INFERRING ORTHOLOGOUS RELATIONSHIPS AND GENE TRANSFER IN MICROBIAL GENOMES AND METAGENOMES

by

Dennis H.-J.Wong

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
April 2018

DEDICATION PAGE

To Susan, Alexander and Oliver, thank you for your love and companionship.

To my parents and siblings, thank you for your unending support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Interest in microbial life and the progress of DNA sequencing technology has led to thousands of sequenced bacterial genomes. In this thesis I develop approaches to identify Lateral Gene Transfer (LGT) in metagenomes, develop fast sequence clustering approaches to create clusters necessary in comparative genomics analyses, and apply them to large data sets.

In chapter two, I identify LGT in two of three metagenomes of phosphorus-removing bacteria in sewage-treatment plants, none in a United States of America community, two in a Danish community and five in an Australian community. Analyses account for the limitations of metagenomic sequence data and focus on gene transfers in energy-related metabolic pathways. These transfers impact pathways associated with the different input carbon feeds for each community, suggesting recent adaptation among community members. This is the first published analysis focusing on the role and direction of transferred genes in a community using metagenomes. In chapter three, I develop two methods to define and refine clusters of homologous sequences from sequenced genomes: ProPhylClust to identify large protein families, and PhyloSubClust to subcluster large protein families based on phylogeny to recover orthologous relationships. ProPhylClust uses a species phylogeny as a guide tree for runtimes with approximately linear scaling relative to the runtimes of all-versus-all homology-search methods that scale quadratically with increasing numbers of genomes. Two different sets of genomes were used, one spanning 24 bacterial phyla and the other sampled from the phylum Proteobacteria. While the sequence comparisons in ProPhylClust make it slower than competing approaches on small genome sets, the hierarchical approach of ProPhyClust yielded equal or faster runtimes on sets with 100 or more genomes. In chapter four, 558 incomplete and complete genomes from the class Clostridia were clustered using ProPhylClust and PhyloSubClust. Of 18 clusters containing toxin proteins and their regulators from Peptoclostridium difficile (toxins A/B), Clostridium botulinum (botulinum toxin) and Clostridium tetani (tetanus toxin), one botulinum-tetani toxin cluster and a toxin A/B cluster, revealed homologous sequences considered non-toxic. Hierarchical clustering of phylogenetic profiles identified potentially toxin-related protein families with unknown function located on the same sequence contig or chromosome, but not in toxin operons.

The computational analysis of large genomic data sets to derive biologically relevant knowledge will continue to be a challenge for years to come. Here, I focused on computational methods relevant to identifying LGT in environmental sequence data, constructing clusters of homologous sequences from genomes, and obtaining functionally associated sequences based on phylogenetic distributions. Promising results were produced for each chapter, with gene transfer events found in phosphorus removing sewage treatment communities, runtimes for cluster construction that are more manageable than other methods with larger data sets, and sequences that possibly are functionally relevant to toxins in *C. botulinum* and *P. difficile*.

LIST OF ABBREVIATIONS AND SYMBOLS USED

ACLAME	A CLAssification of Mobile genetic Elements
AU	Australia
BeT	Best hit
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOcks SUbstitution Matrix
BoNT	Botulinum Neurotoxin
BM	Butanoate Metabolism
CAC	Citric Acid Cycle
CAP	*Candidatus* Accumulibacter phosphatis
COG	Clusters of Orthologous Groups
DIAMOND	Double Index Alignment of Next-Generation Sequencing Data
DK	Denmark
DNA	Deoxyribonucleic Acid
EBPR	Enhanced Biological Phosphorous Removal
EC	Enzyme Commission
eggNOG	evolutionary genealogy of genes: Non-supervised Orthologous Groups
E-value	Expectation Value
GB	Gigabyte
GG	Glycolysis/Gluconeogenesis
GTA	Gene Transfer Agent
HA	Hemagglutinin
HMM	Hidden Markov Model
KEGG	Kyoto Encyclopedia of Genes and Genomes
LG	Le and Gascuel
LGT	Lateral Gene Transfer
MCL	Markov Cluster Algorithm
MGE	Mobile Genetic Element
MSV	Multiple Ungapped Segment Viterbi
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
MySQL	My Structured Query Language
NCBI	National Center for Biotechnological Information
NM	Nitrogen Metabolism
ntnh	Non-toxic Non-haemagglutin
ORF	Open Reading Frame
OS	Operating System
PHA	Poly-b-hydroxyalkanoates
PaLoc	Pathogenicity Locus
PAO	Polyphosphate Accumulating Organism
Pfam	Protein Families (database)
PHOGs	Phylogenetic Orthologous Groups
PM	Propanoate Metabolism
PPP	Pentose Phosphate Pathway
PSI-BLAST	Position-specific Iterative Basic Local Alignment Search Tool

RAxML	Randomized Axelerated Maximum Likelihood
RAM	Random Access Memory
RAPsearch2	Reduced Alphabet Based Protein Similarity Search
RefSeq GI	Reference Sequence GenInfo Identifier
RBH	Reciprocal Best Hit
RDP	Ribosomal Database Project
RITA	Rapid Identification of Taxonomic Assignments
rRNA	Ribosomal Ribonucleic Acid
RNA	Ribonucleic Acid
SH	Shimodaira-Hasegawa
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
USA	United States of America
WAG	Whelan Goldman
▶	P. difficile 630
⯃	P. difficile R20291
■	P. difficile BI1
◆	P. difficile CD196
★	Node has a bootstrap node support of 65-90%
⯃	Node has a support value of 90-100%
X	All-versus-all BLAST searche node for ProPhylClust
★	Sequence-versus-cluster (HMM and/or consensus) node for ProPhylClust
♦	Cluster-versus-cluster searche (HMM and/or consensus) node for ProPhylClust

ACKNOWLEDGEMENTS

I would like to thank the following people and institutions. My committee members: Dr. Joeseph Bielawski, Dr. Christian Blouin, Dr. Norbert Zeh for their guidance. My supervisor, Dr. Robert Beiko for his patience, guidance and support. Special thanks to my external Dr. Gabriel Moreno-Hagelsieb for his time. Dalhousie University for providing an academic environment. The Tula Foundation, Genome Canada, The Centre for Comparative Genomics and Evolutionary Bioinformatics, and the Natural Sciences and Engineering Research Council of Canada for funding. Members, both past and present, of the Beiko lab, Blouin lab and Zeh lab for their academic input and stimulating conversations.

CHAPTER 1 INTRODUCTION

1.1 DNA AND MICROBES

It is not incorrect to say it is a microbial world. Microorganisms, such as single-celled eukaryotes, Bacteria, Archaea and viruses are found in a wide variety of environments, for example, in the air we breathe (e.g. Shaffer and Lighthart 1997), in and on our bodies (e.g. Costello et al. 2009), the ocean (e.g. DeLong and Karl 2005), and in what we consider to be inhospitable environments (e.g. Rothschild and Mancinelli 2001). Microbes exist in communities, where they differ by type and abundance and interact with each other and their environment. Microbial communities are extremely diverse and are often more diverse than communities of multicellular organisms by several orders of magnitude (Haegman et al. 2013). The term "microbes" typically refers to microscopic Bacteria, Eukaryotes and Archaea. From this point on in this thesis, I will focus on Bacteria, and when I refer to microbes, I am referring to Bacteria. We often focus our attention on individual species, or strains of species, since they may play an important role in anthropomorphically valuable enterprises, such as a positive or negative impact on human health or a metabolic function that can clean a pollutant from an environment. The progress and convergence of deoxyribonucleic acid (DNA) sequencing technology, computer science, and laboratory and molecular techniques have facilitated a massive expansion of our knowledge about the diversity of types and the roles played by microbial life in their environments.

Biological information is encoded in genomes, the complete set of DNA in a living organism, where it is organised into chromosomes and inherited from ancestors to descendants. DNA is coded in a series of four different nucleotide bases: adenine (A) and guanine (G), thymine (T) and cytosine (C). A genome encodes the information needed to build and maintain an organism, and genes, sequences of DNA in the genome, have functional significance for an organism. The first genome of a free-living organism, the bacterium *Haemophilus influenzae* was sequenced in the mid 1990's (Fleischmann et al. 1995). As of August 12th, 2017, there were 102,271 bacterial genome projects in the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov), where the number of nucleotide bases, sequences, and genomes increased, and should continue to increase (Figure 1.1, see also e.g. Wetterstrand 2016). As more genomes have been sequenced, novel lineages of bacteria are being discovered, where some of those lineages have defined new phyla (e.g. Tamaki et al. 2011). This ever-expanding volume of data presents opportunities to enlighten our understanding of Bacteria, but also introduces substantial computational challenges that need to be addressed (e.g. Muir et al. 2016, Schatz et al. 2010, Wall et al. 2010).

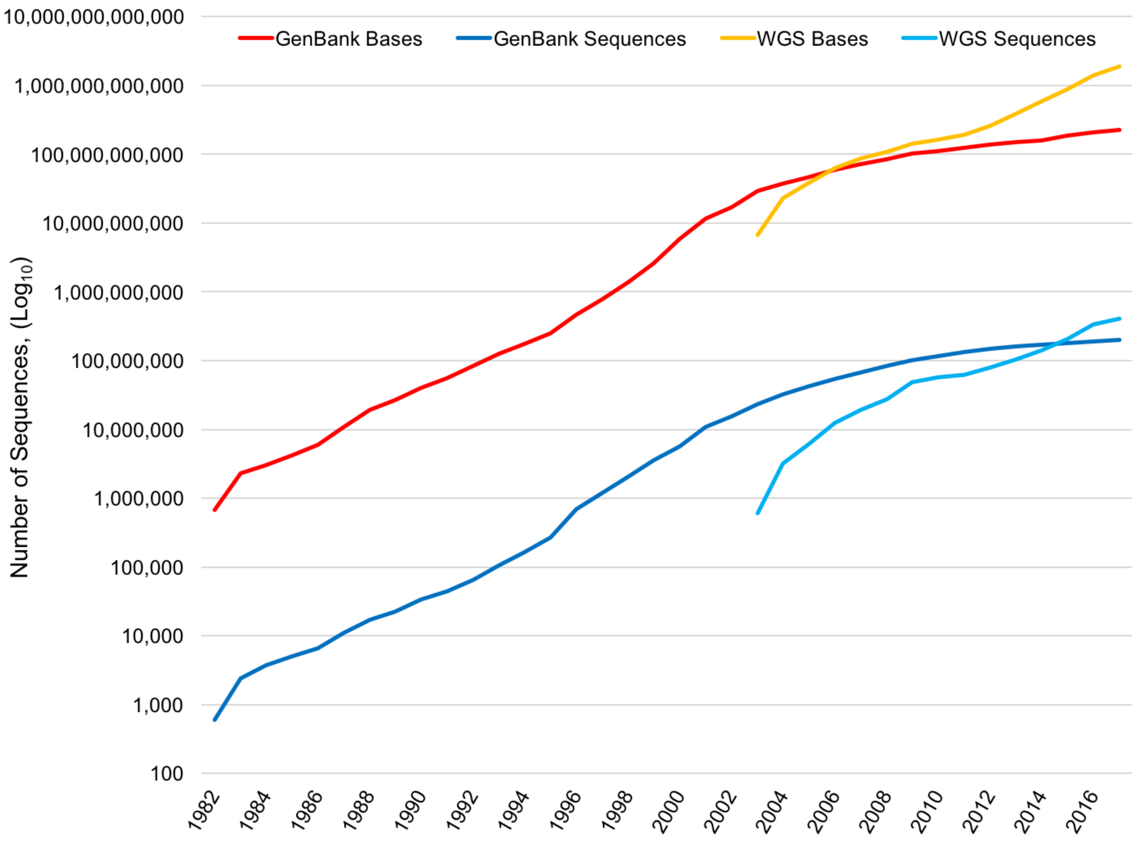


Figure 1.1. Growth of the National Center for Biotechnology and Information's DNA sequence databases from 1982 to 2017, with data downloaded from NCBI. Bases is total number of nucleotides, and sequences is a contiguous string of sequence. Whole Genome Sequencing (WGS) refers to nucleotide bases or contiguous sequences from sequenced genomes.

In this thesis, I develop and apply methods to identify biologically relevant processes in bacterial communities and develop an algorithm to decrease the computational time to create clusters of homologous sequences, sequences which descend from a common ancestor. In chapter two of this thesis, I present techniques to identify how communities of Bacteria share genes in a specific environment. This is the first published study to identify the specific role of genes transferred between organisms in multiple environments using environmental shotgun sequence (i.e. metagenomic) data. In chapter three, I introduce an algorithm to cluster sequences from bacterial genomes, ProPhylClust, that can appreciably decrease runtimes for large data sets, and compare the results of the algorithm to other algorithms that are often used to cluster genomes. In chapter four, I use ProPhylClust to cluster sequences from a large number of *Clostridia* genomes, a diverse class of Bacteria with a wide range of roles in environments, perhaps best known as pathogens in humans. I explore the taxonomic and phylogenetic distribution of genes related to virulence and identify sequence clusters with hypothetical function associated to toxin sequences.

1.2 MICROBES IN THEIR ENVIRONMENT

The diversity of microbes is vast and difficult to determine, with estimates considered to be much higher than what can be currently measured (e.g. Haegeman, 2013, Quince et al. 2008, Whitman et al. 1998). Microbes are important contributors to global biogeochemical cycles (e.g. Arrigo 2005, Rousk et al. 2014) and drive plant diversity and productivity (e.g. Heijden et al. 2007). Human enterprises such as healthcare, bioremediation, and waste treatment are also heavily influenced by microbes. The sequencing of genomes has revolutionized our understanding of bacteria, and with improved technologies making sequencing faster, more affordable, and accurate, interest in their roles will continue.

Specific bacterial species are known to have important impacts on human health. *Peptoclostridium dfficile* (also known as *Clostridium difficile* or *Clostridioides difficile*) is a bacterium often associated with antibiotic resistance, where an antibiotic-resistant *P. difficile* gastrointestinal tract infection causes various gastrointestinal symptoms. Bacteria are not just pathogens. As commensal organisms, they can also be beneficial for human health. Oral application of non-pathogenic strains of *Escherichia coli* to pre-mature infants has been shown to increase the number of generalized antibodies for bacteria, which would enhance immune response in the event of a bacterial infection (Conway and Cohen 2015, Cukrowska et al. 2002). In mutualistic relationships, the human body acts as a stable environment for microbes, and in return microbes provide nutrients for absorption by the host (e.g. Bäckhed et al. 2005). *Bacteroides thetaiotaomicron* can be found in the human gastrointestinal tract and is well known for its ability to degrade carbohydrates which can then be used by its host, and its adaptations that allow it to survive in the human gastrointestinal tract (Xu et al. 2003).

More-recent studies of human-associated bacterial communities, the human microbiome, are starting to highlight the importance of communities rather than individual species of Bacteria in human health. For example, twin studies have shown that low-diversity gastrointestinal communities are associated with obesity (Turnbaugh et al. 2008). An increasing body of research is also starting to highlight the importance of the newborn and infant microbiome in relation to development (e.g. Tamburini et al. 2016).

Microbial communities are also important for remediation of contaminated environments, where they can utilize pollutants in biochemical reactions (metabolism) for growth and reproduction. For example, sites that are contaminated with hydrocarbons (i.e. petroleum) are sometimes inhabited by a variety of strains and species of Bacteria that are capable of metabolizing toxic substances in polluted ground water (Dojka et al. 1998, Holmes et al. 2006). These bacteria have been harnessed for use in bioremediation efforts at other polluted sites where they do not occur naturally (Hood et al. 2008). Microbial metabolism has also been harnessed in sewage treatment. Communities of various microbes are assembled and maintained in engineered environments, known as bioreactors, to treat sewage (Seviour et al. 2003). Like all engineered systems, sewage-treatment systems can fail, and understanding them will make them more reliable.

Furthering our knowledge about how microbes function in their environments (ecology), and how they change through time in the environments they inhabit (evolution) is a necessary step in utilizing them for human enterprises. DNA sequence information provides considerable biological data to study known and unknown members of bacterial communities.

1.3 GENOMICS

1.3.1 Genes, Genomes and Proteins

A bacterial genome is comprised of protein-coding and non-protein-coding regions; gene sequences are continuous strings of nucleotides in coding regions that are basic units of heredity and encode proteins. The percentage of a genome that codes for protein content can vary from 40 to 97%, with an average of 88% (Land et al. 2015). Some genes are important for basic cellular function. For example, ribosomal genes encode components of the ribosome, a molecular complex that is used for the translation of RNA sequence to protein sequences. Ribosomal genes are highly conserved, with low rates of sequence change: if mutation of the ribosomal gene resulted in change of the structure of the ribosome, translation of RNA sequences to protein sequences would most likely fail and would be fatal to an organism. This high degree of conservation makes ribosomal gene sequences useful for taxonomic purposes, and the 16S ribosomal RNA is a gene often used for the classification of Bacteria (Woese et al. 1990).

Translation of genes into proteins is a dynamic process, as not all genes are actively translated and transcribed at all times;in many cases, gene expression can depend on the microbe's interaction with their environment. The differential translation and transcription of genes to proteins result in characteristics and traits known as phenotype. For example, in sewage treatment plants, *Candidatus* “Accumilibacter phosphatis” (CAP) is typically a phosphate-metabolizing organism with some strains also being able to metabolize nitrogen (e.g. Carvalho et al. 2007, Oehman et al. 2010). The nitrogen-metabolizing phenotype becomes more prevalent depending on the amount of nitrogen and phosphorous in the sewage-treatment environment, where genes associated with nitrogen metabolism are more actively expressed and then translated into proteins for nitrogen metabolism (He et al. 2010, Wilmes et a. 2008). Proteins therefore have considerable functional significance for an organism and their analysis can provide insights into the potential metabolism and function in an environment an organism may have.

Determining protein function is a non-trivial task, and while biochemical experimentation is the gold standard, but given the overwhelming volume of available sequence information creates a need for automated functional annotation. Protein sequences can comprise one or more domains, which are functional units within a protein that contribute to the function of the whole protein. Alternatively, “moonlighting” proteins have mutilple functions in various parts of a cell (Henderson and Martin 2011, Jeffery 1999). Often, several different proteins are required to perform a function. An operon is a set of genes clustered together in a genome such that they are translated together under the control of single a promoter sequence. Function can be a difficult to describe, as it can be dependent upon the context in which it is used. A protein may participate in multiple metabolic reactions in different parts of a cell, with those reactions contributing to different cellular processes. Hierarchical classification schemes have been developed to allow for a more comprehensive functional description of a protein (e.g. Ashburner et al. 2000, Tatusov et al. 1997, Kanehisa and Goto 2000). For many classification schemes, the classification of a gene or protein sequence can belong to multiple parents in the hierarchy. One such hierarchical classification scheme is Clusters of Orthologous Groups (COG), one of the first databases that group and functionally classify protein sequences. COG is a dual-tiered classification scheme with 23 functional categories constituting the top level, and the lower level is composed of clusters of sequences that are considered to be homologs, each with a worded functional description (Tatusov et al. 1997). Some COG clusters belong to more than one of the 23 functional categories, which speaks to the difficulty in creating concise and simple functional classifications.

1.3.2 Sequence Homology and its Relationship to Function

Entities that share common ancestry are said to be homologous. At the DNA and protein level, sequence similarity is often used to infer homology. Despite the difficulties in establishing clear definitions of homology and the various types of homologous sequences, such as orthologs and paralogs (Fitch 2000), the identification of homologous sequences is a prerequisite to understanding the evolutionary relationships among genes and genomes, and can be used to infer function..

Genes are orthologs if they diverge at speciation events (Fitch 1970). Paralogs result from a sequence duplication event (Fitch 1970), where the complete sequence of a gene is copied. Once a duplication event occurs the daughter and the original sequences continue to evolve on parallel paths, which allows the paralogous sequence the opportunity to diverge from the parent sequence. If both paralogs and orthologs of a gene are included in a phylogenetic analysis, the phylogeny will reflect the evolutionary history of the gene, but may not reflect the phylogeny of the organisms the genes belong to. Absent of artefacts of phylogenetic inference, the phylogeny of orthologous sequences should share the same phylogeny as the phylogeny of the species (Fitch 1970). However, paralogous sequences may not share the same branching structure as the phylogeny depending on the number and timing of duplication events. Defining orthology and paralogy is therefore an important distinction in understanding the evolutionary history of genes.

Some have suggested expansion and refinement of the ortholog and paralog definitions (e.g. Gogarten 1994, Koonin 2005). Two definitions related to the timing of duplication events are useful for delineating types of paralogs. An inparalog refers to a duplication event that occurred within a lineage subsequent to any identified speciation events, whereas an outparalog is a duplication event preceding a speciation event (Remm et al. 2001, Sonnhammer and Koonin 2002).

Although the functions of DNA and protein sequences are not part of the definition of homology,sequence similarity (and the corresponding implied homologous relationship) is often used to infer functional similarity. The ortholog conjecture states orthologs tend to exhibit more-strongly conserved functions than do paralogs (Koonin 2005), but he degree to which this is true is still a debated topic (e.g. Altenhoff et al. 2012, Chen and Zhang 2012, Dunn et al. 2018). Functional changes are often considered to be more likely with paralogs since the original gene (i.e. ortholog) retains its function and functional conservation of the paralog is no longer required. Research suggests orthologs generally have higher functional conservation than paralogs (e.g. Attenhoff et al. 2012, Fang et al. 2010, Forslund et al. 2011). Due to timing of duplication, it is assumed that the resulting inparalogs have higher sequence conservation and both copies may still retain their original function (e.g. Forslund et al. 2011, Notebaart et al. 2005). Since the duplication event took place outside of the current lineage, outparalogs are assumed to have lower sequence conservation and more likely to have functionally diverged (e.g. Forslund et al. 2011).

1.3.3 Metagenomics

Genome sequencing historically depended on the availability of single organisms in pure culture. However, pure culture is not an accurate representation of the lifestyle of Bacteria. Culturing bacteria is a labour-intensive process, and recent efforts have been made to sequence bacteria in their environments, without the need for pure culture. Metagenomics (Handelsman et al. 1998), is culture-independent sequencing of DNA from a population of microorganisms in an environmental sample. A metagenome is typically obtained by isolating microorganisms from an environmental sample followed by extraction and fragmentation of DNA from organisms into short strands, followed by DNA sequencing, a process known as shotgun sequencing. The result after sequencing is many sequence "reads", which can be ten to several hundred nucleotides long, depending on the sequencing technology used. The number of times a nucleotide from a sequence is represented in a number of overlapping reads is known as coverage or depth, and when coverage is high, the assembly of reads into contiguous sequences produces contigs of high confidence. If coverage is low, reads may be left unassembled.

From the processed sequence data, or from the sequenced reads, it is possible to do further analysis of metagenome sequences, such as the identification and annotation of genes or taxonomic assignment. Complete genes are predicted in contigs, while incomplete genes can be inferred from a start codon for an open reading frame (ORF). Genes can then be conceptually translated into a predicted protein using the genetic code table and assigned a functional annotation by searching reference databases for homologous (and potentially functionally related) protein sequences. From fully annotated sequence data, one can for example, estimate known and unknown microbial diversity and establish functional roles of taxa in their environment (e.g. Albertsen 2011, Biers et al. 2009, Brazelton and Baross 2009, Delmont et al. 2011, Jovel et al. 2016, Tyson et al. 2004, Zhang et al. 2016).

Metagenomics has sampled uncultured microbes and have expanded our knowledge about the vast diversity of microbial life. Metagenome sequencing of seven samples from the Sargasso Sea identified 1800 species, of which 148 were new to science. Approximately 1.2 million genes (Venter et al. 2004) were sequenced, which at the time was greater than 1/40th the total number of genes in public databases at the time. High genetic variation within the same species has also been revealed by metagenome sequencing of communities such as acid mine drainage (Tyson et al. 2004), cyanobacterial algal blooms (Steffen et al. 2012), enhanced biological phosphorous removal (EBPR) communities (Albertsen et al. 2011, García Martín et al. 2006), and a dechlorinating microbial culture known as KB-1 (Hug et al. 2012). Fully sequenced or draft genomes can be obtained from metagenomes, usually with additional sequencing effort (e.g. García Martín et al. 2006, Kantor et al. 2013, Tyson et al. 2004). However, recent advances in assembly algorithms have dramatically increased the ability to reconstruct draft genomes from metagenomes (Brown 2015, Parks et al. 2017). These studies suggest what can be discovered from sequenced metagenomes, but also illustrate the need for continued improvement of bioinformatic analysis of large metagenomic data sets due to increased computational requirements necessary to process such data sets, and the desire to obtain more biologically relevant information from metagenomes.

1.3.4 Genome Evolution

The genome of an organism is a dynamic entity, where nucleotide sequence, presence or absence of genes, and the order of genes in an organism is constantly changing through time as it is inherited from parent to child. Mutation is a change in the nucleotide sequence of an organism's genome, and can be caused by a wide variety of mechanisms such as nucleotide substitution, gene recombination, gene duplication, gene insertion, gene deletion, gene fusion and fission, and movement of genes within and between genomes. Point mutations are single changes in the DNA sequence of a genome, change from one nucleotide base to another, or insertions or deletions of single nucleotides. Point mutations can be either synonymous or non-synonymous. Synonymous mutations do not change protein sequences, while non-synonymous mutations change protein sequences and can therefore change protein structure, which can be advantageous or disadvantageous to an organism.

Duplication is a type of mutation in which a region of genomic sequence is copied more than once, resulting in two or more copies of the region in the descendant sequence. The region implicated in a duplication event can range in size from a few nucleotides to one or more genes, in some cases encompassing an entire genome (Wolf and Shields 1997). Duplicated sequences within a gene can modify the function of the encoded protein, while duplication resulting in multiple copies of a gene (ie. paralogs) could allow for additional copies of proteins to be produced (e.g. Kugelberg et al. 2006). Changes in genes can result in novel function, and paralogs are theorized to enable the development of genes with new functions, since the original copy still performs the original function, the paralog can proceed on an independent path (Zhang 2003).

Fusion is where genes join, while fission is the splitting of genes, and both contribute to the evolution of multidomain proteins (Pasek et al. 2006). Fusion events allow the physical coupling of proteins that are biologically coupled (Marcotte et al 1999). Fission events are speculated to be an advantage to thermophilic organisms, where shorter genes are less likely to gain errors during DNA replication, and the split proteins form complexes in the event where multiple proteins are required for a function (Snel et al. 2000).

Genetic material often moves within and between genomes. Mobile genetic elements (MGEs) are DNA segments that can move within or between genomes. Some types of MGEs include plasmids, viruses, transposable elements, and genomic islands (e.g. Dobrindt et al. 2004, Binnewies et al 2006). Plasmids are usually circular, self-replicating extra-chromosomal DNA molecules. Viruses are small infectious agents composed of genetic material (DNA or RNA) with a protein coat that can replicate inside other organisms. Transposable elements are DNA that can move locations within a genome. Genomic islands are large chromosomal regions that have flexible gene content usually related to adaptability and versatility of the bacterium. Genes found in genomic islands often contain genes that confer advantages in an environment, such as metabolism, antibiotic resistance, pathogenicity, and symbiosis (Dobrindt et al. 2004). Interestingly, genomic islands have inconsistent taxonomic distribution, often present in some strains of a species, and absent in other strains, suggesting they can be excised from a genome (Langille et al. 2010). When the recipient genome is not a child of the donor, the process is known as horizontal gene transfer (HGT) or lateral gene transfer (LGT). This is in contrast to vertical inheritance, where the recipient genome is a child of the donor. It is believed MGEs are major agents of LGT events (Frost et al. 2005). Both full genes and gene fragments can be transferred (Chan et al. 2009), and even sets of genes such as operons can be transferred (e.g. Mussmann et al. 2005, Pál et al. 2005). LGT has been shown to occur in situ (e.g. Graham and Istock 1978, Jiang and Paul 1998) and can confer a variety of new functions such as antibiotic resistance (Akiba et al. 1960, Ochiai et al. 1959), components for novel metabolism (Springael and Top 2004), and the ability to move substances in or out of the cell through transmembrane transporters (Gefland and Rodionov 2008).

Bacteria occupy a wide variety of environments, each with different community compositions and ecology, it should be expected that a wide variety of mechanisms of LGT exist. Transformation is the uptake of naked DNA from the environment and is known to occur in a wide variety of bacteria (Johnsborg et al. 2007). Bacteriophages are viruses that infect bacteria. They often have a specific host range, but they occasionally can shift hosts (Hyman et al. 2010). Transduction is where bacterial DNA is moved from bacteria to bacteria by a virus, and is known to occur frequently in a wide variety of environments (Miller 2001). Conjugation is the transfer of genetic material, such as a plasmid, through cell-to-cell contact and has been observed in only a limited set of organisms (Claverys et al. 2009), but conjugation genes are widespread in specific taxonomic groups (Weinert et al. 2009). Gene transfer agents (GTAs) are virus-like particles that contain random, short pieces of the genome of the producing cell and have been observed in marine habitats, but it is unclear what selective advantages are conferred by GTAs (Marrs 1974, Lang and Beatty 2007).

It has been estimated that anywhere between 2% to 100% of all genes have been transferred at least once in their history (Dagan and Martin 2007). Evidence suggests LGT events between distantly related prokaryotic groups are infrequent, and unevenly distributed (e.g. Andam and Gogarten 2011, Beiko et al. 2005, Boucher et al. 2003, Tamminen et al. 2012), but specific taxonomic groups such as class *Clostridia* tend to be more promiscuous participants in LGT (Beiko et al. 2005, Meehan and Beiko 2014). Lateral gene transfers on shorter evolutionary time scales are more likely between close relatives (Andam and Gogarten, 2011) and the mechanism of LGT may also be taxonomically dependent. For example, GTAs have only been identified within the class *Alphaproteobacteria* (Lang and Beatty 2007) and several members of the phylum *Spirochaetes* (Hampson and Ahmed 2011). Lateral gene transfer could also be environment-dependent, and driven by the ecology of specific environments (e.g. Rhodes et al. 2011, Smillie et al. 2011, Wiedenbeck and Cohen 2011). Indeed, the possible combined scenarios for mechanisms of LGT with environment and taxonomy is potentially very high. However, for transferred DNA to persist after insertion into the recipient genome, the survival of the recipient organism must not be negatively impacted, or it will be outcompeted by the non-mutated members of the population. It may be the case that not all transfers are equal from the perspective of major evolutionary events. For example, genes related to carbohydrate metabolism have a higher tendency to be transferred than genes related to the processing and modification of RNA (Cohen et al. 2011).

1.4 Comparative Genomics of Bacteria

Comparative genomics is the study of genomic features across multiple genomes. Due to the number of available sequenced genomes, comparative genomics studies are now commonplace, and have led to discoveries about the diversity, evolution, physiology, pathogenicity and ecology of Bacteria. For example, *Heliobacter pylori* is the causal agent of gastric ulcers, and is found in 20-50% of adults in industrialized nations (Suerbaum and Michetti 2002). It has been believed to have spread beyond east Africa 58,000 years ago with genetic diversity decreasing as geographic distance from east Africa increases (Linz et al. 2007). Genome changes have been identified in *H. pylori* during early infection (Colbeck et al. 2006) and during the progression of mild disease symptoms to cancer, genes are gained, while other genes are lost (Oh et al. 2006).

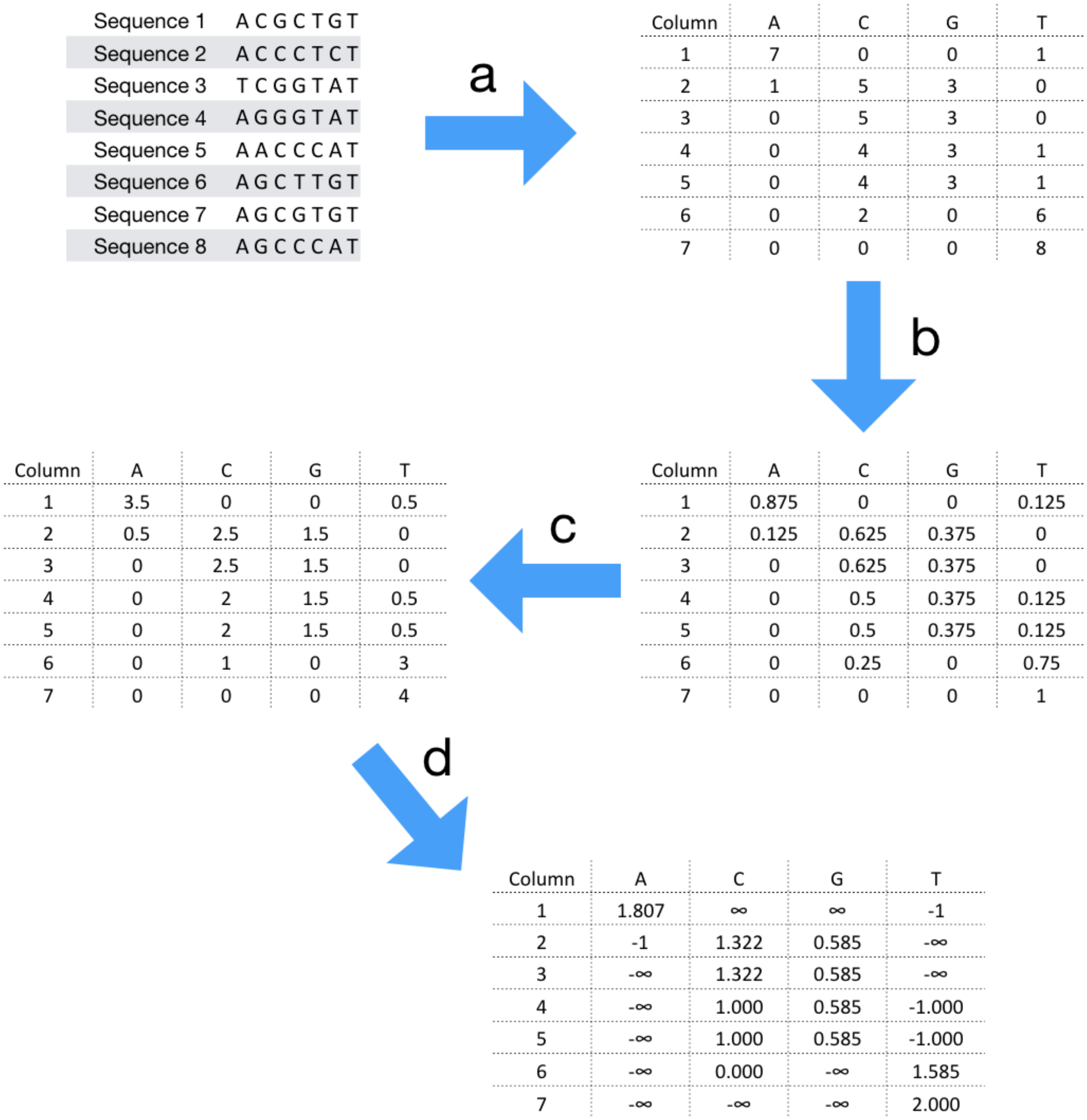
Discoveries in comparative genomics have also challenged our understanding of how we should classify Bacteria, since the diversity of genome content among closely related bacteria may not intuitively suggest they are close relatives. For example, *Escherichia coli*, a common mammalian intestinal bacterium, can be pathogen or mutualist in their human hosts (Savageau, 1983, van Elsas et al. 2011). An analysis of 61 *E. coli* genomes and the closely related, pathogenic genus *Shigella* reveals only 20% of genes are shared across all of their genomes, with the other 80% being variable (Lukjancenko et al. 2010). These variable sections of *E. coli* and *Shigella* genomes tend to be associated with genomic island. However, despite these apparent genomic differences, members of *Shigella* are not clearly delineated from *E. coli* based on 16S rRNA, or other genes that can be used to classify Bacteria (Lukjancenko et al. 2011). Genes are obviously important to the classification and understanding of the roles of Bacteria, and can be classified themselves, either through function or through evolutionary relationships with other genes.

1.4.1 Homology-Search Algorithms

Homology-search algorithms align meaningful regions of a query gene or protein sequence to subject sequences and calculate the statistical significance of those alignments. An alignment that falls within a threshold statistical value are assumed to be homologous to a query sequence. Sequence alignments can be either global (complete sequence) or local (subsequence). Substitution matrices describe the probability at which sequence character states change over time and are used to score the quality of sequence alignments. For a substitution matrix, each row and column represents a nucleotide or amino acid, and each cell in the matrix is a log-odds score that represents the probability of substitution between any two nucleotides or any two amino acids. Substitution matrices are obtained from high-quality sequence alignments from which log-odds are calculated (Henikoff & Henikoff 1992). Different matrices, based on reference alignments with different degrees of divergence, are used based on the degree of divergence between sequences in the alignment, for example the BLOcks SUbstitution Matrix (BLOSUM; Henikoff & Henikoff 1992) comes in several versions based on minimum percent amino acid identity of reference sequences: 45%, 62% and 80% are most commonly used. Using a given substitution matrix, homologous sequences are found by computing local alignments. Several algorithmic approaches are commonly used. Smith-Waterman (Pearson 1995, Shpaer et al. 1996) is a local alignment algorithm that will always find the optimal alignment given a particular soring scheme, but has runtimes that can be prohibitive for large data sets. To decrease runtimes of homology searches for large data sets, heuristics are often employed.

The BLAST algorithm (Altschul et al. 1990) is a frequently used local alignment heuristic in to identify homologous sequences. The BLAST algorithm finds initial matches between a query sequence and potential matches in a reference database by finding seeds, highly similar short matching words, to limit the number of possible alignments that need to be calculated. To become a seed, a threshold value for the quality of each initial word match between the query and subject sequences must be met before extension. After matches meet the threshold value, extension generates local alignments using dynamic programming until an alignment quality score no longer improves based on a number of allowed mismatches. The statistical significance of the alignment score for each sequence is then assessed and if the score is high enough, an e-value is calculated that indicates the number of times an unrelated sequence in the database would obtain a higher score than the subject sequence by chance. The best hit (BeT) is not always the nearest neighbour based on phylogeny, in rare cases 30% of the BeTs are not in the same domain of life (Koski and Golding 2001). For a given substitution matrix and gap penalty, BLAST can also miss distant homologs, find non-homologous proteins relative to Smith-Waterman (Pearson 1995, Shpaer et al. 1996). Smith-Waterman is more computationally expensive, being up to 50 times slower than BLAST. Since only seeds are used to create alignments and are terminated after a certain number of mismatches, BLAST does not always find the same solutions as Smith-Waterman. Although BLAST generally has lower sensitivity and lower specificity, it is seen as a reasonable trade-off in speed and accuracy and is often a close approximation (Korf 2003)

Instead of a set of homologous protein sequences that can continue to grow in size as more homologs are discovered, it is often more desirable to represent such sets of homologs as a model. A profile (Figure 1.2) is a representation of the distribution of nucleotides or amino acids at each position in a sequence alignment of homologs (Grisbskov et al, 1987), and are an important component to some homology search algorithms such as Position-Specific Iterated BLAST (PSI-BLAST, Altschul et al. 1997). At each position, a score is calculated that reflects the degree of sequence conservation with penalties against sequence insertions and deletions. Profile-based methods are capable of finding as many as three times more remote homologs than pairwise search methods (Park et al. 1998).



Sequence 1	A C G C T G T
Sequence 2	A C C C T C T
Sequence 3	T C G G T A T
Sequence 4	A G G G T A T
Sequence 5	A A C C C A T
Sequence 6	A G C T T G T
Sequence 7	A G C G T G T
Sequence 8	A G C C C A T

Column	A	C	G	T
1	7	0	0	1
2	1	5	3	0
3	0	5	3	0
4	0	4	3	1
5	0	4	3	1
6	0	2	0	6
7	0	0	0	8

Column	A	C	G	T
1	0.875	0	0	0.125
2	0.125	0.625	0.375	0
3	0	0.625	0.375	0
4	0	0.5	0.375	0.125
5	0	0.5	0.375	0.125
6	0	0.25	0	0.75
7	0	0	0	1

Column	A	C	G	T
1	3.5	0	0	0.5
2	0.5	2.5	1.5	0
3	0	2.5	1.5	0
4	0	2	1.5	0.5
5	0	2	1.5	0.5
6	0	1	0	3
7	0	0	0	4

Column	A	C	G	T
1	1.807	∞	∞	-1
2	-1	1.322	0.585	-∞
3	-∞	1.322	0.585	-∞
4	-∞	1.000	0.585	-1.000
5	-∞	1.000	0.585	-1.000
6	-∞	0.000	-∞	1.585
7	-∞	-∞	-∞	2.000

Figure 1.2 Sample position specific scoring matrix of a simple nucleotide alignment. a) Nucleotides are counted for each site (i.e. column) in alignment. b) Proportion of each nucleotide at each site. c) Division of each site by background nucleotide frequency, in this case ¼ for each nucleotide. d) Log base 2 conversion for log-odds score

An extension of profiles, a Hidden Markov Model (HMM) profile captures not only the frequency of different nucleotides or amino acids at each position within an alignment, but also allows for deletion or insertion mutations at each site in a protein alignment. Like profile-based methods, HMM profiles are used to represent a group of aligned sequences. For each position in the alignment, a match state, a deletion state, and an insertion state are the possible states for each of the four nucleotides for a DNA sequence alignment, or 20 amino acids for a protein sequence alignment. I will continue describing HMM profiles for amino acids only. The emission probability represents the probability of the three states (match state, deletion state, and insertion state) for each amino acid. For example, a highly conserved amino acid at a certain position would be represented by a high emission probability. The transition probability represents the probability of switching to a different state for a site, and unlike profile methods, is calculated from the frequency of amino acid residues from the alignment (Figure 1.3). For example, a transition probability exists for going from a match to an insertion. HMMs can be used to represent the distributions of amino acids, as well as the probabilities of different insertions and deletions within all sequences in a cluster. Relative to BLAST and other methods such as PSI-BLAST, HMM profile methods can find more distant, possibly functionally similar homologs at a lower error rates (Johnson et al. 2010; Krogh et al. 1994; Park et al. 1998).

HMMER is one of the most popular HMM profile software packages (Eddy 2009), which can search query sequences against a database of HMM profiles for alignments. For each column, there are three states, a match state M, a delete state D, and an insert state I. Sequences are aligned to the profile, and then local alignments are identified, but instead of using an alignment quality score, a probabilistic framework is used to calculate local alignments. This is possible since each position in an alignment is already represented by a state with a probability. The most recent version of HMMER, HMMER3, passes sequences through three filtering steps before full probabilistic analysis is performed with the Forward/Backward algorithms. The first filter is multiple ungapped segment Viterbi (MSV) algorithm), a heuristic of the dynamic-programming Viterbi algorithm (Viterbi 1967), to identify ungapped high scoring alignments. The second filter is the Viterbi filter, which is a dynamic-programming algorithm that calculates gapped optimal alignment scores. The third filter is the full Forward algorithm, which sums over all the possible alignments of the profile to the sequence. If the sequence passes all filters, the Forward/Backward algorithm calculates the the probability of the sequence alignment given the model by calculating the sum probabilities across all paths in the profile (Eddy 2011).

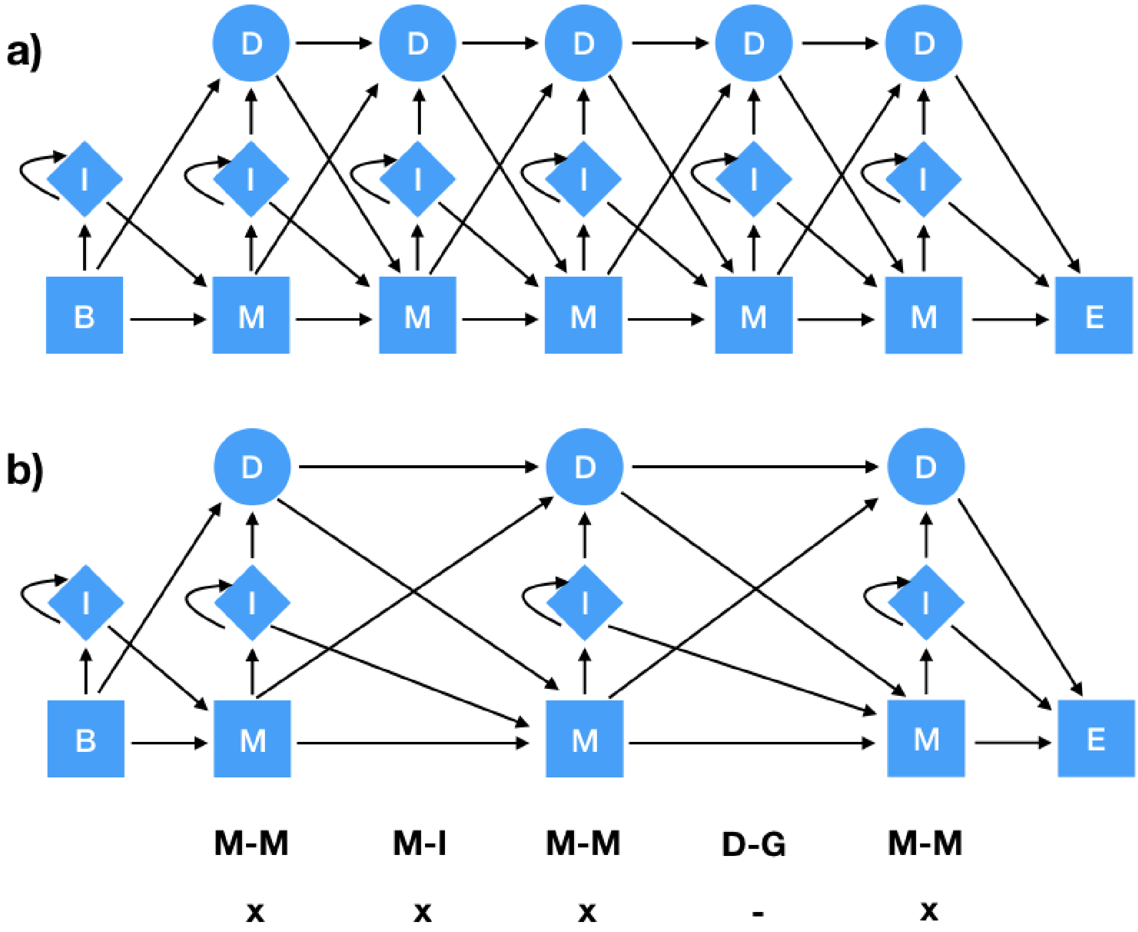


Figure 1.3. Sample HMM profiles. a) B is the start of the HMM, and E is the emitted sequence for the profile. For each column in the HMM, a probilitiy of insertion (I), deletion (D) and a match (M), with arrows representing transition probability and direction. b) An aligned HMM profile to the profile in a). For each column, the alignment type is specified, with “D-G” representing a deletion or gap. Alignment produced is four sequences, denoted as “x”, with a single gap “-”.

HMM profiles can also be searched against one another. HHSearch (Söding 2005) aligns each position in the HMM, where for each position in a pair of aligned HMMs, there are seven possible aligned pair states: match-match (M-M), match-insertion (M-I), insertion-match (I-M), insertion-insertion (I-I), deletion-deletion (D-D), deletion-gap (D-G) and gap-deletion (G-D) (Figure 1.3). To simplify and speed up HHsearch, I-I and D-D states are not considered. For each pair state a dynamic-programming matrix calculates the best alignment, and a log-sum-of-odds score is calculated, which is a generalized version of the log-odds score for the emission of a sequence from sequence to HMM profile comparisons.

1.4.2 Protein Function

The function of a protein can be a contentious issue. Although the name of proteins can describe aspects of function, proteins often act in concert with others in an organism via metabolic pathways or as interaction networks, and they may be localized in different areas in a cell, all of which would play into the function of the protein. Bioinformatic approaches to assign some type of function to proteins have been developed, and classification schemes aim to describe protein function.

There are various ways that gene and protein sequences are assigned a function. The simplest method is “guilt by association” where a function is assigned through association with proteins of known function (e.g. Avarind 2000), through, for example, homology searches against databases of complete sequences or domains. Alternatively, one can use phylogenetic information along with homology information. Phylogenetic profiling (Pellegrini et al. 1999) is a popular “genome context” method (e.g. Kensche et al. 2008), which uses the presence or absence information of all genes from a set of genomes of interest (Table 1.1). One of the principles underlying phylogenetic profiles is that functionally related genes are gained and lost together from genomes during evolution, which will be reflected in the profiles. Similar or identical profiles tend to be functionally linked, and can therefore be used to predict function of uncharacterized proteins (e.g. Basu et al. 2011, Pellegrini et al. 1999, Wu et al. 2003), such as hypothetical proteins that could potentially be linked to pathogenicity (Lin et al. 2011).

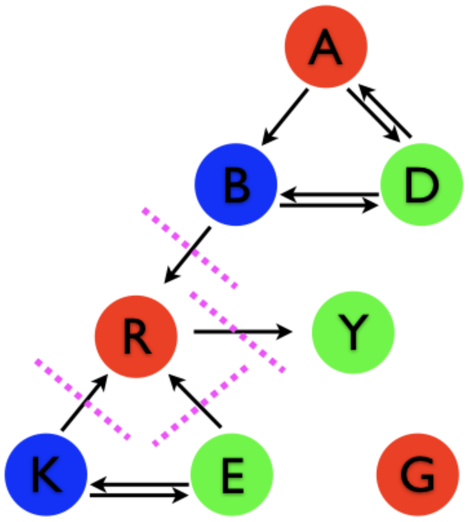
Table 1.1 Sample phylogenetic profile using binary representation of presence of homologs for each protein in each genome for a set of four proteins with GenInfo Identifier (gi or GI) numbers 226948035 226948032 226948031 and 226948030.

Genome	226948035	226948032	226948031	226948030
Clostridium_botulinum_A1_str_CFSAN002368	1	0	0	0
Clostridium_botulinum_A2_str_Kyoto	1	1	1	1
Clostridium_botulinum_A3_str_Loch_Maree	1	1	1	1
Clostridium_botulinum_A_str_ATCC_19397	1	0	0	0

1.4.3 Approaches to Clustering Homologous Sequences

Expectation-value cut-offs for the significance of alignments from any homology search algorithm can be used to define a homologous sequence set, and those sets are the starting point for many cluster construction approaches. The cut-off values can have an effect on the size of the clusters: too relaxed and the clusters are large, too strict and the clusters are small. Even with complete taxonomic sampling, clustering can lead to ORFans (i.e. singletons), proteins that do not cluster with other proteins (Siew and Fischer 2003). If taxon sampling is sparse, homologs may have sequence divergence so great that BLAST or similar algorithms that rely on word extension heuristics may not detect homology. If ORFans are due to distant homologies, HMM profile methods will more likely be able to detect homologous sequences due to their increased sensitivity. Cluster construction approaches can be grouped in three types: graph-based, distance-based, and hybrid (Kuzniar et al. 2008). Many do not have their run times formally characterized, and vary greatly in automation, and degree of manual curation of clusters and their annotations.

The clustering of homologous sequences is an important step in comparative genomics. A variety of clustering algorithms exist, which draw on biological information in different ways. I describe categories of clustering algorithms used by Kuzinar et al. 2008: graph-based, tree-based, and hybrid approaches.



Genome 1

query	hit	genome	e-value
A	B	2	1.7e-120
A	D	3	4.3e-110
R	Y	3	2.1e-52

...

Genome 2

query	hit	genome	e-value
B	D	3	2e-140
B	R	1	1.1e-50
K	R	1	3.9e-72
K	E	3	2.2e-90
K	Y	3	1.3e-60

...

Genome 3

query	hit	genome	e-value
D	A	1	4e-118
D	B	2	3.1e-110
E	R	1	1e-10
E	G	1	1.4e-9
E	K	2	9e-90

...

Figure 1.4 Graph-based approach to cluster sequences using homology search algorithm results, using the best hit for each query to the top hit in each other genome as criteria to draw directed edges. For example, for query “E”, the best hit in genome 1 is “R” and an edge is drawn, while no hit is drawn the second-best hit “G” in genome 1. Dotted lines are cut edges for strongly connected components, in which each nodes has a path to reach all other nodes.

Graph-based approaches represent a set of objects and their inter-relationships by edges that connect nodes, where proteins are nodes and edges are statistically significant relationships between proteins (Figure 1.2) based on alignment statistic thresholds. Edges can be either undirected or directed, the direction of the search (query to hit) is reflected in the direction of the edge if more criteria for drawing edges is desired. First, sequence-homology searches must be performed between sets of sequences, and statistics, such as a maximum e-value or percent identity within specified thresholds, used to filter out spurious hits for a set of queries. If thresholds are too relaxed and spurious hits from chance or conserved and short high-scoring local alignments are included, those edges may join unrelated or distantly related sets of sequences. Therefore, graph-theoretic approaches or additional criteria for selection of homologs should be used to refine clusters. When selecting homologs, best hits (BeTs) for a query sequence are the top subject sequences from each of the other genomes in a sequence database. A reciprocal best hit (RBH) results when two proteins are the corresponding best hits from each other's genomes. Different clustering approaches use either or both of the BeT and RBH criteria to describe homologous relationships. As a way to represent homology, an RBH is a stricter definition as opposed to a BeT and is often used to define orthologous relationships (Wolf and Koonin 2012). Some have noted that RBH may not be sufficient for orthology and further identification of paralogs is necessary (Fang et al. 2010).

One of the first databases of homologous sequences, Clusters of Orthologous Groups (Tatusov et al. 1997) uses a more stringent definition of orthology that relies on the idea of BeTs. A list of BeTs can be created for a protein in a genome by identifying the most closely related homologous protein from all other genomes. The minimum graph relationship to form a COG is three proteins with BeT or RBH relationships such that they form a triangle; in Figure 1.4, sequences the triangle shaped subgraphs A-B-D and R-K-E form COGs. The relationship between three proteins, visualized by a triangle is the most basic representation of a COG (Figure 1.2). Clusters grow in size as more triangles are created and edges and nodes are shared between different triangles. Extensive manual curation of the resulting COGs is required to ensure that COGs are valid, and to separate clusters that are considered non-homologous or of divergent function. The Inparanoid algorithm (Remm et al. 2001) identifies orthologs using RBH, and then considers inparalogs as those sequences that are not the best hit, within a specified threshold. The homology search scores are then compared between each genome and an outgroup genome. The top scores for RBH are kept as orthologs.

Inparalogs that are nearly identical to the orthologs are included, but inparalogs that are dissimilar at a certain threshold are excluded. Although computationally efficient, Inparanoid is designed for eukaryote genomes and is not intended for Bacteria.

The selection of thresholds for drawing edges in graphs is an important aspect to creating clusters of homologous sequences. If thresholds are too relaxed, clusters can grow to untenable sizes. Graph theory-based approaches to subcluster, or split up a larger graph into sub-graphs, can be applied instead of relying on manual curation. Graph connectivity concepts are one possible way to extract subgraphs from larger graphs (e.g. Lechner et al. 2011, Wittkop et al. 2010), or in other words, extracting subclusters from a cluster. A strongly connected component is a portion of a directed graph where between all possible pairs of nodes, a directed path exists (nodes A, B, D and E, K in Figure 1.4). This criterion limits the possibility of distinct sets of very highly related proteins being joined by tenuous one-directional relationships. The main advantage of strongly connected graphs is algorithms that identify them are fast, with a linear run time. The Markov Clustering algorithm (MCL: Stijn van Dongen 2000) is also a fast approach to subclustering that scales well with the size of the graph. In MCL, simulated random walks across a graph are used determine the probability that any two nodes should be connected. The graph is partitioned based on the calculated probabilities calculated. The MCL algorithm is best known in OrthoMCL (Li et al. 2003), and has been implemented in several other clustering approaches (Enright et al. 2002, Harlow et al. 2004). In OrthoMCL, the input graph for MCL is constructed with edge weights applied to decrease the influence of paralogous sequences.

If all possible homologous relationships within cut-off scores are used to construct graphs, the size of the graph could quickly consume significant system resources. For example, estimates of memory usage by the Perl graph package (http://search.cpan.org/~jhi/Graph-0.94/lib/Graph.pod) are 100 bytes for a node, and 400 bytes for an edge, a cluster for 1000 genomes each with 3000 genes and two edges for each protein requires ~2.7GB of system memory. For a real data set, this would be a very conservative estimate, as many more edges would be expected since individual proteins have many homologs. One cannot expect the amount of memory used to scale linearly with additional genomes: the number of nodes increases linearly with the number of sequences, but multiple homologous relationships can exist for each sequence.

Tree-based methods typically require a collection of homologous sequences, and since clusters are typically not *ab initio* constructed, many tree-based methods are post-processing methods. Unlike graph-based methods, tree-based methods use information about phylogeny, and when a reference species tree is first constructed, the topology of the reference tree is compared to the topology of the tree constructed from the cluster of interest. Agreement between the topologies of a reference species tree and a tree from a cluster of interest can inform about possible orthologs. Inparalogs are identified in tree-based methods under the assumption that they are more closely related to the ortholog they were duplicated from. Therefore, in a phylogeny, they should be the sister to the orthologous sequence. Outparalogs would branch with more distant proteins. Tree-based methods are less common since alignment construction and phylogenetic reconstruction can be computationally restrictive, as such several pre-computed databases of orthologs and paralogs from tree-based methods are available (e.g. Altenhoff et al. 2017, Huerta-Cepas et al. 2014, Pryszcz et al. 2011, Waterhouse et al. 2013).

Hybrid methods use both tree-based and graph-based methods to construct clusters of homologs and filter out orthologs. Hybrid methods typically use a species tree to guide the clustering process. As the tree is traversed from leaves to root, clustering of sequences is performed at internal nodes. The Phylogenetic orthologous groups (PHOGs) is an example of a hybrid approach (Figure 1.5), with the intention of obtaining clusters of orthologs (Merkeev and Mironov 2006). Starting from an internal node where all children are leaves, BLAST searches are performed, RBHs are identified, clusters aligned, and a consensus sequence then represents the orthologous cluster. Sequences that are not an RBH, but within a threshold value for a significant BLAST alignment score are then considered to be paralogs. The consensus sequences, paralogs and sequences without any significant alignments then traverse up the tree to the internal node that is ancestral to the current one, to form what is coined a “supergenome”. This supergenome is then treated as a genome, and clustering continues against other supergenomes or genomes from leaf nodes that descend from the ancestral node. When the root is reached, clustering is complete. Although this method is fast, methods that rely strictly on BLAST may not be able to detect remote homologs, and the top BLAST hit may not be the best hit according to phylogeny (Koski and Golding 2001).

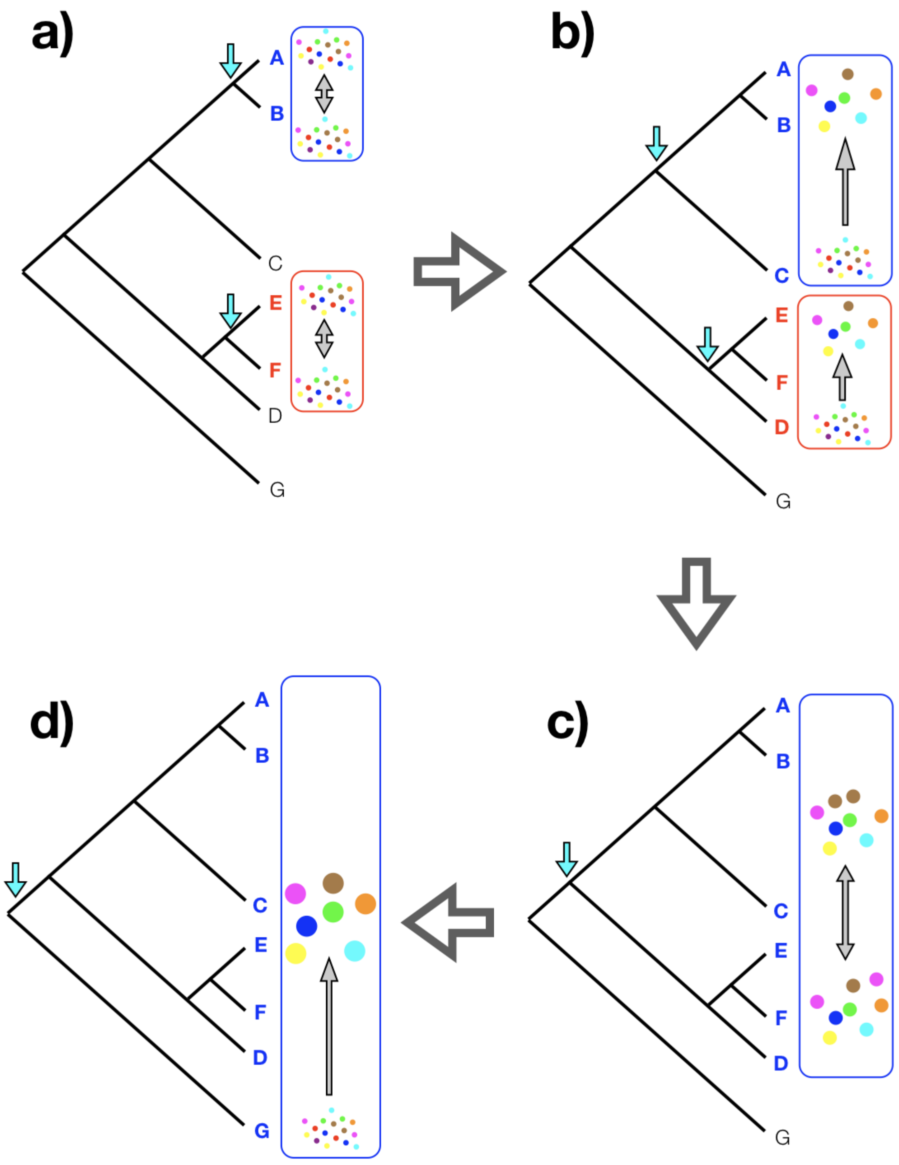


Figure 1.5. Overview of Phylogenetic orthologous groups (PHOGs), an example of a hybrid approach to clustering. Clustering of genomes takes place at internal nodes as the guide tree is progressed from leaves A, B, C, D, E, F and G to root. a) Clustering of leaves is first performed for internal nodes with only leaf nodes (A-B, E-F) as descendants, and clustered sequences form "supergenomes". b) Leaf node sequences are clustered with "supergenomes" (AB-C, EF-D) that descend from internal nodes. c) Supergenome vs supergenome search (ABC-EFD). d) Leaf node vs supergenome search (G-ABCDEF).

Hieranoid is a clustering algorithm to construct clusters of orthologous sequences for Eukaryotes (Schreiber and Sonnhammer 2013). Hieranoid uses a guide tree for clustering like PHOG, but uses the Inparanoid algorithm at internal nodes to identifiy orthologs, and then uses HMM profile searches to amalgamate clusters at internal nodes where at least one child of the node is an internal node. Hidden Markov Model profile searches are more sensitive, and would be more able to detect remote homologs, but the trade-off is slower speed of searches for large data sets. To compensate for the decreased speed due to HMM profile searches, Hieranoid uses BLAST searches on consensus sequences to filter out candidate clusters for HMM profile-HMM profile searches.

Various approaches to benchmark clustering algorithms have been developed (e.g. Bernandes et al. 2015, Chan et al. 2013, Altenhoff et al. 2016). To determine the accuracy of a clustering algorithm, a benchmark data set would have to be established as ground truth. Although one data set is used for orthologs of mainly eukaryotic genomes (Altenoff et al. 2016), the accuracy of clustering algorithms for Bacteria may not be applicable to a ground truth based on eukaryotic sequence clusters. Other approaches to benchmarking compare the similarity of clusters from different methods (e.g. Chan et al. 2013) which compares methods to each other. Comparing methods to each other would be ideal in the event a ground truth cannot be established, or if the idea of establishing a ground truth set of sequence clusters, such as orthologous clusters, is impossible due to the difficulty in defining practical and clearly definitions of homology (Fitch 2000).

1.4.4 Choosing a Clustering Strategy

There is no shortage of approaches, algorithms and databases to obtain clusters of homologous sequence, but clusters should be chosen based on the method that is ideal for the taxonomic affiliation of the organisms of interest and for the type of analysis that is performed. Homologous protein clusters can be composed of orthologs, paralogs and transferred genes along with sequences that share domains as a result of gene fission and fusion. Orthologous clusters are often chosen for functional analysis due to the assumption that orthologs have a conserved function. However, paralogous sequences may still retain similar or identical function, and in the case of inparalogs, it may be difficult to determine which sequence is the ortholog and which is the inparalog. The choice of clusters should also be informed by the type of algorithm used for cluster construction, and whether it is appropriate for the type of data (e.g. Bacteria, eukaryote, complete genomes, draft genomes). It would also be important to choose clusters based on the number of homologs present and taxonomic diversity. The NCBI protein clusters database (Klimke et al. 2009) contains smaller clusters that tend to have limited taxonomic diversity, while other methods vary in their ability to detect remote homologs (Bernardes et al. 2015). Many methods also have difficulty distinguishing LGT and orthologs (Dalquen et al. 2013), so methods that explicitly cluster orthologs in organisms where LGT is common, such as Bacteria, may not produce clusters that are strictly composed of orthologs. It should be expected that clusters composed of bacterial sequences may contain laterally transferred genes, unless the clustering algorithm attempts to detect and filter them out. Due to the ubiquity of gene transfer events in bacterial genomes, it may be more appropriate to avoid filtration of paralogs, since they may also remove LGT events.

1.5 Purpose and Scope of Thesis

Genome sequences of both individual organisms and communities in environments are common and increasingly important sources of data to understand microorganisms and their roles in human health and the environment. The purpose of this thesis is to develop approaches to identify LGT in metagenomes, to develop fast sequence clustering approaches to create clusters necessary in comparative genomics analyses such as those required for LGT, and to apply them to large data sets.

1.5.1 Chapter 2: Metagenomics and LGT

Lateral gene transfer is an important mode in the evolution of organisms, especially bacteria. Often metabolic capabilities are gained through transfer of genes from one organism to another. Metagenomic data provides a snapshot of the genomic content of a community. Enhanced biological phosphorus removal communities are often composed of microorganisms present in local environments, from which, metabolic capability is engineered. Due to their nature as engineered communities with specific metabolic capabilities, EBPR communities are likely candidate communities to identify gene transfer events between organisms that have developed the phosphorous removal phenotype. I develop methods to classify different organisms in three EBPR communities: Maddison, Wisconsin, United States of America; Brisbane, Queensland, Australia; and Aalborg, Denmark. Reference genomes are used, along with databases to identify potentially transferred genes. These methods account for the limitations presented by metagenomics data, specifically uncertainty in taxonomic identification, and uncertainty in the direction of lateral gene transfer.

1.5.2 Chapter 3: ProPhylClust & PhyloSubClust

Obtaining clusters of homologous sequences that were used in Chapter 2 analyses were one of the main challenges that did not present an easy solution. Some cluster databases did not include recent genomes, or did not provide clusters with remote homologs. In this chapter I present two algorithms "ProPhylClust" and "PhyloSubClust" to cluster protein sequences from bacterial genomes, which rely on phylogenies to cluster sequences. Both algorithms can cluster sequences regardless of the presence of LGT events. ProPhylClust is a hybrid method to create large clusters of homologs. PhyloSubClust extracts subtrees from phylogenies constructed from clusters based on the taxonomic content of the subtree relative to the complete tree. I characterize ProPhylClust and PhyloSubClust's runtimes and compare them against other clustering methods, without relying upon benchmark data sets. The intention is to create clusters of homologous sequences that contain orthologous sequences, however clusters may contain paralogs since no attempt is made to filter out paralogous sequences.

1.5.3 Chapter 4: Clustering and Identification of Potential Virulence Factors in Pathogenic *Clostridia*

Clostridia are a class of Bacteria with members that are known to have important functional roles in their environmental and as pathogens, especially several human related pathogens. *Clostridia* are known to widely share genes within their own class, and with other classes of Bacteria (e.g Beiko et al. 2005, Doxey et al. 2008, Meehan and Beiko 2014, Skarin and Segerman 2014). Three members of this class, *Clostridium botulinum* and *Clostridium tetani* and *Peptoclostridium difficile* (also referred to in the literature as *Clostridium difficile* or *Clostridioides difficile*) are notable human pathogens known for being highly virulent, affecting the nervous system (*C. botulinum* and *C. tetani*) or the intestinal tract (*P. difficile*), and all are known to have LGT as a major contributor to the evolution of virulence (e.g. Doxey et al. 2008, Monot et al. 2015, Popoff and Bouvet 2013, Skarin and Segerman 2014). Approximately 25% of protein families in the genus *Clostridium* (including *P. diffiicile*) are of hypothetical function, which ranks 10th out of 22 other pathogenic bacterial genera in the PATRIC database (Wattam et al. 2017), which range from 15% (*Bacillus*) and 40% (*Brucella*). Only 10% of *Clostridium* genomes share conserved protein families, which, along with *Bacillus*, is the lowest in the PATRIC database. Due to the importance of *Clostridium* as pathogens, the lack of conserved protein families across the genus, and the presence of hypothetical sequences, clustering of sequences can help enlighten function and evolutionary relationships of proteins. I use ProPhylClust and PhyloSubClust to cluster protein sequences from 558 draft and fully sequenced genomes of *Clostridia*. From the clusters with toxin related sequences for *P. difficile*, *C. tetani*, and *C. botulinum* I construct phylogenies to gain insight into the evolution of their toxin sequences. To identify hypothetical sequences that are potentially associated with toxins, genome context methods such as phylogentic profiles are implemented for this large data set.

CHAPTER 2 TRANSFER OF ENERGY PATHWAY GENES IN MICROBIAL ENHANCED BIOLOGICAL PHOSPHORUS REMOVAL COMMUNITIES

Published as: Wong, DHJ and Beiko RG. 2015. Transfer of Energy Pathway Genes in Microbial Enhanced Biological Phosphorus Removal Communities. BMC Genomics 16. BMC Genomics: 1–13.

2.1 ABSTRACT

Lateral gene transfer (LGT) is an important evolutionary process in microbial evolution. In sewage treatment plants, LGT of antibiotic resistance and xenobiotic degradation-related proteins has been suggested, but the role of LGT outside these processes is unknown. Microbial communities involved in Enhanced Biological Phosphorus Removal (EBPR) have been used to treat wastewater in the last 50 years and may provide insights into adaptation to an engineered environment. We introduce two different types of analysis to identify LGT in EBPR sewage communities, based on identifying assembled sequences with more than one strong taxonomic match, and on unusual phylogenetic patterns. We applied these methods to investigate the role of LGT in six energy-related metabolic pathways.

The analyses identified overlapping but non-identical sets of transferred enzymes. All of these were homologous with sequences from known mobile genetic elements, and many were also in close proximity to transposases and integrases in the EBPR data set. The taxonomic method had higher sensitivity than the phylogenetic method, identifying more potential LGTs. Both analyses identified the putative transfer of five enzymes within an Australian community, two in a Danish community, and none in a US-derived culture.

Our methods were able to identify sequences with unusual phylogenetic or compositional properties as candidate LGT events. The association of these candidates with known mobile elements supports the hypothesis of transfer. The results of our analysis strongly suggest that LGT has influenced the development of functionally important energy-related pathways in EBPR systems, but transfers may be unique to each community due to different operating conditions or taxonomic composition.

2.2 Background

Enhanced biological phosphorus removal (EBPR) communities are a common form of microbial treatment developed by Banard (1976) that removes phosphorus and occasionally nitrogen from sewage. EBPR is environmentally sustainable and affordable (Oehmen et al. 2007), with microbial communities typically seeded from the local environment or from a seed stock. Considerable effort has been put into understanding EBPR, from community diversity (e.g. He et al. 2006, Mielczarek et al. 2013, Nielsen et al. 2012), to metabolic function (e.g. Oehman et al. 2007, Yuan et al. 2012) and engineering (e.g. Tu and Schuler 2013, Zhang et al. 2005), with the objective of improving efficiency and stability. A substantial amount of work has gone into understanding what organisms are present in EBPR plants (Mielczarek et al. 2013, Nielsen et al. 2012, He and McMahon 2011, Kong et al. 2002, Wong et al. 2005), which organisms tend to be associated with each other (e.g. Mielczarek et al. 2013, Nielsen et al. 2012), their ecology (e.g. Kong et al. 2002, Gonzalez-Gil et al. 2011, He et al. 2008), and how to engineer the EBPR process (e.g. Gonzalez-Gil et al. 2011; Zhang et al. 2011b). Recently, a conceptual ecosystem model (Nielsen et al. 2010) and a core microbiome Nielsen et al. 2012) have been proposed, based mainly on 25 plants in Denmark, revealing a taxonomically broad group of characterized and uncharacterized organisms. However, the majority of EBPR-associated organisms are not found in all EBPR samples.

To develop the EBPR process, a carbon source, typically acetate or propionate, is input to the system, and anaerobic and aerobic conditions are cycled in a bioreactor to select for phosphate accumulating organisms (PAOs). Other organisms perform functions such as fermentation and hydrolysis, and are often referred to as the "flanking community" (e.g. Nielsen et al. 2012, García Martín et al. 2006). Phosphorus uptake occurs during the anaerobic cycle, and carbon and energy-providing polymers are stored as polyhydroxyalkanoates (PHAs). During the aerobic phase, energy stored in the PHAs is used for growth and reproduction. The type of input carbon source is taken up at different rates for different organisms in EBPR, which could affect treatment plant operation (Oehmen et al. 2005). Because anaerobic and aerobic cycling is so important for EBPR community function, emphasis has been placed on metabolic pathways related to PHA metabolism (e.g. Oehmen et al. 2005, Seviour et al. 2003), glycolysis and gluconeogenesis (e.g. García Martín et al. 2006, Oehmen et al. 2005, Lanham et al. 2013), the pentose phosphate pathway (e.g. Oehmen et al. 2007, McIlroy et al. 2013), and the citric acid cycle (e.g. Oehmen et al. 2007, García Martín et al. 2006, Lanham et al. 2013, McIlroy et al. 2013) as a means to understand how EBPR functions, and through usage of a particular metabolic pathway, a way to make EBPR more efficient at removing phosphate (e.g. McIlroy et al. 2013).

Metagenomic sequencing of two lab-scale EBPR enrichment reactors allowed the elucidation of EBPR-relevant metabolism of a major PAO, the Betaproteobacterium *Candidatus* Accumulibacter phosphatis (CAP) Clade IIA strain UW-1 (García Martín et al. 2006), including phosphate uptake and PHA degradation during the aerobic phase, and PHA storage and polyphosphate degradation during the anaerobic stage. The amount of sequence generated, and the technology used (Sanger sequencing, which generates relatively long reads) allowed the eventual assembly of the first complete genome of CAP Clade IIA strain UW-1. Recently, draft genomes of other CAP have been sequenced (Flowers et al. 2013, Mao et al. 2014). Sequencing of a full-scale reactor metagenome from Denmark highlighted an enrichment of genes associated with biofilm and phosphate metabolism, and the taxonomic diversity of full-scale reactor communities (Albertsen et al. 2011). Despite a group of organisms considered to be common in EBPR communities, exact strains and membership can vary considerably between treatment plants (Mielczarek et al. 2013).

Lateral gene transfer (LGT) is a well-established mode of evolution in bacteria that can be studied through a variety of approaches using genome sequences (e.g. Beiko et al. 2005, Koonin and Galperin 1997, Lawrence and Ochman 1997, Ragan 2001a, Ragan 2001b). LGT plays an important role in adaptation, for example, in heavy-metal metabolism (e.g. Sentchilo et al. 2013, Sobecky and Coombs 2009), and in antibiotic resistance (e.g. Sentchilo et al. 2013, Barlow 2009). Transfers tend to take place between close relatives, but many examples of transfer between more distant relatives have been reported as well (e.g. Beiko et al. 2005, Popa et al. 2011). LGT is known to have occurred in sewage treatment plants, impacting antibiotic resistance genes (e.g. Sentchilo et al. 2013, Hong et al. 2014, Ma et al. 2013, Szczepanowski et al. 2008, Zhang et al. 2011a), and xenobiotic degradation (e.g. Schlüter et al. 2007, Top et al. 2002). Many of these transfers are mediated by mobile genetic elements (MGEs) such as plasmids and transposons (e.g. Schlüter et al. 2007, Top et al. 2002). Engineering of treatment plants have used plasmids to bioaugment communities to allow metabolism of xenobiotics (Bathe et al. 2005). Other mechanisms of LGT exist, such as gene transfer agents (e.g. Lang et al. 2012) and transformation (e.g Thomas et al. 2005), but their role in sewage treatment communities is not known.

The metagenomes of two non-EBPR sludge community plasmids were sequenced (Sentchilo et al. 2013), revealing substantial differences in genes from a plant with primarily industrial waste and a plant with primarily household waste. The differences suggested that the prominence of carbohydrate metabolism genes from the industrial waste plant, and the genes related to defense factors in the household waste plant, were the result of selection in each of those communities. Others have noted that transferred plasmids in non-EBPR sludge can have a mosaic of functional genes (Hong et al. 2014). Some evidence of LGT has been identified in PAO genomes (Flowers et al. 2013) but no such events have been proposed from metagenome data thus far.

There are many different bioinformatic approaches for the identification of LGT events (reviewed in Ragan 2001a, Zhaxybayeva 2009), but most rely on whole-genome sequences. Different methods can identify very different sets of genes as putatively acquired via LGT (e,g. Lawrence and Ochman 2002; Ragan 2001a; Ragan et al. 2006). Metagenomic data introduce several challenges that make identification of LGT difficult, in particular, metagenome sequence fragments are short (typically < 1000 nucleotides in length) and of uncertain provenance in the community. Incorrectly assembled chimeric contigs often combine sequences from multiple members of the same genus, species or strain (Charuvaka and Rangwala 2011, Mavromatis et al. 2007, Pignatelli and Moya 2011). Chimeric contigs are more common in more diverse communities (e.g. Mavromatis et al. 2007), and when using short-read sequencing technology with closely related strains (Charuvaka and Rangwala 2011), and can often lead to incorrect classification of contigs.

Despite these challenges, it would be an important step to develop sequence-based approaches to identify LGT within an environment to further our understanding of microbial adaptation. Approaches such as genetic exchange networks (Skippington and Ragan 2011) could identify transfers between multiple taxonomic groups. Here we develop and apply two different analyses to identify candidate LGT events in EBPR metagenomic data for six relevant metabolic pathways. We focus on class-level gene transfers to avoid any errors in assembly at lower taxonomic levels that can affect the accuracy of classification. Our first method, classification discordance, exploits disagreement between taxonomic classifications of genes and longer assemblies. Our second method relies on phylogenetic incongruence. Both are then filtered by homology with known MGEs to identify putative cases of LGT that have been putatively transferred through MGEs.

2.3 METHODS

2.3.1 Sequence Data

The EBPR enrichment culture metagenomes for lab-scale bioreactors in Madison, Wisconsin, United States of America (USA) and Brisbane, Australia (OZ) that comprised the first EBPR metagenome study (García Martín et al. 2006), both sequenced using Sanger sequencing, were downloaded on April 21st 2009 from the Joint Genome Institute (Macdonald et al. 2012). The USA community is composed of 15,866 contigs and assemblies, 25,312,906 nucleotides, with reads an average of 986 nucleotides in length, and the OZ community 11,188 contigs and 24,385,629 nucleotides, with reads an average of 1038 nucleotides in length. The EBPR metagenome for a full-scale bioreactor in Aalborg, Denmark (DK) that performs nitrogen removal in addition to phosphate removal (Albertsen et al. 2011), sequenced using Illumina GAII (2 x 72 paired end), was downloaded from the SEED (http://metagenomics.anl.gov/ metagenomics.cgi?page=MetagenomeOverview& metagen-ome=4463936.3), is composed of 269,385 contigs and 145,725,513 nucleotides of sequence data. We used the assemblies and predicted genes and putative proteins as generated by the original sequencing projects.

Mobile genetic element sequence data consisted of MGEs from the Phast (Zhou et al. 2011) and the ACLAME databases (Leplae et al. 2010). The Phast database is composed primarily of viral sequences and the ACLAME database is composed of plasmids, phage genomes and transposons. We also included the complete NCBI plasmid database, and added other plasmids from NCBI that were not in the plasmid database, but matched the search terms “sewage treatment”, “waste-water” and “wastewater”. In total, this amalgamated database contained 7,584,934 sequences.

2.3.2 Taxonomic and Functional Annotation of Metagenomic Contigs

Class-level taxonomic classification of contigs was done using RITA (Macdonald et al. 2012). RITA uses a reference database to assign a taxonomic classification to sequence data using both homology and nucleotide composition. We used RITA v1.0.1 with a reference data set of over 2986 genomes representing 65 different taxonomic classes (Appendix 1), using USEARCH v4.1.93 (Edgar 2010) for homology searches and FCP v1.0.3 (Parks et al. 2011) for nucleotide composition matching. RITA performs taxonomic classification and assigns sequences to one of four confidence groups based on the strength of evidence in favor of that classification. Sequences with identical taxonomic predictions from both homology and composition were assigned to Group I. Group II comprised sequences where the expectation value for the best-matching genome was at least 10 orders of magnitude smaller than the best-matching genome from a different class. Group III assignments are made when the NB likelihood score for the best-matching genome is at least 1.5 times greater than the NB likelihood for the best-matching genome from another class. Group IV assignments are based only on the best NB likelihood value. Accuracy of classifications increases with longer contigs Macdonald et al. 2012), so only contigs at least 1000 nucleotides in length were used.

Sequences were functionally annotated through a BLAST (version 2.2.23) (Altschul et al. 1990) homology search. Annotations were based on the top hit to a reference data set of microbial proteins from the NCBI Protein Clusters database (Klimke et al. 2009) with a 60% alignment length of the predicted protein with the reference sequence, an expectation value of 1e-5 or smaller, and neither the predicted protein or reference sequence greater than 1.2 times the length of the other. Additional annotations for enzymes were assigned using a publicly available version (58.1) of the KEGG database (Kanehisa and Goto 2000). A subset of KEGG pathways and their enzymes (see Table 2.1 and Appendix2) related to EBPR metabolism during anaerobic and aerobic cycling, carbon feed source, and nitrogen metabolism were subjected to detailed analysis and were annotated with a more recent version (67.1) of KEGG: butanoate metabolism (BM) for EBPR PHA metabolism, citric acid cycle (CAC), glycolysis/gluconeogenesis (GG), pentose phosphate pathway (PPP), propanoate metabolism (PM) for EBPR propionate metabolism (propionate is the propanoate ion), and nitrogen metabolism (NM).

Table 2.1 List of enzymes by Enzyme Commission (EC) number, and common name in text.

EC Number	Name
1.1.1.1	alcohol dehydrogenase
1.2.1.12	glyceraldehyde-3-phosphate dehydrogenase
1.6.5.3	NADH:ubiquinone reductase
1.9.3.1	cytochrome-c oxidase
2.3.1.9	acetyl-CoA C-acetyltransferase
2.7.1.11	6-phosphofructokinase
2.7.1.2	glucokinase
2.7.1.63	polyphosphate-glucose phosphotransferase
2.7.2.3	phosphoglycerate kinase
4.2.1.11	phosphopyruvate hydratase
4.2.1.17	enoyl-CoA hydratase
5.4.2.1	phosphoglycerate mutase
6.3.5.4	asparagine synthase

2.3.3 Identification of Putative LGT Events

Sequenced reference genomes are typically used for the identification of LGT, since obtaining complete genomes from metagenomes may not be possible without an appropriate amount of sequencing effort. However, the complete genome of CAP Clade IIA strain UW-1 was reconstructed from the USA EBPR metagenome after additional sequencing effort was applied. We used this genome to look for initial evidence of LGT in this EBPR community. We performed homology searches, using BLAST, of its genome against itself and 2773 reference genomes, and MGEs used in the EBPR-MGE homology searches. The top hits with a minimum of 60% shared alignment were used as evidence of potential LGT.

We used two complementary approaches to identify putative LGT events in the EBPR metagenomes. The first approach identified strong disagreement between taxonomic classifications (“classification discordance”) of entire contigs and individual genes within those contigs. The second approach considered incongruence in phylogenetic trees as evidence of LGT. LGT identified by the two approaches were then filtered by homology with known MGEs.

2.3.4 Classification Discordance

The taxonomic classification of a whole contig suggests the lineage of the organism from which it was sequenced, but individual protein-coding open reading frames (ORFs) from the contig may differ in their taxonomic assignments. Such disagreements can suggest LGT events with an implied direction of transfer; the donor is the classification of the ORF, and the recipient is the classification of the entire contig. Each predicted ORF was classified at the class level using RITA with the same command-line parameters used above for the contig classifications, with ORFs from group I and group II RITA classifications considered as accurate.

Spuriously classified ORFs originating from classified contigs meeting our length requirements would lead to a questionable inference of LGT. To prevent this, we filtered out candidate transferred ORFs whose best composition-based prediction (i.e., the Naïve Bayes likelihood score) was not at least 15% better than the contig prediction. If this criterion was satisfied, then the contig was considered a transfer recipient of the implicated ORF.

2.3.5 Phylogenetic Incongruence

Phylogenetic methods incorporate models of the evolutionary process, providing a more accurate representation of evolutionary relationships amongst homologous sequences. We first performed all-versus-all BLAST (version 2.2.23) searches within each community to identify clusters of putative homologous proteins. These sets were then compared with 1642 reference prokaryotic genomes to expand and join clusters. Clusters were represented as an undirected graph using the "networkx" python package (1.8.1). In the network, a node represents each sequence, and an undirected edge represents a homologous relationship between two sequences. For an edge to be drawn between two EBPR proteins, they must have 70% sequence identity, and share 60% alignment length with an e-value of 10^{-5} or smaller. This network was expanded by drawing edges between the nodes, the reference genome sequences and EBPR homologs meeting the BLAST similarity requirements. The network was then split into connected components, or a set of nodes that are connected to each other by a path of edges, where each connected component is considered a cluster.

The resulting clusters were often very large (≥1000 sequences), and included distantly related proteins of little use to LGT inference. To obtain sub-clusters, we constructed phylogenies and extracted subtrees. Sequence alignments were constructed from large clusters using MUSCLE (version 3.8.31) (Edgar 2004) with default settings, and trees were constructed using FastTree (version 2.1.4) (Price et al. 2009) with the WAG model of amino acid evolution (Goldman and Whelan 2000). We then manually extracted subtrees where FastTree Shimodaira-Hasegawa (SH)-test-based (Shimodaira and Hasegawa 1999, Goldman et al. 2000) branch support values of at least 70 % denoted clusters of closely related sequences. Subtree extraction, alignment and phylogeny construction was repeated until subtrees comprised a maximum of approximately 200 sequences.

For detecting LGT, phylogenies are typically compared against a reference species tree (e.g. (Beiko et al. 2005)). However, because EBPR community structure can vary over time (e.g. (Slater et al. 2010)) and metagenomes can represent incomplete samples of the total genetic material (Ni et al. 2013), crucial taxa including donors of genetic material may not be present in the sample. We used the DendroPy library (Sukumaran et al. 2010) to calculate the patristic (branch-length) distances between sequences in the same phylogenetic tree, finding for each EBPR sequence the closest EBPR sequence from the same community and the closest reference sequence with an absolute branch length of 0.3 substitutions per site. Sequences with shorter branch lengths should be closest relatives.

2.3.6 Identifying Candidate Mobile Genetic Elements

Potential LGTs from each of the phylogenetic incongruence and classification discordance methods were then filtered by sequences that have homologs, as identified using BLAST with a maximum e-value of 1e-30 against our custom database of MGEs. We included EBPR sequences with hits to MGE sequences of a different taxonomic class from the EBPR sequence, and had an alignment length of at least 60 % of the query EBPR sequence and 60 % of the subject MGE sequence.

2.4 Results

The published CAP genome was used to find evidence of recent LGT, possibly in the context of the EBPR community. Of the 4562 sequences in the CAP genome, 1438 sequences had hits to genomes outside the Betaproteobacteria with the same e-value as the top CAP hit, suggesting the acquisition of many genes by CAP. The high degree of similarity indicates the possibility that many of these transfers occurred very recently. The observation of these recent transfers led us to search for LGT events in all sampled EBPR community members.

Table 2.2 Summary of sequences used in analyses from all communities. Number of retained contigs, open reading frames from retained contigs, and energy pathway related enzymes (butanoate metabolism, citric acid cycle, glycolysis and gluconeogenesis, nitrogen metabolism, pentose phosphate pathway, and propanoate metabolism) from open reading frames annotated as enzymes. Contigs at least 1000 nucleotides in length were retained.

	USA	AU	DK
# (%) of contigs retained	7,610 (47.96%)	7,331 (65.52%)	18,024 (6.69%)
# (%) of ORFs retained	22,894 (66.06%)	25,003 (81.15%)	30,516 (10.14%)
# enzymes in energy pathways (%) of annotated enzymes	645 (22.13%)	714 (22.60%)	524 (22.40%)

Filtering out contigs that were less than 1000 nucleotides in length reduced the size of the Sanger-sequenced datasets to ~48 % (USA) and ~65 % (AU) of their original sizes, while the Illumina-sequenced DK reactor metagenome was reduced to only ~6 % (Table 2.2). This result should be expected given the differences in read length and the expected differences in diversity between lab-scale reactors and full-scale reactors (Wong et al. 2005, Lanham et al. 2013). The DK community had the largest number of taxonomic classes represented in the filtered contigs (63), followed by AU (53) and USA (39; see Appendix3). For all communities, RITA classification Groups I-III accounted for the vast majority of classifications, although the relative proportion of contigs assigned to these groups varied (see Appendix 4). The number of potentially transferred ORFs from the retained contigs also varied by community, analysis type, and the six energy-related pathways.

2.4.1 Classification Discordance

Our first approach to identify putative LGT compared the taxonomic classification of an entire contig with the classification of its predicted ORFs. Of the ORFs that had hits to the metabolic pathways of interest, at least 50% from each community (US: 20 ORFs, 68.9%, OZ: 88 ORFs, 55.7%, and DK: 58 ORFs, 54.2%) satisfied the criteria for discordance. All LGTs suggested by this method had hits to annotated MGEs from our database. The number of inferred transfers, the implicated enzymes and the participating taxonomic groups vary among metabolic pathways and communities (Appendix5). However, some members appear to be more common recipients or donors of gene transfer in all communities and metabolic pathways, with Betaproteobacteria to Gammaproteobacteria (21 transfers) in AU the most common direction of transfer (Appendix6). LGT events with Alphaproteobacteria as donor and Betaproteobacteria as recipient were the only pattern identified in all three communities.

Of the six pathways, the pentose phosphate pathway is the only pathway to not have any detected transfers in the DK community (Appendix5), most likely due to lack of annotated enzymes. Certain pathways have enzymes that appear to have been transferred in all three communities: butanoate metabolism (enoyl-CoA hydratase: EC 4.2.1.17), glycolysis and gluconeogenesis (glucokinase: EC 2.7.1.2), nitrogen metabolism (asparagine synthase: EC 6.3.5.4, cytochrome-c oxidase: EC 1.9.3.1) and propanoate metabolism (EC 4.2.1.17). For example, for butanoate metabolism and propanoate metabolism, enzyme 4.2.1.17 is commonly transferred across all three communities, with directed networks suggesting transfers from Alphaproteobacteria and Betaproteobacteria to Gammaproteobacteria in the AU community, from Betaproteobacteria to Alphaproteobacteria in the USA community, and from Acidobacteria to Deltaproteobacteria (Figure 2.1). These genetic exchange networks suggest that PAOs (e.g. Betaproteobacteria) and competing glycogen accumulating organisms (GAOs) (e.g., from Gammaproteobacteria and Alphaproteobacteria) may be involved in transfers of core metabolic enzymes. There also appears to be parallel transfer of genes between taxonomic groups across communities. For example, in glycolysis and gluconeogenesis, 6-phosphofructokinase (EC 2.7.1.11) shows evidence of transfer from Chloroflexi to the Betaproteobacteria in the USA and AU, but not in DK.

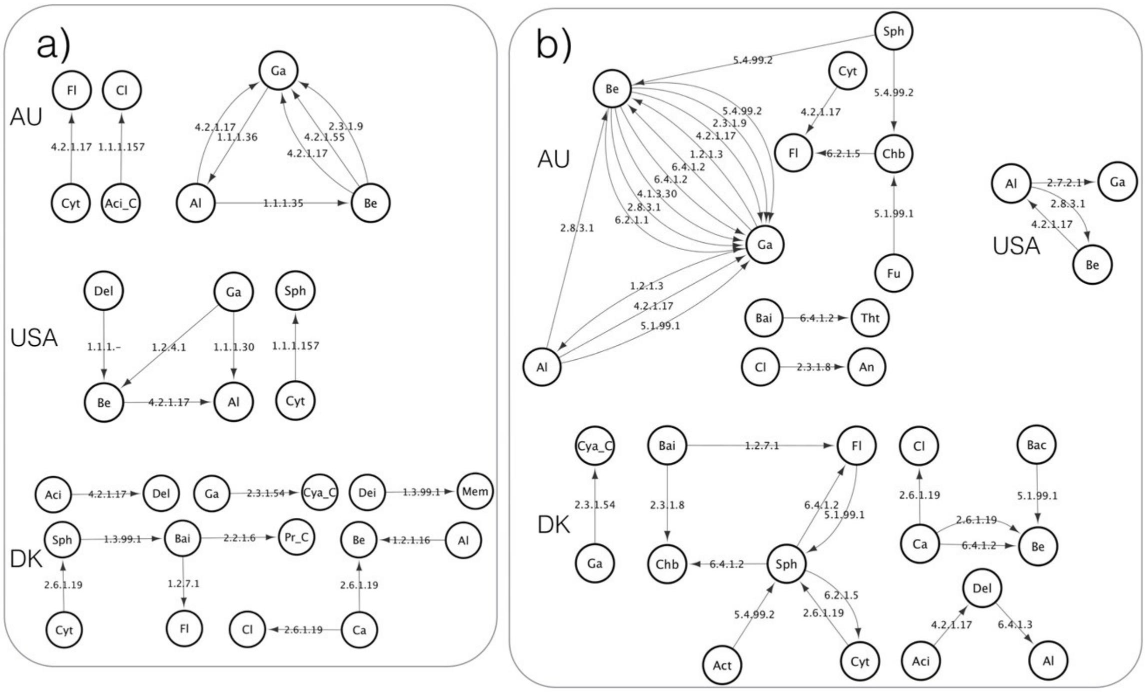


Figure 2.1. Directed transfer of enzymes involved in KEGG a) Butanoate metabolism and b) Propanoate metabolism for the Denmark (DK), Australia (AU) and United States (USA) EBPR communities. Taxonomic groups are nodes, and direction of transfer from donor to recipient is indicated by arrows. See Appendix 18 for taxonomic abbreviation guide.

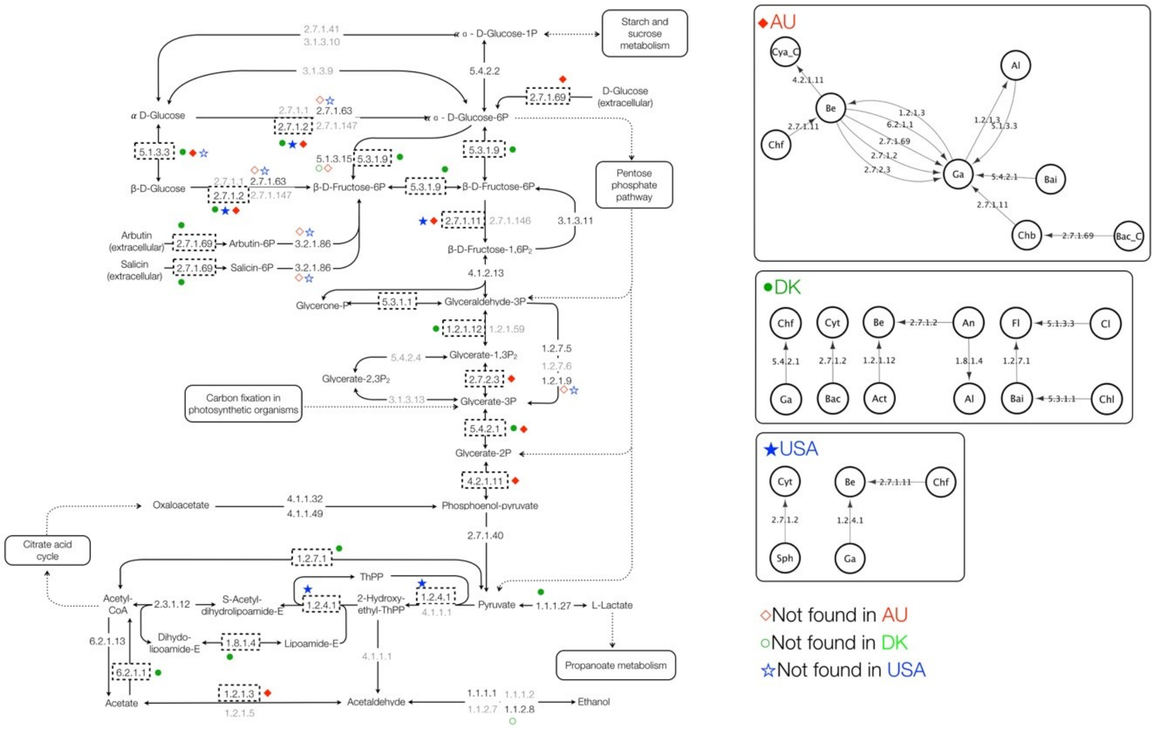


Figure 2.2. KEGG Glycolysis/Gluconeogenesis metabolic pathway and directed LGT for the Denmark (DK), Australia (AU) and United States (USA) EBPR communities. Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating enzymes not inferred to be present in a community. EC numbers in gray correspond to enzymes not found in any community. See Appendix 18 for taxonomic abbreviation guide.

Transfers may be localized at key locations in some pathways, for example, where alternative paths between certain metabolites are not present, suggesting an important role for the transfer in the metabolism of the recipients. For example, in glycolysis and gluconeogenesis, glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) and phosphoglycerate mutase (EC 5.4.2.1) are transferred in the DK community, and phosphoglycerate kinase (EC 2.7.2.3), phosphoglycerate mutase (EC 5.4.2.1) and phosphopyruvate hydratase (EC 4.2.1.11) are transferred in the AU community (Fig. 2.2). These enzymes are involved in a single path for reactions leading from glyceraldehyde-3-phosphate to phosphoenolpyruvate. Missing enzymes in pathways would increase the need for other enzymes to catalyze key reactions. LGT is one way that genes can be acquired by organisms that need specific enzymes for reactions in pathways. In gluconeogenesis and glycolysis, for example, polyphosphate glucokinase (EC 2.7.1.63) is missing in the AU and USA communities, but glucokinase (2.7.1.2) also catalyzes the reaction ß-D-Glucose to ß-D-Fructose-6-phosphate and shows evidence of LGT in all three communities (Fig. 2.2). Figures for the other five pathways, indicating gene transfers and the direction of transfer can be found in Appendices 7, 8, 9, 10, 11, 12, 13, 14 and 15.

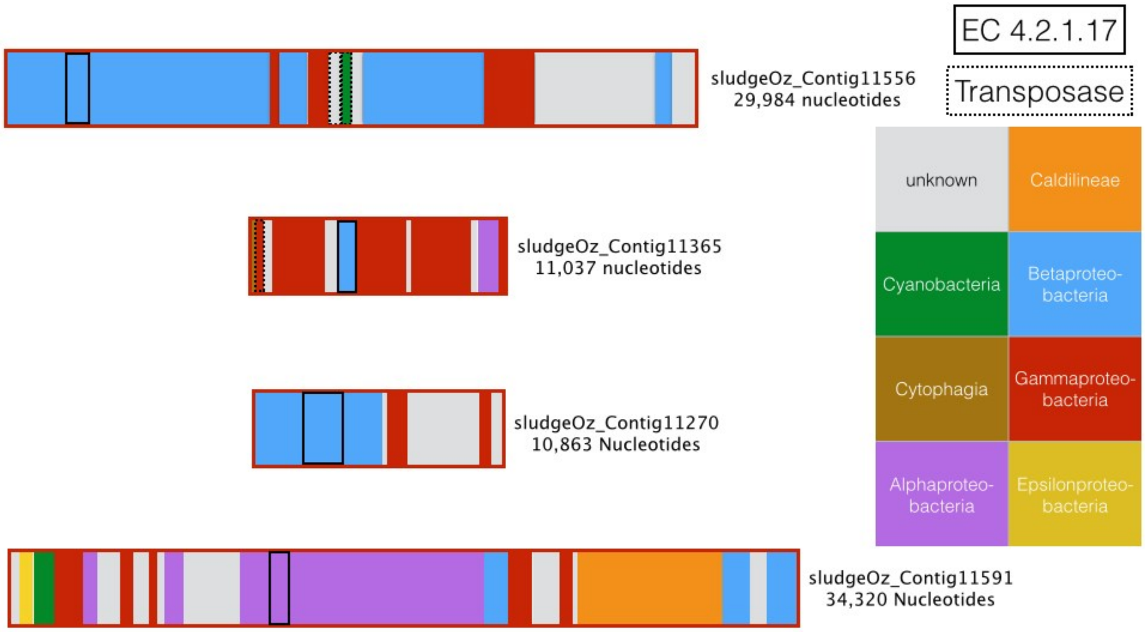


Figure 2.3 Sample contigs from classification discordance. AU contigs classified as having gammaproteobacterial origin but with an inferred transfer of Enoyl-CoA hydratase (EC 4.2.1.17), an enzyme involved in butanoate metabolism and propanoate metabolism, originating from the Betaproteobacteria or Alphaproteobacteria. Transposases are present on two contigs. Colours represent the taxonomic origin of different genes on each contig according to RITA’s naïve Bayes compositional classifier.

Closer scrutiny of the transfers in the directed networks suggests multiple class-level transfers of the same enzyme between specific taxonomic groups. For example, on long contigs, for transfers to the Gammaproteobacteria in the AU community, enoyl-CoA hydratase (EC 4.2.1.17) has been identified as transferred once from the Alphaprotebacteria to the Gammaproteobacteria, and three times from the Betaproteobacteria to three different Gammaproteobacterial contigs. Inspecting the genes on the contig reveals two transposases on one contig and a single transposase on the other

(Fig. 2.3). Classification of sequences in each contig indicates a mixed taxonomic history, suggesting that the present distribution of genes has arisen from a series of independent LGT events.

2.4.2 Phylogenetic Incongruence

A set of 987 trees covering 46,031 proteins from 1622 reference organisms were extracted from an initial set of 981 trees covering 243,031 proteins from 1642 reference organisms. The direction of transfer is difficult to infer as metagenomic sequencing and quality-filtering approaches remove possible within-community donor and recipient lineages, and tree topologies often cannot distinguish which of two implicated lineages is the most likely donor. A total of 14, 27, and 1 (DK, AU, and USA communities, respectively) predicted EBPR proteins differed from a reference sequence or an EBPR protein of a different taxonomic class by less than 0.3 substitutions per site (Appendix16). This represented 4.71 %, 8.84 %, and 0.65 % (DK, AU, USA communities, respectively) of all sequences whose closest relative was a member of a different taxonomic class, but did not meet the 0.3 substitutions per site branch length cutoff.

The recipient of the single proposed transfer within the USA community is classified as Gammaproteobacteria, with predicted function associated with glycolysis and gluconeogenesis (EC 1.1.1.1). The AU community accounted for the majority of transfers, with some transfers identified on the same contig, but not evenly distributed across each metabolic pathway. The DK community had the largest number of inferred transfers in the citric acid cycle and nitrogen metabolism pathway. In the AU community, transfers consistently involved sequences belonging to contigs classified as Gammaproteobacteria and Betaproteobacteria, with Alphaproteobacteria, Bacilli and Chlorobia also implicated in transfer of some of the metabolic pathways.

For the DK community, no common taxonomic groups were shared across metabolic pathways, and no sequences identified as transferred were classified as

Betaproteobacteria. The Cytophagia were implicated in three pathways (butanoate metabolism, citric acid cycle and nitrogen metabolism), while a mixture of the Alphaproteobacteria, Bacteroidia, Flavobacteriia, Gammaproteobacteria, Methanomicrobia, Sphingobacteria are other classes present in the other three pathways (gluconeogenesis and glycolysis, pentose phosphate pathway, propanoate metabolism).

2.4.3 MGE Homology and a Merged Prediction Set

Each method of LGT detection differs in its ability to identify different types of LGT events. All high-confidence LGT events have homology with sequences in known MGEs. A substantial number of sequences from each community had hits to known MGEs: 11,718 of 30,516 sequences from the DK community, 16,156 of 24,956 sequences from the AU community, and 15,530 of 22,662 sequences from the USA community. Of those MGE homologs, 2097 DK, 824 AU, and 875 USA community sequences are enzymes in KEGG pathways (Appendix17).

Given the very high proportion of metagenomic sequences matching to MGEs, we used additional criteria to support inferences of LGT. To obtain a high-confidence set of transfers, we examined the intersection of the two analyses for each of the six pathways (Fig. 4). Pathways differed by the percent of shared transfers, with each detection method sharing a different percentage of transfers. Up to 55 % of LGT events predicted by the classification discordance approach were shared with the phylogenetic approach. This wide variation in shared LGT events is not correlated to the number of detected LGT events, and illustrates the tendency of each approach to find different types of transfers.

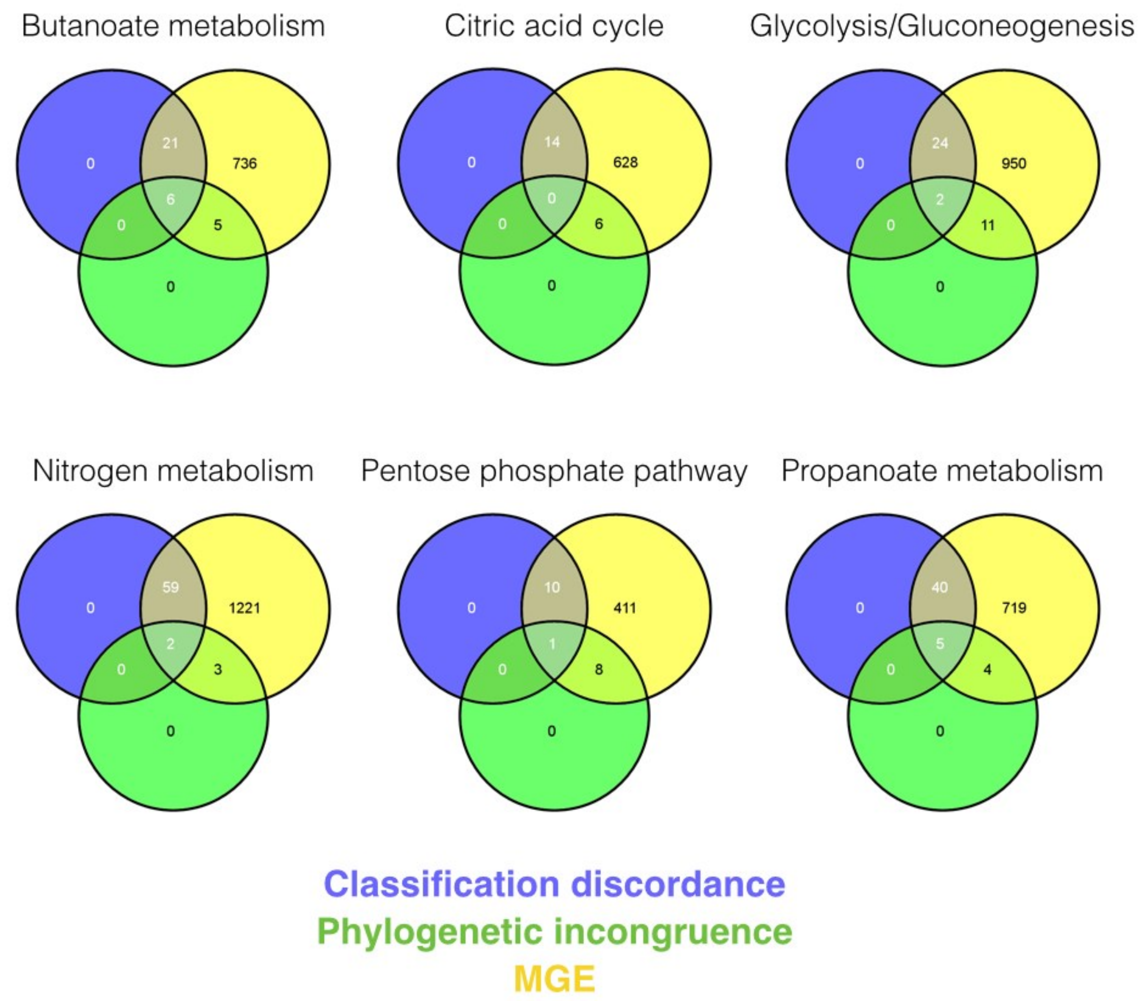


Figure 2.4. Three-way Venn diagram between classification discordance, phylogenetic incongruence and MGE homology filtering for all sequences from all KEGG pathways. Intersections for circles are the number of transferred genes shared between analyses. Remaining genes not in intersections are unique potential LGT events identified by each analysis. All sequences have homologs with known MGEs. Venn diagrams were generated using VENNY (Oliveros 2007).

A total of ten sequences, representing five enzymes, were identified as putatively transferred by the two approaches: enoyl-CoA hydratase (EC 4.2.1.17), acetyl-CoA C-acetyltransferase (EC 2.3.1.9), cytochrome-c oxidase (EC 1.9.3.1), phosphoglycerate kinase (EC 2.7.2.3), and 6-phosphofructokinase (EC 2.7.1.11). Of those ten sequences, eight were identified in the AU community, two in the DK community, and none in the USA community. All of the identified enzymes were present on plasmids in the ACLAME database, suggesting a possible mode of transfer. Both analyses almost always identified the same taxonomic classes as donors or as the top hit. The only transfer in DK, enzyme 1.9.1.3, was associated with nitrogen metabolism. Enzyme 2.7.2.3 was unique to GM. Two enzymes, 4.2.1.17 and 2.3.1.9, are common to BM and PM while 2.7.1.11 is common to glycolysis and gluconeogenesis and the pentose phosphate pathway. No common transfers were found that belonged to the CAC. For AU, six of the eight recipient contigs were classified as Gammaproteobacteria, with the remainder Betaproteobacteria and Chlorobia. For DK, the recipient contigs were classified as Bacilli and Cytophagia.

Table 2.3 Length statistics for contigs with putative LGT events. Predicted gene counts and length for each contig from each community that are at least 1000 nucleotides in length, that have a detected LGT event, and those that have an annotated transposase and integrase.

Community	**Average Contig Length**	**Average Number of Genes**
DK all	1808.84	1.26
DK LGT six pathways	2885.32	1.91
DK LGT transposases & integrases 6 pathways	—	—
AU all	2883.52	2.70
AU LGT six pathways	11793.48	9.80
AU LGT transposases & integrases 6 pathways	25755.10	23.20
USA all	2477.62	2.37
USA LGT six pathways	23000.37	17.21
USA LGT transposases & integrases 6 pathways	68048.67	53.0

Closer inspection of the contigs that contained the transfers from each analysis provides further support for these putative LGT events. In total, ten of 88 contigs from the AU community, none of the 55 from the DK community, and three of the 19 from the USA community had integrases or transposases on contigs that contained transferred genes from both the classification discordance and phylogenetic incongruence methods. This subset of contigs with integrases and transposases are about two times (AU) or three times (USA) longer, and contain more genes: two to three (AU) or three (USA) more than all contigs with LGTs (Table 2.3). The relationship between LGT detection and contig length does indicate that longer contigs are more suitable for identification of LGT, and aid the identification of transposases and integrases. This could explain why the DK community did not have any identified transposases and integrases on contigs with an LGT: the majority of contigs were likely too short.

Since DK reads were not available and USA LGTs were not part of the shared set of transfers from both analyses, we were only able to assess coverage of AU LGTs. Only seven of the eight AU LGTs had matching reads, but all homologous reads had an expectation value of 0.0. Of the 57 reads with at least partial homology to the putatively transferred ORFs, 50 had alignments that extended into adjacent ORFs, suggesting that the inferred events were not due to misassembly. Two putatively transferred ORFs each had an aligned read that spanned the full length of the ORF (Appendix18). One LGT had two reads that did not extend into neighbouring ORFs, and started or ended in intergenic regions. Alignments for the remaining five reads partially covered the putatively transferred ORFs.

2.5 Discussion

Using a series of approaches that are applicable to metagenomic data, we found strong evidence that LGT has impacted six energy metabolism pathways in EBPR communities. Some genes appear to have been independently transferred in more than one community. Although some groups are associated with multiple LGT events, no clear patterns of donor/recipient partners emerged for all three communities. The common set of transfers between the two analyses, and MGE homology filtering, provide the strongest evidence for LGT. The majority of transfers shared by both analyses were identified in the AU community, none in the USA community, and only two genes transferred in the DK community, which were the only shared transfers identified in nitrogen metabolism.

Differences in predicted events across the three communities may represent independent evolutionary trajectories, differences in local community composition, or biases in observation due to incomplete sampling of the metagenome.

Our contig length and ORF taxonomic quality-filtering approaches favored the detection of a relatively small set of high-confidence LGT predictions. Although choosing the class level decreases the number of potential LGTs found and precludes detection of LGT between members of the same class, the long-range transfers we have identified show the strongest evidence for discordance. Our use of contigs in excess of 1000 nucleotides long considerably reduced the proportion of sequences being retained, especially for the DK community, where the average contig length was 504 nucleotides. However, longer contigs are better for detecting LGT (Table 2.3). This could be due to a higher probability that genes from a different source are found on longer contigs, or inaccurate classification due to short contigs. Additionally, longer contigs were needed to identify transposases and integrases in tandem with our genes of interest.

Mapping of metagenomic reads to contigs validated most of our LGT inferences; however, one putatively transferred ORF in our high-confidence set did not have any matching reads. Accuracy of assemblies, including metagenomic assemblies, depends on sequencing technology and the complexity of communities (Sims et al 2014, Mende et al 2012). Less-complex communities (~10 genomes) have the most accurate assemblies with Sanger sequencing, and complex communities (100+ genomes) have the most accurate assemblies with Illumina sequencing (Mende et al 2012). Regardless of assembly accuracy, it is unclear why this ORF should be present in the assembled contigs, while having no corresponding match in the reads used to generate those contigs.

Different methods of detecting LGT are often biased towards finding certain types of transfer events (Lawrence and Ochman 2002, Ragan et al. 2006). Our approaches do not identify transfers at lower taxonomic levels and are biased towards detection of complete genes. Naïve Bayes likelihood ORF filtering should eliminate many dubious classifications, but does not provide any information about the age of the transfer event.

The phylogenetic approach provides information about age of transfers, but identified the fewest candidate LGT events. This is because it requires that the donor lineage in the community or a close relative be sampled, and LGT events that do not appreciably distort the tree will not be detected by this approach.

Different EBPR plants have distinct population characteristics (Mielczarek et al. 2013, He et al. 2011), with different operational parameters between the sampled EBPR communities, and full-scale plants being more complex and dynamic than lab-scale reactors (Kong et al. 2002, Wong et al. 2005, Lanham et al. 2013). All three communities use different carbon feeds: molasses in DK (Albertsen et al. 2011), propionate in AU and acetate in USA (García Martín et al. 2006). Propionate has been shown to be a more desirable carbon source relative to acetate, providing PAOs a selective advantage over competitors, and resulting in a more stable community over time (Gonzalez-Gil and Holliger 2011, Oehmen et al. 2005, Chen et al. 2004, Thomas et al. 2003). The propanoate metabolic pathway, which shows different amounts of evidence for LGT between the three communities, with very few transfers in the acetate-fed USA community, and a large number of transfers in the propionate-fed AU community, especially between the Betaproteobacteria and Gammaproteobacteria. The DK community has an intermediate number of transfers, but with more taxonomic groups implicated than the AU community. The taxonomic composition of EBPR communities is known to change over time (Slater et al. 2010), and with changing carbon sources (Gonzalez-Gil and Holliger 2011); this variability may also manifest through gene exchange between constituents of the community.

Focusing on LGT in energy-related metabolic pathways considered relevant to EBPR function provides context to the role of LGT in EBPR communities. LGT events not in the six energy pathways are also likely to be important in EBPR communities, such as phosphate metabolism, bacteriophage resistance, and flocculation/ biofilm formation. Future analyses should also focus on other metabolic pathways for insights into alternative metabolism, and in broad functional categories for overall community functional aspects of LGT. Additional sequencing of EBPR communities would provide further insight into whether there are common LGT events.

2.6 Acknowledgments

DHJW was supported by the TULA Foundation and Genome Atlantic. RGB acknowledges the support of the Canada Research Chairs program. Computational analyses were supported by a grant from the Canada Foundation for Innovation to RGB. We are also grateful to Christian Blouin, Norbert Zeh, Elizabeth Edwards, Cheryl Devine, and Frances Slater for their feedback and enlightening discussions.

2.7 Competing Interests

The authors declare that they have no competing interests.

2.8 Authors' Contributions

Both authors conceived of the approach, designed the analyses, interpreted results, and wrote the manuscript. DHJW carried out the analyses. Both authors read and approved the final manuscript.

CHAPTER 3 PROPHYLCLUST AND PHYLOSUBCLUST: FAST PROTEIN CLUSTERING AND SUBCLUSTERING USING PHYLOGENY

3.1 Abstract

As the number of sequenced bacterial genomes available continues to increase, the need for tools of homology inference that scale favourably with larger sets of genomes. Methods that can define clusters of sequences composed of or contain orthologs continue to be a goal of developers. Two methods are introduced here, ProPhylClust, which clusters protein sequences using the topology of a phylogeny as guide, and PhyloSubClust, a method to extract subclusters by pruning subtrees from the phylogeny of a cluster. ProPhylClust is compared to all-versus-all BLAST-based graph methods (undirected graphs, directed graphs) to create homologous clusters, while PhyloSubClust is compared to methods typically promoted to subcluster resulting in orthologs (Reciprocal Best Hits, OrthoMCL). Runtimes, cluster size distributions, cluster compositions, and the stability of clusters as genome sets increase in size, were compared between methods. ProPhylClust clustered the most sequences and achieved shorter runtimes with 200 sequenced genomes, than undirected and directed graph methods, while PhyloSubClust was slower than RBH, but had considerably shorter runtimes than OrthoMCL. ProPhylClust can also be run without HMM searches, instead relying on consensus-sequence searches, resulting in the shortest runtimes of all methods,.

3.2 Introduction

Inference of homology is of central importance to many tasks in bioinformatics, including phylogenetic analysis of genes and genomes, identification of critical sequence variants, and functional prediction of novel genes. While pairwise homology search between a query sequence and a reference database is relatively straightforward, the inference of homologous or orthologous groups of genes, or clusters, from multiple genomes is considerably more difficult. A common approach to identification of these sets of genes is to first perform pairwise comparisons, then refine the resulting set of relationships to produce a robust set of genes that share defined common ancestry properties. However, processing these relationships in a consistent and appropriate way is non-trivial, as clusters can be composed of genes with different evolutionary histories, and their sequences may be only partially homologous, for example due to transferred domains, or gene fusions. It is therefore relevant to have methods that can identify genes in clusters based on evolutionary relationships. Coupled with the problem of definitions is the computational time required to identify homologous sequence clusters. Clustering methods such as hierarchical clustering or all-versus-all methods often have time complexities that are quadratic or worse as data sets increase in size (Lechner et al. 2011, Li et al. 2003, Matias Rodrigues and Von Mering 2014). Clusters are often too large to be of use, requiring additional subclustering (Lechner et al. 2011, Li et al. 2003). New clustering methods need to be introduced to address increasing computational times due to ever-increasing volumes of bacterial genomic data. Here we introduce two algorithms that decrease the runtimes needed to create clusters of homologous sequences.

Sequence-clustering algorithms can make use of graphs, trees, or both (Kuzinar et al. 2008). Graph-based approaches typically use all-versus-all sequence homology search results from algorithms such as BLAST (Altschul et al. 1990) to construct a graph, where proteins are nodes, and connections between nodes (i.e., edges) correspond to putatively homologous relationships between the two connected nodes. The choice of homology-search algorithm is important as it can influence the speed, sensitivity and specificity of the analysis (Johnson et al. 2010, Sonhammer et al. 2014, Ward et al. 2014,). Naïve approaches that compare all members of a set of *n* genomes against one another have runtimes and memory use that increases quadratically (Sonhammer et al. 2014). As *n* increases into the thousands of genomes, many algorithms become infeasible on even the most powerful computers, and the majority of proteins can converge into a single "blob" (Harlow and Gogarten 2004). Graph-based methods can also construct large clusters that contain remote or non-homologous sequences due to transitivity, (Bino and Sali 2004, Bolten et al. 2001, Park et al. 1997). *Directed* graphs have asymmetric relationships and have been used in the past to construct clusters of homologs (e.g. Meyerguz et al. 2007), and can be less affected by spurious edges if the graph is purged of connections between nodes that are not bidirectional. Markov clustering (MCL, van Dongen 2000, Enright et al. 2002) is a popular approach to subdividing overly large clusters. OrthoMCL (Li et al. 2003) implements the MCL algorithm and uses a relational database to store information about the graph. Network edges in OrthoMCL are typically based on BLAST matches with a default threshold for drawing edges between nodes at a relaxed e-value of 1e-5, without any consideration of edge direction. Reciprocal best hits (RBH) limits edges to those that connect genes or proteins to one another, requiring each protein to be the best match to the other in its genome. RBHs are frequently used to identify orthologs in Bacteria (e.g. Wolf and Koonin 2012), but may miss orthologs in genomes with high rates gene duplication (Dalquen et al. 2013). Graph-based methods, due to their reliance on all-versus-all homology searches, scale quadratically with the number of sequences (e.g. Sonnhammer et al. 2014), and do not include any information about vertical inheritance of sequences during searches.

In tree-based methods, the topology of a guide tree can be used to identify orthologs in the phylogeny of a cluster. A "species"-level phylogeny can also be used to constrain the homology searches performed between genomes, and therefore limit the total number of homology searches that need to be performed. However, the topology of the guide tree would restrict the order of searches as the guide tree is traversed from leaves to root. This could result in homologs in different lineages failing to cluster if opportunities are not available for clusters to amalgamate with other clusters, or single sequences to join pre-existing clusters. The scaling of this heuristic approach is linear with *n*-1 nodes

(Schreiber and Sonnhammer 2013) where sequence clustering is performed during node traversal. Beiko (2011) developed an approach that identified representative homologs from different taxonomic groups to speed up the orthology reconstruction process for a set of 1080 genomes, but this approach had an unacceptably high rate of false negatives. Hieranoid (Schreiber and Sonnhammer 2013) is an extension of a tree-based method, inParanoid, and uses phylogeny and hidden Markov model (HMM) profiles for cluster construction. As internal nodes of a guide tree are traversed, homology searches between child-node clusters can occur, and homologous clusters are amalgamated using sequence profile searches. However, due to the cluster amalgamation process, the potential exists for clusters to grow to undesirable sizes.

Clustering approaches that incorporate phylogenies connect more directly with the phylogenetic principles of homology, and, as such, may be expected to produce more reliable sets of orthologs and paralogs, where paralogs can be identified in clades in a phylogeny where multiple sequences from one species are present (Kristensen et al. 2011). Given that properties such as homology and orthology have evolutionary definitions, our view is that a phylogenetic method is necessary to obtain correctly defined clusters. Since phylogenetic analysis of "blob" clusters containing hundreds of sequences is infeasible, here we propose a two-step clustering approach that uses phylogenetic information in two distinct ways. In the first step we use the ProPhylClust program to build large homologous sets of proteins, represented as an undirected graph, based on a guide tree of genomes. Methods that create large homologous sets, such as ProPhylClust, without explicit methods to identify othologs will often be referred to as creating "inclusive clusters." ProPhylClust constructs clusters that comprise one or more orthologous sets of proteins. The second step uses the PhyloSubClust program, which uses phylogenetic analysis and tree cutting to extract clusters with putative orthologs from the output generated by ProPhylClust. We compare ProPhylClust to two other clustering methods to create inclusive clusters that rely on all-versus-all sequence homology searches to construct undirected and directed graphs. We compare PhyloSubClust to three methods that create clusters of orthologs, two that partition graphs of homologs, RBH and OrthoMCL, and one hybrid method, Hieranoid. The taxonomic origin of genomes appears to affect clustering with RBH (Dalquen et al. 2013, Wolf and Koonin 2012). To gain some insights into the effect of taxonomic composition on runtimes and clustering with ProPhylClust and PhyloSubClust, we use two different genome sets, one diverse, composed of membership across several phyla, and a less diverse set of genomes from the phylum *Proteobacteria*. To reveal how runtimes and clusters change as genomes are added, serial subsampling of each of those taxonomic sets to create genome sets of increasing size.

3.3 METHODS

ProPhylClust is written in Python version 2.7 (www.python.org), and requires the DendroPy (version 4.2.0) phylogenetic computing library (Sukumaran and Holder 2010) and the network "NetworkX" (version 1.11, Hagber et al. 2008) modules. Sequence data are amino acid sequences in fasta file format. Optionally, Genbank ".gbk" files can be included for annotations. Sequences and optional sequence annotations are stored in an SQLite database (https://www.sqlite.org) for access during ProPhylClust clustering. Descriptions for ProPhylClust and PhyloSubClust are in the following two sections: 3.3.1 and 3.3.2.

3.3.1 ProPhylClust

ProPhylClust uses the topology of a rooted phylogenetic tree to guide clustering of sequences. A rooted, possibly multifurcating phylogeny is required for post-order node traversal. Figure 3.1 shows a rooted species tree of 20 proteobacterial genomes with the order of node traversal numbered and the type of clustering strategy for each topology. A pseudocode version of the ProPhylClust algorithm can be found in Appendix 22. Clustering proceeds progressively as internal nodes of the guide tree are traversed in post order, from the tips to the root, where all children of a node are visited before the node itself. As internal nodes are traversed, the sequences or clusters that descend from the branches of that node are clustered based on homology searches of the two or more descendant branches. Clusters propagate up the tree in this fashion until the root is reached. Unclustered singleton sequences progress up the tree for possible clustering at higher nodes. Clusters are represented as undirected networks with no RBH requirement, where nodes represent sequences and edges indicate homology between the two incident nodes.

Prior to progressive clustering, a first pass is performed to simplify individual genomes. To decrease the number of identical and highly similar sequences in genomes at internal nodes with a leaf as a descendant, sequences with high similarity are grouped together and represented by a single sequence. A directed graph is constructed based on strict thresholds, and strongly connected components are identified. Threshold values to define similar or identical sequences are user specified, with four variables considered: sequence length, alignment length, e-value and percent identity. Directed edges are be drawn between a candidate query to subject if sequence length is within three percent, alignment length is 97%, e-value is less than 1e-90, and percent identify is 97%. After all clustering is complete, these duplicate sequences are entered back into the clusters.

Depending upon the different types of topology at an internal node, three clustering scenarios are applied. The first scenario is applied when only leaves subtend an internal node (Figure 3.1). In that case, all-versus-all homology searches are performed using BLAST and based on the BLAST results, graphs are created to represent clusters. The second scenario is applied if an internal node has exactly one internal node and one or more leaf nodes as children. In this scenario, sequences from the leaf (or leaves) are first added to pre-existing clusters from the sister internal node using either HMMER (Johnson et al. 2010) or BLAST. Remaining leaf-node sequences are then clustered against themselves and singletons from the sister internal node using BLAST. The third scenario is applied when an internal node has at least two internal nodes as children. If any leaf-node children are also present, attempts at clustering them are first made as in the second clustering scheme. At an internal node, if singletons are present, sequence-to-cluster homology searches associate the singleton to pre-existing clusters that were clustered in nodes that descend from the internal node using HMMER, and failing that, attempts to cluster them with singletons in all-versus-all BLAST searches. The clusters from each internal node are then searched against each other using HHsearch (Söding et al. 2005), and clusters are then amalgamated into a single cluster.

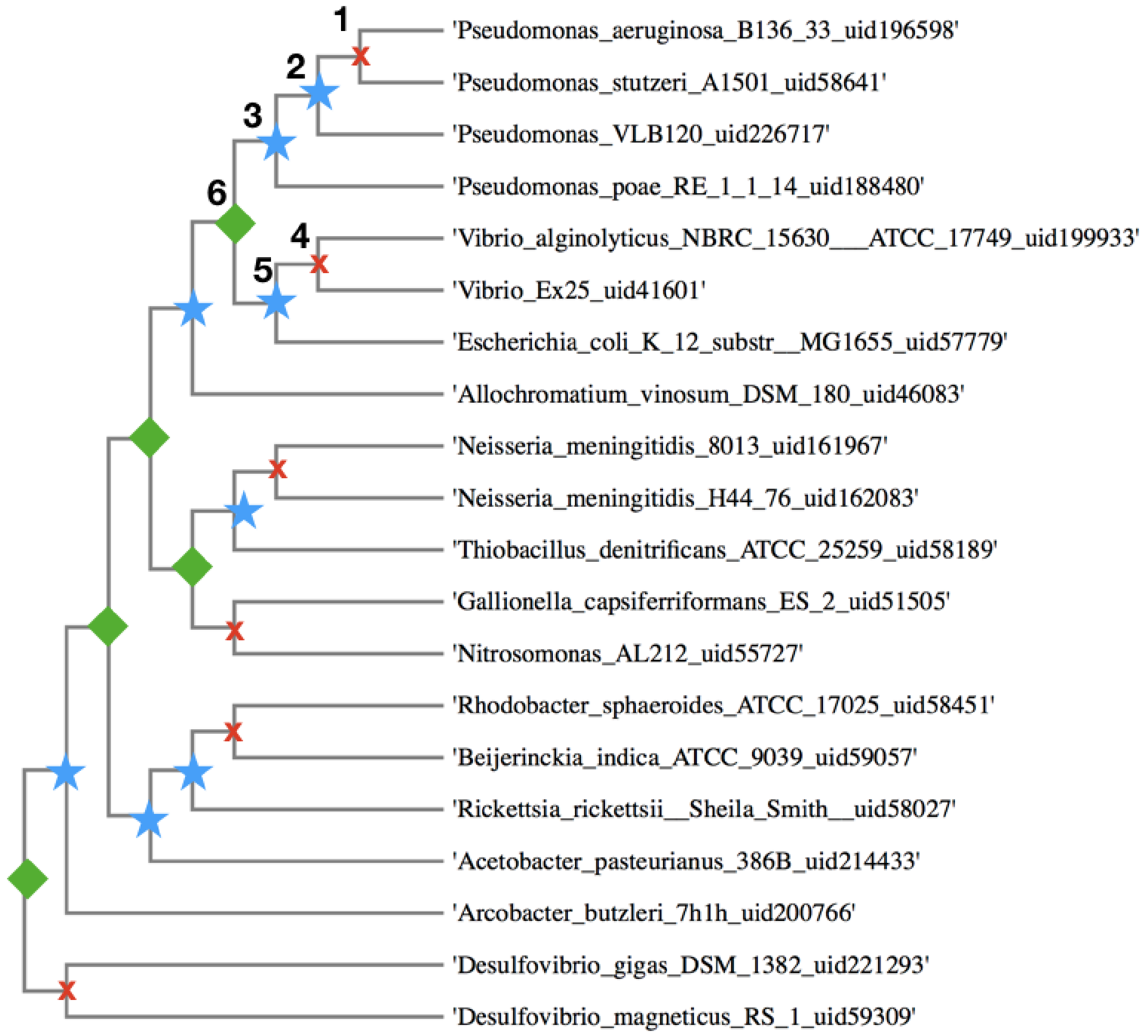

1
2
3
6
4
5
'Pseudomonas_aeruginosa_B136_33_uid196598'
'Pseudomonas_stutzeri_A1501_uid58641'
'Pseudomonas_VLB120_uid226717'
'Pseudomonas_poae_RE_1_1_14_uid188480'
'Vibrio_alginolyticus_NBRC_15630___ATCC_17749_uid199933'
'Vibrio_Ex25_uid41601'
'Escherichia_coli_K_12_substr__MG1655_uid57779'
'Allochromatium_vinosum_DSM_180_uid46083'
'Neisseria_meningitidis_8013_uid161967'
'Neisseria_meningitidis_H44_76_uid162083'
'Thiobacillus_denitrificans_ATCC_25259_uid58189'
'Gallionella_capsiferriformans_ES_2_uid51505'
'Nitrosomonas_AL212_uid55727'
'Rhodobacter_sphaeroides_ATCC_17025_uid58451'
'Beijerinckia_indica_ATCC_9039_uid59057'
'Rickettsia_rickettsii__Sheila_Smith__uid58027'
'Acetobacter_pasteurianus_386B_uid214433'
'Arcobacter_butzleri_7h1h_uid200766'
'Desulfovibrio_gigas_DSM_1382_uid221293'
'Desulfovibrio_magneticus_RS_1_uid59309'

Figure 3.1 Sample ProPhylClust guide tree. Sample post-order node traversal is numbered 1 through 6. At each internal node, symbols X denote all-versus-all BLAST searches, ★ denotes sequence-versus-cluster (HMM and/or consensus) and a ♦ denotes cluster-versus-cluster searches (HMM and/or consensus).

3.3.2 PhyloSubClust

Many of the clusters constructed by ProPhylClust are likely to contain sequences that are too distantly related for use in orthology reconstruction, functional prediction, or phylogenetics. In such cases the cluster must be subdivided. Approaches such as RBH and OrthoMCL use graph-clustering techniques to perform this refinement; although effective, these approaches are not based on phylogeny. By contrast, PhyloSubClust starts with clusters that are typically smaller and tractable for phylogenetic tree reconstruction. From constructed phylogenies, PhyloSubClust recursively extracts sub-trees of putatively orthologous protein sequences from this tree. If needed, PhyloSubClust can create both alignments using MUSCLE (Edgar 2004) and phylogenies using FastTree (Price et al. 2009) from clusters. PhyloSubClust is written in Python 2.7 and uses the DendroPy (version 4.2.0) phylogenetic computing library. Pseudocode for PhyloSubClust can be found in Appendix 23.

PhyloSubClust is a stand-alone sub-clustering algorithm that adapts BranchClust (Poptsova and Gogarten 2007), an algorithm that uses the genome composition of a subtree to select subtrees within phylogenies as the phylogeny is traversed. Figure 3.2 is a sample PhyloSubClust extraction of subtrees from a phylogeny of 29 leaves from nine genomes. Given an input tree covering *n* genomes, PhyloSubClust first attempts to extract complete clusters, clusters where at least one sequence from all *n* genomes are present, and then incomplete clusters with a number of genomes $< n$. Incomplete subclusters can be extracted based on two criteria, first is a sister subtree also contains a threshold number of representative genomes, known as the “MANY”/”FEW” threshold, or if an extracted subtree is reached during node traversal. PhyloSubClust starts cluster selection at leaves that connect directly to the same internal node and have the shortest average leaf-to-leaf branch lengths (the *patristic* distance), which guarantees cluster membership for closest relatives (blue dot, Figure 3.2a). Phylogenies must be rooted for internal nodes to be traversed from the starting node. PhyloSubClust chooses the furthest node, by number of internal nodes, from the starting node for subcluster extraction as a root node (red dot, Figure 3.2a). Internal nodes are then traversed from the starting node, and at each node the number of genomes represented at the internal node is evaluated and classified as either “incomplete”, or “complete”. Incomplete subtrees are classified as either “MANY” or “FEW”, where a user-defined percentage, *p*, of the total number of genomes in a cluster serves as the boundary, *b*: “MANY” >= *b* > “FEW”. For the example in Figure 3.2, nine genomes are present with *p* = 80%, and *b* = seven (rounded to the nearest integer). For each incomplete subtree, the number of genomes represented in the sister subtree is determined to establish whether an incomplete subtree is extracted. When the cluster from the subtree of interest is “MANY” and the cluster from the sister subtree subtending the same parent node is “MANY”, the subtree of interest is pruned for an incomplete cluster (pink dotted box, Figure 3.2b). A stopper node then replaces and represents the location of an extracted subtree (pink leaf, Figure 3.2c), and then the tree is rerooted on the new stopper node (Figure 3.2d). Subclustering then restarts by finding the most distant internal node, by internal node count, from the newly rooted stopper node (blue dot, Figure 3.2d). If a subtree represents sequences from all *n* genomes, it is a complete cluster (orange box, Figure 3.2d), and the subtree is then pruned and replaced by a stopper node (orange leaf, Figure 3.2e). Subclustering then continues with the tree rerooted on the new stopper (Figure 3.2f) If a stopper node is reached during node traversal, the subtree is pruned as a cluster regardless if incomplete or complete (green dotted box, Figure 3.2f), and as with other pruned subtrees, is replaced by a stopper node (green highlighted leaf, Figure 3.2g) and rerooted on the new stopper node (3.2h). In the example in Figure 3.2, a single leaf node remains, and is pruned as a singleton. In the rare event of ties for the starting node, full sub-cluster extraction is performed for each possible start node, choosing the run of extractions that produces the single largest cluster.

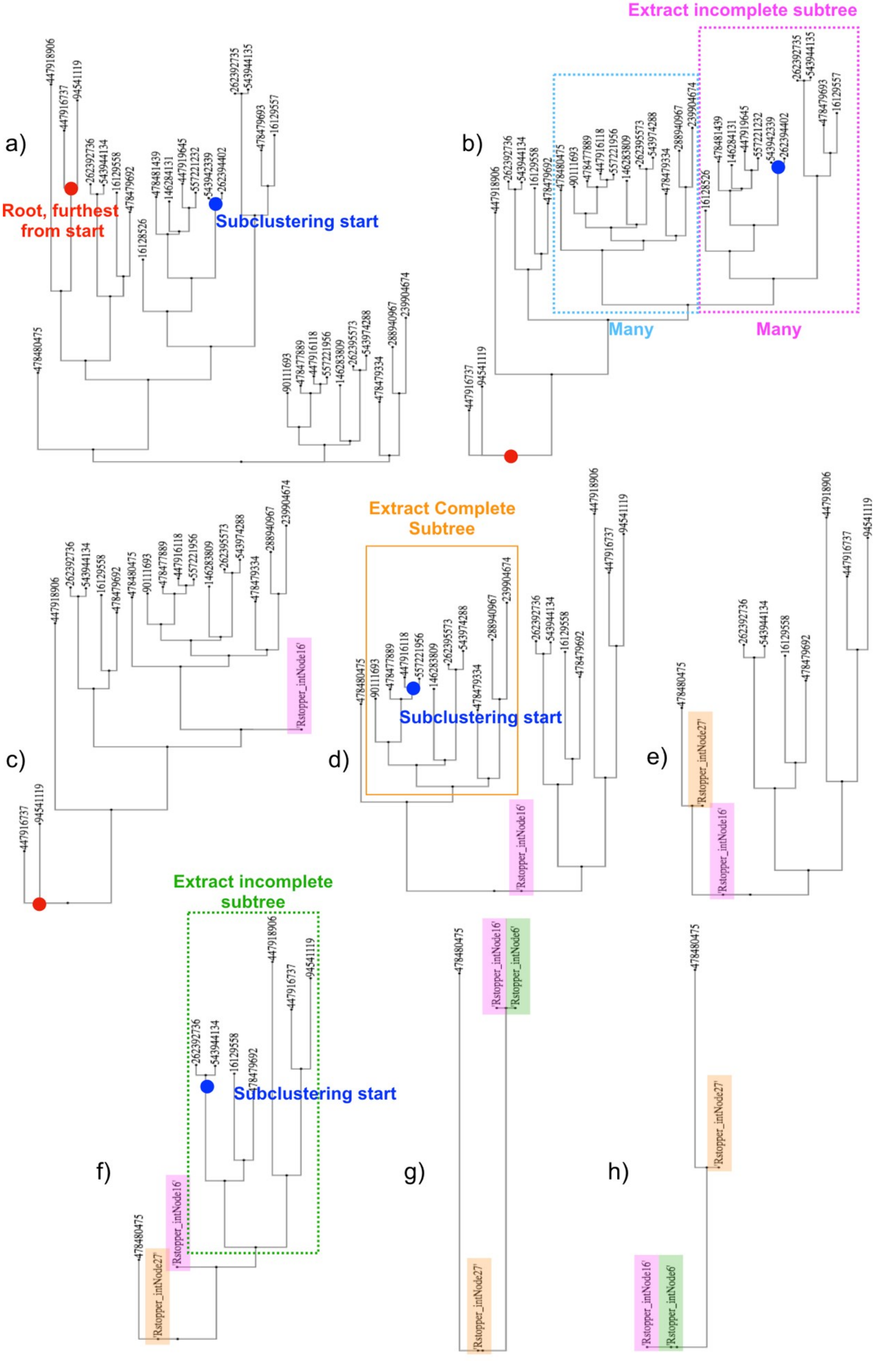

a)
447918906
447916737
94541119
262392736
543944134
16129558
478479692
16128526
478481439
146284131
447919645
557221232
543942339
262394402
262392735
543944135
478479693
16129557
478480475
90111693
478477889
447916118
557221956
146283809
262395573
543974288
478479334
288940967
239904674
Root, furthest from start
Subclustering start
b)
Extract incomplete subtree
Many
Many
c)
'Rstopper_intNode16'
d)
Extract Complete Subtree
Subclustering start
'Rstopper_intNode16'
e)
'Rstopper_intNode27'
'Rstopper_intNode16'
f)
Extract incomplete subtree
Subclustering start
'Rstopper_intNode16'
'Rstopper_intNode27'
g)
'Rstopper_intNode16'
'Rstopper_intNode6'
'Rstopper_intNode27'
h)
'Rstopper_intNode27'
'Rstopper_intNode16'
'Rstopper_intNode6'

Figure 3.2 PhyloSubClust extraction of a cluster generated using ProPhylClust. Phylogeny has 29 leaves from nine genomes with a "FEW" to "MANY" threshold of p = 80% of nine genomes where b = seven genomes. a) A root and starting point are chosen based on the shortest leaves and the internal node most distant from those leaves. b) Internal nodes are traversed and for each node the number of genomes present are determined for the current lineage and the sister lineage. An incomplete subcluster composed of seven genomes and sequnces leaves (dotted pink box) is extracted due to "MANY" criteria for both current and sister lineages (dotted light blue box) from the internal node. c) Extracted incomplete cluster is replaced with a stopper node "Rstopper_intNode16" (pink), and sister cluster (light blue) continues on for future clustering. d) Tree is re-rooted at stopper node, and the most distant internal node (by number of internal nodes) from the root is chosen as starting point for subclustering. A complete cluster is extracted. e) "Rstopper_intNode27" (orange) stopper node is placed in to mark extracted complete cluster. f) Tree is re-rooted on new stopper node and clustering begins from most distant internal node. An incomplete cluster is extracted once a stopper node is reached. g) Cluster is replaced by stopper node "Rstopper_intNode16" (green). h) Tree is re-rooted on new stopper node, with only one sequence remaining.

3.3.3 Genome Sequence Data and Guide Tree Construction

Our tests were based on two sets of 100 finished genomes, one drawn from across 24 different bacterial phyla (Appendix 19 for a list of genomes) which we term the "PanPhyla" data set), and the other a less-diverse set drawn from phylum *Proteobacteria* (Appendix 20 for list of genomes). Genomes were retrieved from the NCBI. For each genome set, we created subsamples of size 20, 40, 60, and 80, in addition to using the whole 100-genome dataset. We also combined both sets into a set of 191 genomes (due to redundancy of some genomes between the two data sets), and added nine additional genomes from class *Clostridia* to bring the total count to 200 (Appendix 21).

Guide trees for each bacterial data set and the subsampled data sets were constructed using 16S rRNA gene sequence data from the downloaded genomes. The 16S ribosomal sequences were aligned against the Ribosomal Database Project's (Cole et al. 2014) curated 16S alignments using the "Aligner" tool. Phylogenies were then constructed using RAxML version 8.2.4 (Stamatakis 2014), using the general time reversible model of nucleotide substitution with gamma-distributed rate variation among sites. We use the best tree as the guide tree. Taxa used to root each guide were selected according to the tree in Hug et al. (2016).

3.3.4 Homology Search, Sequence Alignment and Phylogeny Construction

Homology searches are parallelized using the “multiprocessing” and “subprocessing” modules in Python. To optimize processor usage so that all available processor cores are being used, query sequences are split into separate files, and the number of searches is equivalent to the number of available processor cores. The type of homology search performed depends on the type of comparison. Comparisons between pairs of unclustered protein sequences are performed using BLAST version 2.2.31+ (Camacho et al. 2009), although alternative search programs such as rapsearch2 (Zhao et al. 2012) or DIAMOND (Buchfink et al. 2015) could also be used. It would be inefficient to compare all sequences in a cluster to the target sequence or cluster, and we represent the sequences within a cluster using either a consensus sequence or an HMM. HMMs can effectively represent the variation in many protein sequences, and increase sensitivity in homology searches (Johnson et al. 2010, Söding et al. 2005). Due to their complexity the computational cost of using HMMs can be high. Sequence-vs-cluster searches are performed using HMMER version 3.1b (Johnson et al. 2010), while for cluster-versus-cluster searches we use HMM profile vs HMM profile searches implemented in HHsearch version 2.0.15 (Söding et al. 2005). HMMs are constructed from sequence alignments built using MUSCLE version 3.8.1 (Edgar 2004). Consensus-sequence searches offer a faster but less-sensitive alternative to HMMs. Consensus sequences are also created from MUSCLE alignments, where for each column, the majority rules assign the consensus amino acid, and if more than 50% of entries in a column are gaps, a gap is assigned. In the rare case where no amino acid is the majority, a random amino acid is chosen from those that are observed at the corresponding position in the alignment. In our test runs below, we consider two alternative approaches. In the first approach, we use consensus sequence searches followed by HMMs; if a match is found using consensus sequences, no HMM search is required. The second approach eliminates the use of

HMMs entirely, which is expected to accelerate the runtime at the cost of diminished sensitivity.

To ensure e-values were comparable between BLAST runs with different database sizes, all BLAST searches used a database size "–dbsize" set to 1e9. For graph-based methods, a maximum e-value was set at 1e-5 for homology searches, graphs were then constructed by filtering results of BLAST searches (see section 3.3.5 for e-value and alignment length thresholds for graph construction). For ProPhylClust, most parameters for BLAST searches, HMMER and HHsearch searches are specified in a configuration file, with e-value being the primary parameter value to restrict search results. For BLAST and HHsearch, in addition to e-values, sequence alignment length is provided in tab-delimited/abbreviated results, while for HMMER searches we use full sequence e-values. We also use the sequence lengths/alignment lengths as parameters in clustering, where the relative lengths of a query sequence/alignment to a subject sequence/alignment must fall within a range of user-specified values.

When comparing sequences against sequences and clusters against clusters, we required all sequence pairs to differ by no greater than 40%, be within 0.6 and 1.4 times the length of each other, have a minimum e-value of 1e-10, and have a minimum alignment length that is at least 50% of the length of the sequence, consensus sequence or profile. Since HMMER does not provide alignment length in search results, we did not apply these criteria when mapping individual sequences to HMM clusters.

Sequence alignments for HMM-construction purposes in ProPhylClust and for phylogenetic reconstruction in PhyloSubClust were created using MUSCLE version 3.8.1, with the parameter "-maxiters 2" to optimize alignments with thousands of sequences. When amalgamating a sequence to a homologous cluster or a cluster to a homologous cluster, MUSCLE's profile alignment option was used. During HMMER searches, multiple query sequences may find a single subject HMM profile as the best hit. Those query sequences were used to created and alignment using MUSCLE, and then aligned to the alignment from the best hit HMM profile using MUSCLE's profile alignment option. FastTree version 2.1.10 (Price et al. 2009) is used for phylogenetic construction for PhyloSubClust using the LG model of protein evolution (Le and Gascuel 2008).

3.3.5 Graph Construction

We aimed to compare the performance of PhyloSubClust to graph-based (OrthoMCL version 2.0, RBH) and hybrid (Hieranoid version 1.0) approaches. Network representations and RBH were implemented in Python using the NetworkX package version 1.11. Undirected and RBH graph-based methods define a cluster as a set of nodes that are connected either directly or indirectly via other nodes. The directed graph-based method defines clusters by strongly connected components where a path of edges must exist between any two nodes in both directions. We used e-value thresholds of 1e-30 with a minimum of 50% alignment length for undirected, directed edge and RBH graph methods, which prevented clusters from being composed of sequences of spurious homology. Only e-values can be used as criteria for drawing edges in OrthoMCL, and we used OrthoMCL's default of 1e-5 and 1e-30 to match the other graph-based methods. The last clustering method we attempted, Hieranoid, failed to complete for any of our genome sets and are not included in the results. OrthoMCL often crashed while running the "orthomclPairs" script with the 80 and 100 size Proteobacteria, and 200 size genome sets and had to be re-started using the "=startAfter" option.

3.3.6 Cluster Assessment and Comparisons

All methods were run on a 2 x 2.8 GHz Quad Core Intel Xeon Mac Pro with 12 GB of RAM running Mac OS X version 10.11.6, with usage of all cores for ProPhylClust and BLAST searches. Clock times are reported, as CPU time could not be recorded for ProPhylClust. Undirected graph connected components, directed graph strongly connected components, RBH connected components, OrthoMCL, and PhyloSubClust were run on single processors. It is expected that users supply a guide tree. The runtimes for creation of guide trees are therefore excluded, as methods and sequences to create a guide trees vary between users of ProPhylClust, and is also the case for Hieranoid and its original publication. We also excluded runtimes for creation and population of the SQLite databases for ProPhylClust, and MySQL databases for OrthoMCL, since population of databases is prerequisite step to running ProPhylClust and OrthoMCL. For the three network-based approaches we implemented, and for OrthoMCL, all-versus-all BLAST searches are included in runtimes.

It is difficult to establish non-trivial “ground-truth” data sets to evaluate the accuracy of clustering (Apatoff et al. 2006, Bolten et al. 2001, Dessimoz et al. 2012, Emms and Kelly 2015, Enright et al. 2002, Wiwi et al. 2015). We instead adopted a set of principles and approaches for cluster evaluation and comparison. The simplest measure is the percentage of sequences that were clustered. Sequences that are unclustered are considered to be singletons. We also examine the distribution of cluster sizes for each method to observe average and median cluster sizes, but also maximum cluster sizes, and whether subclustering can reduce the frequency of the largest clusters.

A desirable property of orthology reconstruction is stability of clusters as more sequences are added. Although in some cases the addition of sequences may correctly split up a cluster that was previously thought to be orthologous, in general adding new genomes to a data set should grow existing clusters. We examine the stability of clusters for each clustering method, where we determine the percentage of sequences in clusters that remain identical, become subsets, supersets or are unique when more sequences are added. Three-way proportional Venn diagrams across methods were used to quantify clusters that remain identical between methods. Methods are more similar to each other should share more identical sequence clusters.

3.4 Results and Discussion

3.4.1 Runtimes

Overall, runtimes were dependant on genome set size, genome set type (*Proteobacteria* or PanPhyla), and whether ProPhylClust HMM profile searches were implemented. All runtimes are listed in table 3.1. Although all-versus-all BLAST in itself is not a clustering approach, we include runtimes to contrast methods that require all-versus-all BLAST: directed and undirected graphs, RBH and OrthoMCL. I describe runtimes for PanPhyla and then for the *Proteobacteria* genome sets. Figure 3.3 shows the increases in runtimes for all methods as more genomes are added, and for each genome set type, runtimes for the 200 genome set are included.

For the PanPhyla set (Figure 3.3a and Table 3.1), all runtimes increased as the genome set size increased, regardless of clustering method. Runtimes for ProphylClust and PhyloSubClust with HMMs were consistently greater than all other methods as genome set size increased. Without HMMs, ProPhylClust and PhyloSubClust were faster than all other methods.

For the *Proteobacteria* set (Figure 3.3b and Table 3.1), all runtimes increased as the genome set size increased, regardless of clustering method. Runtimes for ProPhylClust and PhyloSubClust with HMMs were consistently greater than all other methods as genome set size increased until 100 genomes, where runtimes were roughly equivalent to graph-based methods. Without HMMs, ProPhylClust and PhyloSubClust were faster than all other methods. OrthoMCL runtimes for the Proteobacteria set were the longest of all methods at 162,448 and 151,719 seconds for 1e-5 and 1e-30 thresholds, respectively, surpassing runtimes for ProPhylClust and PhyloSubClust with HMMs between 60 and 80 genomes. OrthoMCL's considerable increase in runtimes appears to be due to the calculation of paralog and ortholog tables in MySQL.

Table 3.1 Runtimes in seconds for inclusive clustering and subclustering methods for the PanPhyla, Proteobacteria and 200 genome sets. ProPhylClust is PPC, PhyloSubClust is PSC and Reciprocal Best Hits is RBH. BLAST runtimes are included as a reference point for runtimes of graph-based methods.

		PanPhyla					Proteobacteria					
		20	**40**	**60**	**80**	**100**	**20**	**40**	**60**	**80**	**100**	**200**
Inclusive Clustering	**PPC HMMs**	10,317	28,801	49,156	75,820	111,940	14,260	36,664	60,385	84,365	115,754	270,483
	PPC no HMMs	1,710	7,337	15,099	20,880	40,950	2,451	9,225	17,821	32,471	47,269	152,537
	BLAST	1,890	8,865	20,880	37,260	61,830	4,005	18,225	37,125	72,405	113,985	440,940
	Undirected Graph	1,992	8,934	21,028	37,527	63,286	4,085	18,417	38,088	73,701	116,616	444,724
	Directed Graph	1,927	8,946	21,030	37,529	62,192	4,071	18,424	37,530	73,256	116,663	446,925
Subclustering	**PSC HMMs**	10,699	29,919	50,743	78297	11,5241	15,031	38,353	62,632	87,653	120,437	280,458
	PSC no HMMs	1,968	7,954	15,969	22,190	42,649	3,107	10,155	19,277	34,554	49,929	157,749
	RBH	1,911	8,943	21,050	37,558	62,608	4,057	18,429	37,508	74,967	118,895	447,505
	OrthoMCL 1e-5	2,233	10,481	26,719	47,069	74,985	4,432	21,286	45,753	105,190	162,448	543,875
	OrthoMCL 1e-30	2,011	9,441	25,095	44,887	72,034	4,614	20,735	55,173	96,570	151,719	542,537

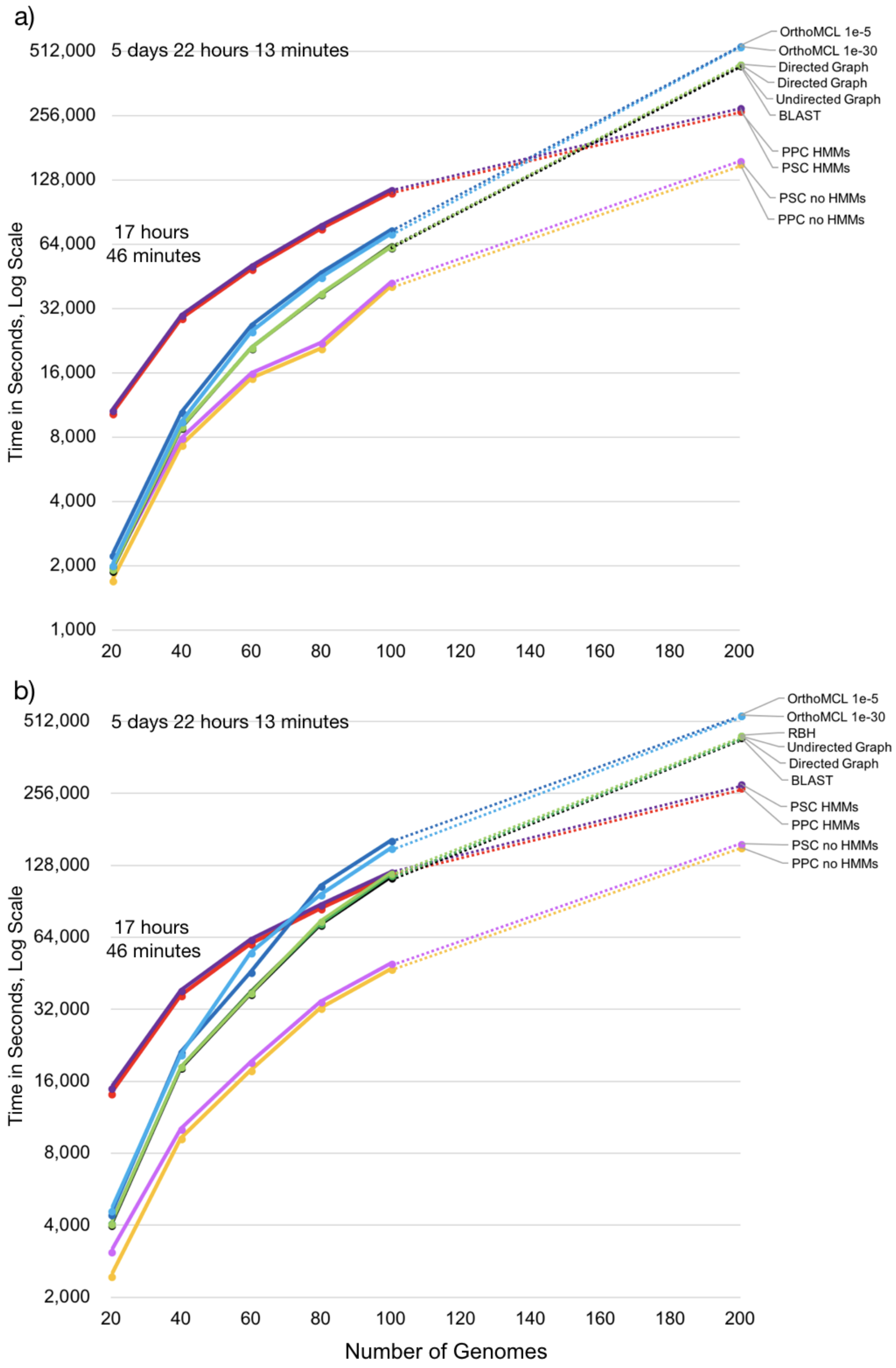

a)
512,000
5 days 22 hours 13 minutes
256,000
128,000
64,000
17 hours
46 minutes
32,000
16,000
8,000
4,000
2,000
1,000
Time in Seconds, Log Scale
20
40
60
80
100
120
140
160
180
200
OrthoMCL 1e-5
OrthoMCL 1e-30
Directed Graph
Directed Graph
Undirected Graph
BLAST
PPC HMMs
PSC HMMs
PSC no HMMs
PPC no HMMs
b)
512,000
5 days 22 hours 13 minutes
256,000
128,000
64,000
17 hours
46 minutes
32,000
16,000
8,000
4,000
2,000
Time in Seconds, Log Scale
20
40
60
80
100
120
140
160
180
200
Number of Genomes
OrthoMCL 1e-5
OrthoMCL 1e-30
RBH
Undirected Graph
Directed Graph
BLAST
PSC HMMs
PPC HMMs
PSC no HMMs
PPC no HMMs

Figure 3.3 Runtimes plotted on a log base 2 scale for all methods for the a) Pan-phyla datasets and the b) *Proteobacteria* datasets. PPC is ProPhylClust and PSC is PhyloSubClust.

The reversal in trends for ProPhylClust/PhyloSubClust runtimes with HMMs compared to BLAST and directed and undirected graph methods are most likely due to the number of potential homologs that are present in the data set. Relative to the *Proteobacteria* genome set, the PanPhyla genome set would most likely have fewer homologs, and therefore fewer searches that BLAST needs to perform. This is reflected in the proportion of sequences clustered (Figure 3.4), where for each genome size set the PanPhyla had a smaller proportion of sequences clustered relative to the equivalent *Proteobacteria*-sized genome set, regardless of clustering method.

For graph-based methods, OrthoMCL (1e-5 and 1e-30), undirected graph, directed graph and RBH runtimes are longer than BLAST runtimes, as expected, since additional algorithms were run after the BLAST runs were finished, and in the case of OrthoMCL, ortholog and paralog filtering is used before subclustering with MCL. OrthoMCL's dependence on a dedicated relational database to access BLAST results to create networks, find duplicates and protein (orthologous/paralogous) pairs impacts its performance with the 200-genome set. It may be that system resources and configuration are a limiting factor with OrthoMCL with very large genome sets due to its dependence on relational database software to store and process BLAST search results. ProPhylClust's use of a guide tree decreases the total number of searches that need to be performed despite the added runtime penalties of HMMER and HHsearch. Memory usage appears to be a limiting factor with undirected and directed graphs and RBH. When an e-value threshold of 1e-5 for drawing edges between nodes was tested, all 12 GB of available system memory was used to store the graphs and runs had to be terminated.

3.4.2 Proportion of Sequences Clustered and Basic Cluster Statistics

Within each genome set, PanPhyla and *Proteobacteria*, the proportion of sequences clustered increased as the size of genome sets increased, regardless of method (Figure 3.4). The *Proteobacteria* data sets consistently had the highest proportion of sequences clustered, with proportions ranging from 0.79 (20 genomes) to 0.86 (100 genomes) for ProPhylClust with HMMs. The PanPhyla, had a smaller proportion of sequences clustered, with proportions ranging from 0.58 (20 genomes) to 0.72 (100 genomes) for ProPhylClust with HMMs. The 200-genome combined data set was intermediate between the two 100-genome sets, with the proportion of sequences clustered under ProPhylClust with HMMs at 0.82, with variation in proportion of sequences clustered exists between other methods.

OrthoMCL with the low 1e-5 threshold consistently clustered the most sequences. It clustered 10% more sequences than the next method, ProPhylClust across all panPhyla genome sets, and approximately 1% more sequences than all *Proteobacteria* sets. Proportion of sequences clustered varied between 0.69 and 0.81 for PanPhyla 20 and 100, respectively and 0.81 and 0.9 for *Proteobacteria* 20 and 100, respectively. A lower threshold of 1e-30 resulted in a smaller proportion of sequences clustered, about 0.52 to 0.66 for the 20 to 100 PanPhyla, respectively and 0.70 to 0.83 for the *Proteobacteria* 20 to 100, respectively. After OrthoMCL with a threshold of 1e-5, ProPhylClust with HMMs clustered the next most number of sequences. With HMMs 3-5% more sequences were clustered than without HMMs. PhyloSubClust subclustering did not drastically reduce the proportion of clustered sequences, often leaving the proportion clustered the same. This was not the case without HMMs, as more singletons resulted after subclustering with PhyloSubClust, often resulting in 3-5% fewer sequences clustered. Undirected graphs, directed graphs and RBH, shared the same e-value threshold as OrthoMCL at a threshold of 1e-30, where undirected and directed graphs had an almost identical proportion of sequences clustered. RBH had the smallest proportion of sequences clusters, often 3% fewer clustered sequences than directed, undirected and OrthoMCL with a threshold of 1e-30.

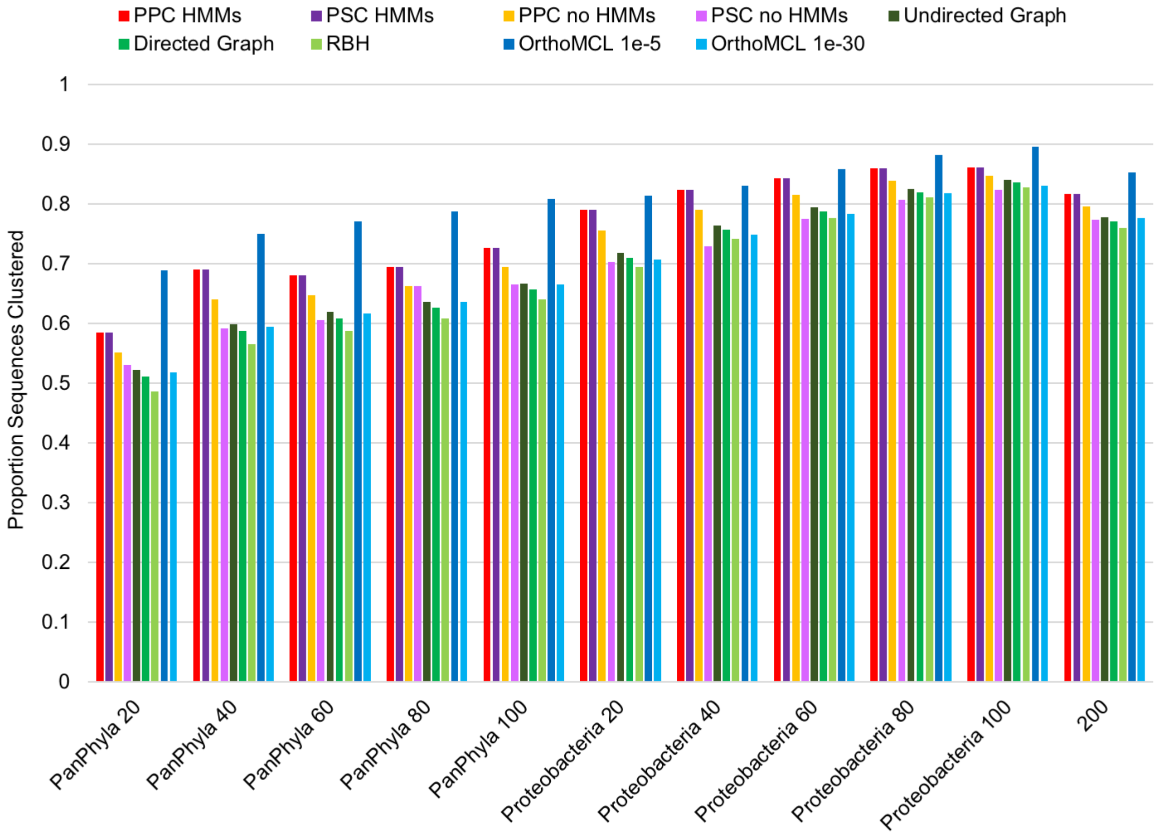


Figure 3.4 Proportion of sequences clustered for each data set for each clustering method. PPC is ProPhylClust and PSC is PhyloSubClust.

The differences between methods are most clear with the 200-genome set, with basic statistics for cluster distributions listed in Table 3.4. Methods that generated larger clusters produced fewer clusters. Mean cluster size for ProPhylClust with HMMs was 18.42 sequences, relative to directed graph and undirected graphs, which had 11.85 and 12.01 sequences, respectively. ProPhylClust with HMMs had the fewest clusters at 36,317, relative to directed graph and undirected graphs, with 44,740 and 43,916 clusters, respectively. The undirected graph, directed graph and RBH approaches produced the largest clusters of all methods, where maximum cluster sizes for the 200 size genome sets were 12,043, 12,632 and 9,270 sequences respectively, while ProPhylClust with HMMs produced the next largest cluster at 3,584 sequences. Across all methods, the minimum cluster size was two and median cluster size was three, with the exception of PhyloSubClust without HMMs with two. ProPhylClust with and without HMMs and OrthoMCL (1e-5) and PhyloSubClust with HMMs had the fewest singletons, as reflected in the proportion of sequences clustered and in the basic statistics (Table 3.4). OrthoMCL had the fewest singletons, most likely due to the highly relaxed clustering threshold of 1e-5, as the stricter threshold of 1e-30 resulted in a number of singletons similar to directed and undirected graph methods.

Basic cluster statistics for all genome sets and methods are listed in Appendix 24. Unsurprisingly, as datasets increased in size, the number of clusters and the mean size of clusters increased with the addition of genomes. As genome set size increased, the *Proteobacteria* genome set had larger mean cluster sizes and more clusters than the PanPhyla genome sets. Mean cluster sizes also increased with increasing number of genomes across all methods, with *Proteobacteria* clusters tending to be composed of more sequences than the PanPhyla data set. Distributions of cluster sizes are long tailed with the majority of clusters consisting of fewer than three sequences and then very large clusters at the long end of the tail. Twenty-fifth percentile (q1) cluster sizes were at two sequences per cluster. Seventy fifth percentile cluster sizes varied considerably, and do not increase in size with more genomes. The Proteobacteria genome sets had the largest 75th percentile (q3), greater than the PanPhyla and the 200 genome sets. The 75th percentile for the 100 genome *Proteobacteria* data set is 7-11 genomes, while the percentiles for the 100 genome PanPhyla data set is four to eight sequences, and 5-9 sequences for the 200 genome data set. The 75th percentiles OrthoMCL, ProPhylClust and PhyloSubClust tended to be largest across all methods and data sets, 6-9 sequences for the 200-genome set versus five to six sequences for undirected, directed and RBH.

The disparity in cluster sizes between ProPhylClust and PhyloSubClust with and without HMMs does emphasize the importance of HMMs in amalgamating homologous clusters as opposed to using consensus sequences only. Comparing ProPhylClust with HMMs to directed graph and undirected graphs, ProPhylClust clusters more sequences into fewer clusters, and based on 75th percentile counts, were more spread across the distribution of cluster sizes. This is unsurprising as HMMs are more sensitive than local alignment methods such as BLAST (Johnson et al. 2010, Söding et al. 2005), and profile-profile methods are extremely effective at finding remote homologs (Bernardes et al. 2015). In addition, ProPhylClust did not produce excessively large maximum sized clusters when compared to graph-based inclusive methods. PhyloSubClust subclustering of ProPhylClust clusters does appear to shift the distribution of clusters away from the long tail, and produces fewer clusters that RBH and OrthoMCL, but still manages to keep, on average clusters larger than both. Although PhyloSubClust does not produce as many clusters as OrthoMCL, others have noted that OrthoMCL can create more subclusters than is desirable (Emms and Kelly 2015).

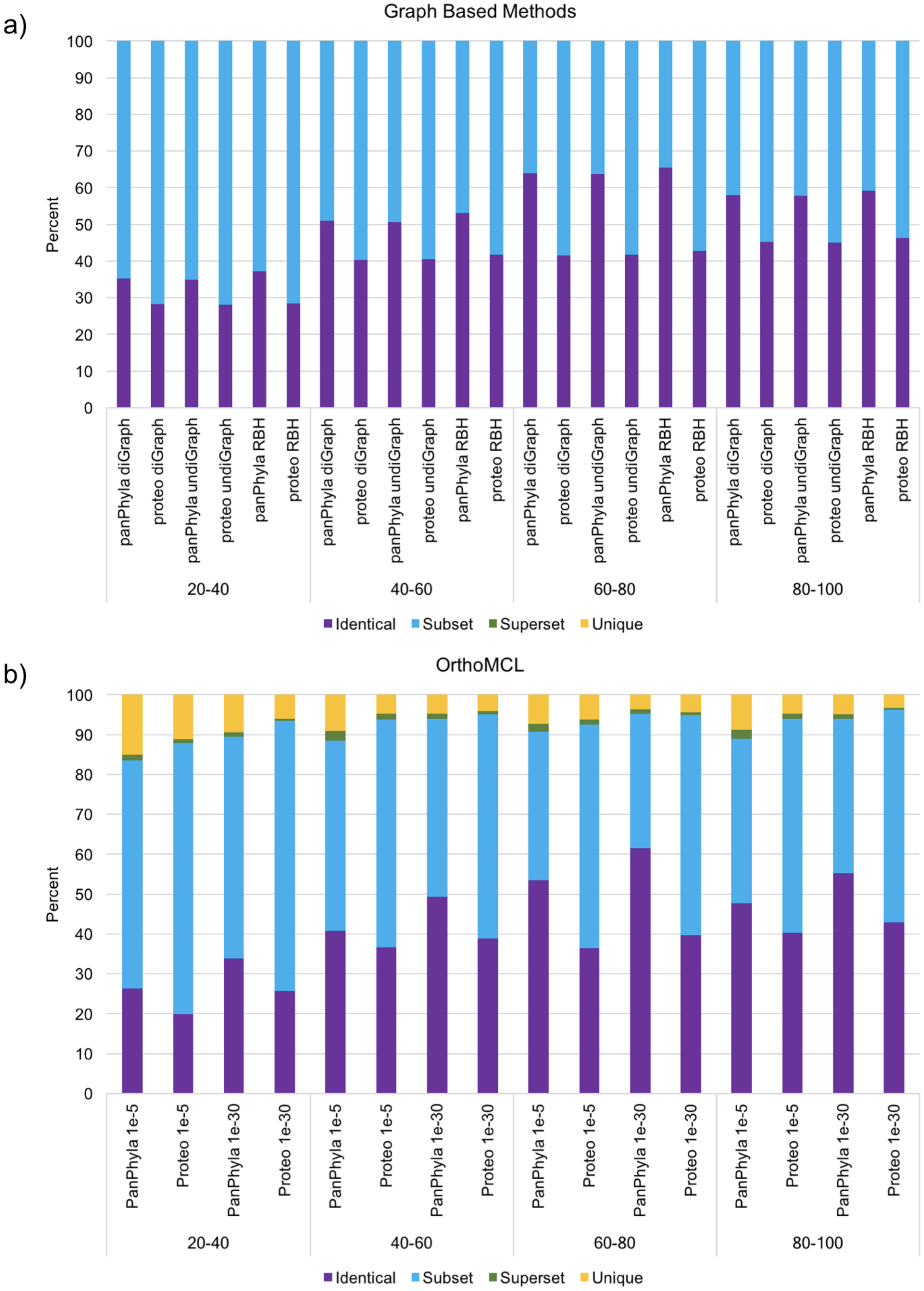
3.4.3 Stability of Clustering Across Data Sets

Methods such as OrthoMCL and PhyloSubClust attempt to partition sets of orthologous sequences, with some degree of tolerance for paralogs, from larger clusters. The addition of genomes should allow for novel cluster formation in addition to increasing the size of pre-existing clusters, and in general, should not subdivide existing clusters of homologous sequences for methods that seek to create, such as ProPhylClust and directed and undirected graph-based methods. The addition of sequences to existing clusters, which we term stability, could be considered a desirable property of clustering, as cluster composition would remain predictable as new genomes are added. To determine how cluster composition varies for each algorithm as genomes are added, we quantified the change in cluster composition as data set size increases. A cluster from a larger data set can be (i) identical to one in a smaller data set, (ii) a proper superset or (rarely), (iii) a proper subset, and (iv) unique (i.e., a new cluster). A unique cluster can share some sequences with clusters (i.e. an intersection) from a smaller or larger genome set, or be composed of sequences from new genomes in the larger set.

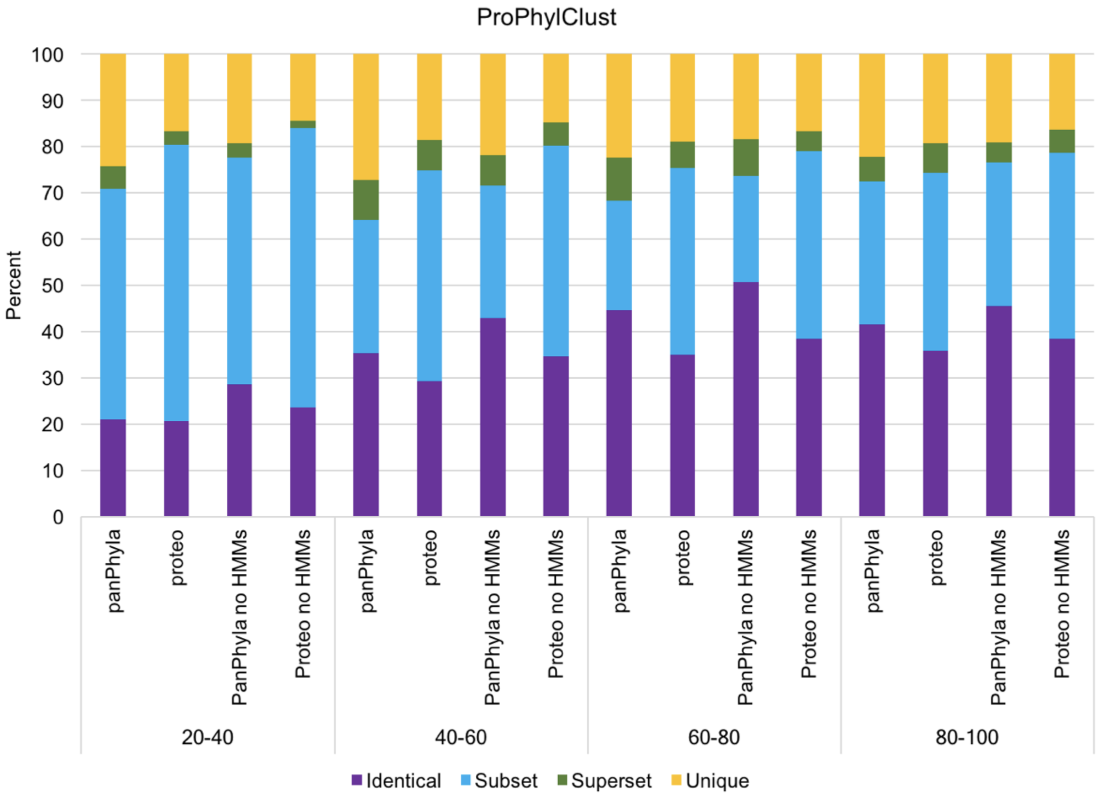
Across all methods, the percentage of identical clusters in the PanPhyla genome set is typically greater than the *Proteobacteria* genome set. New genomes would not typically added to a phylum as genome sets increased in size, limiting opportunities for cluster composition to change. Undirected, directed and RBH graph-based methods resulted in the most stable graphs when moving up a genome set size. Relative to the Proteobacteria genome set, PanPhyla sets displayed approximately 8%, 10%, 20%, and 10% more identical sequences for the 20-40, 40-60, 60-80 and 80-100 genome set comparisons (Figure 3.5a). The number of identical OrthoMCL clusters for the PanPhyla was consistently 5% or more than the *Proteobacteria* set for each genome size set increase. This was also the case and for the 1e-5 ande 1e-30 e-value threshold (Figure 3.5b).

Due to the presence of unique and superset clusters, ProPhylClust has a smaller percentage of identical (up to 25%) or subset clusters (up to 1-9%) compared to graph-based methods (Figure 3.5c). With HMMs, approximately 5% fewer identical clusters were present as the number of genomes increased. The percentage of identical clusters is consistently less by typically 5% for PhyloSubClust relative to ProPhylClust (Figure 3.5d). For ProPhylClust with HMMs the percentage of identical clusters was typically 1-5% lower than with HMMs. The PanPhyla genome sets also had a higher percentage of identical clusters than *Proteobacteria* genome sets, regardless of whether HMMs were implemented, typically between 5-12% identical clusters. No consistent trends were present in the PhyloSubClust comparisons.

No unique clusters or supersets exist for the three graph-based methods (Figure 3.5a), which should be expected given no subclustering is performed. OrthoMCL with an e-value of 1e-5 (Figure 3.5b) has the lowest percentage of unique clusters (1-5%) relative to a 1e-30 (approximately 5-10%). OrthoMCL had no more than 2-3% superset clusters across all genome sets, Fewer unique and superset clusters should be expected for the more stringent e-value threshold of 1e-30, since the MCL algorithm extracts subclusters from smaller starting clusters. ProPhylClust with HMMs has a greater percentage of unique clusters relative to PhyloSubClust by several percent (Figures 3.5c and d). The percentage of superset clusters increased by as much as two-fold with PhyloSubClust, sometimes constituting as much as 15% of all clusters. The addition of genomes and changes to guide tree topology introduced homologs that should be clustered together, splitting up clusters. Although it may be counter-intuitive to have supersets as genomes are added, this may be a valuable component to ProPhylClust. Relative to graph-based methods, the addition of sequences and re-clustering based on the topology of a phylogeny may allow homologous clusters to form and maintain a more cohesive membership, which would be maintained by more the more sensitive HMMER/HHsearch searches.


a)
Graph Based Methods
Percent
100
90
80
70
60
50
40
30
20
10
0
panPhyla diGraph
proteo diGraph
panPhyla undiGraph
proteo undiGraph
panPhyla RBH
proteo RBH
20-40
40-60
60-80
80-100
■ Identical
■ Subset
■ Superset
■ Unique
b)
OrthoMCL
Percent
PanPhyla 1e-5
Proteo 1e-5
PanPhyla 1e-30
Proteo 1e-30
20-40
40-60
60-80
80-100
■ Identical
■ Subset
■ Superset
■ Unique

c)


ProPhylClust
Percent
100
90
80
70
60
50
40
30
20
10
0
panPhyla
proteo
PanPhyla no HMMs
Proteo no HMMs
20-40
40-60
60-80
80-100
Identical
Subset
Superset
Unique

d)

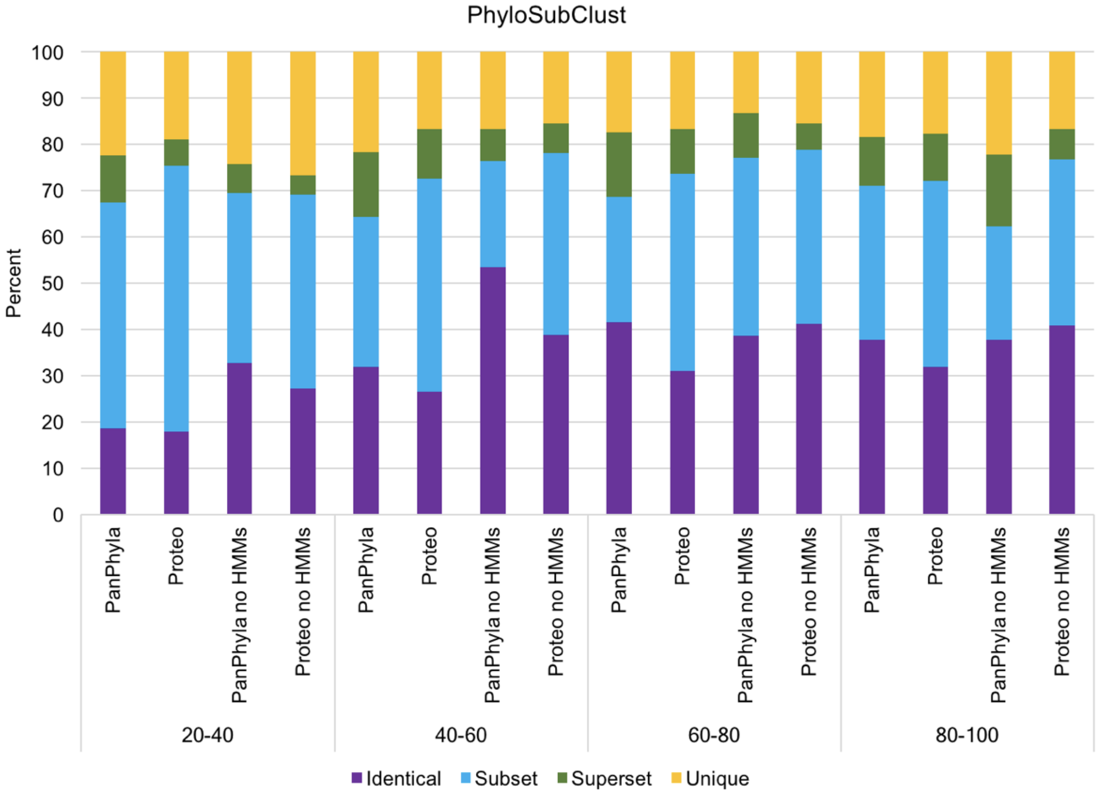

PhyloSubClust
Percent
100
90
80
70
60
50
40
30
20
10
0
PanPhyla
Proteo
PanPhyla no HMMs
Proteo no HMMs
20-40
40-60
60-80
80-100
Identical
Subset
Superset
Unique

Figure 3.5 Stability analysis of clusters for a) Graph-based methods, b) OrthoMCL, c) ProPhylClust and d) PhyloSubClust. Percentage of clusters for a genome set size that are identical, a subset, a superset, and unique. Unique clusters only contain sequences from newly added genomes or share some sequences with other clusters (i.e. an intersection).

3.4.4 Pan-Method Cluster Comparison

We compare ProPhylClust and PhyloSubClust with and without HMMs, we then compare methods to create inclusive type clusters: undirected graphs, directed graphs and ProPhylClust, and methods to partition clusters: PhyloSubClust, RBH and OrthoMCL. We only use the 200-genome sets, which display the clearest patterns. Inclusive clusters and partitioning of clusters are important distinctions, since inclusive methods obtain larger clusters, while partitioning methods seek to obtain clusters of orthologs or contain orthologs. It is therefore important to examine whether inclusive clustering methods and methods to partitioning clusters each produce identical clusters as different classes of clustering methods. The number of common sequences between methods are visualized as counts using three-way Venn diagrams. For inclusive type clusters we compare undirected graphs, directed graphs and ProPhylClust (Figures 3.6a and b), and for methods that partition clusters we compare PhyloSubClust, RBH and OrthoMCL at an e-value threshold of 1e-30 (Figures 3.6c and d). Here, we focus on results from the 200-genome set as a representative of the overall trends seen across all results.

For ProPhylClust with HMMs 65.5% of clusters (19,679/30,040) were shared with ProPhylClust without HMMs, while 43% of clusters (19,679/45,722) without HMMs were shared with ProPhylClust with HMMs. Of those clusters shared, 58% contain more than two sequences, which represents 92% of all clusters with more than two sequences in the ProPhylClust with HMMs cluster set. This does suggest that ProPhylClust without HMMs is a reasonable clustering approach to quickly obtain homologous clusters with more than two sequences. However, for clusters of size two, only 47% from the ProPhylClust with HMMs cluster set are shared.

The percentage of shared ProPhylClust or PhyloSubClust clusters with other methods are presented in Table 3.2. For inclusive clustering methods, the undirected and directed graph-based methods revealed considerable overlap between each other, at approximately 90%. ProPhylClust shares approximately 42% of its clusters with at least one, and approximately 40% with both of the directed and undirected graph-based methods. regardless of whether HMMs are used (Figures 3.6a and b). For methods that partition graphs, the percentage of clusters PhyloSubClust shared with other methods is less than inclusive clustering methods. The increase in the percentage of identical PhyloSubClust clusters from an OrthoMCL e-value threshold of 1e-5 to 1e-30 is most likely a reflection of the initial set of clusters created by OrthoMCL before MCL partitioning subclusters, where similar clustering thresholds resulted in more clusters of similar composition, despite subclustering.

Table 3.2 Percentage of identical clusters shared between ProPhylClust (PPC) and PhyloSubClust (PSC) with and without HMM profile searches and other clustering methods.

	Undirected	**Directed**	**RBH**	**OrthoMCL 1e-5**	**OrthoMCL 1e-30**
PPC w/HMMs	42.3	40.8	-	-	-
PPC w/o HMMs	43.1	41.4	-	-	-
PSC w/HMMs	-	-	33.4	29.1	31.6
PSC w/o HMMs	-	-	32.1	20.5	30.8

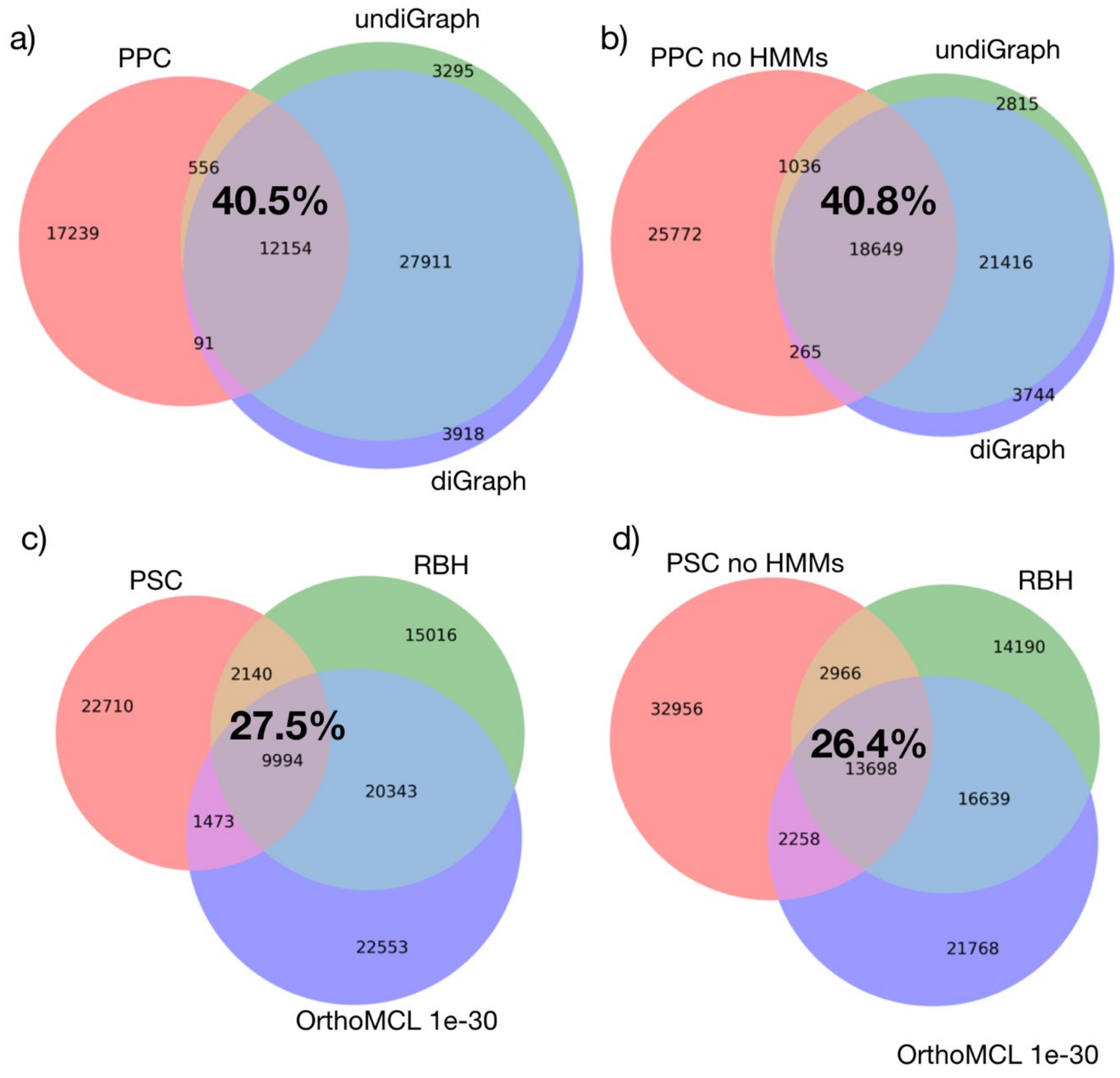


Figure 3.6 Three-way Venn diagrams between clustering methods for clusters from the combined 200 size genome set. PPC is ProPhylClust and PSC is PhyloSubClust. Percentages in bold indicate the percentage of either PPC or PSC sequences shared in the intersection of all three methods. For the inclusive approaches: undirected graphs, directed graphs and ProPhylClust a) with HMMs and b) without HMMs. For the subclustering approaches: c) RBH and OrthoMCL at 1e-30 and PhyloSubClust with HMMs, and d) PhyloSubClust without HMMs and OrthoMCL at 1e-30 and RBH.

3.4.5 Drug Resistance Sample Cluster

The ProPhylClust (with HMMs) sample cluster illustrated in Figure 3.1 features 29 sequences from nine different genomes. It was split by PhyloSubClust into three different clusters: one complete and two incomplete, and one singleton table, 3.5. Twenty of the 29 are drug-resistance related transporters, while three are annotated as hypothetical proteins. In all cases, the proteins are located within the main chromosome of the organism. Almost all are *Gammaproteobacteria*, with the exception of one *Deltaproteobacteria* sequence, which was present in the complete cluster.

Sequences in “complete cluster 4550” all had BLAST e-values < 1e-5. By contrast, for the incomplete clusters, 39/121 and 28/49 sequence pairs did not match one another at this e-value threshold. Sequences from these clusters were identified across two different OrthoMCL clusters, but with some additional sequences. No single protein annotation was the majority for each of the OrthoMCL and ProPhylClust clusters, however it is noted that the evolutionary history of multi and drug specific drug transporters involves frequent functional conversion (Saier et al. 1998). Each of the two OrthoMCL clusters are composed of sequences from different classes of *Proteobacteria*.

Although it is not possible to verify which clusters from OrthoMCL or PhyloSubClust, are closer to or are true clusters, the phylogenies generated from PhyloSubClust clusters can provide some insights into whether sequence membership of partitioned clusters is justified. Of clusters that contained sequences from cluster 4550, one OrthoMCL cluster contained eight sequences together, seven were clustered into incomplete cluster 4550.1 (557221232, 146284131, 447919645, 478481439, 16128526, 239905915, and 543942339), and the remaining sequence, 478479334, was clustered into complete cluster 4550 by PhyloSubClust. That single sequence 478479334 is clearly a member of the sister clade of the other sequences (Figure 3.1b), suggesting that the OrthoMCL cluster should not contain 478479334. The phylogenies of small drug resistance proteins also suggest that proteins cluster by taxonomy (Bay et al. 2008).

Table 3.3 ProPhylClust cluster 4550 composed of sequences from 9 genomes. Subclustered by PhyloSubClust into three clusters, one complete and two incomplete and a singleton.

Refseq GI	Product	Genome
Complete cluster 4550 (9 genomes represented)		
543974288	molecular chaperone SugE	Vibrio_alginolyticus_NBRC_15630___ATCC_17749_uid199933
478477889	transporter	Pseudomonas_aeruginosa_B136_33_uid196598
288940967	small multidrug resistance protein	Allochromatium_vinosum_DSM_180_uid46083
557221956	multidrug efflux pump	Pseudomonas_VLB120_uid226717
146283809	SugE protein	Pseudomonas_stutzeri_A1501_uid58641
262395573	molecular chaperone SugE	Vibrio_Ex25_uid41601
447916118	quaternary ammonium compound-resistance protein	Pseudomonas_poae_RE_1_1_14_uid188480
90111693	multidrug efflux system protein	Escherichia_coli_K_12_substr__MG1655_uid57779
239904674	small multidrug resistance protein	Desulfovibrio_magneticus_RS_1_uid59309
478479334	transporter	Pseudomonas_aeruginosa_B136_33_uid196598
Incomplete cluster 4550.1 (7 genomes represented)		
16128526	DLP12 prophage; multidrug resistance protein	Escherichia_coli_K_12_substr__MG1655_uid57779
447919645	SMR family multidrug resistance protein	Pseudomonas_poae_RE_1_1_14_uid188480
557221232	small multidrug resistance protein	Pseudomonas_VLB120_uid226717
543944135	putative multidrug transmembrane resistance signal peptide protein	Vibrio_alginolyticus_NBRC_15630___ATCC_17749_uid199933
478481439	SMR multidrug efflux transporter	Pseudomonas_aeruginosa_B136_33_uid196598
543942339	quaternary ammonium compound-resistance protein	Vibrio_alginolyticus_NBRC_15630___ATCC_17749_uid199933
146284131	multidrug efflux SMR transporter	Pseudomonas_stutzeri_A1501_uid58641
262394402	quaternary ammonium compound-resistance protein	Vibrio_Ex25_uid41601
478479693	multidrug efflux system protein MdtI	Pseudomonas_aeruginosa_B136_33_uid196598
262392735	spermidine export protein mdtI	Vibrio_Ex25_uid41601
16129557	multidrug efflux system transporter	Escherichia_coli_K_12_substr__MG1655_uid57779
Incomplete cluster 4550.2 (5 genomes represented)		
447916737	hypothetical protein	Pseudomonas_poae_RE_1_1_14_uid188480
16129558	multidrug efflux system transporter	Escherichia_coli_K_12_substr__MG1655_uid57779

Refseq GI	Product	Genome
94541119	undecaprenyl phosphate-alpha-L-ara4N exporter; flippase ArnEF subunit	Escherichia_coli_K_12_substr__MG1655_uid57779
543944134	hypothetical protein	Vibrio_alginolyticus_NBRC_15630___ATCC_17749_uid199933
478479692	putative drug efflux transporter	Pseudomonas_aeruginosa_B136_33_uid196598
447918906	hypothetical protein	Pseudomonas_poae_RE_1_1_14_uid188480
262392736	spermidine export protein mdtJ	Vibrio_Ex25_uid41601
Singleton		
478480475	Transporter	Pseudomonas_aeruginosa_B136_33_uid196598

3.5 Conclusion

We show that the taxonomic composition of clusters can affect runtimes, the percentage of sequences that are clustered, and distributions of cluster sizes. ProPhylClust with HMM profile searches successfully clustered homologous sequences, creating clusters with a larger median and average cluster size than other clustering methods, and clustered a greater percentage of sequences than other methods. PhyloSubClust did not substantially add to runtimes, and also successfully split larger clusters, reducing the maximum observed cluster sizes and decreasing average and median cluster sizes. Although slower compared to other methods with smaller genome sets, ProPhylClust with PhyloSubClust has runtime advantages with the *Proteobacteria* as the number of genomes reached 100 and with the 200 genome set.

Larger genome sets (e.g. 200) reveal that ProPhylClust and PhyloSubClust cluster distributions with HMMs, while long tailed towards large clusters, are not as long as directed and undirected graph methods with the maximum cluster size always smaller than graph-based methods, with the exception of OrthoMCL. This may create a more desirable distribution of cluster sizes, since extremely large clusters may have to be further partitioned to be useful in other analyses.

Runtimes for ProPhylClust and PhyloSubClust without HMMs were almost twice as fast than with HMMs. ProPhylClust and PhyloSubClust had similar runtimes regardless of whether the genome set was *Proteobacteria* or PanPhyla, with similar runtimes for genome sets with HMMs and similar runtimes for genome sets without HMMs. This was not the case with other methods, as their runtimes were longer with the *Proteobacteria* genome set relative to their runtimes with the PanPhyla genome set.

Although the vast majority of clusters with more than two sequences (92%) were obtainable without the use of HMMs in ProPhylClust, less than half were obtained (46%). If one's goal is to quickly obtain clusters with more than two sequences, then that the use of HMM searches may not be necessary, and consensus sequence searches may be sufficient. However, to cluster as many sequences as possible, HMM profile searches should be implemented. Our results with and without HMMs indicate that HMMER and HHsearch increase the sequences clustered (80 versus 82% for 200 genomes) while resulting in fewer (30,040 versus 45,722 for 200 genomes) and larger clusters (18.42 versus 11.81 mean sequences per cluster for 200 genomes).

The main advantage of PhyloSubClust is its use of phylogeny to subcluster clusters, which is an essential component for the orthology definition. Despite lack of phylogeny for inference, RBH is often considered to be a reliable method to obtain orthologs, but not in the presence of numerous gene duplication events (Dalquen et al. 2013. PhyloSubClust clusters are more likely to contain orthologous sequences due to phylogenetic subclustering, but has no explicit step designed to filter out paralogs. PhyloSubClust clusters should be considered to be homologs defined by phylogeny, as no ground truth set of orthologous clusters exist. Some have set out to establish benchmark data sets of orthologs (e.g. Attenhoff et al. 2016), however, without an established ground truth set of orthologs it is not possible to know the absolute accuracy of clustering algorithms.

Future improvements to ProPhylClust include increasing speed and exploring alternative search schemes for cluster to cluster searches. Fast homology search algorithms such as RAPSearch2 (Zhao et al. 2012) could be easily implemented. The usage of consensus sequences to decrease runtimes, although effective for decreasing runtimes, is not an ideal solution and homology searches using consensus sequences may not find all potentially similar clusters. Algorithms such as PSI-BLAST (Altschul et al. 1997) could also be used to speed up sequence-profile searches and if developed, profile-profile searches with a reasonable trade-off in sensitivity relative to HMM profile searches.

CHAPTER 4 PHYLOGENETICS, PHYLOGENETIC PROFILING AND EXPLORATION OF TOXIN-RELATED GENES IN *Peptoclostridium difficile, Clostridium tetani* AND *Clostridium botulinum*

4.1 ABSTRACT

Clostridium botulinum, *Clostridium tetani*, and *Peptoclostridium difficile* are virulent bacteria that have a considerable impact on human health. Here, 558 complete and draft *Clostridia* genomes are used to examine the evolution of toxin-related sequences from 38 *C. botulinum*, two *C. tetani*, and 73 *P. difficile* genomes, and potential toxin-related hypothetical proteins. Clustering with ProPhylClust and PhyloSubClust produced *P. difficile* toxin clusters and *C. botulinum* toxins clustered with *C. tetani* toxins. For each cluster, toxin sequences often grouped in clades with non-toxin sequences, suggesting diverse evolutionary relationships and multiple evolutionary origins for toxins. To obtain potential toxin-related hypothetical sequences, hierarchical clustering of phylogenetic profiles was used to obtain a hierarchy of related clusters. Clusters within a Hamming distance of 0.3 from toxins in the hierarchy highlighted portential toxin-related sequence clusters, all of hypothetical function. These sequences are often found on the same contiguous sequence as toxin sequences. Although these sequences could not be conclusively assigned a functional annotation, they represent potential targets for future inquiry as sequences that may be important for toxin function.

4.2 INTRODUCTION

The class *Clostridia* is a highly diverse group of bacteria which includes many pathogenic organisms, accounting for approximately 20% of known bacterial toxins (Popoff and Bouvet 2013). *Clostridium botulinum*, *Clostridium tetani*, and

Peptoclostridium difficile are three *Clostridia* pathogens with toxins that have known and substantial impacts on human health (e.g. Bruggemann et al. 2003, Elliott et al. 2017, Popoff and Bouvet 2013). Lateral gene transfer (LGT) is a known means of genome evolution in members of *Clostridia* (e.g. Beiko et al. 2005, Meehan and Beiko 2012, Sebaihia et al. 2006) that can transform non-pathogenic organisms into pathogens (e.g. Brouwer et al. 2013, Lacey et al. 2017, Skarin and Segerman 2014). Toxin regulatory proteins and quorum-sensing proteins, as well as "hypothetical" proteins of unknown or putative function should also be considered in addition to toxins when evaluating what contributes to virulence (e.g. Carter et al. 2005, Connan et al. 2013, Martin-Verstraete et al. 2016). Typically, 30-40% of all gene sequences from sequenced bacterial genomes lack an assigned function (e.g. Bork 2000, Galperin and Koonin 2004) and represent a substantial volume of sequence data that needs to be accounted for in pathogenic bacteria. Comparative-genomic approaches such as phylogenetic profiles have been utilized to propose functions for hypothetical proteins (Kensche et al. 2008, Pellegrini et al. 1999), and could be used to highlight candidate hypothetical proteins that are highly likely to be associated to toxins and pathogenesis.

A substantial amount of effort has been directed to understanding the evolution of toxins and their related proteins in *C. botulinum*, *C. tetani* and *P. difficile*. Botulinum neurotoxin (*BoNT*) is the deadliest toxin known (Arnon et al. 2001) and causes foodborne, infant and wound-associated botulism. *BoNT* is well studied, and is typically found in *C. botulinum*, with numerous toxinotypes (A, B, C, D, E, F, G) organized into four groups based on phenotypic and ribosomal differences (Collins and East 1998): I (A, B and F), II (B, E and F), III (C and D), and IV (G). Botulinum neurotoxin and is always associated with and requires non-toxic non-hemagglutinin (*ntnH*) for toxicity. Depending on the phenotypic group of *C. botulinum*, *BoNT* is associated to one of two operons, HA or OrfX, and when expressed, non-*BoNT* sequences from each operon are non-toxic. The OrfX and HA operons are each considered non-toxic components of *BoNT* toxin-protein complexes, with different gene arrangements and evolutionary history, including recombination, insertion events and LGT both within and between species on chromosomes, plasmids and phages (e.g. Dineen et al. 2003, Hill et al. 2009, Marshall et al. 2007, Smith et al. 2007, Williamson et al. 2016). *Clostridium tetani* tetanus toxin (*tetX*) is a distant homolog of *BoNT*, but unlike *BoNT*, only requires one protein to form a toxin and is not believed to be typically transferred between organisms (Brüggemann et al. 2003). *Peptoclostridium difficile* (Yutin et al. 2013), formerly named *Clostridium difficile* and recently renamed to *Clostridioides difficile* (Lawson et al. 2016), has two toxins named A and B whuch are encoded by the *tcdA* and *tcdB* genes, respectively. The toxins, and all other toxin-related proteins are located in a pathogenicity locus (PaLoc) operon on the main chromosome. Although evidence for recent LGT is uncommon (Brouwer et al. 2013), it is proposed that PaLoc was obtained through an LGT event as part of a genetic element, despite partial sequence homology to known mobile genetic elements (Braun et al. 1996).

Various attempts have been made to automate the assignment of functions to hypothetical proteins (e.g. Shahbaaz et al. 2013, Osterman and Overbeek 2003, Pellegrini et al. 1999). Bioinformatic approaches to predict the function of hypothetical proteins in pathogenic bacteria typically involve retrieval sequenced genomes and annotation using homology searches against various sequence databases. Results are then inspected for proteins with annotations that may be associated to pathogenicity (e.g. Alam and Dwiveldi 2016, Mishra et al. 2014). Phylogenetic profiling is a method that uses information about the presence and absence of all sequences across a set of genomes according to homology searches (e.g. Aravind 2000, Pellegrini et al. 1999, Wu et al. 2003), and groups homologous sets of proteins based on the similarity of their corresponding presence / absence patterns. Correlations in the phylogenetic distribution of profiles can imply similar functions for proteins and provide phylogenetically associated sequences of interest for further investigations (Kensche et al. 2008). Others have developed methods to hierarchically cluster phylogenetic profiles based on distance measures. Psomopoulos and Mitkas (2012) created sequence clusters from phylogenetic profiles based on all distances between profiles. Their algorithm creates an initial set of clustered profiles based on a threshold distance value, the centroid profile of each cluster and a threshold distance value is then used for further recursive clustering of clusters. Hierarchical clustering of phylogenetic profiles has also been used to annotate the function of sequences. Liu (2016) used maximum likelihood to hierarchically cluster profiles, and then applied a hierarchy of functional annotations to the clusters based on the branch lengths of the dendrogram of clusters. However, maximum likelihood is computationally intensive and restricts the number of genomes that can be analysed using this technique. Due to the continuously growing volume of sequence data, it is important to develop methods that can be applied to large genomic data sets that are not computationally restrictive.

Here we use a large data set of complete, draft and high-quality assemblies of 558 *Clostridia* genomes to examine the distribution and evolutionary relationships of toxin and toxin-related proteins of *P. difficile*, *C. botulinum* and *C. tetani*. We define homologous relationships among proteins associated with *BoNT*, *TetX*, and *tcdA* and *tcdB*, and assess their phylogenetic distribution. To avoid *a priori* chosen sets of toxins and their homologs, clusters of homologs are created using ProPhylClust and PhyloSubClust, and phylogenetic relationships are examined for insights into the evolutionary history of toxins. Potentially important toxin-related proteins are then identified through phylogenetic profiling, with a focus on hypothetical proteins that are not yet highlighted as functionally relevant to pathogenicity.

4.3 METHODS

4.3.1 Sequence Data and Selection of Candidate Virulence Proteins

A total of 558 *Clostridia* high-quality assemblies, draft, or complete genomes with 16S ribosomal RNA sequences were downloaded from NCBI in 2015, representing a total of 1,550,457 encoded protein sequences. Of the 558 genomes, 284 were assemblies, 143 were draft genomes and 131 were complete genomes (see Appendix 25 for a complete list), of which, 73 *P. difficile* genomes, two *C. tetani* and 38 *C. botulinum* genomes were present. Protein sequences of interest were selected *a priori* based on specific type strains that are known pathogens. ProPhylClust and PhyloSubClust (see Chapter 3) were used to create clusters of homologous proteins, predicted proteins and open reading frames.

Clusters were then identified if they contained toxin and toxin associated proteins from the specific *a priori* type strains.

Genomes of *Clostridium botulinum* A str. ATCC 19397, *C. botulinum* A3 str. Loch Maree, *C. botulinum* B1 str. Okra, *C. botulinum* C str. Eklund, *C. botulinum* E3 str. Alaska E43, and *C. botulinum* F str. 230613 were used as representative strains and represented toxin types A, B, C, E and F (Doxey et al. 2008, Peck et al. 2017). The HA operon (ha33, ha17, ha70) is typically associated with *BoNT* toxinotypes B, C, D, and G, whereas the OrfX operon (OrfX1, OrfX2 and OrfX3) is typically linked to *BoNT* toxinotypes A2, A3, A4, E and F (Popoff and Bouvet 2013). The non-toxic non-haemagglutin (*ntnH*) component is common to all *BoNT*, and forms a toxin complex with the protein products of either the OrfX operon or the HA operon (Popoff and Bouvet 2013). The sigma factor *botR* is required for expression of *BoNT* and the non-toxic components of the toxin complex, and is found in *BoNT* toxinotypes A1, A2 and C and D (Popoff and Bouvet 2013). *Clostridium tetani* 12124569, and *C. tetani* E88 were the only available genomes of *C. tetani*, with only two proteins of importance to virulence, *tetX* and the sigma factor *tetR*.

Peptoclostridium. difficile strain 630, the first *P. difficile* genome sequenced (Sebaihia et al. 2006), and strain R20291 was used to represent typical *P. difficile* genomes. The pathogenicity locus PaLoc for *P. difficile* is composed of *tcdA* and *tcdB*; *tcdR* (an alternative RNA polymerase sigma factors for expression of PaLoc), *tcdC*; (negative regulator of toxin gene expression) for transcriptional regulation (e.g. Popoff and Bouvet 2013); and lastly *tcdE* (holin-like pore-forming protein) which facilitates excretion of *tcdA* and *tcdB* (e.g. Govind and Dupuy 2012, Martin-Verstraete et al. 2016).

4.3.2 Cluster Construction Using ProPhylClust and PhyloSubClust

Sequence to consensus sequence, and consensus sequence to consensus sequence BLAST searches were implemented in ProPhylClust to filter sequence to HMM profile, and HMM profile to HMM profile searches. For guide-tree construction, 16S rRNA gene sequences were aligned using the Ribosomal Database Project's "aligner" tool (Cole et al. 2014). The guide tree was then constructed using RAxML version 8.2.4 (Stamatakis 2014), using the general time reversible model of nucleotide substitution with gamma distributed rate variation among sites. The organisms *Orenia marismortui* DSM 5156, *Halonatronum saccharophilum* DSM 13868, *Halobacteroides halobius* DSM 5150, *Acetohalobium arabaticum* DSM 5501, *Halothermothrix orenii* H 168, *Halanaerobium hydrogeniformans*, and *Halanaerobium praevalens* DSM 2228, were chosen as the root of the tree as the most basal members of *Clostridia* present in the data set based on a recent "tree of life" (Hug et al. 2016). For ProPhylClust's initial reduction of identical or similar sequences, the e-value threshold for clustering identical or redundant sequences was ≤ 1e-90 with sequence percent identity and alignment lengths differing by three percent or less. For clustering, we required all sequence pairs to differ by no greater than 40%, be within 0.6 and 1.4 times the length of each other, have a maximum e-value of 1e-10, and have a minimum alignment length that is at least 50% of the length of the sequence, consensus sequence or profile. Subclusters were extracted using PhyloSubClust with a many/few boundary of 80%. Clusters were then given an annotation based on the majority rule of sequence product annotations supplied by NCBI, where at least 50% of sequences must share the same annotation. Although simple, this annotation approach is intended to identify clusters where the majority of sequences are hypothetical proteins so further investigations can be carried out. The distribution of clusters across the 16S rRNA gene phylogeny was then visualized with a heatmap using the Interactive Tree of Life (https://itol.embl.de, Letunic and Bork 2006).

4.3.3 Annotation of Genomic Islands

IslandViewer (Bertelli et al. 2017) was used to identify genes that are members of genomic islands. For genomes already present in the IslandViewer database, genes that were found partly or completely within a genomic island were considered as being a member of a genomic island. Low quality contigs, draft genomes and assemblies were filtered out from contigs, draft genomes, and assemblies that were not present in the IslandViewer database based on recommendations by Bertelli (personal communication).

Low-quality contigs were less than 1000 nucleotides long, and if the number of contigs in a draft genome or assembly exceeded 300, the genome was excluded. In order for a draft genome to be annotated by IslandViewer, a reference genome already present in IslandViewer must be provided; for each draft genome, the IslandViewer reference genome present in the 16S rRNA tree with the shortest patristic distance was chosen as the reference. A total of 222 genomes (18 complete, 45 drafts, 159 assemblies) were submitted to IslandViewer for annotation, which included *P. difficile* (65 assemblies), *C. botulinum* (two drafts, seven assemblies) and *C. tetani* (two complete) genomes.

4.3.4 Annotation of Plasmids

Potential plasmids were classified from draft *Clostridia* contigs that were not defined as a "complete genome" or a "plasmid". If the majority of a contig is homologous to a portion or a complete plasmid sequence, it most likely represents sequence from a plasmid. Global to local alignment homology searches between contigs of minimum 2000 nucleotides in length and 163 fully sequenced clostridial plasmids downloaded from NCBI were performed using "glsearch36" from the Fasta suite of applications, version 36.3 (Pearson and Lipman 1988). If a contig had an e-value less than 1e-80 with an alignment length of at least 90% the length of the contig to a fully sequenced plasmid, it was considered as a potential plasmid.

4.3.5 Phylogenetic Profiling and Hierarchical Clustering of Profiles

Phylogenetic profiles were constructed from the output clusters after applying PhyloSubClust, based on the presence or absence of each of the 558 *Clostridia* genomes in each cluster. If a gene from a genome is present in a given genome, it is coded as "1" in the profile, otherwise the encoding is "0" for that genome. The result is a binary vector for each cluster, where each element in the vector represents the presence or absence of the gene in the corresponding genome. This differs from typical phylogenetic profiles, where the profile is a binary vector of presence or absence for all genes in the data set, which are based directly on the results of all-versus-all homology searches (e.g. Pellegrini et al, 1999). The use of cluster membership will drastically decrease profile size and therefore memory usage and runtimes for downstream analyses.

After phylogenetic profiles were constructed, hierarchical clustering analysis (HCA) was used to identify clusters that contain sequences that are potentially associated with *BoNT*, *tcdA*, *tcdB*, and *tetX* toxin clusters and their regulatory sequences. Closest neighbours to the clusters of interest in the hierarchy are those that have similar or identical phylogenetic profiles.

The Hamming distance was calculated between all pairs of profiles to obtain a distance matrix of size *i* clusters by *i* clusters. To decrease the size of the distance matrix, the number of clusters was reduced to include only those clusters that contained sequences from any of the *P. difficile*, *C. tetani*, and *C. botulinum* genomes. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm for hierarchical clustering was chosen due to its computational speed and memory efficiency. A custom Python script was used to calculate the Hamming distance for each profile, and the Python script "upgma_cluster.py" from QIIME version 1.9.1 (Caporaso et al. 2010) was used for UPGMA clustering and construction of a UPGMA tree in Newick format. For each of the clusters of interest, the UPGMA tree was used to identify neighbouring clusters that share identical binary taxonomic profiles. To explore other potential functionally valid clusters, which may have different taxonomic content due to missing (unsequenced) genes or gene loss/gain, the Hamming distance threshold was relaxed to 0.1, 0.2 and 0.3 from cluster of interest for clusters with ten or more represented genomes. Clusters are given a singular annotation based on the product annotation from genome ".gbk" files, where the annotation is chosen based on majority rules, where at least half of all sequence product annotations are "hypothetical protein". If a PhyloSubClust cluster wass labelled as a hypothetical protein, PSORTb version 3.02 (Yu et al. 2010) was then used to annotate and provide a cellular localization prediction, and Pfam version 31.0 (e.g. Finn et al. 2016) used to assign a functional annotation.

4.3.6 Sequence Alignment and Phylogenetic Analysis

Using sequences from the *a priori* type strains, a total of 18 toxin and toxin related clusters were identified from the clusters output by PhyloSubClust. Muscle version 3.8.1 was used to create alignments for the 18 clusters. Phylogenies were then constructed using RAxML version 8.2.4 under a general time reversible model of amino acid substitution, with the fast bootstrap option with 100 bootstrap replicates. Nodes with bootstrap values less than 70% were collapsed. All phylogenies are visualized using the Interactive Tree of Life, where sequence product name, contig and chromosome type (unknown, annotated plasmid, plasmid, genome) and genomic island annotations are featured.

4.4 RESULTS

4.4.1 Homologous Clustering and Hierarchical Clustering of Protein Families

The clustering of protein families from 558 genomes using ProPhylClust and PhyloSubClust took approximately six and a half days to complete across 14 Intel® Xeon® X7350 CPUs running at 2.93GHz each with 64 gigabytes of RAM. A total of 1,404,716 sequences (90.6% of all sequences) were assigned to 68,749 clusters. Average cluster size was 20.4 protein sequences per cluster, with a median cluster size of three squences, a minimum of two sequences and maximum of 2401 sequences. Clusters that contained sequences with known GIs of interest were identified, producing 18 clusters associated to the *BoNT* locus, and the HA, OrfX, and PaLoc operons. In particular, I focus on clusters containing *BoNT*, *ntnH* and *tetX*, and *tcdA* and *tcdB*. This produced four clusters of interest which I focus on interpretation of two phylogenies, a *BoNT* and *tetX* cluster "BoNTA-BoNTF-BoNTE-ntnH-Tetx", and a toxinAB_tcdD cluster for t*cdA* and t*cdB*. The other two clusters, a BoNTC cluster and a ntnHC cluster only contained two sequences each, and no phylogeny was constructed.

4.4.2 16S rRNA Guide Tree Topology

Few large pan-*Clostridia* phylogenies exist in the literature (e.g. Collins et al. 1994, Gupta and Gao 2009, Yutin and Galperin 2013). The best 16S rRNA gene tree reveals hypothetical phylogenetic relationships of all 558 *Clostridia* genomes and a heat map indicates proportional representation of genes in each of the 18 clusters from each genome (Figures 4.1 and 4.2). Polyphyletic relationships split *P. difficile* into two clades in separate parts of the 16S rRNA tree (Figure 4.1). Polyphyletic relationships were also observed with *C. botulinum*, which was split into three clades (Figure 4.2). The two *C. tetani* genomes form a monophyletic group, contained in a paraphyletic clade which includes all *C. botulinum* genomes (Figure 4.2). The association of *C. tetani* and some *C. botulinum* with *Clostridium novyi* and *Clostridium sporogenes* as neighbours was observed in other 16S rRNA analyses (Sasaki et al. 2001).

C. botulinum is classified into four groups based on physiology (I, II, III, and IV; Smith and Williams 1975), of genetically diverse organisms. In the 16S rRNA gene tree, the *C. botulinum* genomes are split into clades I, II and III, which correspond to groups I, II and III based on a phylogenetic analysis of 179 orthologous genes (Weigand et al. 2015). However, exact phylogenetic relationships within each group differ. Strain types C and D are more closely related to *C. novyi* and *Clostridium haemolyticum* to form *C. novyi sensu lato*, where they are known to transfer a collection of approximately 61 plasmids (Skarin et al. 2014).

P. difficile was spread across three divergent clades: two distinct clades, labelled clade "A" and "B", and a third clade "C" comprising a single genome (Figure 4.1). The relative position of Clade A is consistent with other studies (e.g. Pereira et al. 2016), and is associated with *Peptostreptococcaceae*, the family that was the basis for the genusname refinement of *Clostridium difficile* to *Peptoclostridium difficile* (Yutin and Galperin 2013). Clade B is composed of 17 genomes associated with *Peptococcaceae* and is unassociated with clade A. Clade C is a single genome that is external to clade B, near the base of the tree, and is associated with *Eubacterium*. It is unknown to what degree the phylogenetic relationships displayed in the 16S rRNA tree represent actual genomic relationships, since 16S rRNA is a single gene, and phylogenetic studies that have sampled as many *P. difficile* isolates are non-existant. However, the phylogenetic relationships among *P. difficile* has recently suggested an affiliation, as the genus *Peptoclostridium*, with the family *Peptostreptococcaceae*, and it is suggested that *P. difficile* may actually comprise as many as four genera (Yutin and Galperin 2013).

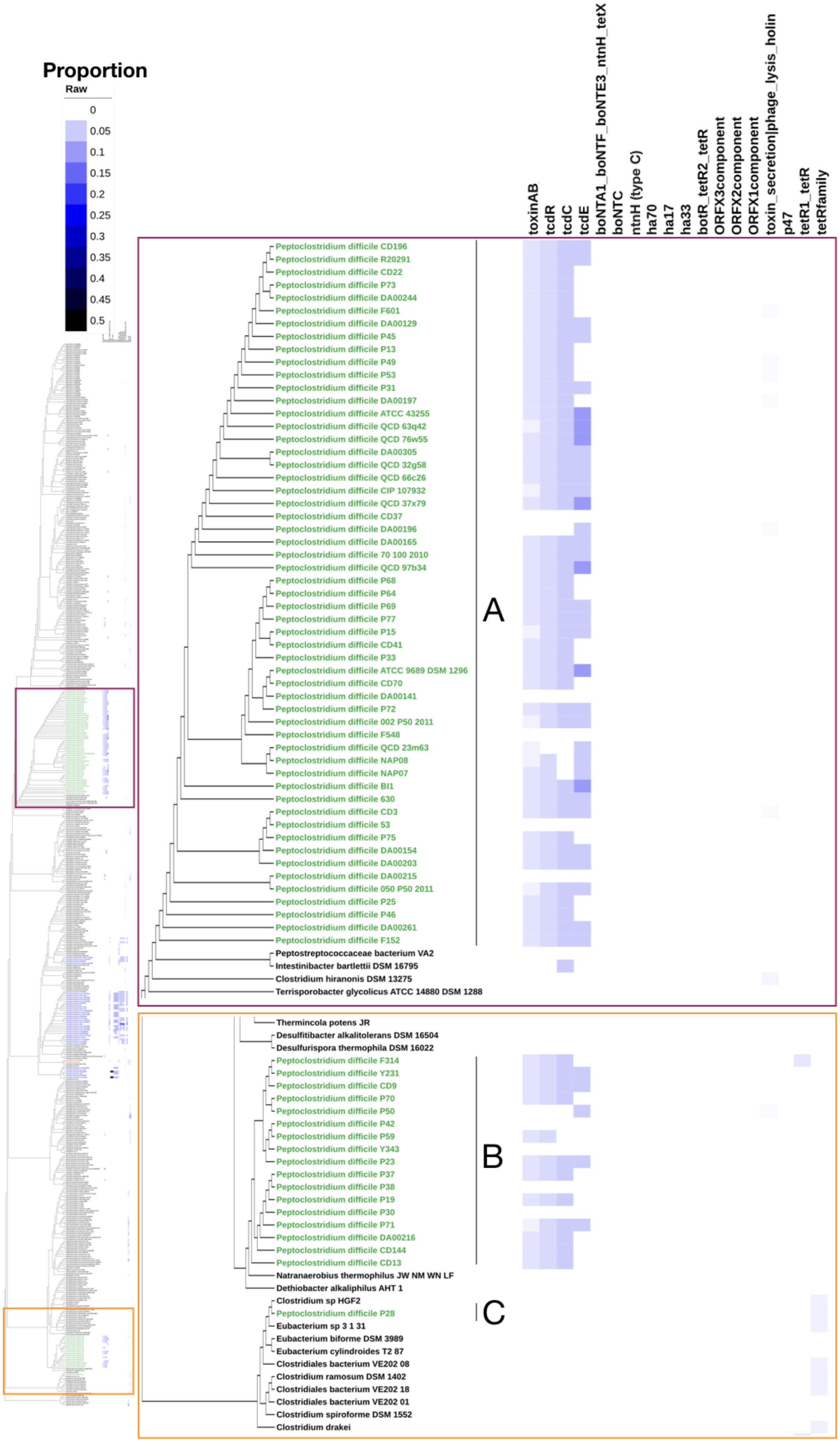

Proportion
Raw
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
toxinAB
tcdR
tcdC
tcdE
boNTA1_boNTF_boNTE3_ntnH_tetX
boNTC
ntnH (type C)
ha70
ha17
ha33
botR_tetR2_tetR
ORFX3component
ORFX2component
ORFX1component
toxin_secretion|phage_lysis_holin
p47
tetR1_tetR
tetRfamily
Peptoclostridium difficile CD196
Peptoclostridium difficile R20291
Peptoclostridium difficile CD22
Peptoclostridium difficile P73
Peptoclostridium difficile DA00244
Peptoclostridium difficile F601
Peptoclostridium difficile DA00129
Peptoclostridium difficile P45
Peptoclostridium difficile P13
Peptoclostridium difficile P49
Peptoclostridium difficile P53
Peptoclostridium difficile P31
Peptoclostridium difficile DA00197
Peptoclostridium difficile ATCC 43255
Peptoclostridium difficile QCD 63q42
Peptoclostridium difficile QCD 76w55
Peptoclostridium difficile DA00305
Peptoclostridium difficile QCD 32g58
Peptoclostridium difficile QCD 66c26
Peptoclostridium difficile CIP 107932
Peptoclostridium difficile QCD 37x79
Peptoclostridium difficile CD37
Peptoclostridium difficile DA00196
Peptoclostridium difficile DA00165
Peptoclostridium difficile 70 100 2010
Peptoclostridium difficile QCD 97b34
Peptoclostridium difficile P68
Peptoclostridium difficile P64
Peptoclostridium difficile P69
Peptoclostridium difficile P77
Peptoclostridium difficile P15
Peptoclostridium difficile CD41
Peptoclostridium difficile P33
Peptoclostridium difficile ATCC 9689 DSM 1296
Peptoclostridium difficile CD70
Peptoclostridium difficile DA00141
Peptoclostridium difficile P72
Peptoclostridium difficile 002 P50 2011
Peptoclostridium difficile F548
Peptoclostridium difficile QCD 23m63
Peptoclostridium difficile NAP08
Peptoclostridium difficile NAP07
Peptoclostridium difficile BI1
Peptoclostridium difficile 630
Peptoclostridium difficile CD3
Peptoclostridium difficile 53
Peptoclostridium difficile P75
Peptoclostridium difficile DA00154
Peptoclostridium difficile DA00203
Peptoclostridium difficile DA00215
Peptoclostridium difficile 050 P50 2011
Peptoclostridium difficile P25
Peptoclostridium difficile P46
Peptoclostridium difficile DA00261
Peptoclostridium difficile F152
Peptostreptococcaceae bacterium VA2
Intestinibacter bartlettii DSM 16795
Clostridium hiranonis DSM 13275
Terrisporobacter glycolicus ATCC 14880 DSM 1288
A
Thermincola potens JR
Desulfitibacter alkalitolerans DSM 16504
Desulfurispora thermophila DSM 16022
Peptoclostridium difficile F314
Peptoclostridium difficile Y231
Peptoclostridium difficile CD9
Peptoclostridium difficile P70
Peptoclostridium difficile P50
Peptoclostridium difficile P42
Peptoclostridium difficile P59
Peptoclostridium difficile Y343
Peptoclostridium difficile P23
Peptoclostridium difficile P37
Peptoclostridium difficile P38
Peptoclostridium difficile P19
Peptoclostridium difficile P30
Peptoclostridium difficile P71
Peptoclostridium difficile DA00216
Peptoclostridium difficile CD144
Peptoclostridium difficile CD13
Natranaerobius thermophilus JW NM WN LF
Dethiobacter alkaliphilus AHT 1
B
Clostridium sp HGF2
Peptoclostridium difficile P28
Eubacterium sp 3 1 31
Eubacterium biforme DSM 3989
Eubacterium cylindroides T2 87
Clostridiales bacterium VE202 08
Clostridium ramosum DSM 1402
Clostridiales bacterium VE202 18
Clostridiales bacterium VE202 01
Clostridium spiroforme DSM 1552
Clostridium drakei
C

Figure 4.1 Phylogeny of all 558 Clostridia genomes based on 16S ribosomal RNA, and heatmap of gene distributions for each genome across 18 clusters of interst. Heatmap colours are proportion of sequences from each genome in the cluster. Enlarged windows are the subtrees that contain *P. difficile* genomes, split into three clades “A”, “B” and “C”.

4.4.3 Heatmap and Cluster Distribution Across Phylogeny

The heatmap provides an overview of the phylogenetic distribution for each of the 18 clusters across all genomes (Figure 4.1 and 4.2). The proportion of sequences provides an indication of which sequences are present or absent in organisms, and whether they tend to be exclusive to a limited set of genomes (i.e. a high proportion across a few genomes). Correlation between heatmaps and phylogenetic patterns can also inform about the evolution of sequences across different lineages. The heatmap of toxins and related sequences identifies possible cases of sequence loss and gain, and cases where the addition of a toxin protein could cause specific strains to become pathogenic, or if virulence may be different between strains due to missing genes or variation in gene content. However, sequences may also be missing due to incomplete genome sequencing; as such, the distributions observed here are the minimum distribution of sequences.

Toxins A and B are restricted to *P. difficile*, and as observed by others, not all *P. difficile* genomes have *tcdA* and *tcdB* (Munoz et al. 2017), and almost all genomes without *tcdA* and *tcdB* are also missing toxin regulatory sequences *tcdR*, *tcdC* and *tcdE*, suggesting they have either not received (through LGT) or have lost the PaLoc operon (Figure 4.1). Further examination of each of *tcdA* and *tcdB* in the “toxinAB” cluster is in the section 4.4.5 “Sequence Cluster Phylogenies”. The toxin secretion/phase lysis holin cluster is associated with the PaLoc operon, where its function is to lyse cell walls during infection. This cluster was uncommon within *P. difficile*, present only in the assembled genomes from strains F601, P49, P53, DA00197, DA00196, CD3, and P50, with none present in the finished genomes. It was more common among *C. perfringens*, the family *Lachnospiraceae*, and *C. botulinum*. However, others have determined that toxin secretion/phase lysis holin function is also performed by tcdE (Monot et al. 2015).

Seventeen *P. difficile* strains with tcdAB lacked *tcdE* and holin, suggesting lack of virulence, however all are incomplete assemblies. The assembled genome of one *P. difficile* strain, da00196, has *tcdE* and holin, but lacks all other toxin-related sequences, suggesting possible loss of the majority of the PaLoc operon.

Four of the six *BoNT* toxinotypes were represented in the 38 *C. botulinum* genomes. The *BoNT* toxins *BoNTA/A1*, *BoNTF*, *BoNTE*, which cause botulism in humans, were clustered by ProPhylClust (but not subclustered by PhyloSubClust) into the "BoNTA-BoNTF-BoNTE-ntnH-Tetx" cluster. The BoNTA-BoNTF-BoNTE-ntnH-Tetx cluster represents a functionally heterogeneous cluster and only members of group I and II had sequences that were a part of this cluster (Figure 4.2). In addition to *BoNT*, *ntnH* and *tetX*, a number of peptidase M27 sequences, progenitor *ntnH* sequences, BoNToxylisin (unknown type of *BoNT*) sequences and a hypothetical protein are present in the cluster. This cluster most likely represents a homologous cluster of sequences, despite different annotated protein products. Other studies select sequences *a priori* genes to group together for analysis, and as because of clustering using ProPhylClust and PhyloSubClust, this is the first time these sequences have been grouped together as a set of homologous sequences. The presence of *tetX* in this cluster should be expected given homology with *BoNT* A, B and E (Eisel et al. 1986). Non-toxin non-haemagglutinin is a known paralog of *BoNT* (e.g. Bhandari et al. 1997, Mansfield et al. 2015), and peptidase M27 is a paralog of *ntnH* and *BoNT* (Doxey et al. 2008, Mansfield et al. 2015). In fact, the sequence identity between between *ntnH* sequences and *BoNT* toxins range from 69% and 95% (Singh et al. 2014). All *BoNT* regulatory sequences *botR* as well as tetX regulatory sequences, *tetR*, were found in the "botR_tetR_tetR2" cluster, and only one strain, "BoNT E Beluga" (draft genome), in group II, did not have a *botR* regulatory sequence.

The other cluster that contained group III *BoNT* toxins was composed of *BoNTC/D* from two C type genomes: "C str Eklund" (draft genome) and "BKT015925" (complete genome), which are the result of recombination of the *BoNTC* and *BoNTD* toxinotypes (Woudstra et al. 2015). Type C/D *ntnH* also formed a distinct cluster. There is no clear explanation why two *BoNT* containing clusters were formed, although unlike the other *BoNT* toxinotypes, *BoNT*C, *BoNT*D, *BoNT*C/D, and *BoNT*D/C mosaics cause botulism in animals and along with their non-toxic and regulatory sequences, are found on a prophage-containing plasmid. *C. botulinum* strain types C and D are also part of *C. novyi sensu lato* species complex. Phylogenetic analysis of multiple toxinotypes of *BoNT* and *ntnH* revealed that types C/D for *BoNT* and *ntnH* each form distinct clades (Mansfield et al. 2015).

The HA operon and OrfX operons were both represented in group I (Figure 4.2), with some *C. botulinum* genomes having both operons, as noted by others (Carter and Peck 2015). Group III genomes had the HA operon, and group II had the ORFX operon, as observed by others (e.g. Connan et al. 2013, Raffestin et al. 2004). This does suggest that the type of operon associated with *BoNT* toxin depends on group type.

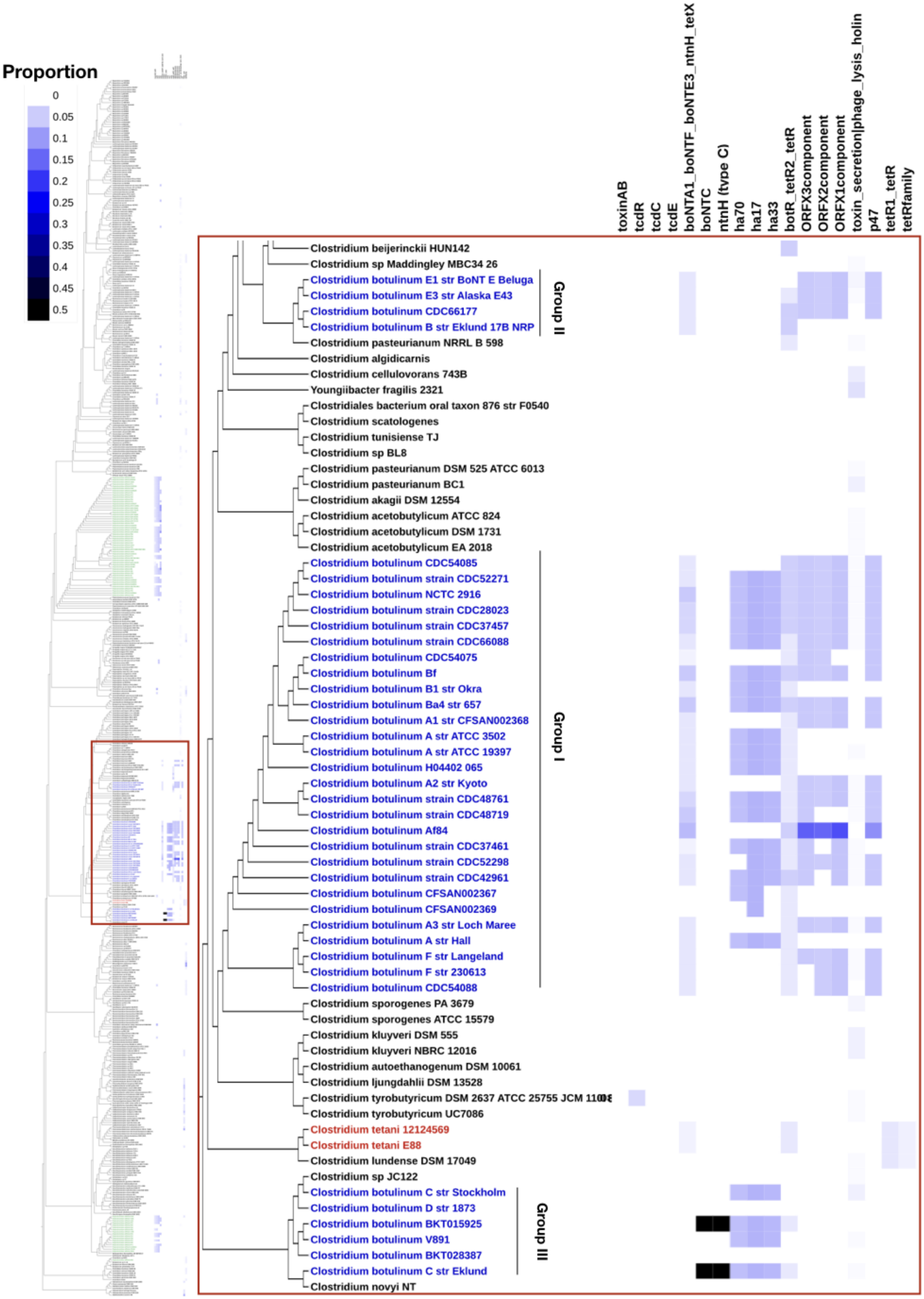

Proportion
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
toxinAB
tcdR
tcdC
tcdE
boNTA1_boNTF_boNTE3_ntnH_tetX
boNTC
ntnH (type C)
ha70
ha17
ha33
botR_tetR2_tetR
ORFX3component
ORFX2component
ORFX1component
toxin_secretion|phage_lysis_holin
p47
tetR1_tetR
tetRfamily
Clostridium beijerinckii HUN142
Clostridium sp Maddingley MBC34 26
Clostridium botulinum E1 str BoNT E Beluga
Clostridium botulinum E3 str Alaska E43
Clostridium botulinum CDC66177
Clostridium botulinum B str Eklund 17B NRP
Clostridium pasteurianum NRRL B 598
Group II
Clostridium algidicarnis
Clostridium cellulovorans 743B
Youngiibacter fragilis 2321
Clostridiales bacterium oral taxon 876 str F0540
Clostridium scatologenes
Clostridium tunisiense TJ
Clostridium sp BL8
Clostridium pasteurianum DSM 525 ATCC 6013
Clostridium pasteurianum BC1
Clostridium akagii DSM 12554
Clostridium acetobutylicum ATCC 824
Clostridium acetobutylicum DSM 1731
Clostridium acetobutylicum EA 2018
Clostridium botulinum CDC54085
Clostridium botulinum strain CDC52271
Clostridium botulinum NCTC 2916
Clostridium botulinum strain CDC28023
Clostridium botulinum strain CDC37457
Clostridium botulinum strain CDC66088
Clostridium botulinum CDC54075
Clostridium botulinum Bf
Clostridium botulinum B1 str Okra
Clostridium botulinum Ba4 str 657
Clostridium botulinum A1 str CFSAN002368
Clostridium botulinum A str ATCC 3502
Clostridium botulinum A str ATCC 19397
Clostridium botulinum H04402 065
Clostridium botulinum A2 str Kyoto
Clostridium botulinum strain CDC48761
Clostridium botulinum strain CDC48719
Clostridium botulinum Af84
Clostridium botulinum strain CDC37461
Clostridium botulinum strain CDC52298
Clostridium botulinum strain CDC42961
Clostridium botulinum CFSAN002367
Clostridium botulinum CFSAN002369
Clostridium botulinum A3 str Loch Maree
Clostridium botulinum A str Hall
Clostridium botulinum F str Langeland
Clostridium botulinum F str 230613
Clostridium botulinum CDC54088
Group I
Clostridium sporogenes PA 3679
Clostridium sporogenes ATCC 15579
Clostridium kluyveri DSM 555
Clostridium kluyveri NBRC 12016
Clostridium autoethanogenum DSM 10061
Clostridium ljungdahlii DSM 13528
Clostridium tyrobutyricum DSM 2637 ATCC 25755 JCM 1108
Clostridium tyrobutyricum UC7086
Clostridium tetani 12124569
Clostridium tetani E88
Clostridium lundense DSM 17049
Clostridium sp JC122
Clostridium botulinum C str Stockholm
Clostridium botulinum D str 1873
Clostridium botulinum BKT015925
Clostridium botulinum V891
Clostridium botulinum BKT028387
Clostridium botulinum C str Eklund
Clostridium novyi NT
Group III

Figure 4.2 Phylogeny of all 558 Clostridia genomes based on 16S ribosomal RNA, and heatmap of gene distributions for each genome across 18 clusters of interest. Heatmap colours are proportion of sequences from each genome in each cluster. Enlarged window is the phylogeny of C. botulinum groups I, II, and III and close relatives.

4.4.4 Annotation of Genomic Islands and Plasmids

A total of 45,909 sequences from 245 genomes were annotated as belonging to genomic islands using IslandViewer, but did not result in any of the *BoNT* locus, HA operon or OrfX operon associated with islands. Likewise, no genomic islands were annotated in relation to PaLoc operon or the *TetX* locus. A total of 137 contigs were annotated as originating from plasmids from 67 genomes, which included sequences for *BoNT*, the HA operon and the OrfX operon and the PaLoc operon. Five genomes had annotated plasmids in the BoNTA-BoNTF-BoNTE-ntnH-tetX cluster, all were peptidase M27. Interestingly, sequences belonging to annotated plasmids were identified in the toxinAB cluster, where six of seven sequences on annotated plasmids were *tcdB*. Those sequences belonged to contigs that shared global to local alignments with *Clostridium perfringens* plasmids, which carry toxins that are related to *tcdA* and *tcdB* (Freedman et al. 2015). The PaLoc operon is not known to be found on plasmids, despite being assumed to be a mobile element, so it is not clear whether these sequences are located on plasmids, or have high sequence similarity to plasmids, or are the result of assembly error.

4.4.5 Sequence Cluster Phylogenies

The annotated phylogenies of all 18 clusters revealed many expected phylogenetic patterns, where sequences with similar or identical functional annotations formed clades. We focus on the two toxin sequence clusters, BoNTA-BoNTF-BoNTE-ntnH-Tetx and toxinAB. For the “toxinAB” cluster (Figure 4.3), *tcdA* and *tcdB* are known homologs (e.g. Bella et al. 2016, Oezguen et al. 2012), but it is unknown if they are orthologs. It should therefore be expected that the two toxins cluster with each other; however, there is a lack of annotated *tcdA* in the phylogeny, with the majority of the sequences labelled either as “peptidase C80” or *tcdB*. Clustering with peptidase C80 should also be expected, given *tcdA* and *tcdB* are members of the peptidase C80 family (Shen 2012). All sequences in the cluster belong to *P. difficile*, and *tcdA* sequences and *tcdB* sequences are polyphyletic. Only five *tcdA* sequences were present. The lack of *tcdA* may be surprising, and could be due to incomplete genome sequencing, but many *tcdA*$^{-}$/*tcdB*$^{+}$ and *tcdA*$^{+}$/*tcdB*$^{+}$ strains exist (e.g. Elliot et al. 2017), with *tcdA*$^{+}$/*tcdB*$^{-}$ only recently discovered (Monot et al. 2015). It should be noted that *tcdA*$^{-}$ and *tcdB*$^{+}$ isolates are routinely isolated from infected patients, but *tcdA*$^{+}$ and *tcdB*$^{-}$ are not (Drudy et al. 2015), suggesting that they are very rare. One *tcdA*$^{+}$/*tcdB*$^{-}$ genome was present, *P. difficile* strain "NAP07", a draft assembly, so it is unknown if it is truly *tcdB*$^{-}$ or is a missing sequence. Not all *P. difficile* genomes contain *tcdA* and *tcdB*, and the PaLoc operon can be transfered from toxin producing to non-toxigenic strains to confer pathogenicity (Brouwer et al. 2013). None of the sequences in the tree were predicted to be associated with a genomic island, but six sequences were annotated as belonging to a plasmid.

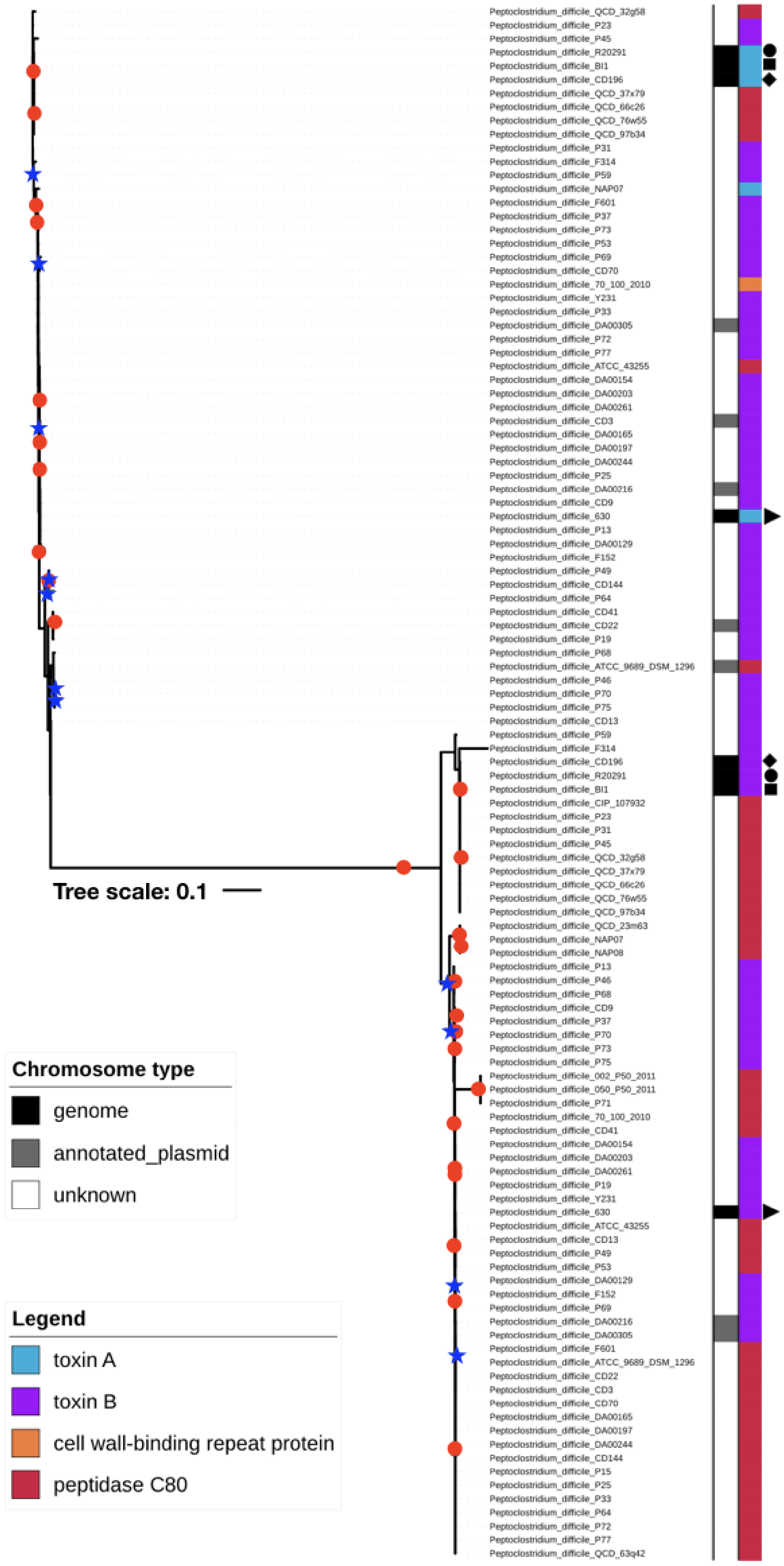

Peptoclostridium_difficile_QCD_32g58
Peptoclostridium_difficile_P23
Peptoclostridium_difficile_P45
Peptoclostridium_difficile_R20291
Peptoclostridium_difficile_BI1
Peptoclostridium_difficile_CD196
Peptoclostridium_difficile_QCD_37x79
Peptoclostridium_difficile_QCD_66c26
Peptoclostridium_difficile_QCD_76w55
Peptoclostridium_difficile_QCD_97b34
Peptoclostridium_difficile_P31
Peptoclostridium_difficile_F314
Peptoclostridium_difficile_P59
Peptoclostridium_difficile_NAP07
Peptoclostridium_difficile_F601
Peptoclostridium_difficile_P37
Peptoclostridium_difficile_P73
Peptoclostridium_difficile_P53
Peptoclostridium_difficile_P69
Peptoclostridium_difficile_CD70
Peptoclostridium_difficile_70_100_2010
Peptoclostridium_difficile_Y231
Peptoclostridium_difficile_P33
Peptoclostridium_difficile_DA00305
Peptoclostridium_difficile_P72
Peptoclostridium_difficile_P77
Peptoclostridium_difficile_ATCC_43255
Peptoclostridium_difficile_DA00154
Peptoclostridium_difficile_DA00203
Peptoclostridium_difficile_DA00261
Peptoclostridium_difficile_CD3
Peptoclostridium_difficile_DA00165
Peptoclostridium_difficile_DA00197
Peptoclostridium_difficile_DA00244
Peptoclostridium_difficile_P25
Peptoclostridium_difficile_DA00216
Peptoclostridium_difficile_CD9
Peptoclostridium_difficile_630
Peptoclostridium_difficile_P13
Peptoclostridium_difficile_DA00129
Peptoclostridium_difficile_F152
Peptoclostridium_difficile_P49
Peptoclostridium_difficile_CD144
Peptoclostridium_difficile_P64
Peptoclostridium_difficile_CD41
Peptoclostridium_difficile_CD22
Peptoclostridium_difficile_P19
Peptoclostridium_difficile_P68
Peptoclostridium_difficile_ATCC_9689_DSM_1296
Peptoclostridium_difficile_P46
Peptoclostridium_difficile_P70
Peptoclostridium_difficile_P75
Peptoclostridium_difficile_CD13
Peptoclostridium_difficile_P59
Peptoclostridium_difficile_F314
Peptoclostridium_difficile_CD196
Peptoclostridium_difficile_R20291
Peptoclostridium_difficile_BI1
Peptoclostridium_difficile_CIP_107932
Peptoclostridium_difficile_P23
Peptoclostridium_difficile_P31
Peptoclostridium_difficile_P45
Peptoclostridium_difficile_QCD_32g58
Peptoclostridium_difficile_QCD_37x79
Peptoclostridium_difficile_QCD_66c26
Peptoclostridium_difficile_QCD_76w55
Peptoclostridium_difficile_QCD_97b34
Peptoclostridium_difficile_QCD_23m63
Peptoclostridium_difficile_NAP07
Peptoclostridium_difficile_NAP08
Peptoclostridium_difficile_P13
Peptoclostridium_difficile_P46
Peptoclostridium_difficile_P68
Peptoclostridium_difficile_CD9
Peptoclostridium_difficile_P37
Peptoclostridium_difficile_P70
Peptoclostridium_difficile_P73
Peptoclostridium_difficile_P75
Peptoclostridium_difficile_002_P50_2011
Peptoclostridium_difficile_050_P50_2011
Peptoclostridium_difficile_P71
Peptoclostridium_difficile_70_100_2010
Peptoclostridium_difficile_CD41
Peptoclostridium_difficile_DA00154
Peptoclostridium_difficile_DA00203
Peptoclostridium_difficile_DA00261
Peptoclostridium_difficile_P19
Peptoclostridium_difficile_Y231
Peptoclostridium_difficile_630
Peptoclostridium_difficile_ATCC_43255
Peptoclostridium_difficile_CD13
Peptoclostridium_difficile_P49
Peptoclostridium_difficile_P53
Peptoclostridium_difficile_DA00129
Peptoclostridium_difficile_F152
Peptoclostridium_difficile_P69
Peptoclostridium_difficile_DA00216
Peptoclostridium_difficile_DA00305
Peptoclostridium_difficile_F601
Peptoclostridium_difficile_ATCC_9689_DSM_1296
Peptoclostridium_difficile_CD22
Peptoclostridium_difficile_CD3
Peptoclostridium_difficile_CD70
Peptoclostridium_difficile_DA00165
Peptoclostridium_difficile_DA00197
Peptoclostridium_difficile_DA00244
Peptoclostridium_difficile_CD144
Peptoclostridium_difficile_P15
Peptoclostridium_difficile_P25
Peptoclostridium_difficile_P33
Peptoclostridium_difficile_P64
Peptoclostridium_difficile_P72
Peptoclostridium_difficile_P77
Peptoclostridium_difficile_QCD_63q42
Tree scale: 0.1
Chromosome type
genome
annotated_plasmid
unknown
Legend
toxin A
toxin B
cell wall-binding repeat protein
peptidase C80

Figure 4.3 Phylogeny of *P. difficile* toxin A (*tcdA*) and toxin B (*tcdB*) along with homologs peptidase C80 and a cell wall-binding repeat protein. Genomes with both *tcdA* and *tcdB* sequences are marked by symbols: *P. difficile* 630▶, *P. difficile* R20291⬣, *P. difficile* BI1 ■, *P. difficile* CD196 ◆. Interior nodes labeled with a ★ have a bootstrap node support of 65-90%, while a ● indicates a support value of 90-100%. The chromosome type "unknown" are contigs from draft or genome assemblies that were not annotated as a plasmid by Fasta glsearch36.

The phylogeny for the BoNTA-BoNTF-BoNTE-ntnH-tetx cluster (Figure 4.4) revealed *BoNT* toxinotypes do not always form clades. Thirteen additional sequences were identified as originating from annotated plasmids. Three sequences were labelled "bontoxilysin A", as unknown *BoNT* toxinotypes. It is known that *BoNT* is found both on chromosomes and on plasmids, and often transferred between different *C. botulinum* strain types (e.g. Skarin and Segerman 2011, Skarin and Segerman 2014, Weigand et al. 2015). Consequently, the 16S rRNA gene phylogeny and the phylogeny based on *BoNT* do not have identical topologies, and it has been recognized for considerable time that *BoNT* is not useful for delineating phylogenetic relationships of different *C. botulinum* strains (e.g. Colins and East, 1998). Sequences from *BoNT* form a clade to the exclusion of *ntnH*. Peptidase M27, a homolog of *ntnH* and *BoNT*, did not form a distinct clade and was often associated with both *BoNT* and *ntnH*, suggesting complicated evolutionary histories for these sequences. Alternatively, this could be an annotation error, and peptidase M27 sequences grouping with *BoNT* sequences could in fact be *BoNT*.

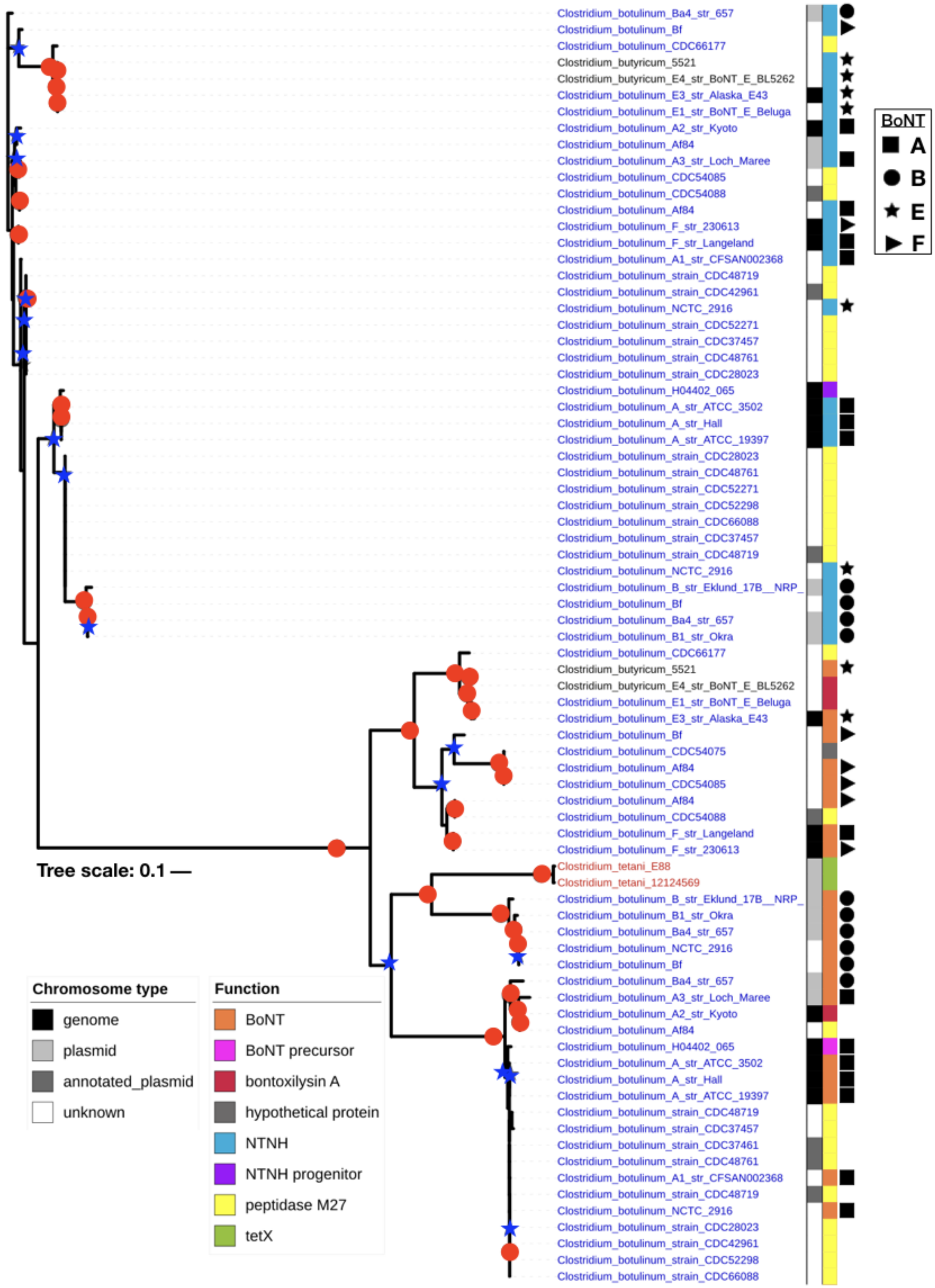

Clostridium_botulinum_Ba4_str_657
Clostridium_botulinum_Bf
Clostridium_botulinum_CDC66177
Clostridium_butyricum_5521
Clostridium_butyricum_E4_str_BoNT_E_BL5262
Clostridium_botulinum_E3_str_Alaska_E43
Clostridium_botulinum_E1_str_BoNT_E_Beluga
Clostridium_botulinum_A2_str_Kyoto
Clostridium_botulinum_Af84
Clostridium_botulinum_A3_str_Loch_Maree
Clostridium_botulinum_CDC54085
Clostridium_botulinum_CDC54088
Clostridium_botulinum_Af84
Clostridium_botulinum_F_str_230613
Clostridium_botulinum_F_str_Langeland
Clostridium_botulinum_A1_str_CFSAN002368
Clostridium_botulinum_strain_CDC48719
Clostridium_botulinum_strain_CDC42961
Clostridium_botulinum_NCTC_2916
Clostridium_botulinum_strain_CDC52271
Clostridium_botulinum_strain_CDC37457
Clostridium_botulinum_strain_CDC48761
Clostridium_botulinum_strain_CDC28023
Clostridium_botulinum_H04402_065
Clostridium_botulinum_A_str_ATCC_3502
Clostridium_botulinum_A_str_Hall
Clostridium_botulinum_A_str_ATCC_19397
Clostridium_botulinum_strain_CDC28023
Clostridium_botulinum_strain_CDC48761
Clostridium_botulinum_strain_CDC52271
Clostridium_botulinum_strain_CDC52298
Clostridium_botulinum_strain_CDC66088
Clostridium_botulinum_strain_CDC37457
Clostridium_botulinum_strain_CDC48719
Clostridium_botulinum_NCTC_2916
Clostridium_botulinum_B_str_Eklund_17B__NRP_
Clostridium_botulinum_Bf
Clostridium_botulinum_Ba4_str_657
Clostridium_botulinum_B1_str_Okra
Clostridium_botulinum_CDC66177
Clostridium_butyricum_5521
Clostridium_butyricum_E4_str_BoNT_E_BL5262
Clostridium_botulinum_E1_str_BoNT_E_Beluga
Clostridium_botulinum_E3_str_Alaska_E43
Clostridium_botulinum_Bf
Clostridium_botulinum_CDC54075
Clostridium_botulinum_Af84
Clostridium_botulinum_CDC54085
Clostridium_botulinum_Af84
Clostridium_botulinum_CDC54088
Clostridium_botulinum_F_str_Langeland
Clostridium_botulinum_F_str_230613
Clostridium_tetani_E88
Clostridium_tetani_12124569
Clostridium_botulinum_B_str_Eklund_17B__NRP_
Clostridium_botulinum_B1_str_Okra
Clostridium_botulinum_Ba4_str_657
Clostridium_botulinum_NCTC_2916
Clostridium_botulinum_Bf
Clostridium_botulinum_Ba4_str_657
Clostridium_botulinum_A3_str_Loch_Maree
Clostridium_botulinum_A2_str_Kyoto
Clostridium_botulinum_Af84
Clostridium_botulinum_H04402_065
Clostridium_botulinum_A_str_ATCC_3502
Clostridium_botulinum_A_str_Hall
Clostridium_botulinum_A_str_ATCC_19397
Clostridium_botulinum_strain_CDC48719
Clostridium_botulinum_strain_CDC37457
Clostridium_botulinum_strain_CDC37461
Clostridium_botulinum_strain_CDC48761
Clostridium_botulinum_A1_str_CFSAN002368
Clostridium_botulinum_strain_CDC48719
Clostridium_botulinum_NCTC_2916
Clostridium_botulinum_strain_CDC28023
Clostridium_botulinum_strain_CDC42961
Clostridium_botulinum_strain_CDC52298
Clostridium_botulinum_strain_CDC66088
BoNT
■ A
● B
★ E
▶ F
Tree scale: 0.1
Chromosome type
■ genome
■ plasmid
■ annotated_plasmid
□ unknown
Function
■ BoNT
■ BoNT precursor
■ bontoxilysin A
■ hypothetical protein
■ NTNH
■ NTNH progenitor
■ peptidase M27
■ tetX

Figure 4.4 Phylogeny of Botulinum neurotoxin (*BoNT*) and its homologs, nontoxic-nonhemaggulutinin (*ntnH*), peptidase M27 and tetanus toxin (*tetX*). *BoNT* toxinotypes for both *BoNT* and *ntnH* are indicated by symbols. Interior nodes labeled with a ★ have a

bootstrap node support of 65-90%, while a ● indicates a support value of 90-100%. The chromosome type "unknown" represents contigs from draft or genome assemblies that were not annotated by Fasta glsearch36.

4.4.6 Protein Families with Hypothetical Functions

Few clusters annotated as "hypothetical function" (i.e. where the majority of sequence products were "hypothetical function") were found to have identical taxonomic profiles as the 18 clusters of interest. However, as the percentage of shared genomes in the taxonomic profiles of clusters decreased, the number of clusters with hypothetical functions increased. I focus on clusters in the hierarchy with a Hamming distance of zero (i.e. identical profile) to 0.1, related to the toxin protein families for *C. botulinum BoNT* and *P. difficile tcdA* and *tcdB*. Four clusters have identical taxonomic profiles to *BoNT*C, which is comprised of two neurotoxin sequences from *C. botulinum* strains BKT015925 and C str Eklund (Table 4.1). Each of the four clusters is composed of two sequences, one from each genome, belonging to the same *C. botulinum* C str Eklund contig and *C. botulinum* BKT015925 plasmid from each genome. Sequences from one of the four clusters, intNode557_47332, were annotated by Pfam as *ParBc* (ParB-like nuclease domain), a protein involved in chromosome partitioning during division; GIs 168187183 and 331271072. Further examination of the surrounding sequences revealed PSORTb annotated sequences from intNode557_47332 as being localized in the cytoplasm of the cell. Other clusters had sequences that were classified as either being localized in the cytoplasm or exported out of the cell (extracellular). Examination of the other proteins on the contig from *C. botulinum* C str Eklund revealed viral related sequence function, suggesting it is most likely belongs to a plasmid encoding a prophage.

Table 4.1 Annotations for four clusters with identical genome profiles as *C. botulinum BoNT*C cluster. Strain C_str_Eklund is a draft genome and BKT015925 is a completely sequenced genome. Sequences from strain C_str_Eklund are located on the same contig, and for strain BKT015925 on the same chromosome as their respective BoNT. Clusters annotated as hypothetical protein based on majority rules.

Cluster ID/Refseq GI	Strain	Product	Contig/ Chromosome	Psortb	Pfam
intNode557_47332					
168187183	C_str_Eklund	hypothetical protein	unknown	Cytoplasmic	ParB-like nuclease domain
331271072	BKT015925	hypothetical protein	plasmid	Cytoplasmic	ParB-like nuclease domain
intNode557_37369					
168187153	C_str_Eklund	hypothetical protein	unknown	Extracellular	unclassified
331271041	BKT015925	hypothetical protein	plasmid	Extracellular	unclassified
intNode557_55110					
168187122	C_str_Eklund	hypothetical phage-related protein	unknown	Cytoplasmic	unclassified
331271014	BKT015925	hypothetical protein	plasmid	Unknown	unclassified
intNode557_47700					
168187118	C_str_Eklund	hypothetical protein	unknown	Unknown	unclassified
331271010	BKT015925	hypothetical protein	plasmid	Extracellular	unclassified

No clusters with identical taxonomic profiles to *P. difficile* cluster toxinAB were observed. The closest cluster annotated as a “hypothetical protein” had a Hamming distance of 0.15, with the addition of *Intestinibacter bartlettii* DSM 16795, *Parvimonas* sp. oral taxon 110 str. F0139, and *Parvimonas micra* A293 (table 4.2). All *P. difficile* sequences in the cluster were located on the same chromosome/contig as the toxin sequences, and annotated by PSORTb as localized in the cytoplasm. The majority of functional annotations are hypothetical proteins (10/20), but others are putative *tcdC* (regulator of *tcdA* and *tcdB* synthesis) variants (6/20) or regulators of gene expression (4/20), suggesting either a distant homolog of *tcdC*, or an alternative/accessory regulator of toxin expression in *P. difficile*.

Table 4.2 Annotations for a single cluster "intnode557_59767", with 85% common taxonomic profile to the *P. difficile* toxin A and B sequence cluster. Strains are *P. difficile* unless stated otherwise. An asterisk (*) indicates belonging to an annotated plasmid. No annotations were identified with Pfam.

Cluster ID/Refsq GI	**Strain**	**Product**	**Contig/ Chromosome**	**Psortb**
intNode557_59767				
544953098*	CD3	putative variant *tcdC*	unknown	Cytoplasmic
544966352	F152	putative variant *tcdC*	unknown	Cytoplasmic
490589310*	70_100_2010	hypothetical protein	unknown	Cytoplasmic
489522058	002_P50_2011	hypothetical protein	unknown	Cytoplasmic
497581513	DA00305	hypothetical protein	unknown	Cytoplasmic
545005943	DA00154	putative variant *tcdC*	unknown	Cytoplasmic
545043482	P13	putative variant *tcdC*	unknown	Cytoplasmic
500187548	CD13	Negative regulator of toxin gene expression	unknown	Cytoplasmic
126698241	630	Negative regulator of toxin gene expression	genome	Cytoplasmic
545018000*	DA00216	putative variant *tcdC*	unknown	Cytoplasmic
545019171	DA00244	putative variant *tcdC*	unknown	Cytoplasmic
648209140	ATCC_43255	Negative regulator of toxin gene expression	unknown	Cytoplasmic
260682357	CD196	hypothetical protein	genome	Cytoplasmic
384359936	BI1	hypothetical protein	genome	Cytoplasmic
497574264	CIP_107932	hypothetical protein	unknown	Cytoplasmic
260685956	R20291	hypothetical protein	genome	Cytoplasmic
648036008	ATCC_9689_DSM_1296	Negative regulator of toxin gene expression	unknown	Cytoplasmic
494497602	Intestinibacter bartlettii DSM_16795	hypothetical protein	unknown	Cytoplasmic
335047586	Parvimonas sp. oral_taxon_110_str_F0139	hypothetical protein	unknown	Unknown
661253301	Parvimonas micra A293	hypothetical protein	unknown	Cytoplasmic

Relaxing the Hamming distance can dramatically increase the number of clusters that may share a functional relationship with a cluster of interest. For toxinAB cluster, relaxing the Hamming distance from 0 to 0.1, 0.2 and 0.3 resulted in an increase from zero, to one, to 293 to 468 clusters. From the clusters within 0.2 and 0.3 Hamming distance, 151/293 and 213/468, respectively, were comprised of proteins with hypothetical function.

4.5 Discussion

The use of 558 *Clostridia* genomes provides a large pool of sequences to explore the diversity of homologs, and novel and informative phylogenetic relationships. The 16S rRNA phylogeny revealed a topology where *P. difficile* and *C. botulinum* each formed polyphyletic relationships. Clustering produced known sequence clusters, and cluster distributions across the phylogeny were consistent with what is generally known about *C. botulinum, C. tetani and P. difficile* toxin sequence distributions. Further analysis of the two clusters containing toxins *tcdA*/*tcdB* and *BoNT*/*tetX* revealed unique phylogenetic relationships between homologous sequences that may be ancestral sequences, convergent evolution, or misannotations. The lack of clustering of some *BoNT*C/D sequences with other *C. botulinum* sequences could reflect a distinctive cluster or could be due to the influence of the topology of the guide tree during clustering with ProPhylClust. The increase in the number of available sequenced genomes has placed an emphasis on the need to develop bioinformatics tools to annotate sequences of hypothetical or unknown function (e.g. Torrieri et al. 2012), and the phylogenetic profiling method developed here has identified candidate proteins of hypothetical function related to toxins.

The taxonomy and evolutionary relationships of *Clostridia* are still in a state of development (Gupta and Gao 2009, Lawson et al. 1993, Yutin and Galperin 2013). The 16S rRNA phylogeny is based on one gene and may not reflect the actual relationships between all genomes present; however, no comparable 16S rRNA *Clostrida* phylogenies with such a wide sampling of 16S rRNA sequences exist in the literature to draw comparisons. Although monophyletic groups were not formed for each of all *C. botulinum* and *P difficile* genomes, distinct clades were formed, which often showed similar topology as others who have used 16S rRNA for phylogenetic reconstructions (e.g. Kurka et al 2014, Weigand et al. 2015).

Although not common in phylogenetic analyses, inclusion of peptidases in BoNTA-BoNTF-BoNTE-ntnH-tetX and toxinAB clusters is a reflection of the autoproteolytic activity of each of the *BoNT*, *tetX tcdA*, and *tcdB* toxins (Lebeda et al. 2010, Shen 2010). The phylogeny of the cluster for *tcdA* and *tcdB* included numerous peptidase C80 sequences with very short branch lengths to *tcdA* and *tcdB* sequences. This suggests possible ancestral relationships with *tcdA* and *tcdB* descending from peptidase C80, but may also represent a misannotation of sequence products. The phylogeny of cluster BoNTA-BoNTF-BoNTE-ntnH-tetX has extremely short branch lengths between peptidase M27 with *BoNT* and *ntnH*. This suggests several possibilities such as ancestral relationships and sequence convergence, but may also be due to misannotation of sequence product.

Recombination has led to chimeric sequences of group I and II *ntnH* (East et al. 1996), which could be problematic for phylogenetic reconstruction. For the BoNTA-BoNTF-BoNTE-ntnH-tetX cluster, it does not appear to affect *BoNT*, as they are members of a single clade (Figure 4.4). However, the exact relationships of *ntnH* sequences could be affected if recombination is frequent.

Group III toxinotypes of *BoNT* and its toxin-related sequences did not cluster with group I and II sequences. Members of *C. botulinum* group III are phylogenetically and physiologically distinct from other *C. botulinum*, and their inclusion with *C. botulinum* is questionable given group III's association with *C. novyi sensu lato* (Skarin et al. 2011). *Clostridium botulinum* group III are only known to cause animal (avian and bovine) botulism. They are known to require bacteriophage for pathogenicity (Ecklund et al. 1971, 1972), and are suspected to have a slower rate of genomic change than other *C. botulinum* (Woudstra et al. 2015). Lateral gene transfer is common in group III *C. botulinum* via plasmids and phages with high-copy mobile elements (Skarin et al. 2011). Due to high divergence of group III *C. botulinum* with other subtypes, it could be expected that ProPhylClust does not cluster the BoNTC or ntnHC clusters with other *BoNT* or *ntnH* sequences. Alternatively, the lack of clustering could be due to an artefact of ProPhylClust where topological constraints and lack of homologs prevented clustering at internal nodes. The BoNTA-BoNTF-BoNTE-ntnH-tetX is a large and diverse sequence cluster, and the consensus sequence created by ProPhylClust may have been too divergent from the consensus sequences for the BoNTC and ntnHC clusters during the BLAST filtering steps for HMM profile-HMM profile searches at the internal nodes between groups I and II with group III. The lack of clustering of *BoNT*C with *ntnH* type C is also unexplained. Only two studies have focused on the distant evolutionary relationships of *BoNT* toxins and related toxin complex sequences and peptidase M27 is considered a homolog (Doxey 2008, Mansfield 2015). However, peptidase M27 was not present in *C. botulinum* group III genomes. Their presence in group I and II may have added additional sequence variation to the consensus sequences and HMMs to join the ntnH and BoNT clusters, but at the internal node with group I, II, and III, sequences may have diverged to the point where e-values did not meet the required threshold for clustering.

Only 18 of the available 68,749 clusters were used in this analysis, leaving a large proportion of the data set that could be used for the analyses of other sequences. It has been recognized that the addition of sequence data increases the accuracy of functional predictions from phylogenetic profiles, but at the cost of computational time (Škunca and Dessimoz 2015). Although there are sampling strategies for reducing the number of genomes needed for phylogenetic profiling (e.g. Simonsen et al. 2012), one of the main objectives was to examine evolutionary relationships of toxin sequences, and taxonomic sampling of closely related genomes is necessary.

As the number of sequenced genomes continues to increase, proteins with hypothetical function continue to accumulate. It is therefore important to develop additional techniques to prioritize proteins with unknown function for bacteria pertinent to human health, such as pathogenic bacteria (e.g. Alam and Dwiveldi 2016, Mishra et al. 2013). Inclusion of additional sequence data set generally improves predictions made by phylogentic profiles (Škuna and Dessimoz 2015). However, our inclusion of incomplete genomes could falsely exclude membership of a genome in a cluster. Our use of cluster membership instead of sequence homology introduces the potential for error into our phylogenetic profiles. For example, if cluster membership varies dramatically due to multiple homologs of different function and includes genes obtained through LGT, the taxonomic membership of the profiles would reflect this. This could be the reason no clusters with identical taxonomic composition were found for the BoNTA-BoNTF-BoNTE-ntnH-tetX cluster.

Although phylogenetic profiling can inform about function and phenotype, it should be combined with additional evidence such as laboratory confirmation to confidently establish protein function (Kensche et al. 2008). It may also be possible to use other genome-context approaches to establish function of hypothetical proteins. Gene neighbourhoods use information from the conserved order of neighbouring sequences on a chromosome to provide a functional context for sequences (Korbel et al. 2004). For example, “integration host factor” was found upstream (~6000 nucleotides) of both sequences in cluster intNode557_47332 (Table 4.1), while “N-acetylmuramoyl-L-alanine amidase”, involved in peptidoglycan biosynthesis and cellular lysis, was found adjacent to sequences in cluster intNode557_37369 (Table 4.1). STRING (Jensen et al. 2009) is an example of a database and web tool that uses gene neighbourhoods that could potentially be informative for annotation of hypothetical sequences. PSORTb did not annotate any membrane-related proteins, suggesting no transporters are among the hypothetical proteins. Sequences annotated as being extracellular suggests they are excreted out of the cell and although no protein function for these sequences was annotated by Pfam (table 4.1), protein secretion is involved in quorum sensing (Rutherford et al. 2012) and virulence (Lee et al. 2001). In *C. perfringes* quorum sensing is even known to regulate virulence genes (Ohtani et al. 2009).

Future work with phylogenetic profiles would expand upon the use of presence information of genomes to include distance measures that account for the number of representative sequences from each genome in a cluster. Accounting for absence information could be more informative, but caution would have to be exercised due to missing sequences from incomplete genomes. However, such large-scale analyses do appear to hold promise, especially given runtimes can be managed for such large genomic data sets, and hypothetical proteins that are potentially relevant to sequences of interest can be identified.

4.6 CONCLUSION

Clustering placed additional homologs (peptidases) in toxin clusters of *P. difficile* and *C. botulinum*, and *C. tetani* that were not typically included in other analyses and phylogenies revealed extremely short branch lengths with toxins, demonstrating shared ancestry and/or probable misannotations. Hierarchical clustering of phylogenetic profiles based on genome and cluster membership was applied to identify hypothetical sequence clusters that are close to botulinum and tetanus toxin, and *P. difficile* toxin A and B clusters in the hierarchy. Although limited information regarding the function of these proteins are obtained through annotation, they are proteins that is potentially linked to toxin function.

CHAPTER 5 DISCUSSION

Metagenomic sequencing has provided a wealth of genomic data from many environments, and the fragmentary nature of metagenomes creates a significant need for computational approaches to support further biological interpretation. How much information can be confidently extracted from a metagenome? The analysis of EBPR metagenomes in Chapter 2 emphasized the importance of sequence and taxonomic representation in clusters for the phylogenetic analysis of LGT. The NCBI protein clusters database (Klimke et al. 2009) was a candidate source of protein clusters, but the strict definition of orthologous clusters used to build this database often yielded sets with limited taxonomic breadth. The development of ProPhylClust and PhyloSubClust in Chapter 3 was motivated by need to build broader clusters that still adhered to the evolutionary definitions of orthology. The diversity of approaches that have been developed to infer clusters of related proteins reflects the diversity of areas of application. ProPhylClust was intended to create clusters of homologous sequences, and to cluster as many sequences as possible. PhyloSubClust was intended to extract subclusters that contain orthologs from clusters of homologous sequences. Both approaches compared favourably with other clustering methods, and when applied to a large set of *Clostridia* genomes in Chapter 4, they produced clusters of toxins and toxin related sequences from *C. botulinum* and *C. tetani*, and *P. difficile* with homologs that others have typically excluded from phylogenetic analysis. The motivation of applying a hierarchical clustering analysis to phylogenetic profiles was the fact that 30-40% of genes in a genome are assigned an unknown function, and finding potential functionally related genes to *Clostridia* toxins is a way to expand our understanding of them.

5.1 METAGENOMIC SEQUENCING

In silico reconstruction of a microbial community is arguably one of the main goals of metagenomic sequencing efforts. However, metagenomes are often incomplete due to sequencing error, assembly errors, and biases in sequencing technology (e.g. Luo et al. 2012, Sangwan et al. 2016, Tessler et al. 2017, Vollmers et al. 2017) and downstream bioinformatic analyses must account for this (e.g. Hoff 2009). To account for errors introduced from community shotgun sequencing, we restricted our analysis to assembled sequences > 1000 nucleotides in length, which is longer than the average bacterial gene (Koonin and Wolf 2008), and classified our sequences to the class level only (with the exception of *Candidatus* Accumulibacter phosphatis). This restriction of taxonomic precision does affect our ability to identify LGT at lower taxonomic levels, which can be important for those who are interested in lower-level interactions of CAP communities. The emergence of techniques to reconstruct draft genomes from metagenome sequences (e.g. Kang et al. 2015, Nielsen et al. 2014, Parks et al. 2017) offers new opportunities to test our approaches on longer sequences that will contain more phylogenetic signal and improved taxonomic representation.

5.2 Clustering of Homologous and Orthologous Sequences

ProPhylClust and PhyloSubClust are new algorithms developed to improve the scaling of cluster construction from increasingly large genome sets. A crucial advantage of ProPhylClust is its runtimes that scale favorably compared to all-versus-all methods, as demonstrated on the *Proteobacteria* and PanPhyla data sets tested in Chapter 3 (see Figure 3.3). Although BLAST served as the basis for ProPhylClust testing, faster methods such as RAPSearch2 (Zhao et al. 2012) and DIAMOND (Buchfink et al. 2015) can be tested to see if the expected speedups do not come at the expense of clustering accuracy. ProPhylClust's use of HMM profiles enables the clustering of remote homologs with higher sensitivity than homology-search approaches such as BLASTP, and may also be the reason for the high proportion of sequences clustered (Figure 3.4), and clusters that were less long tailed, exemplified by a larger q3 and smaller maximum cluster size (Table 3.1). Increased cluster size and more sequences clustered would suggest that clusters contain more potential homologs, which is important for analyses that intend to examine the evolutionary relationships of homologous sequences. In the absence of a ground-truth set of orthologs, ProPhylClust and PhyloSubClust are arguably preferable given their direct use of phylogeny for the inference of orthologous clusters of sequences.

The refinement of large clusters into putatively orthologous sets is an important goal of methods such as RBH, inParanoid, Hieranoid, or Branchclust. Our novel PhyloSubClust algorithm makes explicit use of phylogenetic trees and taxonomic distributions of genes and proteins to define orthologous relationships. Like ProPhylClust, PhyloSubClust is written in Python is intended for ease of installation. This is in contrast to OrthoMCL which has not been recently updated and requires a properly configured installation of MySQL or Oracle, and Hieranoid, which required custom software packages that were not readily available. Compared to OrthoMCL, PhyloSubClust had considerably shorter runtimes (Figure 3.3), which is mainly due to OrthoMCL's dependence on an SQL database. The MCL algorithm has short runtimes, however it loads the complete graph of homologous relationships into RAM, and with larger homologous sets, the size of the graph can easily use all available RAM.

A common motivation for constructing homologous and orthologous protein clusters is to label these proteins with functional categories defined by resources such as COG, eggNOG (Jensen et al. 2007), and KEGG orthology (Kanehisa et al. 2016). Virulence factors, explored in Chapter 4, are genes that encode proteins which allow organisms to invade hosts, evade destruction by the host immune system, suppress the host immune system, obtain nutrients from a host and colonize a host (e.g. Cross 2008). The origin of toxins may lie in non-toxinogenic sequences, which highlights the limitations of assigning uniform labels to protein clusters without deeper investigation. The tetanus/botulinum toxin cluster recovered by ProPhylClust and PhyloSubClust contains proteins labelled as virulence-related toxins as well as non-toxic sequences such as *ntnh* and peptidase M27 that are unlikely to be toxins, and are possibly missanotated sequences. In fact, a flagellin protein found upstream of *BoNT* strain A sequences is considered to be an ancestral, non-toxigenic variant (Doxey et al. 2008) of *BoNT*. This example highlights the disconnect between orthology, which is defined in evolutionary terms, and function (e.g. Dalquen et al. 2013, Trachana 2011), which can change substantially with even a small number of amino-acid changes (Ng and Henikoff 2006, Studer et al. 2013).

Relative to alternative methods, such as all-versus-all homology searches, ProPhylClust's use of a phylogeny as a guide provides two advantages to the clustering of homologous sequences. First, this progressive approach reduces the number of times homology searches are performed for n genomes from n^2 to n - 1. This results in much more favourable runtimes with very large genome sets. The second advantage is that the phylogeny provides a biologically meaningful ordering to the clustering process. A phylogeny is a hypothesis of the lines of descent from a common ancestor to a set of extant organisms. The hierarchical clustering that ProPhylClust performs is a method that defines the sequence clusters of the most closely related homologs in sister lineages first, and then continues to add homologs or create new sequence clusters from more distant relatives the tree is traversed to the root of the tree. If HMM profile searches are implemented and HMM profiles are calculated, the model generated should be a more refined representation of the homologous sequence cluster as sequences are added to the alignment while the tree is traversed. Graph-based methods, in contrast, do not use any hierarchy to guide the ordering of clustering and produce “flat” homologous clusters.

The evolution of bacterial sequences is often non-tree-like and the sequence content of genomes can vary for multiple reasons, such as deletion or LGT (Bapteste et al. 2009). Not all sequences will have a phylogeny with a similar topology to the phylogeny used as a guide tree. Hieranoid, which also uses a guide tree for clustering, does not attempt to cluster singleton sequences at internal nodes, and relies on initially constructed clusters between sister leaves in the tree, which makes it unsuitable where genome-sequence evolution is not tree-like or does not share the topology of the guide tree. Singleton versus singleton homology searches, sequence to HMM profile/consensus sequence, and HMM profile/consensus sequence to HMM profile/consensus sequence searches at internal nodes in ProPhylClust were developed to ensure the topological limitations of a guide tree to represent the evolutionary relationships of all genome sequence content is not restrictive during clustering. They also allow for homologous sequences to be clustered with draft genomes and if the phylogeny is poorly resolved.

The definition of orthology is phylogenetic, and PhyloSubClust relies on constructed phylogenies to define orthologous relationships in homologous sequence clusters. Unlike methods that rely on partitioning a graph, agreement between the topology of the cluster's phylogeny and the topology of the species phylogeny can be used to define orthology. However, complete agreement between the topology of the species phylogeny and the cluster phylogeny is complicated by poorly resolved species and cluster phylogenies, and variable sequence membership of the clusters due to the effect of LGT and gene deletion and insertion. PhyloSubClust's criteria for subclustering is not part of the orthology definition, since it initiates clustering at the shortest branch length between and then relies on taxonomic membership of subtrees, instead of referencing the topology of a species tree. Although not orthology, PhyloSubClust's use of phylogeny for subclustering is a more faithful representation of orthologous relationships than graph-based methods. Relying on taxonomic representation as a criterion in subclustering has the advantage that clusters are not overly partitioned. In addition, the homologous sequence clusters from ProPhylClust represent a more desirable input set of homologs for subclustering since they are hierarchically clustered using phylogeny as a guide. Clusters defined by PhyloSubClust to obtain orthologs should therefore be preferable compared to graph-based subclustering methods that do not rely on phylogeny, such as RBH and OrthoMCL.

The lack of a ground truth set of orthologous sequences precludes validation of clusters from PhyloSubClust. It should be expected that the clusters include inparalogs since PhyloSubClust does not attempt to remove them, and inparalogs can be considered orthologs, because they are duplicated after the speciation event (Remm et al. 2001). The toxinAB (figure 4.4) cluster and BoNTA-BoNTF-BoNTE-ntnH-tetX (figure 4.5) cluster from chapter four provide insights into the clusters produced by ProPhylClust. It is unknown whether *BoNT*, *TetX*, *tcdB* and *tcdA* are each considered to be orthologous sequence clusters, or are considered to be orthologs when associated with other homologs, as seen in the BoNTA-BoNTF-BoNTE-ntnH-tetX and toxinAB clusters with peptidases. BoNTA-BoNTF-BoNTE-ntnH-tetX and toxinAB were not subclustered. This could be due to missing sequences since the majority of genomes in the *Clostridia* genome set are draft genomes. However, based on the topology of the phylogenies of the two clusters, each could be further subdivided into two subclusters, as long branches separate the subclusters from the other subclusters in each. For the BoNTA-BoNTF-BoNTE-ntnH-tetX cluster, one potential subcluster contains *ntnh* and the other contains *BoNT*/*TetX*. The peptidase M27s equences associated with each are most likely missannotated sequences. For the toxinAB cluster the two potential subclusters each have peptidase C80 sequences that are most likely missanotated as either *tcdA* or *tcdB*. From these two example clusters, it is clear that ProPhylClust does create clusters of homologs, but PhyloSubClust does not always extract subtrees that could be clusters based on visual inspection of subtrees and branch lengths. Future versions of PhyloSubClust should include relative branch lengths of subtrees as a criterion for clustering.

Examining 18 toxin related sequence, clusters labelled as p47 and ORFX3, based on the majority of product identifications, are *BoNT* associated sequences that are the only two subclusters from the same parent cluster. It is known they are homologous sequences, and ORFX3 shares higher sequence similarity with p47 than homologs ORFX1 and ORFX2 (Chen et al. 2007). No phylogenetic reconstruction has been performed on these clusters, so it is unknown what the relative branch lengths were for each subtree. However, the P47 and ORFX3 subclusters reveal that PhyloSubClust does successfully extract meaningful, potentially orthologous subclusters from clusters of homologs based on the relative consistency of product identifications in each subcluster. However, labeling of subclusters as orthologs in this approach relies on the product label applied to an extracted subcluster, and visual inspection of phylogenies constructed from subclusters before and after extraction. Relying on the consistency of annotations in a sequence cluster, such as the product name, as a way to determine if a cluster is potentially an orthologous group is not suitable, especially given databases of annotations are error prone (e.g. Jones et al. 2007, Schnoes et al. 2009. Ideally, orthology detection through comparison of subtree and species tree topology could be used as a potentially less biased means to determine if the cluster from a subtree is composed of orthologs in the absence of LGT.

5.3 Clostridia Toxins, Hierarchical Clustering and Phylogenetic Profiling

The distribution of the 18 toxin-related clusters across the 16S rRNA phylogeny of 558 *Clostridia* genomes analysed in Chapter four is similar to known toxin sequence distributions for *C. botulinum*, *C. tetani* and *P. difficile* genomes. However, given the high proportion of draft genomes and assemblies in the data set (427 draft genomes or assemblies), there are likely to be cases where specific genes of interest from an organism are not present in the draft assemblies. Nonetheless, distribution and in particular phylogenetic information found in the ortholgous clusters is informative. The phylogenies of the BoNTA-BoNTF-BoNTE-ntnH-tetX and tcdAB sequence clusters revealed annotated peptidase sequences associated with toxin sequences, often with extremely short branch lengths (Figures 4.3 and 4.4). Given the errors that can arise from genome annotation methods, including genome context methods, (e.g. Kyrpides and Ouzounis 1999, Promponas et al. 2015), it is possible that some of the peptidase sequences may in fact represent misannotated sequences.

Future work with the phylogenetic profiling approach can encompass expanded datasets, different approaches to encoding gene presence / absence (e.g., counts vs presence/absence), or applying statistical methods beyond simple binary Hamming distance to identify correlations between profiles. The addition of genomes within a clade increases the predictive accuracy of phylogenetic profiling (Škunca and Dessimoz 2015), but it is unknown whether the addition of more genomes within a clade that experiences considerable LGT events increases accuracy. If plasmids with toxins are regularly transferred and lost, no consistent pattern may be present for phylogenetic profiling. However, careful selection of genomes or exhaustive sequencing of genomes could reveal overall patterns for phylogenetic profiling.

Sequencing technology and assembly algorithms continue to improve the completion and quality of sequenced genomes, and studies of thousands of isolates of single bacterial species (e.g. Merker et al. 2015) will become more common. The bioinformatic methods developed in this thesis are a contribution towards the goal of providing biologically relevant meaning to the efficient analyses of bacterial genomes and their communities.

REFERENCES

Alam, SI, and P Dwivedi. 2016. Putative function of hypothetical proteins expressed by *Clostridium perfringens* type A strains and their protective efficacy in mouse model. Infection, Genetics and Evolution 44:147–56.

Albertsen M, Hansen LBS, Saunders AM, Nielsen PH and Nielsen KL. 2011. A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. The ISME Journal 6:1094-106.

Altenhoff, AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, et al. 2016. Standardized benchmarking in the quest for orthologs. Nature Methods 13 (5):425-30.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403-10.

Altschul, SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25 (7):3389-3402.

Apatoff A, Kim E, and Kliger Y. 2006. Towards alignment independent quantitative assessment of homology detection. PLoS ONE 1 (1):e113.

Aravind, L. 2000. Guilt by association: contextual information in genome analysis. Genome Research 10 (8):1074–77.

Arnon SS, Schechter R, Inglesby TV et al. 2001. botulinum toxin as a biological weapon: medical and public health management. JAMA 285 (8):1059–70.

Ashburner M, Catherine AB, Judith AB, Botstein D, Butler H, Cherry JM, Davis AP, et al. 2000. Gene Ontology: Tool for the unification of biology. Nature Genetics 25 (1): 25–29.

BäckhedF, Ley RE, Sonnenburg JL, Peterson DA, and Gordon JI. 2005. Host-bacterial mutualism in the human intestine. Science 307 (March): 1915–20.

Banard JL. 1976. A review of biological phosphorus removal in activated sludge. Water South Africa. 2:136-44.

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten P, Franklin-Hall L, Lapointe FJ, et al. 2009. Prokaryotic evolution and the tree of life are two different things. Biology Direct 4: 34.

Barlow M. What antimicrobial resistance has taught us about horizontal gene transfer. In: Gogarten MBB, Gogarten JP, Lorraine O, editors. Horizontal Gene Transfer, Genomes in Flux. Clifton, NJ: Humana Press; 2009. p. 397-411.

Bathe S, Schwarzenbeck N, and Hausner M. 2005. Plasmid-mediated bioaugmentation of activated sludge bacteria in a sequencing batch moving bed reactor using pNB2. Letters in Applied Microbiology 41:242-7.

Bay DC, Rommens KL, and RJ Turner. 2008. small multidrug resistance proteins: a multidrug transporter family that continues to grow. Biochimica et Biophysica Acta - Biomembranes 1778 (9):1814-38.

Beiko RG, Harlow TJ, and Ragan MA. 2005. Highways of gene sharing in prokaryotes. PNAS 102 (40):14332-7.

Beiko, RG. 2011. telling the whole story in a 10,000-genome world. Biology Direct 6 (34)

BellaSD, Ascenzi P, Siarakas S, Petrosillo N, and di Masi A. 2016. *Clostridium difficile* toxins A and B: insights into pathogenic properties and extraintestinal effects. Toxins 8 (5):1–25.

BernardesJS, Vieira FRJ, Costa LMM, and Zaverucha G. 2015. evaluation and improvements of clustering algorithms for detecting remote homologous protein families. BMC Bioinformatics 16:34.

Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, and Brinkman FSL. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Research 45 (W1):W30–35.

Bhandari M, Campbell KD, Collins MD, and East AK. 1997. molecular characterization of the clusters of genes encoding the botulinum neurotoxin complex in *Clostridium botulinum* (*Clostridium argentinense*) Type G and Nonproteolytic *Clostridium botulinum* Type B. Current Microbiology 35 (4):207–14.

Bolten E, Schliep A, Schneckener S, Schomburg D, and Schrader R. 2001. Clustering protein sequences--structure prediction by transitive homology. Bioinformatics 17 (10):935-41.

Bork P. 2000. powers and pitfalls in sequence analysis: the 70% hurdle. Genome Research 10 (4):398–400.

Braun V, Hundsberger T, Leukel P, Sauerborn M, and Eichel-Streiber CV. 1996. Definition of the single integration site of the pathogenicity locus in *Clostridium difficile*. Gene 181 (1–2):29–38.

Brouwer MSM, Roberts AP, Hussain H, Williams RJ, Allan E, and Mullany P. 2013. Horizontal gene transfer converts non-toxigenic *Clostridium difficile* strains into toxin producers. Nature Communications 4:1–6.

Brüggemann H, Bäumer S, Fricke WF, Wiezer A, Liesegang H, Decker I, Herzberg C, et al. 2003. the genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. Proceedings of the National Academy of Sciences 100 (3):1316–21.

Brüggemann H, Brzuszkiewicz E, Chapeton-Montes D, Plourde L, Speck D, and Popoff MR. 2015. Genomics of *Clostridium tetani*. Research in Microbiology 166 (4):326–31.

Buchfink B, Xie C, and Huson D. 2015. fast and sensitive protein alignment using DIAMOND. Nat Methods 12 (1):59-60.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10 (1):421.

Chaoyue L. 2016. Gene clustering based on co-occurrence with correction for common evolutionary history. MSc. dissertation. Dalhousie University.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, et al. 2010. qiime allows analysis of high-throughput community sequencing data. Nature Methods 7 (5):335-336.

Carter GP, Purdy D, Williams P, and Minton NP. 2005. quorum sensing in *Clostridium difficile*: analysis of a luxS-type signalling system. Journal of Medical Microbiology 54 (2):119–27.

Carter, AT, and Peck MW. 2015. genomes, neurotoxins and biology of *Clostridium botulinum* group I and group II. Research in Microbiology 166 (4):303–17.

Charuvaka A, and Rangwala H. 2011. Evaluation of short read metagenomic assembly. BMC Genomics 12 Suppl 2:S8.

Chen Y, Randall AA, McCue T. 2004. The efficiency of enhanced biological phosphorus removal from real wastewater affected by different ratios of acetic to propionic acid. Water Research 38:27-36.

Chen Y, Korkeala H, Aarnikunnas J, and Lindström M. 2007. Sequencing the botulinum neurotoxin gene and related genes in *Clostridium botulinum* type E strains reveals orfx3 and a novel type E neurotoxin subtype. Journal of Bacteriology 189 (23): 8643–50.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, and Tiedje JM. 2014. ribosomal database project: data and tools for high throughput rrna analysis. Nucleic Acids Research 42 (D1).

Collins, MD, Lawson PA, Willems A, Cordoba JJ, Fernandez-Garayzabal J, Garcia P, Cai J, Hippe H, and Farrow JA. 1994. The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. Int J Syst Bacteriol 44 (4):812–26.

Collins, MD, and East AK. 1998. phylogeny and taxonomy of the food-borne pathogen *Clostridium botulinum* and its neurotoxins. Journal of Applied Microbiology 84 (1):5–17.

Connan C, Denève C, Mazuet C, and Popoff MR. 2013. Regulation of toxin synthesis in *Clostridium botulinum* and *Clostridium tetani*. Toxicon 75:90–100.

Conway T, and Cohen PS. 2015. Commensal and pathogenic escherichia coli metabolism in the gut. Microbiology Spectrum 3 (3): MBP-0006-2014.

Cross AS. 2008. What is a virulence factor? Critical Care 12 (6): 197.

Cukrowska B, LodInová-ZádnIková R, Enders C, Sonnenborn U, Schulze J, and Tlaskalová-Hogenová H. 2002. Specific proliferative and antibody responses of premature infants to intestinal colonization with nonpathogenic probiotic *E. coli* strain Nissle 1917. Scandinavian Journal of Immunology 55 (2): 204–9.

Dalquen DA, Altenhoff AM, Gonnet GH, and Dessimoz C. 2013. The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PloS One 8 (2):e56925.

Dessimoz C, Gabaldon T, Roos D, Sonnhammer E, Herrero J, Altenhoff A, Apweiler R, et al. 2012. Toward community standards in the quest for orthologs. Method Biochem Anal 28 (6):900-904.

Dineen, SS., Bradshaw M, and Johnson EA. 2003. Neurotoxin gene clusters in *Clostridium botulinum* type A strains: sequence comparison and evolutionary implications. Current Microbiology 46 (5):345–52.

Doxey AC, Lynch MDJ, Müller KM, Meiering EM, and McConkey BJ. 2008. insights into the evolutionary origins of clostridial neurotoxins from analysis of the *Clostridium botulinum* Strain A neurotoxin gene cluster. BMC Evolutionary Biology 8:316.

Drudy, D, Fanning S, and Kyne L. 2007. Toxin A-negative, toxin B-positive *Clostridium difficile*. International Journal of Infectious Diseases 11 (1):5–10.

Dunn CW, Zapata F, Munro C, Siebert S, and Hejnol A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. Proceedings of the National Academy of Sciences 115 (3): E409–17.

East, AK, Bhandari M, Stacey JM, Campbell KD, and Collins MD. 1996. Organization and phylogenetic interrelationships of genes encoding components of the botulinum toxin complex in proteolytic *Clostridium botulinum* types A, B, and F: evidence of chimeric sequences in the gene encoding the nontoxic nonhemagglutinin component. International Journal of Systematic Bacteriology 46 (4):1105–12.

Eddy SR 2009. A new generation of homology search tools based on probabilistic inference. Genome Informatics Int Conf Genome Informatics 23 (1): 205–11.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Computational Biology 7 (10).

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-7.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-1.

Eisel U, Jarausch W, Goretzki K, Henschen A, Engels J, Weller U, Hudel M, Habermann E, and Niemann H. 1986. Tetanus toxin: primary structure, expression in e. coli, and homology with botulinum toxins. Embo J 5 (10):2495–2502.

Eklund MW, Poysky FT, Reed SM, and Smith CA. 1971. Bacteriophage and the toxigenicity of *Clostridium botulinum* Type C. Science 172 (3982):480-82.

Eklund MW, Poysky FT, and Reed SM. 1972. Bacteriophage and the toxigenicity of *Clostridium botulinum* Type D. Nature 235 (53):16-17.

Elliott, B, Androga GO, Knight DR, and Riley TV. 2017. *Clostridium difficile* infection: evolution, phylogeny and molecular epidemiology. Infection, Genetics and Evolution 49:1–11.

Emms DM., and Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16 (1):157.

Enright A, Dongen S, and Ouzounis C. 2002. an efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30 (7):1575-84.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44 (D1):D279–85.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology 19 (2): 99-113.

Fitch, WM. 2000. Homology a personal view on some of the problems. Trends in Genetics : TIG 16 (5):227-31.

Flowers JJ, He S, Malfatti S, del Rio TG, Tringe SG, Hugenholtz P, et al. 2013. Comparative genomics of two "Candidatus *Accumulibacter*" clades performing biological phosphorus removal. The ISME Journal 7:2301-14.

Freedman, JC, Theoret JR, Wisniewski JA, Uzal FA, Rood JI, and McClane BA. 2015. *Clostridium perfringens* type A-E toxin plasmids. Research in Microbiology 166 (4): 264–79.

Galperin MY, and Koonin EV. 2004. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Research 32 (18):5452–63.

García Martín H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, et al. 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nature Biotechnology 24:1263-9.

Goldman N, Anderson JP, and Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. Systematic Biology 49:652-70.

Goldman N, and Whelan S. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. 2000. Molecular Biology and Evolution 17:975-8.

Gonzalez-Gil G, and Holliger C. 2011. Dynamics of microbial community structure and enhanced biological phosphorus removal of propionate and acetate cultivated aerobic granules. Applied and Environmental Microbiology 77:8041-51.

Govind R, and Dupuy B. 2012. Secretion of *Clostridium difficile* toxins A and B requires the holin-like protein tcdE. PLoS Pathogens 8 (6):1–14.

Gupta RS, and Gao B. 2009. Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus *Clostridium* s*ensu stricto* (cluster I). International Journal of Systematic and Evolutionary Microbiology 59 (2):285–94.

Hagberg AA, Schult DA, and Swart PJ. 2008. Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference, edited by Gaël Varoquaux, Travis Vaught, and Jarrod Millman, 11-15. Pasadena, CA USA.

Harlow TJ, Gogarten PJ, and Ragan MA. 2004. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. BMC Bioinformatics 14:1-14.

He S, Gu AZ, and McMahon KD. 2006. Fine-scale differences between *Accumulibacter*-like bacteria in enhanced biological phosphorus removal activated sludge. Water Science and Technology 54:111-7.

He S, Gu AZ, and McMahon KD. 2008. Progress toward understanding the distribution of Accumulibacter among full-scale enhanced biological phosphorus removal systems. Microbial Ecology 55:229-36.

He S, and McMahon KD. 2011. Microbiology of "*Candidatus* Accumulibacter" in activated sludge. Microbial Biotechnology 4:603-19.

Heijden M, Bardgett R, and Straalen N. 2007. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol Lett 11 (3): 296–310

Hill KK, Xie G, Foley BT, Smith TJ, Munk AC, Bruce D, Smith LA, Brettin TS, and Detter JC. 2009. Recombination and insertion events involving the botulinum neurotoxin complex genes in *Clostridium botulinum* types A, B, E and F and *Clostridium butyricum* type E strains. BMC Biology 7 (1):66.

Hoff KJ. 2009. "The effect of sequencing errors on metagenomic gene prediction." BMC Genomics 10: 1–9.

Hong H, Ko H-J, Choi I-G, Park W. 2014. Previously undescribed plasmids recovered from activated sludge confer tetracycline resistance and phenotypic changes to *Acinetobacter oleivorans* DR1. Microbial Ecology 67:369-79.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, et al. 2016. A new view of the tree of life. Nature Microbiology 1 (5):16048.

Jensen L, Julien P, Kuhn M, Mering C, Muller J, Doerks T, and Bork P. 2007. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36 (Database issue): D250-4.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, et al. 2009. STRING 8 - A global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research 37 (SUPPL. 1): 412–16.

John B and Sali A. 2004. Detection of homologous proteins by an intermediate sequence search. Protein Science 13 (1):54-62.

Johnson SL, Eddy SR, and Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431.

Jones CE, Brown AL, and Baumann U. 2007. Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics 8: 1–9.

Kanehisa M and Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28:27-30.

Kang DD, Froula J, Egan R, and Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3: e1165.

Kensche PR, van Noort V, Dutilh BE, and Huynen MA. 2008. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. Journal of The Royal Society Interface 5 (19):151–70.

Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, et al. 2009. The National Center for Biotechnology Information's protein clusters database. Nucleic Acids Research 37(Database issue):D216-23.

Kong Y, Ong SL, Ng WJ, Liu W-T. 2002. Diversity and distribution of a deeply branched novel proteobacterial group found in anaerobic-aerobic activated sludge processes. Environmental Microbiology 4:753-7.

Koonin EV, Galperin MY. 1997. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Current Opinion in Genetics & Development 7:757-63.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annual Review of Genetics 39 (1): 309–38.

Koonin EV, and Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Research 36 (21): 6688–6719.

Korbel JO, Jensen LJ, Von Mering C, and Bork P. 2004. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nature Biotechnology 22 (7): 911–17.

Kristensen D, Wolf Y, Mushegian A, and Koonin E. 2011. Computational methods for gene orthology inference. Brief Bioinform 12 (5):379-91.

Krogh A, Brown M, Mian IS, Sjölander K, and Haussler D. 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. Journal of Molecular Biology 235(5): 1501-1531.

Kugelberg E, Kofoid E, Reams AB, Andersson DI, and Roth JR. 2006. Multiple Pathways of Selected Gene Amplification during Adaptive Mutation. PNAS 103 (46): 17319–24.

Kurka H, Ehrenreich A, Ludwig W, Monot M, Rupnik M, Barbut F, Indra A, Dupuy B, and Liebl W. 2014. Sequence similarity of clostridium difficile strains by analysis of conserved genes and genome content is reflected by their ribotype affiliation. PLoS ONE 9 (1).

Kuzniar A, van Ham RCHJ, Pongor SN, Pongor S, and Leunissen JAM. 2008. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24 (11):539-51.

Kyrpides NC, and Ouzounis CA. 1999. Whole-genome sequence annotation: 'going wrong with confidence' [Letter]. Molecular Microbiology 32 (4): 886–87.

Lacey JA, Keyburn AL, Ford ME, Portela RW, Johanesen PA, Lyras D, and Moore RJ. 2017. Conjugation-mediated horizontal gene transfer of *Clostridium perfringens* plasmids in the chicken gastrointestinal tract results in the formation of new virulent strains. Applied and Environmental Microbiology 83(24):e01814-17.

Lang AS, Zhaxybayeva O, and Beatty JT. 2012. Gene transfer agents: phage-like elements of genetic exchange. Nature Reviews Microbiology.10:472-82.

Lanham AB, Oehmen A, Saunders AM, Carvalho G, Nielsen PH, and Reis MAM. 2013. Metabolic versatility in full-scale wastewater treatment plants performing enhanced biological phosphorus removal. Water Research 47:7032-41.

Lawrence JG and Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. Journal of Molecular Evolution 44:383-97.

Lawrence JG and Ochman H. 2002. Reconciling the many faces of lateral gene transfer. Trends in Microbiology 10:1-4.

Lawson PA., Citron DM, Tyrrell KL, and Finegold SM. 2016. Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prévot 1938. Anaerobe 40:95–99.

Le SQ, and Gascuel O. 2008. An improved general amino acid replacement matrix. Molecular Biology and Evolution 25 (7):1307-20.

Lebeda, RJ, Cer RZ, Mudunuri U, Stephens R, Singh BR, and Adler M. 2010. The Zinc-Dependent Protease Activity of the Botulinum Neurotoxins. Toxins 2 (5):978–97.

Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, and Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. 12 (1):124.

Lee, VT, and Schneewind O. 2001. Review: Protein secretion and the pathogenesis of bacterial infections. Genes and Development 15 (617): 1725–52.

Leplae R, Lima-Mendez G, and Toussaint A. ACLAME: A CLAssification of Mobile genetic Elements, update. 2010. Nucleic Acids Research 38(Database issue):D57-61.

Li, L, Stoeckert C, and Roos D. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 13 (9):2178-89.

Lin FPY, Lan R, Sintchenko V, Gilbert GL, Kong F, and Coiera E. 2011. Computational bacterial genome-wide analysis of phylogenetic profiles reveals potential virulence genes of *Streptococcus agalactiae*. PLoS ONE 6 (4).

Loganantharaj R, and Atwi M. 2007. Towards validating the hypothesis of phylogenetic profiling. BMC Bioinformatics 8 (Suppl 7): S25.

Luo C, Tsementzi D, Kyrpides N, Read T, and Konstantinidis K. 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample." PLoS ONE 7(2): e30087.

Ma L, Zhang X-X, Zhao F, Wu B, Cheng S, and Yang L. 2013. Sewage treatment plant serves as a hot-spot reservoir of integrons and gene cassettes. Journal of Environmental Biology 34(2 Special Number):391-9.

Macdonald NJ, Parks DH, and Beiko RG. 2012. Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Research 40:e111.

Mansfield MJ., Adams JB, and Doxey AC. 2015. botulinum neurotoxin homologs in non-clostridium species. FEBS Letters 589 (3). Federation of European Biochemical Societies:342–48.

Mao Y, Yu K, Xia Y, Chao Y, and Zhang T. 2014. Genome reconstruction and gene expression of "*Candidatus* Accumulibacter phosphatis" Clade IB performing biological phosphorus removal. Environmental Science & Technology 48:10363-71.

MarshallKM, Bradshaw M, PellettS, and Johnson EA. 2007. Plasmid encoded neurotoxin genes in *Clostridium botulinum* serotype A subtypes. Biochemical and Biophysical Research Communications 361 (1):49–54.

Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MGB, et al. 2015. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nature Genetics 47 (3): 242–49.

Nakamura Y, Itoh T, Matsuda H, and Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 36 (7): 760–66.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, et al. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature Biotechnology 32 (8): 822–28.

Matias Rodrigues JF, and Von Mering C. 2014. HPC-CLUST: Distributed hierarchical clustering for large sets of nucleotide sequences. Bioinformatics 30 (2):287-88.

Martin-Verstraete I, Peltier J, and Dupuy B. 2016. The regulatory networks that control *Clostridium difficile* toxin synthesis. Toxins 8 (5):1–24.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, et al. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4:495-500.

McIlroy SJ, Albertsen M, Andresen EK, Saunders AM, Kristiansen R, Stokholm-Bjerregaard M, et al. 2013. "*Candidatus* Competibacter-"lineage genomes retrieved from metagenomes reveal functional metabolic diversity. The ISME Journal 8:613-24.

Meehan CJ and Beiko RG. 2012. Lateral gene transfer of an ABC transporter complex between major constituents of the human gut microbiome. BMC Microbiol 12:248.

Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, et al. 2012. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS ONE 7:e31386.

Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MGB, et al. 2015. Evolutionary history and global spread of the *Mycobacterium tuberculosis* beijing lineage. Nature Genetics 47 (3): 242–49.

Meyerguz, L, Kleinberg J, and Elber R. 2007. The network of sequence flow between protein structures. Proceedings of the National Academy of Sciences of the United States of America 104 (28):11627-32.

Mielczarek AT, Nguyen HTT, Nielsen JL, and Nielsen PH. 2013. Population dynamics of bacteria involved in enhanced biological phosphorus removal in Danish wastewater treatment plants. Water Research 47:1529-44.

Mishra PK, Sonkar SC, Raj SR, Chaudhry U, and Saluja D. 2014. Functional Analysis of Hypothetical Proteins of *Chlamydia Trachomatis*: A Bioinformatics Based Approach for Prioritizing the Targets. Journal of Computer Science & Systems Biology 7 (1):010–014.

Monot M, Eckert C, Lemire A, Hamiot A, Dubois T, Tessier C, Dumoulard B, et al. 2015. *Clostridium difficile*: new insights into the evolution of the pathogenicity locus. Scientific Reports 5:15023.

Muñoz M, Ríos-Chaparro DI, Patarroyo MA, and Ramírez JD. 2017. Determining *Clostridium difficile* intra-taxa diversity by mining multilocus sequence typing databases. BMC Microbiology 17 (1). BMC Microbiology:62.

Ng PC, and Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annual Review of Genomics and Human Genetics 7 (1): 61–80.

Ni J, Yan Q, and Yu Y. 2013. How much metagenomic sequencing is enough to achieve a given goal? Scientific Reports 3:1968.

Nielsen PH, Mielczarek AT, Kragelund C, Nielsen JL, Saunders AM, Kong Y, et al. 2010. A conceptual ecosystem model of microbial communities in enhanced biological phosphorus removal plants. Water Research 44:5070-88.

Nielsen PH, Saunders AM, Hansen AA, Larsen P, and Nielsen JL. 2012. Microbial communities involved in enhanced biological phosphorus removal from wastewater-a model system in environmental biotechnology. Current Opinion in Biotechnology 23:452-9.

Oehmen A, Lemos PC, Carvalho G, Yuan Z, Keller J, Blackall LL, et al. 2007. Advances in enhanced biological phosphorus removal: from micro to macro scale. Water Research 41:2271-300.

Oehmen A, Yuan Z, Blackall LL, Keller J. 2005. Comparison of acetate and propionate uptake by polyphosphate accumulating organisms and glycogen accumulating organisms. Biotechnology and Bioengineering 91:162-8.

Oezguen N, Power TD, Urvil P, Feng H, Pothoulakis C, Stamler JS, Braun W, and Savidge TC. 2012. Clostridial toxins: sensing a target in a hostile gut environment. Gut Microbes 3 (1):35–41.

Ohtani K, Yuan Y, Hassan S, Wang R, Wang Y, and Shimizu T. 2009. Virulence gene regulation by the Agr system in *Clostridium Perfringens*. Journal of Bacteriology 191 (12): 3919–27.

Oliveros J.C. VENNY. 2007. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html.

Osterman A, and Overbeek R. 2003. Missing genes in metabolic pathways: a comparative genomics approach. Current Opinion in Chemical Biology 7 (2):238–251.

Park J, Teichmann SA, Hubbard T, and Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences. Journal of Molecular Biology 273 (1):349-54.

Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, and Chothia C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284 (4): 1201–10.

Parks DH, Macdonald NJ, and Beiko RG. 2011. Classifying short genomic fragments from novel lineages using composition and homology. BMC Bioinformatics 12:328.

Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, and Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology 2: 1533–1542.

Pearson, WR and Lipman DJ. 1988. Improved Tools for biological sequence comparison. PNAS 85 (8):2444–48.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci 96 (8):4285–88.

Pereira FL, Oliveira Júnior CA, Silva ROS, Dorella FA, Carvalho AF, Almeida GMF, Leal CAG, Lobato CFC, and Figueiredo HCP. 2016. Complete genome sequence of *Peptoclostridium difficile* strain Z31. Gut Pathogens 8 (1):11.

Pignatelli M and Moya A. 2011. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PLoS ONE 6:e19984.

Popa O, Hazkani-Covo E, Landan G, Martin W, and Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Research 21:599-609.

Popoff MR and Bouvet P. 2013. Genetic characteristics of toxigenic Clostridia and toxin gene evolution. Toxicon 75:63–89.

Poptsova MS and Gogarten PJ. 2007. BranchClust: a phylogenetic algorithm for selecting gene families. BMC Bioinformatics 8 (1):1-16.

Promponas VJ, Iliopoulos I, and Ouzounis CA. 2015. Annotation inconsistencies beyond sequence similarity-based function prediction - phylogeny and genome structure. Standards in Genomic Sciences 10 (1): 1–5.

Psomopoulos FE., and Mitkas PA. 2012. Multi-level clustering of phylogenetic profiles. International Journal on Artificial Intelligence Tools 21 (5):1240023.

Price M, Dehal P, and Arkin A. 2009. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26 (7):1641-50.

Raffestin S, Marvaud CM, Cerrato R, Dupuy B, and Popoff MR. 2004. Organization and regulation of the neurotoxin genes in *Clostridium botulinum* and *Clostridium tetani*. Anaerobe 10 (2):93–100.

Ragan MA. 2001a. Detection of lateral gene transfer among microbialgenomes. Current Opinion in Genetics & Development 11:620-6.

Ragan MA. 2001b. On surrogate methods for detecting lateral gene transfer. FEMS Microbiology Letters 201:187-91.

Ragan MA, Harlow TJ, and Beiko RG. 2006. Do different surrogate methods detect lateral genetic transfer events of different relative ages? Trends in Microbiology 14:4-8.

Rutherford ST, and Bassler BL. 2012. Bacterial Quorum Sensing: Its role in virulence and possibilities for its control. Cold Spring Harbor Perspectives in Medicine 2 (11): 1–25.

Saier MH, Paulsen IT, Sliwinski MK, Pao SS, Skurray RA, and Nikaido H. 1998. Evolutionary origins of multidrug and drug-specific efflux pumps in bacteria. The FASEB Journal 12 (3):265-74.

Sangar V, Blankenberg DJ, Altman N, and Lesk AM. 2007. Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics 8 (1): 294-309.

Sangwan N, Xia F, and Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. Microbiome 4: 1–11.

Sasaki Y, Takikawa N, Kojima A, Norimatsu M, Suzuki S, and Tamura Y. 2001. Phylogenetic positions of *Clostridium novyi* and *Clostridium haemolyticum* based on 16S rDNA sequences. International Journal of Systematic and Evolutionary Microbiology 51 (3):901–4.

Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, et al. 2006. The multidrug-resistant human pathogen *Clostridium difficile* Has a Highly Mobile, Mosaic Genome. Nature Genetics 38 (7):779–86.

Schreiber F, and Sonnhammer ELL. 2013. Hieranoid: hierarchical orthology inference. Journal of Molecular Biology 425 (11):2072-81.

Sentchilo V, Mayer AP, Guy L, Miyazaki R, Green Tringe S, and Barry K, et al. 2013. Community-wide plasmid gene mobilization and selection. The ISME Journal 6:1173-86.

Seung-Seok C, Sung-Hyuk C, and Tappert CC. 2010. A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics & Informatics 8 (1): 43–48.

Seviour RJ, Mino T, and Onuki M. 2003. The microbiology of biological phosphorus removal in activated sludge systems. FEMS Microbiology Reviews 27:99-127.

Schlüter A, Krahn I, Kollin F, Bönemann G, Stiens M, Szczepanowski R, et al. 2007. IncP-1-beta plasmid pGNB1 isolated from a bacterial community from a wastewater treatment plant mediates decolorization of triphenylmethane dyes. Applied and Environmental Microbiology 73:6345-50.

Shahbaaz M, Hassan MdI, and Ahmad F. 2013. Functional annotation of conserved hypothetical proteins from *Haemophilus influenzae* Rd KW20. PLoS ONE 8 (12).

Shen A. 2012. *Clostridium difficile* toxins: mediators of inflammation. Journal of Innate Immunity 4 (2):149–58.

Shimodaira H and Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution 16:1114-6.

Schnoes A, Brown S, Dodevski I, and Babbitt P. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PloS Computational Biology 5 (12): e1000605.

Simonsen M, Maetschke SR, and Ragan MA. 2012. Automatic selection of reference taxa for protein-protein interaction prediction with phylogenetic profiling. Bioinformatics 28 (6):851–57.

Sims D, Sudbery I, Ilott NE, Heger A, and Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics 15:121-32.

Singh BR, Chang T-W, Kukreja R, and Shuowei C. 2014. The botulinum neurotoxin complex and the role of ancillary proteins. In: Edited by Keith A. Foster. Molecular Aspects of Botulinum Neurotoxin. New York, NY: Springer New York. p. 69-102.

Skarin H, Håfström T, Westerberg J, and Segerman B. 2011. *Clostridium botulinum* group III: a group with dual identity shaped by plasmids, phages and mobile elements. BMC Genomics 12 (1):185.

Skarin H and Segerman B. 2011. Horizontal gene transfer of toxin genes in Clostridium b*otulinum*: involvement of mobile elements and plasmids. Mobile Genetic Elements 1 (3):213–15.

Skarin H and Segerman B. 2014. Plasmidome interchange between *Clostridium botulinum, Clostridium novyi* and *Clostridium haemolyticum* converts strains of independent lineages into distinctly different pathogens. PLoS ONE 9 (9).

Skippington E and Ragan MA. 2011. Lateral genetic transfer and the construction of genetic exchange communities. FEMS Microbiology Reviews 35:707-35.

Škunca N, and Dessimoz C. 2015. Phylogenetic profiling: how much input data is enough? PLoS ONE 10 (2):1-13.

Slater FR, Johnson CR, Blackall LL, Beiko RG, and Bond PL. 2010. Monitoring associations between clade-level variation, overall community structure and ecosystem function in enhanced biological phosphorus removal (EBPR) systems using terminal-restriction fragment length polymorphism (T-RFLP). Water Research 44:4908-23.

Smith TJ, Hill KK, Foley BT, Detter JC, Munk AC, Bruce DC, Doggett NA, et al. 2007. Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3, /Ba4 and /B1 clusters are located within plasmids. PLoS ONE 2 (12).

Sobecky PA and Coombs JM. Horizontal gene transfer in metal and radionuclide contaminated soils. In: Gogarten MBB, Gogarten JP, Lorraine O, editors. Horizontal Gene Transfer, Genomes in Flux. Clifton, NJ: Humana Press; 2009. p. 455-72.

Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21 (7):951-60.

Sonnhammer ELL, Gabaldón T, da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas BD, Dessimoz C, and Quest for orthologs consortium. 2014. Big data and other challenges in the quest for orthologs. Method Biochem Anal 30 (21):2993-98.

Sukumaran J and Holder MT. 2010. DendroPy: A Python library for phylogenetic computing. Bioinformatics 26:1569-71.

Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30 (9):1312-13.

Studer RA and Robinson-Rechavi M. 2009. How confident can we be that orthologs are similar, but paralogs differ? Trends in Genetics 25 (5):210-16.

Studer RA, Dessailly BH, and Orengo CA. 2013. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochemical Journal 449 (3): 581–94.

Szczepanowski R, Bekel T, Goesmann A, Krause L, Krömeke H, Kaiser O, et al. 2008. Insight into the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to antimicrobial drugs analysed by the 454-pyrosequencing technology. Journal of Biotechnology 136:54-64.

Tatusov RL, Koonin EV, and Lipman DJ. 1997. A genomic perspective on protein families. Science 278 (5338): 631–37.

Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, Segovia BT, et al. 2017. Large-scale differences in microbial biodiversity discovery between 16s amplicon and shotgun sequencing. Scientific Reports 7 (1): 1–14.

Thomas M, Wright P, Blackall L, Urbain V, Keller J. 2003. Optimisation of Noosa BNR plant to improve performance and reduce operating costs. Water Science and Technology 47:141-8.

Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology 3:711-21.

Top EM, Springael D, and Boon N. 2002. Catabolic mobile genetic elements and their potential use in bioaugmentation of polluted soils and waters. FEMS Microbiology Ecology 42:199-208.

Torrieri R, Oliveira FS, Oliveira G, and Coimbra RS. 2012. Automatic assignment of prokaryotic genes to functional categories using literature profiling. PLoS ONE 2012 7 (10):e47436.

Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, and Bork P. 2011. Orthology prediction methods: a quality assessment using curated protein families. BioEssays 33 (10): 769–80.

Tu Y and Schuler AJ. 2013. Low acetate concentrations favor polyphosphate accumulating organisms over glycogen-accumulating organisms in enhanced biological phosphorus removal from wastewater. Environmental Science & Technology 47:3816-24.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE T Inform Theory 13 (2): 260–69.

Vollmers J, Wiegand S, and Kaster AK. 2017. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! PLoS ONE 12 (1):1-31.

Ward N and Moreno-Hagelsieb G. 2014. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: How Much Do We Miss? PLoS ONE 9 (7):1-6.

Weigand MR, Pena-Gonzalez A, Shirey TB, Broeker RG, Ishaq MK, Konstantinidis KT, and Raphael BH. 2015. Implications of genome-based discrimination between *Clostridium botulinum* Group I and *Clostridium sporogenes* strains for bacterial taxonomy. Applied and Environmental Microbiology 81 (16):5420–29.

Williamson CHD, Sahl JW, Smith TJ, XieG, Foley BT, Smith LA, Fernández RA, et al. 2016. Comparative genomic analyses reveal broad diversity in botulinum-toxinproducing Clostridia. BMC Genomics 17 (1).

Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, Böcker S, Stoye J, and Baumbach J. 2010. Partitioning Biological Data with Transitivity Clustering. Nature Methods 7 (6): 419–20.

Wiwie C, Baumbach J, and Röttger R. 2015. Comparing the performance of biomedical clustering methods. Nature Methods 12 (11):1033-40.

Wolfe KH and Shields DC. 1997. Molecular Evidence for an ancient duplication of the entire yeast genome. Nature 387 (6634): 708–13.

Wolf YI and Koonin EV. 2012. A Tight Link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biology and Evolution 4 (12):1286-94.

Wong M-T, Mino T, Seviour RJ, Onuki M, and Liu W-T. 2005. In situ identification and characterization of the microbial community structure of full-scale enhanced biological phosphorous removal plants in Japan. Water Research 39:2901-14.

Woudstra C, Maréchal CL, Souillard R, Bayon-Auboyer MH, Anniballi F, Auricchio B, De Medici D, et al. 2015. Molecular gene profiling of *Clostridium botulinum* group III and its detection in naturally contaminated samples originating from various european countries. Applied and Environmental Microbiology 81 (7):2495–2505.

Wu, J, Kasif S, and DeLisi C. 2003. Identification of functional links between genes using phylogenetic profiles. Bioinformatics 19 (12):1524–30.

Xu J, Bjursell MK, Himrod J,Deng S, Carmichael LK, Chiang HC, Hooper LV, and Gordon JI. 2003. A genomic view of the human-bacteroides thetaiotaomicron symbiosis. Science 299 (5615): 2074-76.

Yu, NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, et al. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26 (13):1608-15.

Yuan Z, Pratt S, and Batstone DJ. 2012. Phosphorus recovery from wastewater through microbial processes. Current Opinion in Biotechnology 23:878-83.

Yutin N, and Galperin MY. 2013. A genomic update on clostridial phylogeny: gram-negative spore formers and other misplaced Clostridia. Environmental Microbiology 15 (10):2631-41.

Zhang J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18 (6): 292-98.

Zhang T, Liu Y, and Fang HHP. 2005. Effect of pH change on the performance and microbial community of enhanced biological phosphate removal process. Biotechnology and Bioengineering 92:173-82.

Zhang T, Zhang XX, and Ye L. 2011a. Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. Plos One 6 (10): e26041.

Zhang T, Shao M-F, and Ye L. 2011b. 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. Isme J 6 (6): 1137-47.

Zhang J, Wang X, Huo D, Li W, Hu Q, Xu C, Liu S, and Li C. 2016. Metagenomic approach reveals microbial diversity and predictive microbial metabolic pathways in yucha, a traditional li fermented food. Scientific Reports 6 (April): 1-9.

Zhao Y, Tang H, and Ye Y. 2012. RAPSearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics 28 (1):125-26.

Zhaxybayeva O. Detection and quantitative assessment of horizontal gene transfer. In: Gogarten MBB, Gogarten JP, Lorraine O, editors. Horizontal Gene Transfer, Genomes in Flux. 2009. Clifton, NJ: Humana Press. p. 195-213.

Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. Nucleic Acids Research 39(Web Server issue):W347-52.

Appendix 1 Count of RITA reference genomes by taxonomic class.

Taxonomic Class	Count
Gammaproteobacteria	630
Alphaproteobacteria	432
Bacilli	392
Actinobacteria	247
Betaproteobacteria	238
Clostridia	142
Spirochaetia	133
Cyanobacteria (class)	106
Halobacteria	64
Deltaproteobacteria	60
Epsilonproteobacteria	57
Mollicutes	46
Thermoprotei	41
Flavobacteriia	36
Deinococci	32
Bacteroidia	30
Chlamydiia	30
Methanococci	23
Cytophagia	21
Methanomicrobia	20
Negativicutes	15
Thermococci	14
Thermotogae	14
Bacteroidetes (class)	13
Aquificae	12
Chlorobia	12
Sphingobacteriia	11
Fusobacteriia	10
Acidobacteriia	9
Methanobacteria	9
Chloroflexi	8
Planctomycetia	7
Deferribacteres	6

Taxonomic Class	**Count**
Dehalococcoidetes	6
Archaeoglobi	5
Elusimicrobia (class)	4
Synergistia	4
Thermomicrobia	4
Acidobacteria (class)	3
Archaea (class)	3
Bacteria (class)	3
Thermoplasmata	3
Dictyoglomia	2
Euryarchaeota (class)	2
Nitrospira	2
Opitutae	2
Phycisphaerae	2
Thaumarchaeota (class)	2
Thermodesulfobacteria	2
Viruses (class)	2
Anaerolineae	1
Caldilineae	1
Chrysiogenetes	1
Elusimicrobia	1
Erysipelotrichi	1
Fibrobacteria	1
Gemmatimonadetes	1
Gloeobacteria	1
Korarchaeota (class)	1
Methanopyri	1
Nanoarchaeota (class)	1
Proteobacteria (class)	1
Solibacteres	1
Verrucomicrobia (class)	1
Verrucomicrobiae	1
Total	2986

Appendix 2 List of enzymes for each KEGG pathway and their names.

Excel file of all enzymes for each metabolic pathway. See attached file “Appendix2.xlsx”

Appendix 3 Classes and counts of contigs classified at the class level greater than 1000 nucleotides long for each community.

Classes shared between each community are also listed. Not all classes could be classified to the class level, and are identified by the taxonomic group they are classified to, followed by '(class)'.

Taxonomic Class	USA	AU	DK
Acidobacteria (class)	10	28	147
Acidobacteriia	24	48	337
Actinobacteria	95	168	523
Alphaproteobacteria	529	607	537
Anaerolineae		7	588
Aquificae	1	6	19
Archaea (class)			1
Archaeoglobi			6
Bacilli	22	71	734
Bacteria (class)		8	16
Bacteroidetes (class)	34	36	86
Bacteroidia	24	97	408
Betaproteobacteria	3989	2309	3073
Caldilineae	2	18	73
Chlamydiia	4	11	110
Chlorobia	19	131	194
Chloroflexi	3	172	233
Chrysiogenetes			3
Clostridia	39	246	925
Cyanobacteria (class)	19	69	302
Cytophagia	222	205	1343
Deferribacteres	2	12	32
Dehalococcoidetes			11
Deinococci	2	31	37
Deltaproteobacteria	61	160	812
Dictyoglomia		2	4
Elusimicrobia		1	5

Taxonomic Class	**USA**	**AU**	**DK**
Elusimicrobia (class)		5	
Epsilonproteobacteria	6	25	99
Erysipelotrichi			1
Euryarchaeota (class)		1	6
Fibrobacteria		2	35
Flavobacteriia	464	464	1668
Fusobacteriia		54	107
Gammaproteobacteria	1121	1669	1317
Gemmatimonadetes	2	6	26
Gloeobacteria		6	12
Halobacteria	1	1	11
Korarchaeota (class)			1
Methanobacteria	1	11	50
Methanococci	1	7	15
Methanomicrobia	3	46	210
Methanopyri			2
Mollicutes		2	16
Negativicutes			8
Nitrospira	12	28	69
Opitutae	21	67	143
Phycisphaerae			1
Planctomycetia	16	83	132
Proteobacteria (class)	3	3	7
Solibacteres	11	15	46
Sphingobacteriia	837	258	3192
Spirochaetia	3	43	81
Synergistia		1	3
Thaumarchaeota (class)	1	1	14
Thermococci		2	19
Thermodesulfobacteria	1		5
Thermomicrobia	3	4	34
Thermoplasmata		1	5
Thermoprotei			18
Thermotogae	1	41	91
Verrucomicrobia (class)		3	1
Verrucomicrobiae		38	1
Viruses (class)	1	2	19

Appendix 4 Count of contigs classified under each RITA classification method.

Group 1 is when homology and nucleotide composition both agree. Group 2 is where the USEARCH expectation value for the best-matching genome was at least 10 orders of magnitude smaller than the best-matching genome from a different class. Group 3 assignments are made when the NB likelihood score for the best-matching genome is at least 1.5 times greater than the NB likelihood for the best-matching genome from another class. Group 4 assignments are based only on the best NB likelihood value. See attached file "Appendix4.xlsx"

Appendix 5 Potential LGTs from taxonomic discordance analysis.

Taxonomic discordance analysis (Naïve Bayes filtered and unfiltered) predicted proteins identified as transferred for each pathway, their EC annotation, class-level taxonomic classification of source (ORF) and recipient (contig). See attached file “Appendix5.xlsx”

Appendix 6 Direction and number of transfers in classification discordance analysis.

Source to Recipient	**DK**	**USA**	**AU**
Acidobacteria (class)			
Clostridia			1
Acidobacteriia			
Deltaproteobacteria	1		
Actinobacteria			
Betaproteobacteria	1		1
Clostridia			1
Gammaproteobacteria			1
Sphingobacteriia	1		
Alphaproteobacteria			
Betaproteobacteria	1	3	2
Epsilonproteobacteria			1
Gammaproteobacteria		1	7
Anaerolineae			
Alphaproteobacteria	2		
Betaproteobacteria	1		
Bacilli			
Chlorobia	1		
Gammaproteobacteria			2
Flavobacteriia	1		
Proteobacteria (class)	1		
Thermotogae			1
Bacteroidetes (class)			
Bacilli			1
Chlorobia			1
Methanomicrobia	1		1
Bacteroidia			
Betaproteobacteria	2		

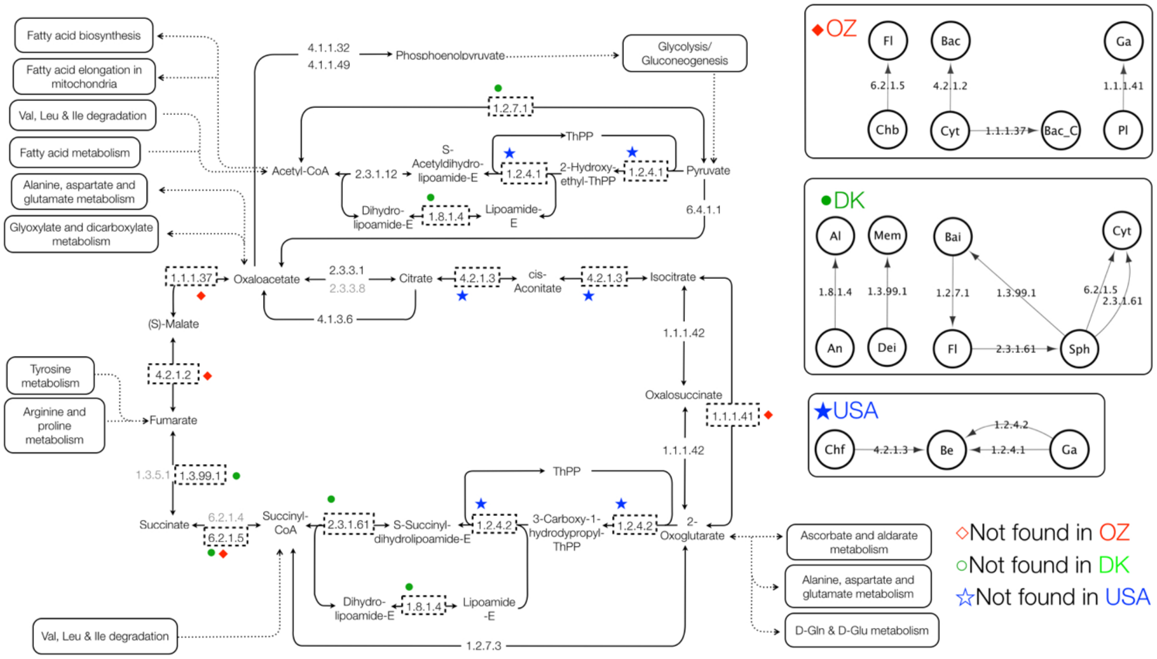
Source to Recipient	**DK**	**USA**	**AU**
Cytophagia	1		
Betaproteobacteria			
Anaerolineae	1		
Alphaproteobacteria		1	2
Chlorobia			1
Cyanobacteria (class)			1
Gammaproteobacteria	1		21
Caldilineae			
Betaproteobacteria	2		
Clostridia	1		
Chlamydiia			
Bacilli	1		
Chlorobia			
Clostridia			4
Deltaproteobacteria			1
Flavobacteriia			3
Gammaproteobacteria			2
Sphingobacteriia			1
Chloroflexi			
Betaproteobacteria		3	1
Methanomicrobia			1
Clostridia			
Anaerolineae			1
Betaproteobacteria			2
Chlorobia			1
Deltaproteobacteria			1
Flavobacteriia	1		
Gammaproteobacteria			1
Sphingobacteriia	1	1	
Cyanobacteria (class)			
Flavobacteriia			1
Cytophagia			
Bacilli	1		

Source to Recipient	**DK**	**USA**	**AU**
Bacteroidetes (class)			1
Bacteroidia	1		1
Flavobacteriia			2
Sphingobacteriia	1	1	
Deinococci			
Methanomicrobia	1		
Deltaproteobacteria			
Alphaproteobacteria	1		
Anaerolineae	1		
Betaproteobacteria		1	
Gammaproteobacteria			2
Flavobacteriia			
Cytophagia	1		1
Gammaproteobacteria			
Sphingobacteriia	3		
Fusobacteriia			
Chlorobia			1
Epsilonproteobacteria			
Alphaproteobacteria		1	
Betaproteobacteria		1	1
Gammaproteobacteria			
Alphaproteobacteria		1	4
Betaproteobacteria		3	6
Chloroflexi	1		
Cyanobacteria (class)	1		
Methanomicrobia			
Chloroflexi	1		
Opitutae			
Gammaproteobacteria			1
Planctomycetia			
Gammaproteobacteria			2
Sphingobacteriia			
Acidobacteriia	1		

Source to Recipient	**DK**	**USA**	**AU**
Alphaproteobacteria	1		
Bacilli	1		
Bacteroidia	1		
Betaproteobacteria			1
Chlorobia	1		1
Cytophagia	5	1	
Flavobacteriia	1	1	
Gammaproteobacteria	1	1	
Spirochaetia			
Deltaproteobacteria			1
Verrucomicrobia (class)			
Betaproteobacteria			1

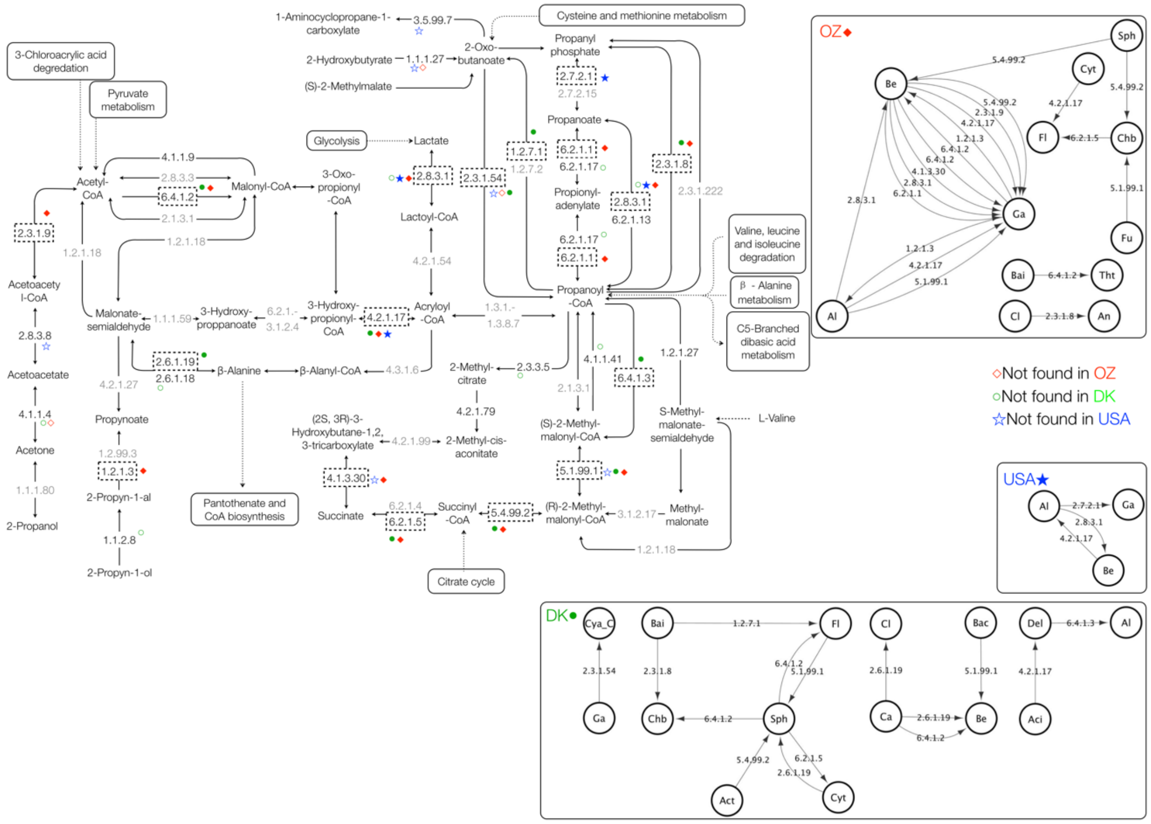
Appendix 7 KEGG citrate cycle pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities.

Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.



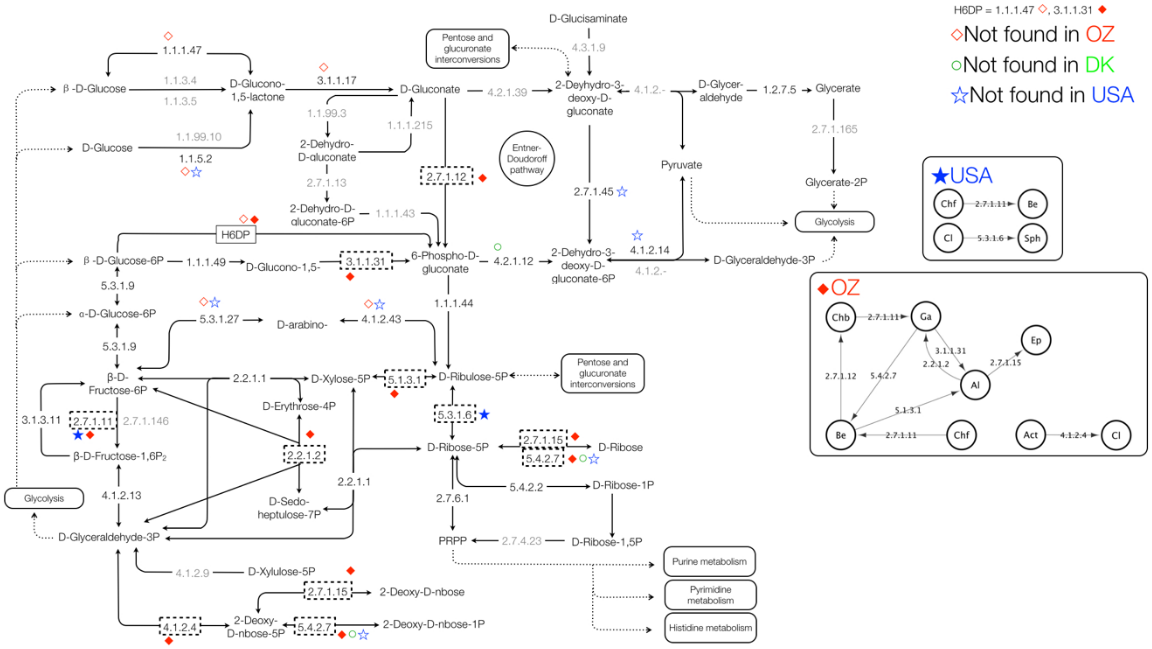
Appendix 8 KEGG propanoate metabolism pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities.

Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.



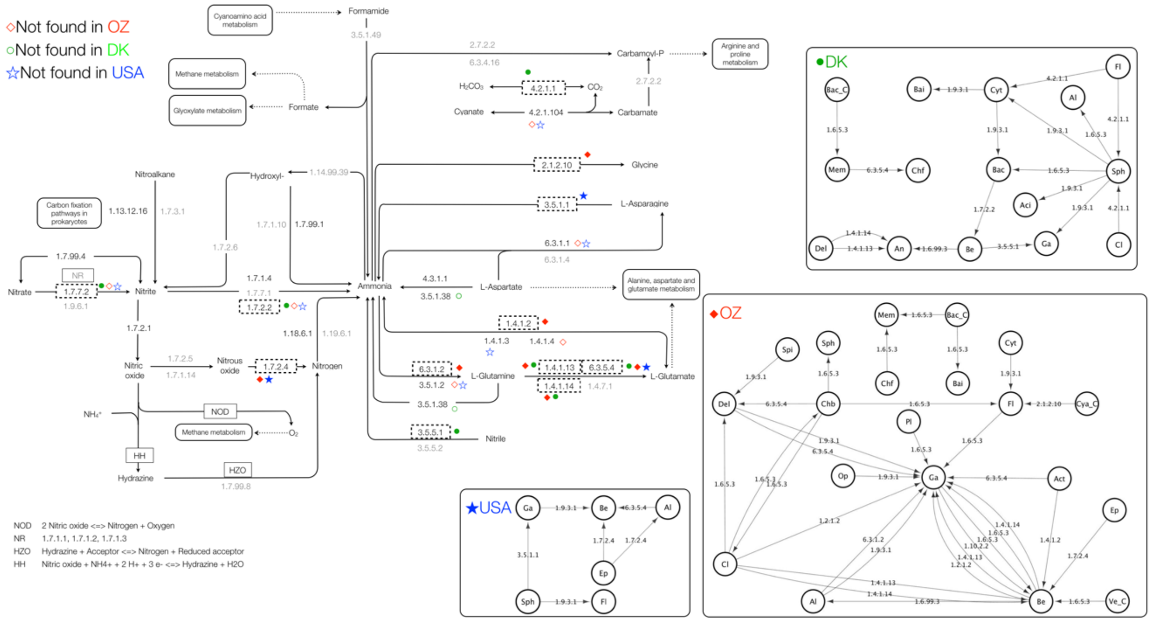
Appendix 9 KEGG pentose phosphate pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities.

Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.



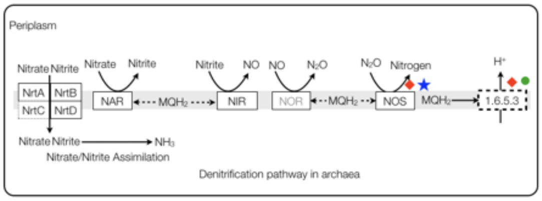
Appendix 10 KEGG nitrogen metabolism pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities

Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.

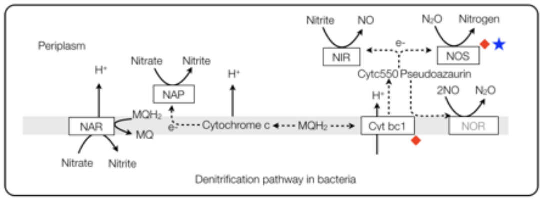


Appendix 11 KEGG partial nitrogen metabolism pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities in the inner membranes of *Nitromonas europaea*, archaea and bacteria.

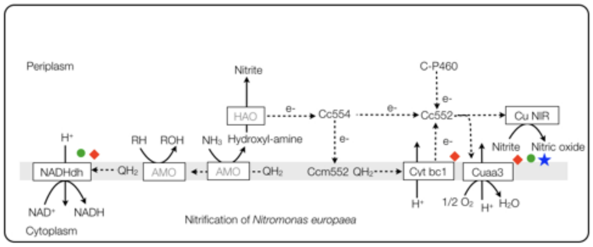
Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.



NrtA	nitrate/nitrite transport system substrate-binding protein
NrtB	periplasmic nitrate reductase NapA [EC:1.7.99.4]
NrtC	nitrate/nitrite transport system ATP-binding protein [EC:3.6.3.-]
NrtD	nitrate/nitrite transport system ATP-binding protein
NAR	1.7.99.4
NIR	1.7.2.1
NOR	1.7.2.5
NOS	1.7.2.4



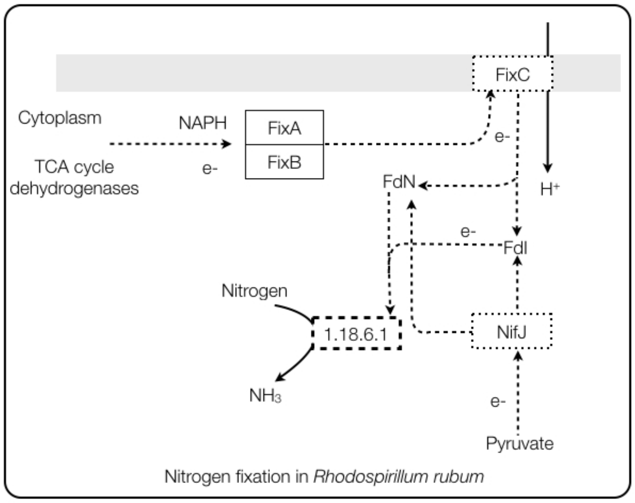
NAR	1.7.99.4
NAP	1.7.99.4
NIR	1.7.2.1
Cyt bc1	1.10.2.2
NOS	1.7.2.4
NOR	1.7.2.5



NADHdh	1.6.5.3, 1.6.99.3, 1.6.99.5 (No LGT)
AMO	1.14.99.39
HAO	1.7.2.6
Cytbc1	1.10.2.2
Cuaa3	1.9.3.1
Cu NIR	1.7.2.1

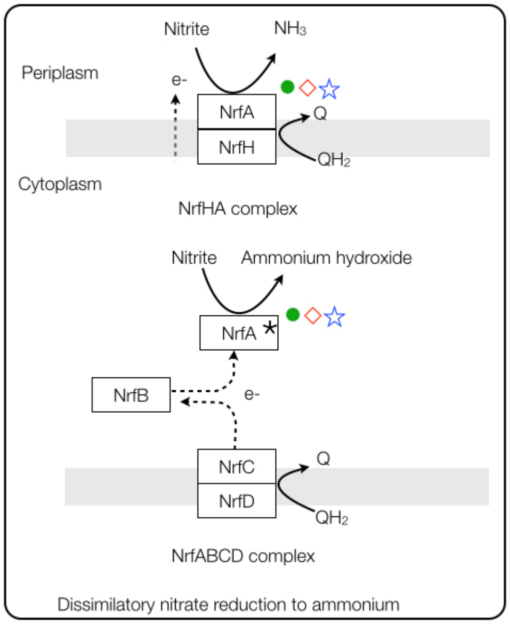
Appendix 12 KEGG partial nitrogen metabolism pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities in *Rhodospirillum rubum*.

Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.



Nitrogen fixation in *Rhodospirillum rubum*

FixA	electron transfer flavoprotein beta subunit
FixB	electron transfer flavoprotein alpha subunit
FixC	electron transfer flavoprotein-quinone oxidoreductase 1.5.5.-
NifJ	putative pyruvate-flavodoxin oxidoreductase 1.2.7.-

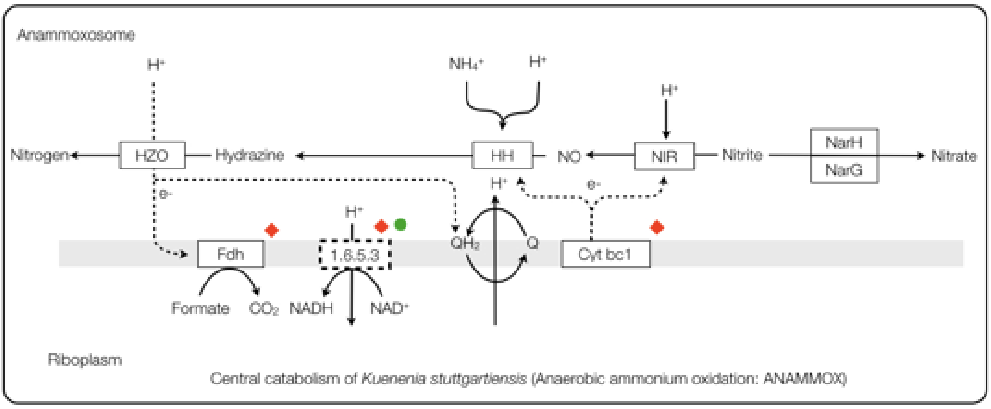


Dissimilatory nitrate reduction to ammonium

NrfA	1.7.2.2
NrfH	cytochrome c nitrite reductase small subunit
NrfB	cytochrome c-type protein
NrfC	protein NrfC
NrfD	protein NrfD

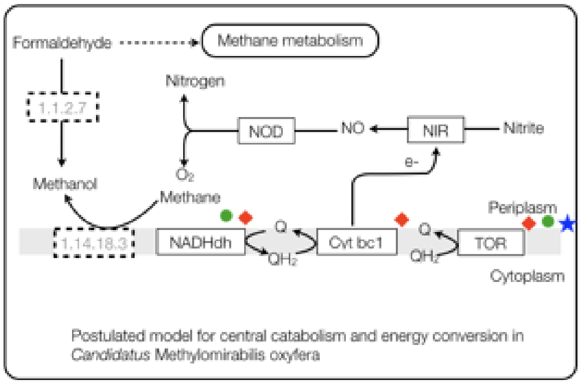
Appendix 13 KEGG partial nitrogen metabolism pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities in *Kuenenia stuttgartiensis*, *Candidatus* Methylomairabilis oxyfera, *Klebsiella oxytoca* and *Synechococcus PCC7942*.

Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.



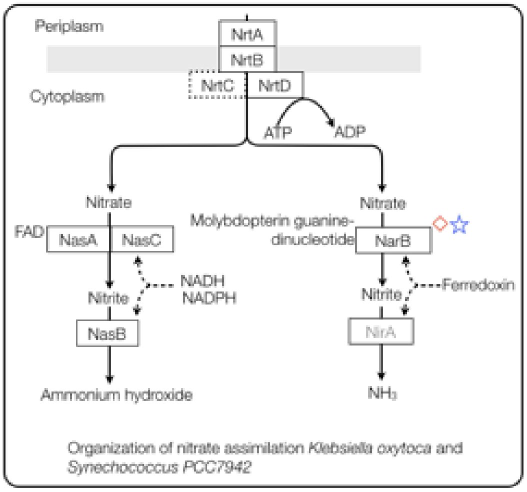
Central catabolism of *Kuenenia stuttgartiensis* (Anaerobic ammonium oxidation: ANAMMOX)

HZO	hydrazine:acceptor oxidoreductase
HH	Nitric oxide + NH4+ + 2 H+ + 3 e- <=> Hydrazine + H2O
Fdh	1.2.1.2
NIR	1.7.2.1
NarH, NarG	1.7.99.4
Cyt bc1	1.10.2.2



Postulated model for central catabolism and energy conversion in *Candidatus* Methylomirabilis oxyfera

NOD	2 Nitric oxide <=> Nitrogen + Oxygen
NIR	1.7.2.1
NADHdh	1.6.5.3, 1.6.99.3, 1.6.99.5 (No LGT)
Cyt bc1	1.10.2.2
TOR	1.9.3.1

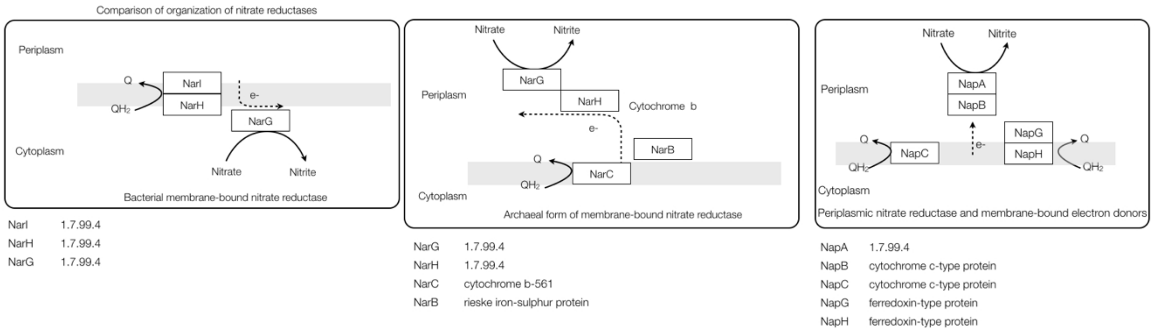


Organization of nitrate assimilation *Klebsiella oxytoca* and *Synechococcus PCC7942*

NrtA	nitrate/nitrite transport system substrate-binding protein
NrtB	nitrate/nitrite transport system permease protein
NrtC	3.6.3.-
NrtD	nitrate/nitrite transport system ATP-binding protein
NasA, NasC	1.7.99.4
NasB	1.7.1.4
NarB	1.7.7.2
NirA	1.7.7.1

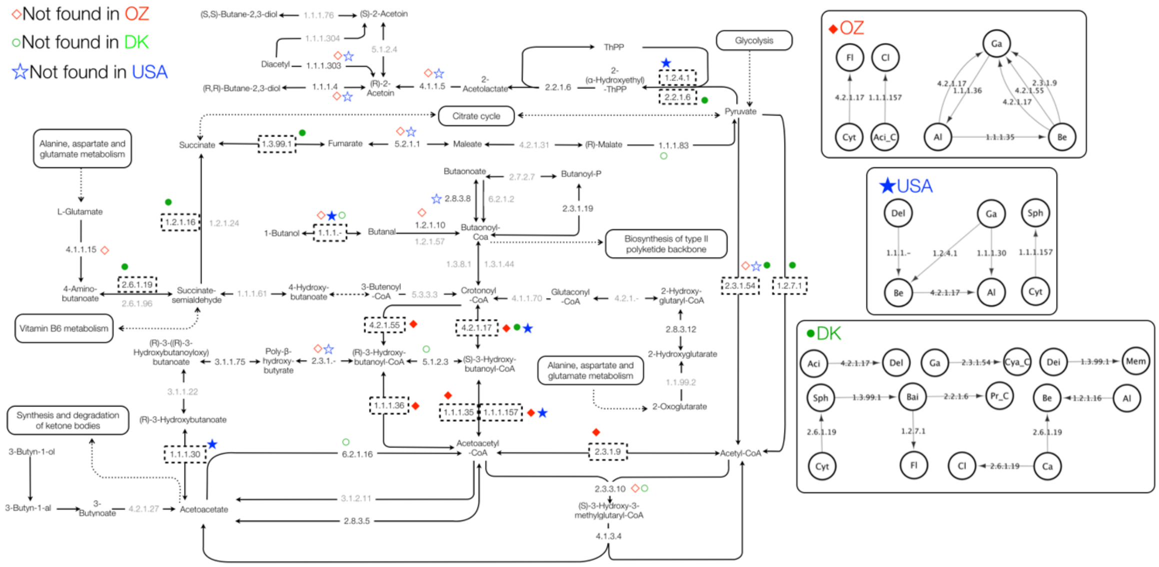
Appendix 14 KEGG partial nitrogen metabolism pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities.

Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.


Comparison of organization of nitrate reductases
Periplasm
Q
QH2
NarI
NarH
e-
NarG
Cytoplasm
Nitrate
Nitrite
Bacterial membrane-bound nitrate reductase
NarI 1.7.99.4
NarH 1.7.99.4
NarG 1.7.99.4
Nitrate
Nitrite
NarG
Periplasm
NarH
Cytochrome b
e-
NarB
Q
QH2
NarC
Cytoplasm
Archaeal form of membrane-bound nitrate reductase
NarG 1.7.99.4
NarH 1.7.99.4
NarC cytochrome b-561
NarB rieske iron-sulphur protein
Nitrate
Nitrite
Periplasm
NapA
NapB
e-
NapG
NapH
Q
QH2
NapC
Q
QH2
Cytoplasm
Periplasmic nitrate reductase and membrane-bound electron donors
NapA 1.7.99.4
NapB cytochrome c-type protein
NapC cytochrome c-type protein
NapG ferredoxin-type protein
NapH ferredoxin-type protein

Appendix 15 KEGG butanoate metabolism pathway and directed LGT for the Denmark (DK), Australian (AU) and United States (USA) EBPR communities.

Dashed boxes indicate LGT, with solid symbols indicating LGT predicted within a community, and hollow symbols indicating missing enzymes in a community. Greyed out enzymes are not found in any community. See Table 2.1 for enzyme names and Appendix 18 for taxonomic abbreviation guide.



Appendix 16 Potential LGT events from phylogenetic analysis.

Phylogenetic analysis predicted proteins identified as transferred for each pathway, their EC annotation, and class-level taxonomic classification of the contig they are located. BM = Butanoate Metabolism, CAC = Citric Acid Cycle, GM = Glycolysis/Gluconeogenesis Metabolism, PM = Propanoate Metabolism, PPP = Pentose Phosphate Pathway

Sequence	Contig Classification	EC	Pathway
2000096700_sludgeUSA_Contig14439	Gammaproteobacteria	1.1.1.1	GM
2000426560_sludgeOz_Contig11426	Betaproteobacteria	1.1.1.35	BM
2000544330_sludgeOz_Contig5118	Alphaproteobacteria	1.1.1.35	BM
2000424640_sludgeOz_Contig11412	Gammaproteobacteria	1.2.1.12	GM
2000588520_sludgeOz_Contig7421	Betaproteobacteria	1.2.1.12	GM
2000445730_sludgeOz_Contig11529	Gammaproteobacteria	1.2.1.3	GM
2000513410_sludgeOz_Contig3469	Alphaproteobacteria	1.2.1.3	GM
2000462440_sludgeOz_Contig11597	Gammaproteobacteria	1.6.5.3	NM
2000610430_sludgeOz_Contig8413	Betaproteobacteria	1.6.5.3	NM
2000540670_sludgeOz_Contig4927	Gammaproteobacteria	1.8.1.4	CAC
2000410250_sludgeOz_Contig11289	Gammaproteobacteria	2.3.1.9	BM;PPP
2000452010_sludgeOz_Contig11556	Gammaproteobacteria	2.3.1.9	BM;PPP
2000465430_sludgeOz_Contig11607	Betaproteobacteria	2.3.1.9	BM;PPP
2000422830_sludgeOz_Contig11399	Gammaproteobacteria	2.7.1.11	
2000527690_sludgeOz_Contig4229	Chlorobia	2.7.1.11	
2000424650_sludgeOz_Contig11412	Gammaproteobacteria	2.7.2.3	GM
2000424660_sludgeOz_Contig11412	Gammaproteobacteria	4.1.2.13	GM
2000609960_sludgeOz_Contig8392	Bacilli	4.1.2.13	GM
2000361210_sludgeOz_Contig10422	Gammaproteobacteria	4.1.3.30	PPM
2000454070_sludgeOz_Contig11564	Gammaproteobacteria	4.1.3.30	PPM
2000408210_sludgeOz_Contig11270	Gammaproteobacteria	4.2.1.17	BM
2000452020_sludgeOz_Contig11556	Gammaproteobacteria	4.2.1.17	BM
2000460640_sludgeOz_Contig11591	Gammaproteobacteria	4.2.1.17	BM
2000374110_sludgeOz_Contig10726	Betaproteobacteria	4.2.1.3	CAC
2000424740_sludgeOz_Contig11412	Gammaproteobacteria	5.4.2.2	GM
2000441420_sludgeOz_Contig11512	Betaproteobacteria	5.4.2.2	GM

Sequence	Contig Classification	EC	Pathway
2000589970_sludgeOz_Contig7493	Gammaproteobacteria	6.4.1.2	
2000647250_sludgeOz_Contig9815	Betaproteobacteria	6.4.1.2	PPM
144965_54_1067_minus_danish144965	Bacteroidia	1.2.1.12	GM
142342_1368_3942_plus_danish142342	Alphaproteobacteria	1.2.1.2	NM
144688_1209_1625_minus_danish144688	Flavobacteriia	1.2.1.3	GM;PPP
146289_1_1239_plus_danish146289	Cytophagia	1.3.99.1	BM;CAC
149437_1_1680_plus_danish149437	Cytophagia	1.8.1.4	CAC
156099_521_1086_plus_danish156099	Gammaproteobacteria	1.8.1.4	CAC
141966_3134_5011_minus_danish141966	Cytophagia	1.9.3.1	NM
19791_4533_5707_plus_danish19791	Bacilli	1.9.3.1	NM
146462_101_1413_minus_danish146462	Methanomicrobia	3.1.3.11	
151070_1_776_minus_danish151070	Sphingobacteriia	3.1.3.11	
151799_1_974_minus_danish151799	Alphaproteobacteria	4.1.2.13	GM
180612_1_689_minus_danish180612	Flavobacteriia	4.1.3.4	BM
245921_1_1136_plus_danish245921	Clostridia	1.1.1.42	CAC
198229_1_704_plus_danish198229	Cytophagia	4.1.3.4	BM

Appendix 17 Sequences from each community with homologs to known mobile genetic elements.

EBPR sequences with homologs to mobile genetic elements for each pathway, their EC annotation, and class-level taxonomic classification of the contig they are located. See attached file "Appendix17.xlsx"

Appendix 18 List of taxonomic classes and their abbreviations.

Class	Abbreviation
Acidobacteria (class)	Aci_C
Acidobacteriia	Aci
Actinobacteria	Act
Alphaproteobacteria	Al
Anaerolineae	An
Aquificae	Aq
Archaeoglobi	Ar
Bacilli	Bai
Bacteria (class)	Baa_C
Bacteroidetes (class)	Bac_C
Bacteroidia	Bac
Betaproteobacteria	Be
Caldilineae	Ca
Chlamydiia	Chl
Chlorobia	Chb
Chloroflexi	Chf
Clostridia	Cl
Cyanobacteria (class)	Cya_C
Cytophagia	Cyt
Deferribacteres	Def
Dehalococcoidetes	Deh
Deinococci	Dei
Deltaproteobacteria	Del
Dictyoglomia	Di
Elusimicrobia (class)	El_C
Elusimicrobia	El
Epsilonproteobacteria	Ep
Fibrobacteria	Fi
Flavobacteriia	Fl
Fusobacteriia	Fu
Gammaproteobacteria	Ga
Gemmatimonadetes	Ge
Gloeobacteria	Gl
Halobacteria	Ha
Methanobacteria	Meb
Methanococci	Mec
Methanomicrobia	Mem
Mollicutes	Mo
Negativicutes	Ne
Nitrospira	Ni
Opitutae	Op

Class	**Abbreviation**
Planctomycetia	Pl
Proteobacteria (class)	Pr_C
Solibacteres	So
Sphingobacteriia	Sph
Spirochaetia	Spi
Synergistia	Sy
Thaumarchaeota (class)	Tha_C
Thermococci	Thc
Thermomicrobia	Thm
Thermotogae	Tht
Unclassified	U
Verrucomicrobia (class)	Ve_C
Verrucomicrobiae	Ve

Appendix 19 Table of 100 PanPhyla genomes for ProPhylClust

PanPhyla Genomes	Phylum, Class
ACIDAMINOCOCCUS_FERMENTANS_DSM_20731_UID43471	Firmicutes,Negativicutes
ACIDIMICROBIUM_FERROOXIDANS_DSM_10331_UID59215	Actinobacteria,Actinobacteria
ACIDOBACTERIUM_CAPSULATUM_ATCC_51196_UID59127	Acidobacteria,Acidobacteria
ACIDOBACTERIUM_MP5ACTX9_UID50551	Acidobacteria,Acidobacteria
AMINOBACTERIUM_COLOMBIENSE_DSM_12261_UID47083	Synergistetes,Synergistia
ANABAENA_CYLINDRICA_PCC_7122_UID183339	Cyanobacteria,Cyanobacteria
ANAEROLINEA_THERMOPHILA_UNI_1_UID62245	Chloroflexi,Anaerolineae
AQUIFEX_AEOLICUS_VF5_UID57765	Aquificae,Aquificae
BACILLUS_CEREUS_ATCC_14579_UID57975	Firmicutes,Bacilli
BACILLUS_SUBTILIS_XF_1_UID189187	Firmicutes,Bacilli
BACTEROIDES_FRAGILIS_638R_UID84217	Bacteroidetes,Bacteroidetes
BIFIDOBACTERIUM_ANIMALIS_LACTIS_ATCC_27673_UID222803	Actinobacteria,Actinobacteria
BORRELIA_BURGDORFERI_B31_UID57581	Spirochaetes,Spirochaetes
BURKHOLDERIA_PSEUDOMALLEI_668_UID58389	Proteobacteria,BetaProteobacteria
CALDITERRIVIBRIO_NITROREDUCENS_DSM_19672_UID60821	Deferribacteres,Deferribacteres
CAMPYLOBACTER_JEJUNI_M1_UID159535	Proteobacteria,EpsilonProteobacteria
CHLAMYDIA_PECORUM_PV3056_3_UID221290	Chlamydiae,Chlamydiae
CHLAMYDIA_TRACHOMATIS_UID196778	Chlamydiae,Chlamydiae
CHLOROBACULUM_PARVUM_NCIB_8327_UID59185	Chlorobi,Chlorobia
CHLOROBIUM_TEPIDUM_TLS_UID57897	Chlorobi,Chlorobia
CHLOROFLEXUS_AURANTIACUS_J_10_FL_UID57657	Chloroflexi,Chloroflexia
CHLOROHERPETON_THALASSIUM_ATCC_35110_UID59187	Chlorobi,Chlorobia
CLOSTRIDIUM_BOTULINUM_A_ATCC_3502_UID61579	Firmicutes,Clostridia
CLOSTRIDIUM_DIFFICILE_CD196_UID41017	Firmicutes,Clostridia
CORYNEBACTERIUM_DIPHTHERIAE_31A_UID84309	Actinobacteria,Actinobacteria

PanPhyla Genomes	Phylum, Class
CYANOTHECE_ATCC_51142_UID59013	Cyanobacteria,Cyanobacteria
CYTOPHAGA_HUTCHINSONII_ATCC_33406_UID57651	Bacteroidetes,Cytophagia
DEFERRIBACTER_DESULFURICANS_SSM1_UID46653	Deferribacteres,Deferribacteres
DEHALOCOCCOIDES_ETHENOGENES_195_UID57763	Chloroflexi,Dehalococcoidia
DESULFITOBACTERIUM_DEHALOGENANS_ATCC_51507_UID82553	Firmicutes,Clostridia
DESULFOTALEA_PSYCHROPHILA_LSV54_UID58153	Proteobacteria,DeltaProteobacteria
DESULFOVIBRIO_VULGARIS__MIYAZAKI_F__UID59089	Proteobacteria,DeltaProteobacteria
ELUSIMICROBIUM_MINUTUM_PEI191_UID58949	Elusimicrobia,Elusimicrobia
ERYSIPELOTHRIX_RHUSIOPATHIAE_SY1027_UID206518	Firmicutes,Erysipelotrichi
ESCHERICHIA_COLI_K_12_SUBSTR__MG1655_UID57779	Proteobacteria,GammaProteobacteria
FERVIDOBACTERIUM_NODOSUM_RT17_B1_UID58625	Thermotogae,Thermotogae
FIBROBACTER_SUCCINOGENES_S85_UID41169	Fibrobacteres,Fibrobacteres
FLAVOBACTERIUM_COLUMNARE_ATCC_49512_UID80731	Bacteroidetes,Flavobacteriia
FUSOBACTERIUM_NUCLEATUM_ATCC_25586_UID57885	Fusobacteria,Fusobacteria
GEMMATIMONAS_AURANTIACA_T_27_UID58813	Gemmatimonadetes,Gemmatimonadetes
GRANULICELLA_MALLENSIS_MP5ACTX8_UID49957	Acidobacteria,Acidobacteria
HELICOBACTER_PYLORI_UID159983	Proteobacteria,EpsilonProteobacteria
HYDROGENOBACULUM_Y04AAS1_UID58857	Aquificae,Aquificae
ILYOBACTER_POLYTROPUS_DSM_2926_UID59769	Fusobacteria,Fusobacteria
LACTOBACILLUS_ACIDOPHILUS_30SC_UID63605	Firmicutes,Bacilli
LEIFSONIA_XYLI_CTCB07_UID57759	Actinobacteria,Actinobacteria
LEPTOSPIRA_INTERROGANS_SEROVAR_LAI_IPAV_UID161957	Spirochaetes,Spirochaetes
LEPTOSPIRILLUM_FERROOXIDANS_C2_3_UID158171	Nitrospirae,Nitrospira
LISTERIA_MONOCYTOGENES_UID43671	Firmicutes,Bacilli
MARINITHERMUS_HYDROTHERMALIS_DSM_14884_UID65783	Deinococcus-Thermus,Deinococci
MEIOTHERMUS_SILVANUS_DSM_9946_UID49485	Deinococcus-Thermus,Deinococci
MESOPLASMA_FLORUM_L1_UID58055	Tenericutes,Mollicutes
MESOTOGA_PRIMA_MESG1_AG_4_2_UID52599	Thermotogae,Thermotogae
METHYLACIDIPHILUM_INFERNORUM_V4_UID59161	Verrucomicrobia,Verrucomicrobia

PanPhyla Genomes	Phylum, Class
MICROCYSTIS_AERUGINOSA_NIES_843_UID59101	Cyanobacteria,Cyanobacteria
MYCOBACTERIUM_TUBERCULOSIS_UID185758	Actinobacteria,Actinobacteria
MYCOPLASMA_GALLISEPTICUM_F_UID162001	Tenericutes,Mollicutes
MYCOPLASMA_PNEUMONIAE_309_UID85495	Tenericutes,Mollicutes
NEISSERIA_MENINGITIDIS_8013_UID161967	Proteobacteria,BetaProteobacteria
NITROBACTER_HAMBURGENSIS_X14_UID58293	Proteobacteria,AlphaProteobacteria
NITROSOSPIRA_MULTIFORMIS_ATCC_25196_UID58361	Nitrospirae,Nitrospirae
NOCARDIA_FARCINICA_IFM_10152_UID58203	Actinobacteria,Actinobacteria
NOSTOC_PCC_7107_UID182932	Cyanobacteria,Cyanobacteria
ONION_YELLOWS_PHYTOPLASMA_OY_M_UID58015	Tenericutes,Mollicutes
OPITUTUS_TERRAE_PB90_1_UID58965	Verrucomicrobia,Opitutae
PARACHLAMYDIA_ACANTHAMOEBAE_UV7_UID68335	Chlamydiae,Chlamydiae
PEDOBACTER_HEPARINUS_DSM_2366_UID59111	Bacteroidetes,Sphingobacteria
PHYCISPHAERA_MIKURENSIS_NBRC_102666_UID157331	Planctomycetes,Planctomycetes
PLANCTOMYCES_BRASILIENSIS_DSM_5305_UID60583	Planctomycetes,Planctomycetes
PORPHYROMONAS_GINGIVALIS_ATCC_33277_UID58879	Bacteroidetes,Bacteroidetes
PREVOTELLA_DENTALIS_DSM_3688_UID184818	Bacteroidetes,Bacteroidetes
PROCHLOROCOCCUS_MARINUS_CCMP1375_UID57995	Cyanobacteria,Cyanobacteria
PROPIONIBACTERIUM_ACNES_ATCC_11828_UID162177	Actinobacteria,Actinobacteria
PROSTHECOCHLORIS_AESTUARII_DSM_271_UID58151	Chlorobi,Chlorobia
RICKETTSIA_RICKETTSII__SHEILA_SMITH__UID58027	Proteobacteria,AlphaProteobacteria
SALMONELLA_ENTERICA_SEROVAR_4_5_12_I__08_1736_UID212969	Proteobacteria,GammaProteobacteria
SELENOMONAS_RUMINANTIUM_LACTILYTICA_TAM6421_UID157247	Firmicutes,Negativicutes
SPIROCHAETA_THERMOPHILA_DSM_6192_UID53037	Spirochaetes,Spirochaetes
SPIROCHAETA_THERMOPHILA_DSM_6578_UID162041	Spirochaetes,Spirochaetes
STAPHYLOCOCCUS_AUREUS_71193_UID162141	Firmicutes,Bacilli
STREPTOBACILLUS_MONILIFORMIS_DSM_12112_UID41863	Fusobacteria,Fusobacteria
STREPTOCOCCUS_PNEUMONIAE_670_6B_UID52533	Firmicutes,Bacilli
STREPTOMYCES_CATTLEYA_NRRL_8057___DSM_46488_UID77117	Actinobacteria,Actinobacteria
STREPTOMYCES_SCABIEI_87_22_UID46531	Actinobacteria,Actinobacteria

PanPhyla Genomes	Phylum, Class
SYNECHOCOCCUS_ELONGATUS_PCC_7942_UID58045	Cyanobacteria,Cyanobacteria
SYNECHOCYSTIS_PCC_6803_UID57659	Cyanobacteria,Cyanobacteria
THERMODESULFATATOR_INDICUS_DSM_15286_UID68285	Thermodesulfobacteria,Thermodesulfobacteria
THERMODESULFOBACTERIUM_OPB45_UID68283	Thermodesulfobacteria,Thermodesulfobacteria
THERMODESULFOVIBRIO_YELLOWSTONII_DSM_11347_UID59257	Nitrospirae,Nitrospira
THERMOMICROBIUM_ROSEUM_DSM_5159_UID59341	Chloroflexi,Thermomicrobia
THERMOTOGA_MARITIMA_MSB8_UID57723	Thermotogae,Thermotogae
THERMOVIRGA_LIENII_DSM_17291_UID77129	Synergistetes,Synergistia
THERMUS_THERMOPHILUS_HB8_UID58223	Deinococcus-Thermus,Deinococci
TREPONEMA_PALLIDUM_PERTENUE_CDC2_UID87051	Spirochaetes,Spirochaetes
TROPHERYMA_WHIPPLEI_TWIST_UID57705	Actinobacteria,Actinobacteria
TRUEPERA_RADIOVICTRIX_DSM_17093_UID49533	Deinococcus-Thermus,Deinococci
UREAPLASMA_PARVUM_SEROVAR_3_ATCC_700970_UID57711	Tenericutes,Mollicutes
UREAPLASMA_UREALYTICUM_SEROVAR_10_ATCC_33699_UID59011	Tenericutes,Mollicutes
VIBRIO_FISCHERI_ES114_UID58163	Proteobacteria,GammaProteobacteria
WADDLIA_CHONDROPHILA_WSU_86_1044_UID49531	Chlamydiae,Chlamydiae

Appendix 20 Table of 100 *Proteobacteria* genomes for ProPhylClust

Proteobacteria Genomes	Class
Acetobacter_pasteurianus_386B_uid214433	Alphaproteobacteria
Acidiphilium_multivorum_AIU301_uid63345	Alphaproteobacteria
Acinetobacter_ADP1_uid61597	Gammaproteobacteria
Acinetobacter_baumannii_ATCC_17978_uid58731	Gammaproteobacteria
Acinetobacter_calcoaceticus_PHEA_2_uid83123	Gammaproteobacteria
Acinetobacter_oleivorans_DR1_uid50119	Gammaproteobacteria
Agrobacterium_fabrum_C58_uid57865	Alphaproteobacteria
Agrobacterium_radiobacter_K84_uid58269	Alphaproteobacteria
Allochromatium_vinosum_DSM_180_uid46083	Gammaproteobacteria
Arcobacter_butzleri_7h1h_uid200766	Epsilonproteobacteria
Azoarcus_BH72_uid61603	Betaproteobacteria
Beijerinckia_indica_ATCC_9039_uid59057	Deltaproteobacteria
Burkholderia_mallei_ATCC_23344_uid57725	Betaproteobacteria
Burkholderia_pseudomallei_668_uid58389	Betaproteobacteria
Campylobacter_coli_15_537360_uid226113	Epsilonproteobacteria
Campylobacter_fetus_82_40_uid58545	Epsilonproteobacteria
Campylobacter_hominis_ATCC_BAA_381_uid58981	Epsilonproteobacteria
Campylobacter_jejuni_81116_uid58771	Epsilonproteobacteria
Campylobacter_jejuni_M1_uid159535	Epsilonproteobacteria
Dechloromonas_aromatica_RCB_uid58025	Betaproteobacteria
Desulfatibacillum_alkenivorans_AK_01_uid58913	Deltaproteobacteria
Desulfocapsa_sulfexigens_DSM_10523_uid189952	Deltaproteobacteria
Desulfococcus_oleovorans_Hxd3_uid58777	Deltaproteobacteria
Desulfotalea_psychrophila_LSv54_uid58153	Deltaproteobacteria
Desulfovibrio_gigas_DSM_1382_uid221293	Deltaproteobacteria
Desulfovibrio_magneticus_RS_1_uid59309	Deltaproteobacteria
Desulfovibrio_vulgaris__Miyazaki_F__uid59089	Deltaproteobacteria
Escherichia_blattae_DSM_4481_uid165043	Deltaproteobacteria
Escherichia_coli_K_12_substr__MG1655_uid57779	Gammaproteobacteria
Escherichia_coli_O157_H7_EC4115_uid59091	Gammaproteobacteria
Escherichia_fergusonii_ATCC_35469_uid59375	Gammaproteobacteria
Gallionella_capsiferriformans_ES_2_uid51505	Gammaproteobacteria

Proteobacteria Genomes	Class
Geobacter_sulfurreducens_PCA_uid57743	Betaproteobacteria
Helicobacter_cetorum_MIT_99_5656_uid162215	Deltaproteobacteria
Helicobacter_cinaedi_PAGU611_uid162219	Epsilonproteobacteria
Helicobacter_pylori_B38_uid59415	Epsilonproteobacteria
Helicobacter_pylori_SNT49_uid159615	Epsilonproteobacteria
Legionella_longbeachae_NSW150_uid46099	Epsilonproteobacteria
Legionella_pneumophila_ATCC_43290_uid86885	Gammaproteobacteria
Legionella_pneumophila_uid170534	Gammaproteobacteria
Methylocella_silvestris_BL2_uid59433	Gammaproteobacteria
Myxococcus_fulvus_124B02	Deltaproteobacteria
Neisseria_gonorrhoeae_FA_1090_uid57611	Betaproteobacteria
Neisseria_gonorrhoeae_TCDC_NG08107_uid161097	Betaproteobacteria
Neisseria_meningitidis_8013_uid161967	Betaproteobacteria
Neisseria_meningitidis_H44_76_uid162083	Betaproteobacteria
Nitrosomonas_AL212_uid55727	Betaproteobacteria
Nitrosomonas_europaea_ATCC_19718_uid57647	Betaproteobacteria
Pseudomonas_aeruginosa_B136_33_uid196598	Gammaproteobacteria
Pseudomonas_brassicacearum_NFM421_uid66303	Gammaproteobacteria
Pseudomonas_denitrificans_ATCC_13867_uid195459	Gammaproteobacteria
Pseudomonas_entomophila_L48_uid58639	Gammaproteobacteria
Pseudomonas_fluorescens_A506_uid165185	Gammaproteobacteria
Pseudomonas_fulva_12_X_uid67351	Gammaproteobacteria
Pseudomonas_mendocina_NK_01_uid66299	Gammaproteobacteria
Pseudomonas_monteilii_SB3078_uid232252	Gammaproteobacteria
Pseudomonas_ND6_uid167583	Gammaproteobacteria
Pseudomonas_poae_RE_1_1_14_uid188480	Gammaproteobacteria
Pseudomonas_putida_GB_1_uid58735	Gammaproteobacteria
Pseudomonas_resinovorans_NBRC_106553_uid208671	Gammaproteobacteria
Pseudomonas_stutzeri_A1501_uid58641	Gammaproteobacteria
Pseudomonas_syringae_B728a_uid57931	Gammaproteobacteria
Pseudomonas_TKP_uid232248	Gammaproteobacteria
Pseudomonas_VLB120_uid226717	Gammaproteobacteria
Rhodobacter_capsulatus_SB_1003_uid47509	Alphaproteobacteria
Rhodobacter_sphaeroides_ATCC_17025_uid58451	Alphaproteobacteria
Rhodobacter_sphaeroides_KD131_uid59277	Alphaproteobacteria
Rickettsia_canadensis_CA410_uid88063	Alphaproteobacteria

Proteobacteria Genomes	**Class**
Rickettsia_prowazekii_Breinl_uid196851	Alphaproteobacteria
Rickettsia_rickettsii__Sheila_Smith__uid58027	Alphaproteobacteria
Roseobacter_denitrificans_OCh_114_uid58597	Alphaproteobacteria
Roseobacter_litoralis_Och_149_uid54719	Alphaproteobacteria
Salmonella_bongori_NCTC_12419_uid70155	Gammaproteobacteria
Salmonella_enterica_arizonae_serovar_62_z4_z23__uid58191	Gammaproteobacteria
Salmonella_enterica_serovar_4_5_12_i__08_1736_uid212969	Gammaproteobacteria
Salmonella_enterica_serovar_Typhimurium_798_uid158047	Gammaproteobacteria
Salmonella_typhimurium_DT104_uid223287	Gammaproteobacteria
Shigella_boydii_CDC_3083_94_uid58415	Gammaproteobacteria
Shigella_dysenteriae_1617_uid229875	Gammaproteobacteria
Shigella_flexneri_2a_301_uid62907	Gammaproteobacteria
Shigella_sonnei_53G_uid84383	Gammaproteobacteria
Thiobacillus_denitrificans_ATCC_25259_uid58189	Gammaproteobacteria
Thiocystis_violascens_DSM_198_uid74025	Gammaproteobacteria
Thiomicrospira_crunogena_XCL_2_uid58183	Gammaproteobacteria
Vibrio_alginolyticus_NBRC_15630___ATCC_17749_uid199933	Gammaproteobacteria
Vibrio_anguillarum_775_uid68057	Gammaproteobacteria
Vibrio_cholerae_IEC224_uid89389	Gammaproteobacteria
Vibrio_EJY3_uid83161	Gammaproteobacteria
Vibrio_Ex25_uid41601	Gammaproteobacteria
Vibrio_fischeri_ES114_uid58163	Gammaproteobacteria
Vibrio_furnissii_NCTC_11218_uid82347	Gammaproteobacteria
Vibrio_harveyi_ATCC_BAA_1116_uid58957	Gammaproteobacteria
Vibrio_nigripulchritudo_SnF1_uid222819	Gammaproteobacteria
Vibrio_parahaemolyticus_BB22OP_uid184822	Gammaproteobacteria
Vibrio_splendidus_LGP32_uid59353	Gammaproteobacteria
Vibrio_vulnificus_MO6_24_O_uid62243	Gammaproteobacteria
Wolinella_succinogenes_DSM_1740_uid61591	Epsilonproteobacteria
Yersinia_enterocolitica_8081_uid57741	Gammaproteobacteria
Yersinia_pestis_A1122_uid158119	Gammaproteobacteria
Yersinia_pseudotuberculosis_PB1__uid59153	Gammaproteobacteria

Appendix 21 Additional *Clostrdia* genomes for 200 size genome set

Additional 200 Size Dataset Genomes	Class
Acetobacterium_woodii_DSM_1030	Clostridia
Butyrivibrio_Proteoclasticus_B316	Clostridia
Clostridiales_genomosp__BVAB3_str_UPII9_5	Clostridia
Ethanoligenens_harbinense_YUAN_3	Clostridia
Peptoclostridium_difficile_630	Clostridia
Roseburia_intestinalis_XB6B4	Clostridia
Ruminococcus_torques_L2_14	Clostridia
Symbiobacterium_thermophilum_IAM_14863	Clostridia
Thermincola_potens_JR	Clostridia

Appendix 22 ProPhylClust Pseudocode

Description: ProPhylClust is a script written in Python 2.7.x that uses a rooted phylogeny with or without multifurcations as a guide tree to cluster genes belonging to genomes that represent leaves in the tree. As the tree is traversed using post-order node traversal (children of an internal node are visited before their parents), genes from genomes are clustered together. Three types of clustering are performed: sequence versus sequence, sequence versus cluster, and cluster versus cluster. Individual sequences are clustered with other sequences typically during leaf versus leaf homology searches. Individual sequences are clustered to preexisting clusters, represented as a consensus sequence, typically during leaf versus internal node homology searches. Clusters are clustered together, typically during internal node versus internal node homology searches, where clusters are represented as consensus sequences.

Definitions:

Dereplication: leaf self-homology search to de-replicate sequences. Directed graph, and clustering using strongly connected components

HMM: Hidden Markov Model

Singleton sequence: Sequence without homolog

Optional Parameters:

-Sequence versus consensus sequence searches (SvsCS), homology searches between sequences and consensus sequences

-Sequence versus HMMs search (SvsHMM), sequence versus HMM searches

-Consensus sequence versus consensus sequence searches (CSvsCS), consensus sequence versus consensus sequence homology searches

-HMMs versus HMMs search (HMMvHMM), HMM versus HMM searches

Optional Parameter Combinations:

-SvsCS = all possible sequence versus consensus sequence homology searches

-SvsHMM = all possible searches sequence versus HMM homology searches

-SvsCS+SvsHMM = sequence versus consensus sequence filtered sequence versus HMM searches

-CSvsCS = all possible consensus sequence versus consensus sequence homology searches

-HMMvsHMM = all possible HMM versus HMM homology searches

-CSvsCS+HMMvsHMM = consensus sequence versus consensus filtered HMM versus HMM searches

Additional Information:

HMMs are used to create consensus sequences

No HMM vs HMM self searches

```
Start: Post-order node traversal: For each internal node $A$
    If children of $A$: count(Leaves, $l_n$) >= 1 and count(internal nodes, $n_n$) == 0
        All vs. All searches: $l_n$ vs. $l_n$
            Sequence dereplication
        Create undirected graph of sequences
        Cluster graph using connected components
        Create new clusters, alignments
        Output: clusters and alignments and singleton sequences to $A$

    If children of $A$: count(Leaves, $l_n$) >= 1 and count(internal nodes, $n_n$) >= 1
        for $l_i$ in $l_n$
            $l_i$ sequences vs. $n_n$ clusters: SvsCS or SvsHMM or SvsCS+SvsHMM
            Add sequences to cluster, create alignments
        Update clusters, alignments
        Output: clusters and singleton sequences to $A$
        If count(internal Nodes, $n_n$) == 1
            Homology search: All ($A$ & $n$) singleton vs. All ($A$ & $n$) singleton searches
                Sequence dereplication
            Create undirected graph of sequences
            Cluster graph using connected components
            Create new clusters, alignments
            Output: clusters and alignments and singleton sequences to $A$

            If $A$ == root node:
                For each cluster:
                    Append duplicate sequences
                    Move all clusters to root node
    If children of $A$: count(internal Nodes, $n_n$) >= 1
        Homology search: $n_n$ singleton vs. $n_n$ singleton searches
            Sequence dereplication
        Create undirected graph of sequences
        Cluster graph using connected components
        Create new clusters, alignments
        Output: clusters and alignments and singleton sequences to $A$

        If count(Leaves, $l_n$) == 0
            Homology search: All ($A$ & $n_n$) singleton vs. All ($A$ & $n_n$) singleton searches
                Sequence dereplication
            Create undirected graph of sequences
            Cluster graph using connected components
            Create new clusters
```

```
            Output: clusters and alignments and singleton sequences to $A$
        $A$ clusters vs $n_n$ clusters, $n_n$ clusters vs $A$ clusters, $n_n$ clusters vs $n_n$ clusters: CSvsCS or HMMvHMM or CSvsCS+HMMvHMM
            Create directed graph of clusters
            Cluster graph using strongly connected components
            Amalgamate clusters, create new alignemnts
            Output: clusters, alignments to $A$

    If $A$ is root node:
        Append duplicate sequences to clusters
        Create singleton file
```

Appendix 23 PhyloSubClust Pseudocode

Description: PhyloSubClust is a script written in Python 2.7.x to extract subtrees (i.e. clusters) from phylogenies. Pseudocode is commented denoted by '#'. Clusters can be either 'complete', with all genomes in the phylogeny represented in the cluster, or 'incomplete', where some of the genomes in the phylogeny are represented in the cluster. The starting node is the internal node with the shortest average patristic distance between descendant leaf nodes, ensuring sequences that should be clustered together are clustered. A root, the most distant internal node from the starting node, is then chosen. The tree is then traversed from the starting node, extracting leaves until the cluster is complete, meets a threshold number of represented genomes, or a previously extracted cluster was extracted. When a subtree is pruned (i.e. cluster is extracted), it is replaced with a marker node, the tree is re-rooted on the marker node, and clustering restarts on the node furthest by number of internal nodes from the root.

Definitions:

FEW: count of genomes in leaves descending from node is less than user specified minimum

MANY: count of genomes in leaves descending from node is greater than user specified minimum

node1: node with leaves with shortest average patristic distance

node2: parent of node1

Complete Cluster: cluster where genome count is equal to the number of genomes in phylogeny

Incomplete Cluster: cluster where genome count is less than number of genomes in phylogeny

mostDistant: furthest internal node from root by number of internal nodes

Start with a set *P*, of unrooted phylogenies *p*, each with or without multifurcations and more than 3 leaves and iterate through each.

```
Calculate patristic distances for all leaves, $l_n$ in $p$
$G$ = Count representative genomes in $p$

#get starting point
$s$ = parent node of most closely related sister leaves
$d$ = node of most distant to $s$
$SD$ = set of $sd$ pairs, if ties

*start PhyloSubClust for $p$ in $P$
root tree on $d$
node1 = $s$

#start extraction of subtrees
While count of leaves in $p > 0$:
    node2 = node1 parent node
    $cg1$ = count genomes node1
    $cg2$= count genomes node2 (without genomes unique to node1)
    $cg12$ = count of genomes in node2 (including node1)

    If stopper in node2
        Extract node1 as Incomplete Cluster
        Stopper for node1, reroot at node2
        node1 = parent node of most distant leaf from root
    else:
        if $cg1 == G$
            Extract node1 as Complete Cluster
            Stopper for node1, reroot at node2
            node1 = parent node of most distant leaf from root
        else if cg1< G
            if (cg1is not MANY) and (cg2is not MANY)
                if cg12 < G
                    node1 = node2
                else if cg12 > G
                    Extract node2 as Complete Cluster
                    Stopper node2, reroot at parent of node2
                    node1 = mostDistant
            if (cg1 is MANY) and (cg2 is MANY)
                Extract node1 as Incomplete Cluster
                Stopper for node1, reroot at node2
                node1 = mostDistant
```

In event of *SD*, repeat for each *sd* pair in *SD* for phylogeny *p* at ***, retain run with largest number of Complete Clusters

Appendix 24 Descriptive statitistics of cluster distributions.

PPC is ProPhylClust and PSC is PhyloSubClust.

Type	clusters	mean	median	q1	q3	min	max	singletons
PanPhyla20 diGraph 1e-30	4482	5.72	3	2	5	2	717	24441
PanPhyla20 undiGraph 1e-30	4948	4.92	2	2	4	2	444	25717
PanPhyla20 PPC 1e-10	4548	5.76	3	2	5	2	738	23892
PanPhyla20 PPC no HMMs 1e-10	6178	5.58	3	2	7	2	91	15590
PanPhyla20 RBH 1e-30	5396	4.82	3	2	5	2	66	24097
PanPhyla20 OrthoMCL 1e-5	3796	7.72	3	2	8	2	343	20785
PanPhyla20 OrthoMCL 1e-30	4864	5.68	3	2	6	2	147	22470
PanPhyla20 PSC 1e-10	4385	6.65	3	2	7	2	326	20933
PanPhyla20 PSC no HMMs 1e-10	5242	5.05	2	2	5	2	137	23619
PanPhyla40 diGraph 1e-30	10074	6.40	3	2	5	2	1340	45109
PanPhyla40 undiGraph 1e-30	11111	5.58	2	2	4	2	997	47587
PanPhyla40 PPC 1e-10	10129	6.48	3	2	5	2	1343	43947
PanPhyla40 PPC no HMMs 1e-10	12830	6.40	3	2	6	2	106	27387
PanPhyla40 RBH 1e-30	12362	5.26	3	2	4	2	106	44452
PanPhyla40 OrthoMCL 1e-5	7659	9.87	3	2	9	2	642	33962
PanPhyla40 OrthoMCL 1e-30	10418	6.74	3	2	6	2	255	39344
PanPhyla40 PSC 1e-10	8976	8.38	3	2	8	2	234	34317
PanPhyla40 PSC no HMMs 1e-10	11226	5.75	2	2	5	2	187	44986
PanPhyla60 diGraph 1e-30	14027	7.27	3	2	5	2	2033	65461
PanPhyla60 undiGraph 1e-30	15494	6.36	2	2	4	2	1678	68981
PanPhyla60 PPC 1e-10	14050	7.38	3	2	5	2	2062	63756
PanPhyla60 PPC no HMMs 1e-10	17630	7.33	3	2	6	2	174	38236
PanPhyla60 RBH 1e-30	17512	5.90	3	2	5	2	159	64061
PanPhyla60 OrthoMCL 1e-5	10510	10.84	3	2	8	2	1058	53480
PanPhyla60 OrthoMCL 1e-30	15046	7.20	3	2	6	2	239	59075
PanPhyla60 PSC 1e-10	12188	9.32	3	2	7	2	758	53926
PanPhyla60 PSC no HMMs 1e-10	16155	6.25	2	2	5	2	214	66479
PanPhyla80 diGraph 1e-30	18032	7.78	3	2	5	2	2894	83499
PanPhyla80 undiGraph 1e-30	19682	6.92	2	2	4	2	2451	87549

Type	clusters	mean	median	q1	q3	min	max	singletons
PanPhyla80 PPC 1e-10	18038	7.90	3	2	5	2	2903	81270
PanPhyla80 PPC no HMMs 1e-10	21850	8.07	3	2	6	2	240	47492
PanPhyla80 RBH 1e-30	22436	6.34	3	2	5	2	228	81435
PanPhyla80 OrthoMCL 1e-5	13242	11.75	3	2	7	2	1354	68200
PanPhyla80 OrthoMCL 1e-30	19466	7.61	3	2	6	2	391	76019
PanPhyla80 PSC 1e-10	21339	6.92	3	2	5	2	368	76019
PanPhyla80 PSC no HMMs 1e-10	15354	10.10	3	2	7	2	627	68698
PanPhyla100 diGraph 1e-30	22892	8.20	3	2	5	2	3782	98174
PanPhyla100 undiGraph 1e-30	25038	7.31	2	2	4	2	3461	102988
PanPhyla100 PPC 1e-10	22887	8.33	3	2	5	2	3792	95273
PanPhyla100 PPC no HMMs 1e-10	26903	8.59	3	2	6	2	313	54831
PanPhyla100 RBH 1e-30	28885	6.59	3	2	5	2	304	95752
PanPhyla100 OrthoMCL 1e-5	16079	12.92	3	2	8	2	1761	78232
PanPhyla100 OrthoMCL 1e-30	23951	8.29	3	2	6	2	457	87299
PanPhyla100 PSC 1e-10	19051	10.87	3	2	7	2	627	78912
PanPhyla100 PSC no HMMs 1e-10	26154	7.25	2	2	5	2	457	96269
Proteo20 diGraph 1e-30	8685	5.88	3	2	5	2	737	20887
Proteo20 undiGraph 1e-30	9330	5.36	3	2	5	2	452	21951
Proteo20 PPC 1e-10	8648	5.97	3	2	5	2	740	20302
Proteo20 PPC no HMMs 1e-10	10551	5.55	3	2	6	2	58	13364
Proteo20 RBH 1e-30	10613	4.79	3	2	5	2	51	21105
Proteo20 OrthoMCL	6395	8.89	3	2	8	2	679	15079
Proteo20 OrthoMCL 1e-30	8365	6.50	3	2	6	2	456	17600
Proteo20 PSC 1e-10	7732	7.32	3	2	8	2	415	15358
Proteo20 PSC no HMMs 1e-10	9361	5.38	2	2	6	2	415	21598
Proteo40 diGraph 1e-30	14122	8.13	3	2	6	2	1908	36757
Proteo40 undiGraph 1e-30	15322	7.34	3	2	6	2	1437	39158
Proteo40 PPC 1e-10	14079	8.23	3	2	6	2	1923	35736
Proteo40 PPC no HMMs 1e-10	16612	7.80	4	2	8	2	396	26499
Proteo40 RBH 1e-30	18154	6.44	3	2	6	2	260	39106
Proteo40 OrthoMCL 1e-5	9929	12.94	4	2	11	2	1412	27574
Proteo40 OrthoMCL 1e-30	13712	8.99	4	2	8	2	757	32725
Proteo40 PSC 1e-10	11818	10.83	4	2	10	2	453	28021

Type	**clusters**	**mean**	**median**	**q1**	**q3**	**min**	**max**	**singletons**
Proteo40 PSC no HMMs 1e-10	15297	7.41	2	2	7	2	503	42686
Proteo60 diGraph 1e-30	17428	9.91	3	2	7	2	2858	46367
Proteo60 undiGraph 1e-30	18764	9.06	3	2	7	2	2314	48979
Proteo60 PPC 1e-10	17360	10.02	3	2	7	2	2865	45058
Proteo60 PPC no HMMs 1e-10	20157	9.52	4	2	9	2	396	31679
Proteo60 RBH 1e-30	22445	7.81	4	2	8	2	260	48280
Proteo60 OrthoMCL 1e-5	12052	15.64	4	2	12	2	1633	34949
Proteo60 OrthoMCL 1e-30	16942	10.77	4	2	9	2	580	41112
Proteo60 PSC 1e-10	14621	12.86	4	2	10	2	522	35539
Proteo60 PSC no HMMs 1e-10	19207	9.00	3	2	8	2	503	50709
Proteo80 diGraph 1e-30	23179	10.90	3	2	8	2	4322	55424
Proteo80 undiGraph 1e-30	24789	10.08	3	2	7	2	3706	58213
Proteo80 PPC 1e-10	22988	11.06	3	2	8	2	4328	53850
Proteo80 PPC no HMMs 1e-10	25616	10.76	4	2	10	2	397	36991
Proteo80 RBH 1e-30	29551	8.66	3	2	8	2	260	56613
Proteo80 OrthoMCL 1e-5	16184	16.60	4	2	10	2	2285	43889
Proteo80 OrthoMCL 1e-30	22700	11.56	3	2	9	2	1047	50185
Proteo80 PSC 1e-10	19283	13.90	4	2	10	2	1296	44567
Proteo80 PSC no HMMs 1e-10	25431	9.89	2	2	8	2	554	61099
Proteo100 diGraph 1e-30	27170	12.01	3	2	8	2	5633	63963
Proteo100 undiGraph 1e-30	29026	11.14	3	2	7	2	4894	67062
Proteo100 PPC 1e-10	26956	12.18	3	2	8	2	5639	62127
Proteo100 PPC no HMMs 1e-10	29839	11.86	4	2	11	2	415	40807
Proteo100 RBH 1e-30	33759	9.71	4	2	9	2	340	66936
Proteo100 OrthoMCL 1e-5	18374	18.51	4	2	11	2	2400	54550
Proteo100 OrthoMCL 1e-30	26141	12.80	3	2	10	2	867	60184
Proteo100 PSC 1e-10	22001	15.43	4	2	11	2	2395	55322
Proteo100 PSC no HMMs 1e-10	29523	10.99	3	2	9	2	730	70174
200 diGraph 1e-30	44074	11.85	3	2	6	2	12043	155664
200 undiGraph 1e-30	43916	12.01	3	2	6	2	12632	150298
200 PPC 1e-10	30040	18.42	3	2	9	2	3584	124549
200 PPC no HMMs 1e-10	45722	11.81	3	2	8	2	943	137848
200 RBH 1e-30	47493	10.84	3	2	5	2	9270	162907
200 OrthoMCL 1e-5	46486	12.44	3	2	8	2	884	99598
200 OrthoMCL 1e-30	54363	9.30	3	2	6	2	552	151252

Type	clusters	mean	median	q1	q3	min	max	singletons
200 PSC 1e-10	36317	15.68	3	2	8	2	1517	125650
200 PSC no HMMs 1e-10	50632	10.33	2	2	7	2	689	154589

Appendix 25 558 *Clostridia* genomes and degree of completion

Organisms	Genome
Acetivibrio_cellulolyticus_CD2	draft
Acetobacterium_dehalogenans_DSM_11527	assembly
Acetobacterium_woodii_DSM_1030	complete
Acetohalobium_arabaticum_DSM_5501	complete
Alkaliphilus_metalliredigens_QYMF	complete
Alkaliphilus_oremlandii_OhILAs	complete
Alkaliphilus_transvaalensis_ATCC_700919	assembly
Anaerococcus_hydrogenalis_ACS_025_V_Sch4	draft
Anaerococcus_hydrogenalis_DSM_7454	draft
Anaerococcus_lactolyticus_ATCC_51172	draft
Anaerococcus_prevotii_ACS_065_V_Col13	draft
Anaerococcus_prevotii_DSM_20548	complete
Anaerococcus_sp__PH9	draft
Anaerococcus_tetradius_ATCC_35098	draft
Anaerococcus_vaginalis_ATCC_51170	draft
Anaerofustis_stercorihominis_DSM_17244	draft
Anaerostipes_caccae_DSM_14662	draft
Anaerostipes_hadrus_DSM_3319	draft
Anaerostipes_sp__3_2_56FAA	draft
Anaerotruncus_colihominis_DSM_17241	draft
Anaerotruncus_sp__G3_2012_	assembly
Blautia_hansenii_DSM_20583	draft
Blautia_hydrogenotrophica_DSM_10507	draft
Blautia_producta_ATCC_27340_DSM_2950	assembly
Blautia_wexlerae_AGR2146	assembly
Blautia_wexlerae_DSM_19850	assembly

Organisms	**Genome**
Butyricicoccus_pullicaecorum_1_2	assembly
Butyrivibrio_crossotus_DSM_2876	draft
Butyrivibrio_fibrisolvens_AB2020	assembly
Butyrivibrio_fibrisolvens_FE2007	assembly
Butyrivibrio_fibrisolvens_MD2001	assembly
Butyrivibrio_fibrisolvens_ND3005	assembly
Butyrivibrio_fibrisolvens_WTE3004	assembly
Butyrivibrio_fibrisolvens_YRB2005	assembly
Butyrivibrio_hungatei_NK4A153	assembly
Butyrivibrio_proteoclasticus_B316	complete
Butyrivibrio_proteoclasticus_FD2007	assembly
Butyrivibrio_proteoclasticus_P6B7	assembly
Butyrivibrio_sp__AC2005	assembly
Butyrivibrio_sp__AD3002	assembly
Butyrivibrio_sp__AE2005	assembly
Butyrivibrio_sp__AE2015	assembly
Butyrivibrio_sp__AE3003	assembly
Butyrivibrio_sp__AE3004	assembly
Butyrivibrio_sp__AE3006	assembly
Butyrivibrio_sp__AE3009	assembly
Butyrivibrio_sp__FC2001	assembly
Butyrivibrio_sp__FCS006	assembly
Butyrivibrio_sp__FCS014	assembly
Butyrivibrio_sp__LB2008	assembly
Butyrivibrio_sp__LC3010	assembly
Butyrivibrio_sp__MB2005	assembly
Butyrivibrio_sp__MC2013	assembly
Butyrivibrio_sp__MC2021	assembly
Butyrivibrio_sp__NC2007	assembly
Butyrivibrio_sp__NC3005	assembly

Organisms	Genome
Butyrivibrio_sp__VCB2001	assembly
Butyrivibrio_sp__VCB2006	assembly
Butyrivibrio_sp__VCD2006	assembly
Butyrivibrio_sp__WCD2001	assembly
Butyrivibrio_sp__WCD3002	assembly
Butyrivibrio_sp__XBB1001	assembly
Butyrivibrio_sp__XPD2002	assembly
Butyrivibrio_sp__XPD2006	assembly
Caldanaerobacter_subterraneus_subsp__yonseiensis_KB_1	assembly
Caldanaerobius_polysaccharolyticus_DSM_13641	assembly
Caldicellulosiruptor_acetigenus_DSM_7040	assembly
Caldicellulosiruptor_bescii_DSM_6725	complete
Caldicellulosiruptor_hydrothermalis_108	complete
Caldicellulosiruptor_kristjanssonii_I77R1B	complete
Caldicellulosiruptor_kronotskyensis_2002	complete
Caldicellulosiruptor_lactoaceticus_6A	complete
Caldicellulosiruptor_obsidiansis_OB47	complete
Caldicellulosiruptor_owensensis_OL	complete
Caldicellulosiruptor_saccharolyticus_DSM_8903	complete
Caldicoprobacter_oshimai_DSM_21659	assembly
Caloramator_sp__ALD01	assembly
Carboxydothermus_ferrireducens_DSM_11255	assembly
Carboxydothermus_hydrogenoformans_Z_2901	complete
Clostridiaceae_bacterium_L21_TH_D2	assembly
Clostridiales_bacterium_9401234	draft
Clostridiales_bacterium_NK3B98	assembly
Clostridiales_bacterium_VE202_01	assembly
Clostridiales_bacterium_VE202_03	assembly
Clostridiales_bacterium_VE202_04	assembly
Clostridiales_bacterium_VE202_06	assembly

Organisms	Genome
Clostridiales_bacterium_VE202_07	assembly
Clostridiales_bacterium_VE202_08	assembly
Clostridiales_bacterium_VE202_09	assembly
Clostridiales_bacterium_VE202_13	assembly
Clostridiales_bacterium_VE202_14	assembly
Clostridiales_bacterium_VE202_15	assembly
Clostridiales_bacterium_VE202_16	assembly
Clostridiales_bacterium_VE202_18	assembly
Clostridiales_bacterium_VE202_21	assembly
Clostridiales_bacterium_VE202_26	assembly
Clostridiales_bacterium_VE202_27	assembly
Clostridiales_bacterium_VE202_29	assembly
Clostridiales_bacterium_oral_taxon_876_str__F0540	assembly
Clostridiales_genomosp__BVAB3_str__UPII9_5	complete
Clostridiisalibacter_paucivorans_DSM_22131	assembly
Clostridium_acetobutylicum_ATCC_824	complete
Clostridium_acetobutylicum_DSM_1731	complete
Clostridium_acetobutylicum_EA_2018	complete
Clostridium_acidurici_9a	complete
Clostridium_akagii_DSM_12554	assembly
Clostridium_algidicarnis	assembly
Clostridium_arbusti_SL206	draft
Clostridium_asparagiforme_DSM_15981	draft
Clostridium_autoethanogenum_DSM_10061	complete
Clostridium_beijerinckii_G117	draft
Clostridium_beijerinckii_HUN142	assembly
Clostridium_beijerinckii_NCIMB_8052	complete
Clostridium_botulinum_A1_str__CFSAN002368	assembly
Clostridium_botulinum_A2_str__Kyoto	complete
Clostridium_botulinum_A3_str__Loch_Maree	complete

Organisms	Genome
Clostridium_botulinum_A_str__ATCC_19397	complete
Clostridium_botulinum_A_str__ATCC_3502	complete
Clostridium_botulinum_A_str__Hall	complete
Clostridium_botulinum_Af84	assembly
Clostridium_botulinum_B1_str__Okra	complete
Clostridium_botulinum_BKT015925	complete
Clostridium_botulinum_BKT028387	assembly
Clostridium_botulinum_B_str__Eklund_17B__NRP_	complete
Clostridium_botulinum_Ba4_str__657	complete
Clostridium_botulinum_Bf	draft
Clostridium_botulinum_CDC54075	assembly
Clostridium_botulinum_CDC54085	assembly
Clostridium_botulinum_CDC54088	assembly
Clostridium_botulinum_CDC66177	draft
Clostridium_botulinum_CFSAN002367	assembly
Clostridium_botulinum_CFSAN002369	assembly
Clostridium_botulinum_C_str__Eklund	draft
Clostridium_botulinum_C_str__Stockholm	draft
Clostridium_botulinum_D_str__1873	draft
Clostridium_botulinum_E1_str__'BoNT_E_Beluga'	draft
Clostridium_botulinum_E3_str__Alaska_E43	complete
Clostridium_botulinum_F_str__230613	complete
Clostridium_botulinum_F_str__Langeland	complete
Clostridium_botulinum_H04402_065	complete
Clostridium_botulinum_NCTC_2916	draft
Clostridium_botulinum_V891	assembly
Clostridium_botulinum_strain_CDC28023	assembly
Clostridium_botulinum_strain_CDC37457	assembly
Clostridium_botulinum_strain_CDC37461	assembly
Clostridium_botulinum_strain_CDC42961	assembly

Organisms	**Genome**
Clostridium_botulinum_strain_CDC48719	assembly
Clostridium_botulinum_strain_CDC48761	assembly
Clostridium_botulinum_strain_CDC52271	assembly
Clostridium_botulinum_strain_CDC52298	assembly
Clostridium_botulinum_strain_CDC66088	assembly
Clostridium_butyricum_5521	draft
Clostridium_butyricum_60E_3	assembly
Clostridium_butyricum_AGR2140	assembly
Clostridium_butyricum_DKU_01	assembly
Clostridium_butyricum_E4_str__BoNT_E_BL5262	draft
Clostridium_cadaveris_AGR2141	assembly
Clostridium_celatum_DSM_1785	draft
Clostridium_cellulolyticum_H10	complete
Clostridium_cellulovorans_743B	complete
Clostridium_cf__saccharolyticum_K10	complete
Clostridium_citroniae_WAL_17108	draft
Clostridium_clariflavum_DSM_19732	complete
Clostridium_clostridioforme_2_1_49FAA	draft
Clostridium_colicanis_209318	assembly
Clostridium_drakei	assembly
Clostridium_hathewayi_DSM_13479	draft
Clostridium_hathewayi_WAL_18680	draft
Clostridium_hiranonis_DSM_13275	draft
Clostridium_hydrogeniformans_DSM_21757	assembly
Clostridium_kluyveri_DSM_555	complete
Clostridium_kluyveri_NBRC_12016	complete
Clostridium_lentocellum_DSM_5427	complete
Clostridium_ljungdahlii_DSM_13528	complete
Clostridium_lundense_DSM_17049	assembly
Clostridium_methylpentosum_DSM_5476	draft

Organisms	**Genome**
Clostridium_novyi_NT	complete
Clostridium_papyrosolvens_DSM_2782	draft
Clostridium_paraputrificum_AGR2156	assembly
Clostridium_pasteurianum_BC1	complete
Clostridium_pasteurianum_DSM_525_ATCC_6013	assembly
Clostridium_pasteurianum_NRRL_B_598	assembly
Clostridium_perfringens_ATCC_13124	complete
Clostridium_perfringens_B_str__ATCC_3626	draft
Clostridium_perfringens_CPE_str__F4969	draft
Clostridium_perfringens_C_str__JGS1495	draft
Clostridium_perfringens_D_str__JGS1721	draft
Clostridium_perfringens_E_str__JGS1987	draft
Clostridium_perfringens_F262	draft
Clostridium_perfringens_JJC	assembly
Clostridium_perfringens_NCTC_8239	draft
Clostridium_perfringens_SM101	complete
Clostridium_perfringens_WAL_14572	draft
Clostridium_perfringens_str__13	complete
Clostridium_ramosum_DSM_1402	draft
Clostridium_saccharobutylicum_DSM_13864	complete
Clostridium_saccharolyticum_WM1	complete
Clostridium_saccharoperbutylacetonicum_N1_4_HMT_	complete
Clostridium_sartagoforme_AAU1	assembly
Clostridium_scatologenes	assembly
Clostridium_scindens_ATCC_35704	draft
Clostridium_sp__12_A_	assembly
Clostridium_sp__7_2_43FAA	draft
Clostridium_sp__7_3_54FAA	draft
Clostridium_sp__ASBs410	assembly
Clostridium_sp__ASF356	assembly

Organisms	**Genome**
Clostridium_sp__ASF502	assembly
Clostridium_sp__ATCC_29733	assembly
Clostridium_sp__ATCC_BAA_442	assembly
Clostridium_sp__Ade_TY	assembly
Clostridium_sp__BL8	assembly
Clostridium_sp__BNL1100	complete
Clostridium_sp__D5	draft
Clostridium_sp__DL_VIII	draft
Clostridium_sp__HGF2	draft
Clostridium_sp__JC122	draft
Clostridium_sp__KLE_1755	assembly
Clostridium_sp__KNHs209	assembly
Clostridium_sp__M62_1	draft
Clostridium_sp__MSTE9	draft
Clostridium_sp__Maddingley_MBC34_26	draft
Clostridium_sp__SS2_1	draft
Clostridium_sp__SY8519	complete
Clostridium_spiroforme_DSM_1552	draft
Clostridium_sporogenes_ATCC_15579	draft
Clostridium_sporogenes_PA_3679	draft
Clostridium_stercorarium_subsp__stercorarium_DSM_8532	complete
Clostridium_sticklandii	complete
Clostridium_symbiosum_WAL_14163	draft
Clostridium_symbiosum_WAL_14673	draft
Clostridium_termitidis_CT1112	draft
Clostridium_tetani_12124569	complete
Clostridium_tetani_E88	complete
Clostridium_tunisiense_TJ	draft
Clostridium_tyrobutyricum_DSM_2637_ATCC_25755_JCM_11008	assembly
Clostridium_tyrobutyricum_UC7086	draft

Organisms	Genome
Clostridium_ultunense_DSM_10521	assembly
Clostridium_ultunense_Esp	draft
Coprococcus_catus_GD_7	complete
Coprococcus_comes_ATCC_27758	draft
Coprococcus_sp__ART55_1	complete
Coprococcus_sp__HPP0048	assembly
Coprococcus_sp__HPP0074	assembly
Coprothermobacter_platensis_DSM_11748	draft
Coprothermobacter_proteolyticus_DSM_5265	complete
Dehalobacter_sp__CF	complete
Dehalobacter_sp__DCA	complete
Dehalobacter_sp__FTH1	draft
Desulfitibacter_alkalitolerans_DSM_16504	assembly
Desulfitobacterium_dehalogenans_ATCC_51507	complete
Desulfitobacterium_dichloroeliminans_LMG_P_21439	complete
Desulfitobacterium_hafniense_DCB_2	complete
Desulfitobacterium_hafniense_DP7	draft
Desulfitobacterium_hafniense_PCP_1	draft
Desulfitobacterium_hafniense_TCP_A	draft
Desulfitobacterium_hafniense_Y51	complete
Desulfitobacterium_metallireducens_DSM_15288	draft
Desulfitobacterium_sp__PCE1	assembly
Desulfosporosinus_acidiphilus_SJ4	complete
Desulfosporosinus_meridiei_DSM_13257	complete
Desulfosporosinus_orientis_DSM_765	complete
Desulfosporosinus_youngiae_DSM_17734	draft
Desulfotomaculum_acetoxidans_DSM_771	complete
Desulfotomaculum_alcoholivorax_DSM_16058	assembly
Desulfotomaculum_carboxydivorans_CO_1_SRB	complete
Desulfotomaculum_gibsoniae_DSM_7213	complete

Organisms	**Genome**
Desulfotomaculum_hydrothermale_Lam5_DSM_18033	draft
Desulfotomaculum_kuznetsovii_DSM_6115	complete
Desulfotomaculum_nigrificans_DSM_574	draft
Desulfotomaculum_reducens_MI_1	complete
Desulfotomaculum_ruminis_DSM_2154	complete
Desulfotomaculum_thermocisternum_DSM_10259	assembly
Desulfovirgula_thermocuniculi_DSM_16036	assembly
Desulfurispora_thermophila_DSM_16022	draft
Dethiobacter_alkaliphilus_AHT_1	draft
Dorea_formicigenerans_4_6_53AFAA	draft
Dorea_formicigenerans_ATCC_27755	draft
Dorea_longicatena_AGR2136	assembly
Dorea_sp__5_2	assembly
Dorea_sp__AGR2135	assembly
Epulopiscium_sp__'N_t__morphotype_B'	draft
Ethanoligenens_harbinense_YUAN_3	complete
Eubacterium_biforme_DSM_3989	draft
Eubacterium_brachy_ATCC_33089	assembly
Eubacterium_cellulosolvens_6	draft
Eubacterium_cylindroides_T2_87	complete
Eubacterium_eligens_ATCC_27750	complete
Eubacterium_hallii_DSM_3353	draft
Eubacterium_infirmum_F0142	draft
Eubacterium_limosum_KIST612	complete
Eubacterium_plexicaudatum_ASF492	assembly
Eubacterium_ramulus_ATCC_29099	assembly
Eubacterium_rectale_ATCC_33656	complete
Eubacterium_rectale_DSM_17629	complete
Eubacterium_rectale_M104_1	complete
Eubacterium_saphenum_ATCC_49989	draft

Organisms	**Genome**
Eubacterium_siraeum_DSM_15702	draft
Eubacterium_siraeum_V10Sc8a	complete
Eubacterium_sp__14_2	assembly
Eubacterium_sp__3_1_31	draft
Eubacterium_sp__AB3007	assembly
Eubacterium_xylanophilum_ATCC_35991	assembly
Eubacterium_yurii_subsp__margaretiae_ATCC_43715	draft
Faecalibacterium_cf__prausnitzii_KLE1255	draft
Faecalibacterium_prausnitzii_A2_165	draft
Faecalibacterium_prausnitzii_SL3_3	complete
Filifactor_alocis_ATCC_35896	complete
Finegoldia_magna_ACS_171_V_Col3	draft
Finegoldia_magna_ATCC_29328	complete
Finegoldia_magna_ATCC_53516	draft
Finegoldia_magna_BVS033A4	draft
Finegoldia_magna_SY403409CC001050417	draft
Flavonifractor_plautii_ATCC_29863	draft
Halanaerobium_hydrogeniformans	complete
Halanaerobium_praevalens_DSM_2228	complete
Halobacteroides_halobius_DSM_5150	complete
Halonatronum_saccharophilum_DSM_13868	assembly
Halothermothrix_orenii_H_168	complete
Helcococcus_kunzii_ATCC_51366	draft
Helcococcus_sueciensis_DSM_17243	assembly
Heliobacterium_modesticaldum_Ice1	complete
Intestinibacter_bartlettii_DSM_16795	assembly
Johnsonella_ignava_ATCC_51276	draft
Lachnoanaerobaculum_sp__ICM7	assembly
Lachnoanaerobaculum_sp__OBRC5_5	assembly
Lachnobacterium_bovis	assembly

Organisms	**Genome**
Lachnobacterium_bovis_C6A12	assembly
Lachnoclostridium_phytofermentans_ISDg	complete
Lachnoclostridium_phytofermentans_KNHs212	assembly
Lachnoclostridium_phytofermentans_KNHs2132	assembly
Lachnospira_multipara_ATCC_19207	assembly
Lachnospira_multipara_MC2003	assembly
Lachnospiraceae_bacterium_10_1	assembly
Lachnospiraceae_bacterium_1_1_57FAA	draft
Lachnospiraceae_bacterium_1_4_56FAA	draft
Lachnospiraceae_bacterium_28_4	assembly
Lachnospiraceae_bacterium_2_1_46FAA	draft
Lachnospiraceae_bacterium_2_1_58FAA	draft
Lachnospiraceae_bacterium_3_1	assembly
Lachnospiraceae_bacterium_3_1_46FAA	draft
Lachnospiraceae_bacterium_3_1_57FAA_CT1	draft
Lachnospiraceae_bacterium_3_2	assembly
Lachnospiraceae_bacterium_4_1_37FAA	draft
Lachnospiraceae_bacterium_5_1_57FAA	draft
Lachnospiraceae_bacterium_5_1_63FAA	draft
Lachnospiraceae_bacterium_6_1_63FAA	draft
Lachnospiraceae_bacterium_7_1_58FAA	draft
Lachnospiraceae_bacterium_8_1_57FAA	draft
Lachnospiraceae_bacterium_9_1_43BFAA	draft
Lachnospiraceae_bacterium_A2	assembly
Lachnospiraceae_bacterium_A4	assembly
Lachnospiraceae_bacterium_AB2028	assembly
Lachnospiraceae_bacterium_AC2012	assembly
Lachnospiraceae_bacterium_AC2031	assembly
Lachnospiraceae_bacterium_AC3007	assembly
Lachnospiraceae_bacterium_AD3010	assembly

Organisms	**Genome**
Lachnospiraceae_bacterium_COE1	assembly
Lachnospiraceae_bacterium_M18_1	assembly
Lachnospiraceae_bacterium_MD2004	assembly
Lachnospiraceae_bacterium_NC2004	assembly
Lachnospiraceae_bacterium_NC2008	assembly
Lachnospiraceae_bacterium_NK4A136	assembly
Lachnospiraceae_bacterium_NK4A144	assembly
Lachnospiraceae_bacterium_NK4A179	assembly
Lachnospiraceae_bacterium_P6B14	assembly
Lachnospiraceae_bacterium_V9D3004	assembly
Lachnospiraceae_bacterium_VE202_12	assembly
Lachnospiraceae_bacterium_VE202_23	assembly
Lachnospiraceae_bacterium_YSB2008	assembly
Lachnospiraceae_bacterium_oral_taxon_082_str__F0431	draft
Lachnospiraceae_oral_taxon_107_str__F0167	draft
Mahella_australiensis_50_1_BON	complete
Marvinbryantia_formatexigens_DSM_14469	assembly
Natranaerobius_thermophilus_JW_NM_WN_LF	complete
Orenia_marismortui_DSM_5156	draft
Oribacterium_asaccharolyticum_ACB7	assembly
Oribacterium_parvum_ACB1	assembly
Oribacterium_parvum_ACB8	assembly
Oribacterium_sinus_F0268	draft
Oribacterium_sp__ACB1	draft
Oribacterium_sp__NK2B42	assembly
Oribacterium_sp__oral_taxon_078_str__F0262	draft
Oribacterium_sp__oral_taxon_078_str__F0263	assembly
Oribacterium_sp__oral_taxon_108_str__F0425	draft
Oscillibacter_sp__1_3	assembly
Oscillibacter_sp__KLE_1728	assembly

Organisms	**Genome**
Oscillibacter_sp__KLE_1745	assembly
Oscillibacter_valericigenes_Sjm18_20	complete
Oscillospiraceae_bacterium_VE202_24	assembly
Parvimonas_micra_A293	assembly
Parvimonas_sp__oral_taxon_110_str__F0139	draft
Parvimonas_sp__oral_taxon_393_str__F0440	draft
Pelotomaculum_thermopropionicum_SI	complete
Peptoclostridium_difficile_002_P50_2011	assembly
Peptoclostridium_difficile_050_P50_2011	assembly
Peptoclostridium_difficile_5_3	assembly
Peptoclostridium_difficile_630	complete
Peptoclostridium_difficile_70_100_2010	assembly
Peptoclostridium_difficile_ATCC_43255	assembly
Peptoclostridium_difficile_ATCC_9689_DSM_1296	assembly
Peptoclostridium_difficile_BI1	complete
Peptoclostridium_difficile_CD13	assembly
Peptoclostridium_difficile_CD144	assembly
Peptoclostridium_difficile_CD196	complete
Peptoclostridium_difficile_CD22	assembly
Peptoclostridium_difficile_CD3	assembly
Peptoclostridium_difficile_CD37	assembly
Peptoclostridium_difficile_CD41	assembly
Peptoclostridium_difficile_CD70	assembly
Peptoclostridium_difficile_CD9	assembly
Peptoclostridium_difficile_CIP_107932	assembly
Peptoclostridium_difficile_DA00129	assembly
Peptoclostridium_difficile_DA00141	assembly
Peptoclostridium_difficile_DA00154	assembly
Peptoclostridium_difficile_DA00165	assembly
Peptoclostridium_difficile_DA00196	assembly

Organisms	Genome
Peptoclostridium_difficile_DA00197	assembly
Peptoclostridium_difficile_DA00203	assembly
Peptoclostridium_difficile_DA00215	assembly
Peptoclostridium_difficile_DA00216	assembly
Peptoclostridium_difficile_DA00244	assembly
Peptoclostridium_difficile_DA00261	assembly
Peptoclostridium_difficile_DA00305	assembly
Peptoclostridium_difficile_F152	assembly
Peptoclostridium_difficile_F314	assembly
Peptoclostridium_difficile_F548	assembly
Peptoclostridium_difficile_F601	assembly
Peptoclostridium_difficile_NAP07	assembly
Peptoclostridium_difficile_NAP08	assembly
Peptoclostridium_difficile_P13	assembly
Peptoclostridium_difficile_P15	assembly
Peptoclostridium_difficile_P19	assembly
Peptoclostridium_difficile_P23	assembly
Peptoclostridium_difficile_P25	assembly
Peptoclostridium_difficile_P28	assembly
Peptoclostridium_difficile_P30	assembly
Peptoclostridium_difficile_P31	assembly
Peptoclostridium_difficile_P33	assembly
Peptoclostridium_difficile_P37	assembly
Peptoclostridium_difficile_P38	assembly
Peptoclostridium_difficile_P42	assembly
Peptoclostridium_difficile_P45	assembly
Peptoclostridium_difficile_P46	assembly
Peptoclostridium_difficile_P49	assembly
Peptoclostridium_difficile_P50	assembly
Peptoclostridium_difficile_P53	assembly

Organisms	**Genome**
Peptoclostridium_difficile_P59	assembly
Peptoclostridium_difficile_P64	assembly
Peptoclostridium_difficile_P68	assembly
Peptoclostridium_difficile_P69	assembly
Peptoclostridium_difficile_P70	assembly
Peptoclostridium_difficile_P71	assembly
Peptoclostridium_difficile_P72	assembly
Peptoclostridium_difficile_P73	assembly
Peptoclostridium_difficile_P75	assembly
Peptoclostridium_difficile_P77	assembly
Peptoclostridium_difficile_QCD_23m63	assembly
Peptoclostridium_difficile_QCD_32g58	assembly
Peptoclostridium_difficile_QCD_37x79	assembly
Peptoclostridium_difficile_QCD_63q42	assembly
Peptoclostridium_difficile_QCD_66c26	assembly
Peptoclostridium_difficile_QCD_76w55	assembly
Peptoclostridium_difficile_QCD_97b34	assembly
Peptoclostridium_difficile_R20291	complete
Peptoclostridium_difficile_Y231	assembly
Peptoclostridium_difficile_Y343	assembly
Peptoniphilus_duerdenii_ATCC_BAA_1640	draft
Peptoniphilus_harei_ACS_146_V_Sch2b	draft
Peptoniphilus_indolicus_ATCC_29427	draft
Peptoniphilus_lacrimalis_DSM_7455	draft
Peptoniphilus_rhinitidis_1_13	draft
Peptoniphilus_senegalensis_JC140	assembly
Peptoniphilus_sp__BV3C26	assembly
Peptoniphilus_sp__oral_taxon_375_str__F0436	draft
Peptoniphilus_sp__oral_taxon_836_str__F0141	draft
Peptostreptococcaceae_bacterium_ACC19a	assembly

Organisms	Genome
Peptostreptococcaceae_bacterium_CM2	assembly
Peptostreptococcaceae_bacterium_CM5	assembly
Peptostreptococcaceae_bacterium_VA2	assembly
Peptostreptococcaceae_bacterium_oral_taxon_113_str__W5053	assembly
Peptostreptococcus_anaerobius_VPI_4330_DSM_2949	assembly
Proteocatella_sphenisci_DSM_23131	assembly
Pseudobutyrivibrio_ruminis_AD2017	assembly
Pseudobutyrivibrio_ruminis_CF1b	assembly
Pseudobutyrivibrio_sp__MD2005	assembly
Pseudoramibacter_alactolyticus_ATCC_23263	draft
Robinsoniella_sp__KNHs210	assembly
Roseburia_hominis_A2_183	complete
Roseburia_intestinalis_L1_82	draft
Roseburia_intestinalis_M50_1	complete
Roseburia_intestinalis_XB6B4	complete
Ruminiclostridium_thermocellum_AD2	assembly
Ruminiclostridium_thermocellum_ATCC_27405	complete
Ruminiclostridium_thermocellum_BC1	assembly
Ruminiclostridium_thermocellum_DSM_1313	complete
Ruminiclostridium_thermocellum_DSM_2360	assembly
Ruminiclostridium_thermocellum_JW20	assembly
Ruminiclostridium_thermocellum_YS	assembly
Ruminococcaceae_bacterium_AB4001	assembly
Ruminococcaceae_bacterium_AE2021	assembly
Ruminococcaceae_bacterium_D16	draft
Ruminococcus_albus_7_DSM_20455	complete
Ruminococcus_albus_8	draft
Ruminococcus_albus_AD2013	assembly
Ruminococcus_bromii_L2_63	complete
Ruminococcus_callidus_ATCC_27760	assembly

Organisms	**Genome**
Ruminococcus_champanellensis_18P13_JCM_17042	complete
Ruminococcus_flavefaciens_AE3010	assembly
Ruminococcus_flavefaciens_ATCC_19208	assembly
Ruminococcus_flavefaciens_FD_1	draft
Ruminococcus_flavefaciens_MA2007	assembly
Ruminococcus_flavefaciens_ND2009	assembly
Ruminococcus_gauvreauii_DSM_19829	assembly
Ruminococcus_lactaris_ATCC_29176	draft
Ruminococcus_lactaris_CC59_002D	assembly
Ruminococcus_obeum_A2_162	complete
Ruminococcus_sp__5_1_39BFAA	draft
Ruminococcus_sp__FC2018	assembly
Ruminococcus_sp__JC304	draft
Ruminococcus_sp__NK3A76	assembly
Ruminococcus_sp__SR1_5	complete
Ruminococcus_torques_L2_14	complete
Shuttleworthia_satelles_DSM_14600	draft
Stomatobaculum_longum	draft
Subdoligranulum_sp__4_3_54A2FAA	draft
Subdoligranulum_variabile_DSM_15176	draft
Symbiobacterium_thermophilum_IAM_14863	complete
Syntrophobotulus_glycolicus_DSM_8271	complete
Syntrophomonas_wolfei_subsp__wolfei_str__Goettingen_G311	complete
Syntrophothermus_lipocalidus_DSM_12680	complete
Tepidanaerobacter_acetatoxydans_Re1	complete
Terrisporobacter_glycolicus_ATCC_14880_DSM_1288	assembly
Thermacetogenium_phaeum_DSM_12270	complete
Thermincola_potens_JR	complete
Thermoanaerobacter_brockii_subsp__finnii_Ako_1	complete
Thermoanaerobacter_ethanolicus_CCSD1	draft

Organisms	**Genome**
Thermoanaerobacter_ethanolicus_JW_200	draft
Thermoanaerobacter_indiensis_BSB_33	assembly
Thermoanaerobacter_italicus_Ab9	complete
Thermoanaerobacter_mathranii_subsp__mathranii_str__A3	complete
Thermoanaerobacter_pseudethanolicus_ATCC_33223	complete
Thermoanaerobacter_siderophilus_SR4	draft
Thermoanaerobacter_sp__A7A	assembly
Thermoanaerobacter_sp__X513	complete
Thermoanaerobacter_sp__X514	complete
Thermoanaerobacter_sp__X561	draft
Thermoanaerobacter_tengcongensis_MB4	complete
Thermoanaerobacter_thermocopriae_JCM_7501	assembly
Thermoanaerobacter_thermohydrosulfuricus_WC1	assembly
Thermoanaerobacter_wiegelii_Rt8_B1	complete
Thermoanaerobacterium_saccharolyticum_JW_SL_YS485	assembly
Thermoanaerobacterium_thermosaccharolyticum_DSM_571	complete
Thermoanaerobacterium_thermosaccharolyticum_M0795	complete
Thermoanaerobacterium_xylanolyticum_LX_11	complete
Thermodesulfobium_narugense_DSM_14796	complete
Thermosediminibacter_oceani_DSM_16646	complete
Tyzzerella_nexilis_DSM_1787	assembly
Youngiibacter_fragilis_232_1	assembly

Appendix 26 Representative sequences used for identifying 18 clusters analysed.

Organism/Refseq GI	**Gene**	**Genome**
Peptoclostridium difficile		
126698240	tcdA	Peptoclostridium_difficile_630
126698238	tcdB	Peptoclostridium_difficile_630
126698237	tcdR	Peptoclostridium_difficile_630
126698241	tcdC	Peptoclostridium_difficile_630
126698239	tcdE	Peptoclostridium_difficile_630
260685955	tcdA	Peptoclostridium_difficile_R20291
260685953	tcdB	Peptoclostridium_difficile_R20291
260685954	tcdE	Peptoclostridium_difficile_R20291
260685953	tcdD	Peptoclostridium_difficile_R20291
Clostridium botulinum A		
153932893	ha70	Clostridium_botulinum_A_str_ATCC_19397
153933825	ha17	Clostridium_botulinum_A_str_ATCC_19397
153932677	ha33	Clostridium_botulinum_A_str_ATCC_19397
153931567	BoNT/A1	Clostridium_botulinum_A_str_ATCC_19397
153931687	botR	Clostridium_botulinum_A_str_ATCC_19397
153932404	ntnH	Clostridium_botulinum_A_str_ATCC_19397
169834787	ORF-X3	Clostridium_botulinum_A3_str_Loch_Maree
169835047	ORF-X2	Clostridium_botulinum_A3_str_Loch_Maree
169834940	ORF-X1	Clostridium_botulinum_A3_str_Loch_Maree
169834772	botR	Clostridium_botulinum_A3_str_Loch_Maree
169834914	ntnh	Clostridium_botulinum_A3_str_Loch_Maree

Organism/Refseq GI	**Gene**	**Genome**
170759234	toxin secretion/phage lysis holin	Clostridium_botulinum_A3_str_Loch_Maree
Clostridium botulinum B		
169834701	ntnh	Clostridium_botulinum_B1_str_Okra
169834581	botR	Clostridium_botulinum_B1_str_Okra
169834594	ha70	Clostridium_botulinum_B1_str_Okra
169834608	ha33	Clostridium_botulinum_B1_str_Okra
169834716	ha17	Clostridium_botulinum_B1_str_Okra
Clostridium botulinum C		
168188051	botR	
169338115	BoNT/C	Clostridium_botulinum_C_str_Eklund
168188047	ntnhC	Clostridium_botulinum_C_str_Eklund
168188048	ha33	Clostridium_botulinum_C_str_Eklund
168188049	ha17	Clostridium_botulinum_C_str_Eklund
168188050	ha70	Clostridium_botulinum_C_str_Eklund
168188051	botR	Clostridium_botulinum_C_str_Eklund
Clostridium botulinum E		
188588266	botR	Clostridium_botulinum_E3_str_Alaska_E43
188590132	BoNT/E3	Clostridium_botulinum_E3_str_Alaska_E43
188589186	ntnH	Clostridium_botulinum_E3_str_Alaska_E43
188587537	p47	Clostridium_botulinum_E3_str_Alaska_E43
188590332	ORF-X1	Clostridium_botulinum_E3_str_Alaska_E43
188589416	ORF-X2	Clostridium_botulinum_E3_str_Alaska_E43
188588926	ORF-X3	Clostridium_botulinum_E3_str_Alaska_E43
Clostridium botulinum F		

Organism/Refseq GI	**Gene**	**Genome**
384461191	ORF-X2	Clostridium_botulinum_F_str_230613
384461192	ORF-X1	Clostridium_botulinum_F_str_230613
384461193	BotR	Clostridium_botulinum_F_str_230613
384461194	p47	Clostridium_botulinum_F_str_230613
384461195	ntnH	Clostridium_botulinum_F_str_230613
384461196	BoNT/F	Clostridium_botulinum_F_str_230613
Clostridium tetani		
557606997	tetX	Clostridium_tetani_12124569
557606998	tetR_2	Clostridium_tetani_12124569
557604037	tetR_1	Clostridium_tetani_12124569
557606602	TR (transcription regulator)	Clostridium_tetani_12124569
28209852	tetR	Clostridium_tetani_E88
28212055	tetR	Clostridium_tetani_E88
28373188	tetX	Clostridium_tetani_E88
28373189	tetR	Clostridium_tetani_E88

Appendix 27 Copyright Permission Letter

April 18th 2018

Transfer of energy pathway genes in microbial enhanced biological phosphorus removal communities

BMC Genomics

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

Transfer of Energy Pathway Genes in Microbial Enhanced Biological Phosphorus Removal Communities. Dennis H.-J. Wong and Robert G. Beiko. BMC Genomics. 16:1-13. 2015.

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

[redacted]

Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: ROBERT BEIKO Title: PROFESSOR

Signature: [redacted] Date: 20 APR 2015