NZ_CP030831.1_171	-	pathogen	a gall forming pathogen[294]
Neorhizobium sp. NCHU2750	NZ_CP030831.1_15	-	pathogen	a gall forming pathogen[294]

NCBI species name	Gene ID	Isolation source	Evidence of pgpb or pathogen	Justification
Nocardia gamkensis NBRC 108242	NZ_BDBM01000003.1_180	-	PGPB	exhibited moderate antibiotic activity against Mycobacterium[295]
Nocardia vaccinii NBRC 15922	NZ_BDCC01000100.1_7	-	pathogen	cause galls on blueberry[296]
Nocardia vinacea NBRC 16497	NZ_BAGN01000129.1_63	-	unknown	cause nocardiosis in human[297]
Nostoc sp. PCC 7120 = FACHB-418	NC_003272.1_223	-	PGPB	nitrogen fixing bacteria[298]
Pectobacterium betavasculorum	NZ_JQHM01000003.1_363	-	pathogen	responsible for vascular necrosis of sugar beet[299]
Pseudomonas psychrotolerans	NZ_CP018758.1_2073	-	PGPB	produce turnerbactin that contributes to nitrogen fixation[300]
Ralstonia solanacearum	NZ_CP014703.1_1080	-	pathogen	cause bacterial wilt disease[301]
Ralstonia solanacearum	NZ_CP022482.1_1232	-	pathogen	cause bacterial wilt disease[301]
Rhizobacter sp. Root1221	NZ_LMDI01000045.1_268	root	PGPB	plant symbiotic bacteria[302]
Rhizobium sp. rho-6.2	NZ_SCFK01000035.1_2	-	pathogen	a causative agent of crown gall on rhododendron[303]
Rhizobium tumorigenes	NZ_PCDQ01000093.1_36	-	pathogen	isolated from cane gall tumors on thornless blackberry[273]
Rhizobium tumorigenes	NZ_PCDQ01000085.1_1	-	pathogen	isolated from cane gall tumors on thornless blackberry[273]
Rhodococcus fascians A44A	NZ_JMEX01000009.1_76	-	pathogen	initiate the le[304]
Rhodococcus fascians NBRC 12155 = LMG 3623	NZ_BCWW01000017.1_35	-	pathogen	initiate the leafy gall formation[304]
Rhodococcus sp. 06-462-5	NZ_NOYD01000054.1_35	-	pathogen	pathogenic isolates[166]
Rhodococcus sp. 14-2483-1-1	NZ_NPFX01000024.1_78	-	pathogen	pathogenic isolates[166]
Streptomyces antibioticus	NZ_KQ948422.1_1234	soil	PGPB	produce antibiotics[305]
Streptomyces aquilus	NZ_CP034463.1_8229	-	unknown	isolated from a chinese medical plant[306]
Streptomyces aureocirculatus	NZ_JOAP01000046.1_42	soil	unknown	-
Streptomyces bicolor	NZ_JOFS01000067.1_23	soil	PGPB	Streptomyces bicolor contains a fox gene cluster that encodes foxicin, a molecule that act as siderophore and has moderate antibiotic activity[307]
Streptomyces caeruleatus	NZ_KQ948929.1_499	tomato rhizosphere soil sample	PGPB	antifungal activities, reduced the root rot disease of tomato seedlings caused by phytopathogenic Fusarium species significantly[308]
Streptomyces cyaneus	NZ_RQJW01000016.1_154	-	PGPB	Streptomyces cyaneus CGMCC 4.1671 has antibiotic-encoding gene clusters[308]
Streptomyces europaeiscabiei	NZ_LRTR01000929.1_5	-	pathogen	isolated from common scab lesions[309]
Streptomyces fulvoviolaceus	NZ_JOEY01000020.1_101	-	PGPB	it is against Fusarium oxysporum, which are soil-born pathogens of plants[310]
Streptomyces geranii	NZ_PJME01000030.1_62	-	unknown	endophytic actinobacterium[311]

NCBI species name	Gene ID	Isolation source	Evidence of pgpb or pathogen	Justification
Streptomyces glauciniger	NZ_FZOF01000021.1_96	-	PGPB	mesophilic bacterium species; isolated from soil; WICC-A03 is a potential chitinase producer for biocontrol of plant pathogens; chitinase can degrade the chitin polymer in the cell wall of the fungal pathogen[312]
Streptomyces graminilatus	NZ_LIQQ01000089.1_10	Bamboo forest	unknown	isolated from bamboo litter[313]
Streptomyces hainanensis	NZ_SMKI01000157.1_4	-	unknown	isolated from soil in the island Hainan in China[314]
Streptomyces hainanensis	NZ_SMKI01000289.1_14	-	unknown	isolated from soil in the island Hainan in China[314]
Streptomyces ipomoeae	NZ_AEJC01000522.1_3	-	pathogen	cause soil rot disease of sweet potato[315]
Streptomyces milbemycinicus	NZ_MUNE01000113.1_6	-	PGPB	produces milbemycin used as insecticide and an anthelmintic[316]
Streptomyces nanshensis	NZ_LJGW01000377.1_72	-	unknown	-
Streptomyces nanshensis	NZ_LJGW01000252.1_26	-	unknown	-
Streptomyces regalis	NZ_LLZG01000411.1_109	-	PGPB	Streptomyces regalis NRRL 3151 contains a fox gene cluster that encodes foxicin, a molecule that act as siderophore and has moderate antibiotic activity[307]
Streptomyces rhizosphaericus	NZ_MUMC01000754.1_2	-	unknown	-
Streptomyces sp. 8K308	NZ_SMKC01000179.1_7	-	unknown	-
Streptomyces sp. cf124	NZ_FOUV01000016.1_167	-	unknown	isolated from the Populus root microbiome[63]
Streptomyces sp. cf386	NZ_FNHV01000013.1_22	-	PGPB	has a ifq gene cluster encodes for Isofuranonaphthoquinones. IFQ reported to have antibacterial activities[317]
Streptomyces sp. LUP47B	NZ_MJAH01000006.1_593	-	PGPB	promote the growth of the rhizobial partner in vitro and significantly increased nodulation and nitrogen fixation in lucerne plants[318]
Streptomyces sp. NBS 14/10	NZ_NCSM01000005.1_275	-	unknown	isolated from sediment river[319]
Streptomyces sp. NRRL WC-3618	NZ_LGDW01000086.1_5	soil	unknown	-
Streptomyces sp. WM6386	NZ_JXTE01000013.1_17	soil	unknown	-
Streptomyces turgidiscabies Car8	NZ_AEJB01000361.1_396	-	pathogen	produce thaxtomins that cause potato scab[320]
Streptomyces turgidiscabies Car8	NZ_AEJB01000023.1_14	-	pathogen	produce thaxtomins that cause potato scab[320]
Streptomyces viridochromogenes	NZ_LGUR01000002.1_359	-	PGPB	produces the antibiotic streptazolin[321]
Trichormus variabilis	NC_007413.1_2780	-	PGPB	nitrogen fixing bacteria[322]
Trinickia caryophylli	NZ_PNXZ01000006.1_136	-	pathogen	cause bacterial wilt of carnation[323]

NCBI species name	Gene ID	Isolation source	Evidence of pgpb or pathogen	Justification
Xanthomonas albilineans	NZ_MDCB01000005.1_57	-	pathogen	cause leaf scald in sugarcane[324]
Xanthomonas axonopodis	NZ_JFAQ01000171.1_2	-	pathogen	cause cirtrus canker[325]
Xanthomonas bromi	NZ_FLTX01000115.1_8	Bromus carinatus	pathogen	pathogen of bromegrasses[326]
Xanthomonas oryzae pv. oryzae	NZ_CP031697.1_3319	-	pathogen	rice bacterial blight pathogen[327]
Xanthomonas sp. GPE 39	NZ_JZHZ01000007.1_2404	water droplets on sugarcane leaves	unknown	close to sugarcan pathogen X. albilineans[324]
Xanthomonas theicola	NZ_MIGX01000097.1_15	-	pathogen	causal agent of bacterial canker on tea plants[328]
Xanthomonas translucens pv. translucens DSM 18974	NZ_CAPJ01000301.1_7	-	pathogen	cause disease in cereal crops and forage grasses[329]

Supplementary Table 4

Table 4: GTDB species name and NCBI species name of the IPT-containing genomes found in the GTDB database.

Genome accession (GTDB ID)	GTDB species name	NCBI species name
RS_GCF_000009705.1	Trichormus sp000009705	Nostoc sp. PCC 7120
RS_GCF_000016285.1	Allorhizobium vitis A	Agrobacterium vitis
RS_GCF_000092025.1	Agrobacterium fabrum	Agrobacterium fabrum
RS_GCF_000204075.1	Trichormus variabilis A	Trichormus variabilis
RS_GCF_000308835.1	Nocardia vinacea	Nocardia vinacea
RS_GCF_000317595.1	Streptomyces ipomoeae	Streptomyces ipomoeae
RS_GCF_000331005.1	Streptomyces turgidiscabies	Streptomyces turgidiscabies Car8
RS_GCF_000331775.1	Xanthomonas translucens	Xanthomonas translucens
RS_GCF_000332255.1	Chlorogloeopsis sp000332255	Cyanobacterium PCC 7702
RS_GCF_000365305.1	Dickeya dianthicola	Dickeya dianthicola
RS_GCF_000379565.1	Cupriavidus sp000379565	Cupriavidus sp. UYPR2.512
RS_GCF_000406105.1	Dickeya chrysanthemi	Dickeya chrysanthemi
RS_GCF_000406125.1	Dickeya sp000406125	Dickeya sp. NCPPB 569
RS_GCF_000409775.1	Halomonas A anticariensis	Halomonas anticariensis
RS_GCF_000426345.1	Cupriavidus sp000426345	Cupriavidus sp. amp6
RS_GCF_000510685.1	Sinorhizobium aridi	Ensifer aridi
RS_GCF_000519245.1	Paraburkholderia sp000519245	Burkholderia sp. UYPR1.413
RS_GCF_000601485.1	Chondromyces apiculatus	Chondromyces apiculatus
RS_GCF_000688655.1	Lonsdalea quercina	Lonsdalea quercina
RS_GCF_000688695.1	Lonsdalea quercina D	Lonsdalea quercina
RS_GCF_000696095.1	Rhizobium rhizogenes	Agrobacterium rhizogenes
RS_GCF_000718165.1	Streptomyces fulvoviolaceus	Streptomyces fulvoviolaceus
RS_GCF_000719285.1	Streptomyces bicolor	Streptomyces bicolor
RS_GCF_000720475.1	Streptomyces aureocirculatus	Streptomyces aureocirculatus
RS_GCF_000739935.1	Agrobacterium rubi	Agrobacterium rubi
RS_GCF_000749845.1	Pectobacterium betavasculorum	Pectobacterium betavasculorum
RS_GCF_000760735.1	Rhodococcus fascians E	Rhodococcus fascians
RS_GCF_000949895.1	Agrobacterium arsenijevicii	Agrobacterium arsenijevicii
RS_GCF_000955955.1	Lentzea aerocolonigenes A	Lechevalieria aerocolonigenes
RS_GCF_000963005.1	Xanthomonas sp000963005	Xanthomonas sp. GPE 39
RS_GCF_000972385.1	Streptomyces sp000972385	Streptomyces sp. WM6386
RS_GCF_000974445.1	Lentzea aerocolonigenes	Lechevalieria aerocolonigenes
RS_GCF_001189295.1	Chondromyces crocatus	Chondromyces crocatus
RS_GCF_001270495.1	Streptomyces viridochromogenes A	Streptomyces viridochromogenes
RS_GCF_001280005.1	Streptomyces sp001280005	Streptomyces sp. NRRL WC-3618
RS_GCF_001298575.1	Streptomyces sp001298575	Actinobacteria bacterium OK074
RS_GCF_001304695.1	Xanthomonas axonopodis	Xanthomonas axonopodis
RS_GCF_001418565.1	Streptomyces graminilatus	Streptomyces graminilatus
RS_GCF_001425785.1	Rhizobacter sp001425785	Rhizobacter sp. Root1221
RS_GCF_001509475.1	Streptomyces regalis	Streptomyces regalis
RS_GCF_001510895.1	Mesorhizobium loti A	Mesorhizobium loti
RS_GCF_001514065.1	Streptomyces antibioticus	Streptomyces antibioticus
RS_GCF_001514235.1	Streptomyces caeruleatus	Streptomyces caeruleatus
RS_GCF_001541315.1	Agrobacterium tumefaciens	Agrobacterium tumefaciens
RS_GCF_001541345.2	Allorhizobium vitis	Agrobacterium vitis
RS_GCF_001550375.1	Streptomyces europaeiscabiei	Streptomyces europaeiscabiei
RS_GCF_001586135.1	Ralstonia solanacearum A	Ralstonia solanacearum
RS_GCF_001612985.1	Nocardia gamkensis	Nocardia gamkensis

Genome accession (GTDB ID)	GTDB species name	NCBI species name
RS_GCF_001613305.1	Nocardia vaccinii	Nocardia vaccinii
RS_GCF_001746295.1	Streptomyces sp001746295	Streptomyces sp. LUP47B
RS_GCF_001751255.1	Streptomyces nanshensis A	Streptomyces nanshensis
RS_GCF_001758275.1	Allorhizobium vitis B	Agrobacterium vitis
RS_GCF_001894785.1	Rhodococcus fascians	Rhodococcus fascians
RS_GCF_001913135.1	Pseudomonas B psychrotolerans A	Pseudomonas psychrotolerans
RS_GCF_002005205.3	Agrobacterium rhizogenes A	Agrobacterium rhizogenes
RS_GCF_002111555.1	Lonsdalea populi	Lonsdalea populi
RS_GCF_002111585.1	Lonsdalea iberica	Lonsdalea iberica
RS_GCF_002111595.1	Lonsdalea britannica	Lonsdalea britannica
RS_GCF_002150765.1	Lentzea kentuckyensis	Lentzea kentuckyensis
RS_GCF_002154605.1	Streptomyces milbemycinicus	Streptomyces milbemycinicus
RS_GCF_002155885.1	Streptomyces rhizosphaericus	Streptomyces rhizosphaericus
RS_GCF_002224125.1	Streptomyces sp002224125	Streptomyces sp. NBS 14/10
RS_GCF_002258315.1	Rhodococcus sp002258315	Rhodococcus sp. 06-462-5
RS_GCF_002259335.1	Rhodococcus sp002259335	Rhodococcus sp. 14-2483-1-1
RS_GCF_002879875.1	Trinickia caryophylli	Paraburkholderia caryophylli
RS_GCF_002915175.1	Agrobacterium rosae	Agrobacterium rosae
RS_GCF_002939705.1	Xanthomonas albilineans	Xanthomonas albilineans
RS_GCF_002940605.1	Xanthomonas theicola	Xanthomonas theicola
RS_GCF_002943835.1	Agrobacterium deltaense	Agrobacterium tumefaciens
RS_GCF_002954775.1	Streptomyces geranii	Streptomyces sp. A301
RS_GCF_003240565.1	Rhizobium tumorigenes	Rhizobium tumorigenes
RS_GCF_003265345.1	Lentzea terrae	Lentzea terrae
RS_GCF_003269295.1	Lentzea atacamensis	Lechevalieria atacamensis
RS_GCF_003289965.1	Mesorhizobium sp003289965	Mesorhizobium sp. CNPSo 3140
RS_GCF_003597675.1	Neorhizobium sp003597675	Neorhizobium sp. NCHU2750
RS_GCF_003955715.1	Streptomyces sp900206255	Streptomyces sp. GGCR-6
RS_GCF_003999715.1	Ralstonia pseudosolanacearum	Ralstonia solanacearum
RS_GCF_004028635.1	Streptomyces cyaneus	Streptomyces cyaneus
RS_GCF_004136375.1	Xanthomonas oryzae	Xanthomonas oryzae
RS_GCF_004348315.1	Streptomyces sp004348315	Streptomyces sp. 8K308
RS_GCF_004348415.1	Streptomyces hainanensis	Streptomyces hainanensis
RS_GCF_004348675.1	Micromonospora sp004348675	Micromonospora sp. KC606
RS_GCF_005145045.1	Agrobacterium larrymoorei	Agrobacterium larrymoorei
RS_GCF_005221325.1	Agrobacterium sp900012625	Agrobacterium tumefaciens
RS_GCF_007000325.1	Rhizobium sp007000325	Rhizobium sp. rho-6.2
RS_GCF_007002745.1	Agrobacterium rhizogenes B	Agrobacterium rhizogenes
RS_GCF_900092025.1	Xanthomonas bromi	Xanthomonas bromi
RS_GCF_900103585.1	Streptomyces sp900103585	Streptomyces sp. cf386
RS_GCF_900114955.1	Streptomyces sp900114955	Streptomyces sp. cf124
RS_GCF_900188405.1	Streptomyces B glauciniger	Streptomyces glauciniger

Supplementary Table 5

Table 5: Pfam annotations detected five or more than five times in the IPT gene neighborhoods.

Pfam IDs	**Frequency**
PF01745	90-99
PF00005	50-59
PF03641; PF02027	40-49
PF00106; PF01565; PF00753; PF00581; PF09265; PF00528; PF13540; PF08352	20-29
PF00664; PF08240; PF00067; PF01425; PF01032; PF00440; PF00107; PF07690; PF02317; PF01593; PF01548; PF00698; PF00293; PF02583; PF02371; PF13847; PF11127; PF07992; PF01527; PF01210	10-19
PF04324; PF02779; PF02518; PF01925; PF00392; PF00155; PF13561; PF12680; PF03734; PF01497; PF00512; PF00496; PF00415; PF00296; PF13788; PF13556; PF12833; PF12697; PF03009; PF00227; PF13535; PF13459; PF12911; PF11575; PF04055; PF03050; PF02729; PF00456; PF00185; PF00072; PF13817; PF13419; PF13385; PF12802; PF12682; PF12681; PF08305; PF07729; PF06903; PF05957; PF05717; PF05237; PF03466; PF03454; PF03453; PF03358; PF02983; PF02929; PF02837; PF02836; PF02310; PF01679; PF01483; PF01406; PF01266; PF01145; PF01019; PF00994; PF00899; PF00881; PF00781; PF00703; PF00672; PF00665; PF00656; PF00501; PF00126; PF00082; PF00070	5-9