Computational Methods for Large Scale Analysis of Microbial Genome Evolution and Application to Antibiotic Resistant Pathogens

by

Ryan D. Crawford

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2022

Doctoral Committee:

Associate Professor Evan Snitkin, Chair
Associate Professor Peter Freddolino
Assistant Professor Jonathan Golob
Associate Professor Stephen Smith
Professor Jianzhi Zhang

Ryan D. Crawford
rcrawfo@umich.edu
ORCID iD: 0000-0002-6173-6092

DEDICATION

This dissertation is dedicated Joe Jennings.

ACKNOWLEDGMENTS

First I would like to thank Evan Snitkin for being such a great mentor for many years, and teaching me the joy of writing code from my first ”for loop”. Furthermore, I would like to thank all members of the Snitkin lab for creating such a great environment to develop learn and grow as a scientist.

I would like to thank my committee for their guidance.

Thanks to all my collaborators for teaching me so much, especially Eric Martens, Mary Hayden, Vince Young, Melanie Marketon, and Fouad Atouf.

Thanks to the Bioinformatics Training Grant and the Centers for Disease Control for funding.

Finally, I would like to thank my family for their love and support, and Steve and Tina Morris for being such great scientific mentors my entire career.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF APPENDICES

LIST OF ACRONYMS

AMR Antimicrobial Resistant

ARG Antibiotic Resistance Genes

CDC Centers for Disease Control

CRE carbapenem-resistant *Enterobacterales*

EIP Emerging Infections Program

ESBL Extended-spectrum β-lactamases

HGT Horizontal Gene Transfer

ICE Integrative and conjugative element

KPC *Klebsiella Pneumoniae* Carbapenemase

MSA Multiple sequence alignment

ML Maximum Likelihood

PATRIC Pathosystems Resource Integration Center

PUL Polysaccharide-Utilization Locus

ABSTRACT

Horizontal Gene Transfer (HGT) is a powerful force shaping microbial evolution [1]. This constant process by which genes are acquired into and excised from bacterial genomes enables an enormous capacity for rapid phenotypic evolution [2]. HGT enables the dissemination of clinically important genes, including antibiotic resistance genes, genes mediating virulence, environmental persistence genes, and metabolic genes. Acquisition of these genes potentiates phenotypic evolution in several important contexts: increasing the capacity for transmission, enhancing the ability for infection, limiting the efficacy of antibiotic therapies, and facilitating the metabolism of new substrates. Methods to characterize the pathways by which these genes spread through bacterial populations are critical for understanding the evolution of these phenotypes and their implications for public health [3]. In this dissertation, I develop a novel computational approach to generate core gene alignments for large numbers of bacterial genomes and implement two methods to characterize HGT events from bacterial whole-genome sequences. I then apply these methods to understand the dissemination of antimicrobial resistance genes and the evolution of carbohydrate utilization phenotypes in the microbiome.

First, we developed *cognac* (Core Gene Alignment Concatenation), an open-source R package for generating concatenated, core gene alignments for microbial genomes. *cognac* rapidly identifies shared phylogenetic marker genes, creates gene alignments, and concatenates them into a single alignment for downstream phylogenetic analysis. We demonstrate that this method can efficiently handle extremely large whole-genome sequencing datasets of diverse bacterial lineages.

Second, we sought to trace the spread of the KPC gene, a carbapenemase conferring broad-spectrum resistance to commonly used antibiotics for treating infections caused by Enterobacterales. Using comprehensive collections of clinical isolates from regional healthcare networks in three US states, we quantify the role of importation, clonal dissemination, and HGT on the total burden of KPC in these regions. To identify HGT events, we implemented a novel marker gene-based approach that enabled us to track KPC plasmid transfer using short-read data and identify HGT events occurring between circulating strains in the same region. Using this approach, we show that while the horizontal transfer of KPC frequently occurs in all three states, the strains and species involved and the overall contribution to the regional burden of KPC-carrying organisms differ substantially across the three states.

Third, we investigated the role of HGT in common members of the human gut microbiome. We developed a novel method to identify ancestral HGT loci by identifying core genes with significantly greater than expected divergence from the assigned species and greater similarity to the putative donor species. We then characterized HGT loci with conserved synteny and collinearity between donor and recipient species that have enabled pan-genome expansion and evolution of new phenotypes. This approach illustrates that HGT is common between two closely related species of *Bacteroides*, with many loci exhibiting evidence of HGT. These data, in conjunction with molecular data, provide insight into the breadth and complexity of metabolism in the microbiome and the underlying genomic events that enable the evolution of complex phenotypes.

In summary, this body of work establishes computational tools with broad application in computational genomics and genomic epidemiology: enabling phylogenetic analysis of large genomic datasets, identifying recent plasmid-mediated transfer occurring within and across regional healthcare networks, and identification of ancestral HGT loci carried on the chromosome mediating the development of complex phenotypes in the microbiome.

CHAPTER 1

Introduction

1.1 Motivation

HGT enables bacteria to adapt to environmental challenges rapidly. Understanding the mechanistic basis for these processes has important implications for human health. HGT is crucial to understanding the processes shaping microbial community structure and function in commensal microbes and understanding the driving forces shaping genome evolution in pathogenic microbes concerning antibiotic resistance and virulence [4, 2]. Understanding the mechanisms of HGT is especially critical for antibiotic resistance. Antibiotic resistance is a major threat to public health: in the United States, an estimated 2.8 million antibiotic-resistant infections result in an estimated 35,000 deaths annually [5]. The spread of antibiotic resistance genes in microbial communities amplifies the abundance of antibiotic-resistant isolates [6]. The spread of antibiotic resistance heightens the risk that resistance genes will be introduced into pathogenic lineages, resulting in limited treatment options in the event of infection [7]. This dissertation provides a framework for phylogenetic analysis of large whole-genome sequencing datasets and novel approaches to identify HGT in bacterial whole-genome sequencing data. I then apply these methods to track the regional spread of antibiotic resistance plasmids in pathogenic isolates and to examine genomic events underlying metabolic phenotypes in commensal microbes.

1.2 Mechanisms of HGT

Not only do bacterial genomes evolve via point mutations, but also via HGT: the exchange of genetic material between bacterial cells. First described in the 1940s, HGT has since gained widespread recognition for its importance in shaping microbial genome evolution [8]. Many bacterial genomes are highly mosaic, and considerable fractions of the genome are the product of multiple, independent HGT events [9]. Horizontal gene transfer typically happens via three main mechanisms: transformation, transduction, and conjugation, which all contribute to the evolution of bacterial populations. All three of these mechanisms work in concert to generate the mosaic nature of bacterial genomes and are all shared across diverse lineages of bacteria [10]. While most of these HGT events are thought to be neutral or deleterious, there is the potential for selective advantage and propagation within phylogenetically diverse bacterial populations.

1.2.1 HGT via Natural Transformation

Transformation is the process by which free DNA is bound at the cell surface, taken up into the cytoplasm, and recombined into the host genome [10]. The ability for natural transformation is widely distributed across the tree of life and represents a strategy for both nutrient acquisition, and genetic diversification [11]. This process involves the expression of distinct sets of genes required for natural competence and can result in the translocation of both chromosomal and plasmid DNA [10]. Typically, genes mediating natural competence are expressed in response to specific environmental conditions: cellular stress, changes in growth conditions, or starvation. Once DNA uptake occurs, recombination into the host genome is mediated by sequence similarity [12]. The recombination rate decreases substantially as sequence similarity in the donor sequence, and the host genome diverges, resulting in higher rates of recombination in closely related organisms [12, 10].

1.2.2 HGT via Transduction

Transduction is the DNA transfer between a donor and recipient cell via a bacteriophage intermediate [13]. HGT can occur via this process when host DNA is packaged into the bacteriophage capsid in place of or in addition to the bacteriophage genome. Upon release of the bacteriophage particles from an infected cell and infection of a subsequent cell occurs, incorporation of the phage-associated bacterial DNA into the chromosome of the newly infected cell can occur [10]. Bacteriophages are the most abundant organisms on the planet and are major contributors to genetic diversity in bacterial genomes [14]. Transduction has been shown to be a dominant source of genetic variation within strains of *E. coli* in microbial communities, generating genetic diversity at much higher rates relative to that of point mutations, illustrating the potential for rapid evolution via transduction in the microbiome [15].

1.2.3 HGT via Conjugation

Conjugative transfer is a mechanism of HGT in which a pore formation occurs between two cells, and DNA is transported through the pore from donor to recipient [9]. Whole chromosomes have the potential to be transferred via this mechanism; however, this process is rare due to the lengthy time required to transfer this quantity of DNA. Although, large genome transfer events have been observed in clinically relevant bacteria, potentially contributing to epidemic success [16].

Conjugative transfer systems are frequently associated with plasmids [17]. Because of their relatively small size, plasmids are readily transferred via this mechanism [9]. Many plasmids are self-transmissible, encoding the conjugative machinery necessary for transferring the plasmid between cells [18]. Once a plasmid is transferred between distinct cells, the conjugative machinery can then be expressed, and the plasmid can be further propagated. Fitness costs associated with plasmid carriage have been documented; however, this cost can be ameliorated by coevolution of the plasmid and host, resulting in stable relationships [19, 20].

Integrative Conjugal Elements (ICEs) constitute a second class of self-transmissible mobile genetic element [21]. ICEs exist in two states: integrated into the host's chromosome; and a conjugative state, whereby the element is existed from the chromosome to be further propagated to a new host [22]. ICEs carry genes for diverse functions, contributing to genome diversification within species and distributing genes across large phylogenetic distances [23, 22].

1.3 Methods to Detect Horizontal Gene Transfer in Bacterial Genomes

There are several common methods for inferring HGT in microbial genomes. First, phylogenetic approaches take common sequences and construct a phylogeny that can then be compared with the phylogeny of the corresponding genomes [24, 25]. HGT is then inferred by incongruencies between the two trees. Next, there are composition-based approaches that use the compositional structure of sequences to infer different origins [26]. These sequence signatures include: codon usage bias, differences in GC content, or differences in dinucleotide frequencies [27, 28, 29, 30]. This method is limited because the donor sequence and the recipient genome must be significantly divergent to exhibit measurable differences. Finally, an efficient and commonly used heuristic for examining HGT is comparing high identity sequences shared between distantly related organisms [31]. For example, identifying genes with >99% sequence similarity shared between shared genes with $<$ 97.5% similarity in the 16S gene sequence. This approach is limited, as it can only detect HGT between distinct genera. Because HGT is more likely to occur between closely related organisms, this approach is likely unable to detect a majority of HGT events [32].

HGT of plasmids, in particular, has been notoriously difficult to characterize: plasmids are highly recombinant, contain a multitude of gene cassettes that are not highly conserved, often have multiple incompatibility groups, and fusion plasmids resulting from the union of two distinct plasmids into a single molecule have been observed [33]. There has been enormous effort expended on sequencing of bacterial genomes by both short-read sequencing technology (reads typically ranging between 75-300 base-pairs) and increasingly by long-read sequencing technology (with average read lengths of 20,000-50,000 base-pairs) [34]. Short-read sequencing data resolves plasmids poorly due to the high prevalence of repetitive sequences [35]. Assemblies generated by this sequencing technology result in highly fragmented plasmid assemblies, frequently contaminated with chromosomal sequences. Several previous studies of HGT in the healthcare environment have relied on long-read sequencing of a collection of isolates or a subset of isolates to define plasmids and then characterize plasmid content in larger sets of short-read sequenced isolates [36, 37, 38, 39, 40]. This can be accomplished by mapping reads to these reference plasmids or using blast to compare sequence similarity and coverage. However, even with known plasmid sequences, the recombinogenic nature of plasmids makes it challenging to determine if an HGT event took place or represents a shared, conserved ancestral sequence [41]. Especially in the absence of long-read sequencing data and a known plasmid sequence, there are few options for identifying and characterizing plasmids in short-read data. Because short-read sequencing technology is, to this day, the most frequently used platform for microbial genome sequencing and comprises the overwhelming majority of isolates sequenced historically, these data present an underutilized resource for studying the epidemiology and evolution of plasmids.

1.4 Classes of β-lactam Resistance Phenotypes and Genotypes

The discovery of antibiotics revolutionized medicine; however, the rise of Antimicrobial Resistant (AMR) threatens these therapies' utility in clinical practice. AMR has important implications for individuals who suffer from these infections due to the associated increase in morbidity and mortality and for society due to the substantial economic cost to the healthcare system. Of perhaps greatest concern are the gram-negative enteric bacteria from the order *Enterobacterales*, for which there is a dearth of novel antibiotics under development to treat resistant infections effectively [5]. Members of *Enterobacterales* are the causative agent of various infections, including pneumonia, urinary tract infections, bloodstream infections, and intra-abdominal infections [42]. β-lactamases, enzymes with hydrolytic activity against antibiotics containing a β-lactam structural unit, primarily drive antibiotic resistance in *Enterobacterales* [43]. These enzymes convey resistance to antibiotics such as penicillins, cephalosporins, cephamycins, monobactams, and carbapenems. These agents prevent peptidoglycan synthesis by inhibiting penicillin-binding proteins, which results in lytic cell death. Additionally, these organisms frequently carry resistance determinants to other classes of antibiotics, complicating treatment [42].

With each generation of β-lactam antibiotic synthesized, the emergence of resistance has rapidly developed [44]. Carbapenems represent a last line of defense for treating infections caused by *Enterobacterales* [42]. Until recently, carbapenem-resistant *Enterobacterales* (CRE) were rarely observed; however, the widespread dissemination of carbapenemase genes — namely KPC — has increased the prevalence of these organisms. Historically, CRE has most frequently been associated with the healthcare setting, whereby patients requiring long-term care, antibiotic therapy, use of indwelling devices, and the presence of comorbidities have been associated with risk of infection [45, 46]. Recently, there has been an increasing prevalence of these organisms in the community due to highly transmissible AMR strains [46].

1.4.1 Chromosomal and Plasmid-spread AmpC

The AmpC β-lactamase was first identified in 1940 [47]. Several species of *Enterobacterales* carry this enzyme on the chromosome. Initial reports of CRE in the 1990s were attributed to the overproduction of AmpC β-lactamases in conjunction with porin mutations [48, 49]. Outer membrane porins, such as OmpK36 in *K. pneumoniae*, facilitate the transport of substrates, including β-lactam antibiotics across the outer membrane into the periplasm [50]. Mutations resulting in loss of function of these genes result in a less permeable cell membrane [50, 51]. The AmpC carbapenem resistance phenotype depends on two genetic factors: mutations resulting in overexpression of the AmpC gene and mutations in porin genes [50]. In this genetic context, the rate at which carbapenems are transported into the periplasm is sufficiently low that AmpC can effectively confer resistance — despite the low hydrolytic activity for these substrates [50].

While historically AmpC was most frequently observed on the chromosome, related alleles of this gene, closely related to the chromosomal variants, have been transposed onto plasmids [52]. These genes are associated with various mobile genetic elements and have been observed in many phylogenetic backgrounds [50]. Similar to chromosomal AmpC mediated carbapenem resistance, porin deficiency and the presence of AmpC are responsible for the resistance phenotype. Typically, these infections are still treatable with cefepime [53].

1.4.2 Extended-spectrum β-lactamases

Extended-spectrum β-lactamases (ESBL)s are enzymes that hydrolyze β-lactam antibiotics, including penicillins, broad-spectrum cephalosporins, and monobactams, and lack activity against cephamycins and carbapenems [54, 55]. The most frequently observed ESBLs are SHV and TEM types, which have been frequently observed in *Enterobacterales* but are most commonly associated with *E. coli* or *K. pneumoniae* [56]. In K. pneumoniae, SHV-type ESBLs are the most commonly observed on the chromosome, and in *E. coli*, ESBLs are most frequently carried on plasmids. While SHV and TEM are the most frequently observed ESBLs, there are currently over 12 different families of ESBLs with hundreds of variants of these enzymes observed [57].

Extended-spectrum β-lactam antibiotics were first used clinically in western Europe in the 1980s, and this is also where ESBL-producing organisms were first detected [54]. First detected in Germany in 1983, these genes were then rapidly detected in the United States and Europe [58, 59, 60]. Many outbreaks of ESBL-producing organisms have been described on every continent involving diverse species: most frequently *E. coli* or *K. pneumoniae*, but also species of Enterobacter, Citrobacter, and Pseudomonas, among others. [54].

Currently, carbapenems are the preferred agents to treat infections caused by ESBL producing *Enterobacterales* [54]. For example, imipenem, ertapenem, or faropenem have been shown to be effective at treating these infections [54]. As the prevalence of ESBL-producers increases, carbapenems are more frequently relied upon to treat these infections, which in turn provides a selective pressure for the evolution of carbapenem resistance [42]. ESBL-producing organisms can also exhibit the carbapenem-resistant phenotype by a similar mechanism to AmpC mediated carbapenem resistance: overproduction of an ESBL and membrane mutations to outer membrane porins [61].

1.4.3 Carbapenemases

Carbapenemases are the most frequently observed mechanism of carbapenem resistance, and the prevalence of these enzymes is increasing [46]. The first carbapenemases identified were species-specific and chromosomally encoded [62]. Carbapenemases were first identified in gram-positive bacilli; however, plasmid-encoded carbapenemases were identified in *Enterobacterales* in the 1980s. In the 1990s, several plasmid-encoded carbapenemases emerged in multiple species of gram-negatives. Since then, many carbapenemases have been disseminated widely, with examples including the Klebsiella pneumoniae carbapenemase KPC, New Delhi Metallo-β-lactamases (NDM), Verona Integrin-encoded Metallo-β-lactamase (VIM), Imipenemase (IMP), and OXA-48-like carbapenemases [62, 46]. Plasmid-encoded carbapenemases have reframed the problem of carbapenem resistance. What was once a problem of clonal dissemination of specific lineages is now a problem of several carbapenemases, which are readily disseminated throughout diverse bacterial species [62]. Some of these carbapenemase-producing organisms also produce AmpC or ESBLs, complicating treatment [42].

KPC is the most frequently observed carbapenemase in the United States and is widely spread across Europe [61]. KPC is particularly notable for the ability to mobilize into diverse lineages [2]. KPC was first identified in North Carolina in 1996 and, in the following years, was associated with outbreaks across the globe [32]. KPC is typically plasmid-associated and has been observed in the context of diverse plasmids [63]. Many of these plasmids have a broad host range, enabling the spread of KPC into diverse bacterial populations.

1.5 HGT in Beneficial Microbes

Just as HGT is a powerful force shaping the evolution of pathogenic microbes for antibiotic resistance and virulence, so too is HGT a driving force shaping the evolution of commensal microbes in the human microbiota [64, 65]. The human microbiome represents a complex ecosystem composed of microorganisms representing different kingdoms, which profoundly impact human physiology and disease [66]. Genome diversification is an essential adaptive strategy for many species, illustrated by the substantial variability in pan-genomes of gut commensals [2]. By enabling rapid phenotypic evolution, HGT can profoundly impact the structure and function of microbial communities. However, not all possible pathways of gene exchange are equally probable, and many factors govern the HGT network, including the phylogenetic distance between the donor and recipient strains, the environmental niche occupied, and the cellular function of the genes involved [3].

Large-scale analysis of genes identified as being horizontally transferred showed significant enrichment for specific gene functions [67, 64, 31]. In genes showing evidence of HGT, there is significant enrichment for functions related to metabolism and translation [67, 31]. HGT of these metabolic genes may enable adaptation to selective pressures related to diet, providing a competitive advantage for nutrient acquisition [67, 64]. In commensal microbes, HGT of antibiotic resistance genes has also been observed [31, 68]. Large plasmids can facilitate the simultaneous transfer of multiple antibiotic resistance genes, and additionally, genes enhancing pathogenicity may also be carried on these plasmids [68]. HGT of these genes into commensal lineages has the potential to generate novel multidrug resistance strains with enhanced potential for virulence [69].

HGT preferentially occurs among organisms that are closely related phylogenetically [70, 31]. Genetic factors shared between closely related organisms concerning the requisite molecular machinery for expression and maintenance of recently transferred genes, and a greater degree of sequence homology that promotes recombination into the recipient genome likely facilitate transfer [71, 72]. Although it is less common, HGT over vast phylogenetic distances has been observed [70, 73, 31]. For example, genomic analysis of Methanobrevibacter smithii, a common methanogenic archaeon found in the human gut, illustrated that 15% of the coding sequences in the genome were predicted to be of bacterial origin [73].

Organisms sharing an environmental niche are more likely to engage in HGT, and therefore specific mobile genetic elements are exchanged between organisms that share a similar environment [71, 64]. In an analysis of thousands of whole genome sequences, HGT between organisms isolated from humans was shown to be 25 times higher than organisms isolated from non-human sources [71]. Furthermore, this effect was amplified among organisms sharing the same body site, illustrating that specific molecular traits define ecological niches in the microbiome. Although less common, HGT between organisms that occupy very different environments has been known to occur. For example, members of the phylum Bacteroidetes have been observed carrying enzymes that degrade polysaccharides found in algae [74]. Phylogenetic analysis of these genes revealed that the likely source of these genes was HGT from seaweed-associated marine bacteria. Strains of Bacteroidetes carrying these enzymes are common in Japanese populations where seaweed is a common component of the diet and are notably absent in North American populations. This analysis illustrates that specific metabolic functions can disseminate to diverse isolates via HGT and spread broadly in human populations in response to selective pressures.

1.6 HGT of Antibiotic Resistance Genes

HGT of Antibiotic Resistance Genes (ARG) can occur via all three mechanisms described above. Environmental strains of the human pathogen *E. coli* have been shown to be naturally competent and capable of uptake of ARGs from the environment [75]. Additionally, phage are common sources of ARGs in the environment [76]. The prevalence of bacteriophage carrying ARGs has been shown to increase significantly in response to antibiotic treatment [77]. Finally, large conjugative plasmids and ICEs often harbor ARGs conveying resistance to multiple classes of antibiotics [19]. In addition to antibiotics, these elements frequently carry resistance to heavy metals and disinfectants, enabling survival in the hospital environment [20]. These plasmids and ICEs can spread rapidly within microbial communities in the human gut in both commensal and pathogenic isolates [32, 21].

While the evolution of antibiotic resistance can happen spontaneously in bacterial populations via spontaneous mutations, resistance to multiple antibiotics in a single bacterial lineage would take significant time to evolve via this mechanism [78]. HGT enables multiple ARGs conveying resistance to multiple classes of antibiotics to simultaneously be spread in bacterial populations, in place of the slow process of acquiring independent resistance mutations [7]. Therefore, understanding the HGT of ARGs is critical for addressing the problem of antibiotic resistance.

1.7 Epidemic Lineages of Antibiotic-Resistant Pathogens

Many multi-drug resistant pathogenic lineages have disseminated globally and are classified as a critical threat to public health by the Centers for Disease Control (CDC) [5]. By sequential acquisition of resistance to multiple antibiotic classes, some clonal lineages have become resistant to all or nearly all available antibiotics [63]. Many of these lineages have demonstrated substantial capacity for transmission and additionally may spread the resistance genes they carry throughout bacterial populations [79]. The most commonly observed species with epidemic potential are *Escherichia* *coli* and *Klebsiella pneumoniae*; however, species belonging to Enterobacter, Citrobacter, Serratia, and other genera are endemic to certain locations [80, 79, 81, 82, 83, 84].

E. coli is the cause of many infections, including urinary tract infections and bloodstream infections. In the early 1990s and 2000s, strains of *E. coli* producing ESBLs began to emerge. Most commonly, these strains harbored plasmid-born CTX-M enzymes. Molecular epidemiological analysis of isolates revealed multiple outbreaks across the UK, Canada, and the United States of the ST131 clonal lineage of *E. coli*, which all carried CTX-M-15 [82, 85]. Additionally, this strain is frequently resistant to fluoroquinolones due to chromosomal mutations in the gyrA and parC genes [85]. ST131 strains are commonly contracted in both the community and healthcare setting [82]. This clonal lineage accounts for a substantial proportion of bloodstream infections, whereby 23% of infection isolates collected in a three-year-long survey in San Francisco were identified as ST131 [86]. Most ST131 isolates are ESBL producers; however, many carbapenemase-producing isolates have been observed carrying several carbapenemases, including KPC, TEM, and NDM [85, 87, 88, 89].

Many lineages of multi-drug resistant *K. pneumoniae* have disseminated globally [90]. *K. pneumoniae* is notable for the ability to maintain multiple plasmids, which can carry many antibiotic resistance determinants [62, 7]. Many antibiotic resistance genes were first observed in *K. pneumoniae*, including the carbapenemase genes: KPC, OXA-48, and NDM, which were subsequently observed in other pathogens, [91]. In the United States, the expansion of carbapenemase-producing *Enterobacterales* was mainly driven by a single strain: *Klebsiella pneumoniae* ST258 [82]. ST258 has been disseminated globally and documented as the cause of a multitude of outbreaks across multiple continents [45]. This lineage is strongly associated with KPC and is frequently multidrug-resistant. KPC-producing *K. pneumoniae* is associated with high mortality rates estimated to be 40%, with higher rates reported in oncology patients [92].

While less frequently observed, there are several documented outbreaks of clonal lineages of carbapenemase-producing Enterobacter. KPC carrying *Enterobacter cloacae* ST171 was attributed as the cause of outbreaks in multiple states, including Minnesota, Michigan, North Dakota, Pennsylvania, and New York; and *E. cloacae* ST114 in North Carolina [93, 94, 95, 96, 97]. These outbreaks illustrate the risks of the spread of KPC into diverse bacterial lineages and adverse impacts on patient populations.

1.8 Dissertation Outline

This dissertation involves the implementation of methods for the analysis of large whole-genome sequencing datasets and the identification of HGT in bacterial genomes. In chapter two, I define an R package capable of generating alignments for phylogenetic analysis of extremely large whole-genome sequencing datasets. This software is applied to generate alignments for the projects in chapters three and four. In chapter three, I quantify the role of importation, transmission, and horizontal transfer of the KPC gene and associated plasmids across regional healthcare networks in three distinct US states. HGT events are identified using a novel phylogenetic approach to characterize the dissemination of KPC plasmids from short-read sequencing data. In chapter four, I implement a second approach to identify HGT occurring on the chromosome mediated by recombination in core genes. I also discuss how these genomic events shape phenotypes in two species of *Bacteroides*. In chapter five, I discuss the implications of these findings and future directions.

CHAPTER 2

cognac: Rapid Generation of Concatenated Gene Alignments for Phylogenetic Inference from Large, Bacterial Whole-genome Sequencing Datasets

2.1 Preamble

This chapter defines a novel method for generating concatenated gene alignments capable of application to large and diverse datasets of microbial whole-genome sequences. We published this work in BMC Bioinformatics: Crawford, R. D., & Snitkin, E. S. (2021). *cognac*: rapid generation of concatenated gene alignments for phylogenetic inference from large, bacterial whole-genome sequencing datasets. BMC bioinformatics, 22(1), 1-10.

2.2 Introduction

Phylogenetic analysis is becoming an increasingly integral aspect of biological research with applications in population genetics, molecular biology, structural biology, and epidemiology [98]. Generating a quality multiple sequence alignment Multiple sequence alignment (MSA) is fundamental to robust phylogenetic inference. MSA is a foundational tool in many disciplines of biology, which aims to capture the relationships between residues of related biological sequences and therefore facilitate insights into the evolutionary or structural relationships between the sequences in the alignment.

The first analysis incorporating genetic sequences to understand the evolutionary history of an organism was a sample of 11 *Drosophila melanogaster* Adh alleles in 1983 [99]. Since then, there has been a growing interest in using gene sequences to estimate the evolutionary relationships between organisms. However, it was quickly observed that individual gene trees are often inaccurate estimations of the species tree [100]. These incongruencies can arise from errors while building the tree or from biological processes such as incomplete lineage sorting, hidden paralogy, and horizontal gene transfer [101].

One approach for mitigating the incongruence between gene and species trees is the analysis of multiple genes at multiple loci concatenated into a supergene to generate more precise phylogenies [102, 103, 104, 105, 106]. This approach better leverages the large quantity of available data using multiple genes to substantially increase the number of variant sites and minimize the stochastic errors that may be associated with the limited information contained in a single gene [107]. This approach to infer the species tree has also been shown to be accurate under a range of simulated conditions, in spite of the biological processes, which might pose a challenge to accurate phylogenetic inference [108, 109].

Prior selection of a gene or set of genes for a given species is a commonly used strategy for selecting phylogenetic marker genes. The most commonly used marker gene for bacteria for phylogenetic analysis is the 16S rRNA gene [110]. This gene is ubiquitous in bacteria and archaea with highly conserved and variable regions, which makes it a useful marker for estimating the evolutionary relationships between prokaryotes. However, as a catalytic RNA, this gene evolves slowly relative to protein-coding genes, often resulting in few variant positions within a species [111]. Curated methods for selecting marker genes, such as multi-locus sequence typing, expand the number of marker genes for a given species and have led to improved resolution within a species [112]. However, this approach remains limited in that only a small number of curated genes are selected for a specific species, limiting its application to understudied organisms. Recently this concept has been expanded to include 400 marker genes commonly present in bacteria and archaea concatenated into a supergene for phylogenetic analysis of prokaryotes [113]. While these tools have many useful applications, relying on a limited number of predefined genes may limit the number of phylogenetically informative markers contained in a given dataset, which is important in situations where maximizing variation to distinguish closely related isolates is required.

In this work, we present *cognac* (core gene alignment concatenation), a novel data-driven method and rapid algorithm for identifying phylogenetic marker genes from whole-genome sequences and generating concatenated gene alignments, which scales to extremely large datasets of greater than 11,000 bacterial genomes. Our approach is robust when handling data sets with extremely diverse genomes and is capable of creating an alignment with large numbers of variants for phylogenetic inference.

2.3 Implementation

The inputs to *cognac* are fasta files and genome annotations in gff format, which can be obtained via commonly used programs such as, RAST, Prokka, or Prodigal 2.1 [114, 115, 116]. First, the

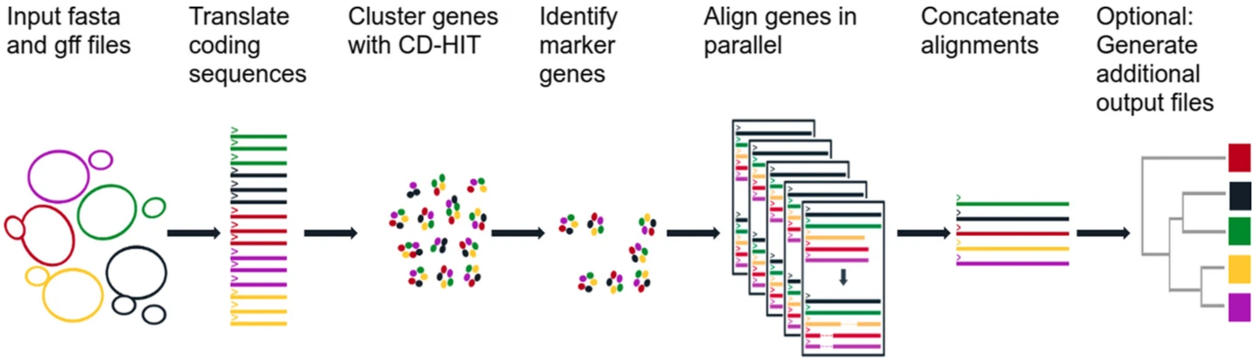

Input fasta and gff files
Translate coding sequences
Cluster genes with CD-HIT
Identify marker genes
Align genes in parallel
Concatenate alignments
Optional: Generate additional output files

Figure 2.1: Overview of the *cognac* algorithm. Whole-genome sequences and gene annotations are input, and the coding sequences are extracted and translated to return the amino acid sequences. The amino acid sequences are clustered to identify orthologous genes, and the single copy core genes are extracted from the dataset. For each core gene, unique alleles are identified and aligned, and the alignment is parsed to represent the aligned sequence for the full dataset. Alignments are then concatenated and are ready for downstream analysis

sequences corresponding to the coding genes are extracted using the coordinates provided in the gff file, and the nucleotide sequence for each gene is translated. To identify phylogenetic marker genes, CD-HIT is then used to cluster the amino acid sequences into clusters of orthologous genes (COGs) by their sequence similarity and length [117]. By default, COGs are defined at a minimum of 70% amino acid identity, and the alignment coverage for the longer sequence is 80% at minimum.

The CD-HIT output file is then parsed, and marker genes within the dataset are selected for inclusion in the alignment [117]. By default *cognac* identifies core genes to a given set of genomes; however, the selection criteria are customizable to allow for flexibility when creating alignments for various applications. The default selection criteria for selecting marker genes are: 1) present in 99% of genomes, 2) present in a single copy in 99.5% of genomes, and 3) ensuring that there is at least one variant position in the gene sequence. Allowing some degree of missingness allows for assembly errors that may arise in large datasets. We also allow the user to input a minimum number of genes to be included, and a minimum fraction of genes that are allowed to be missing, as genomes that don't share a sufficient number of phylogenetic markers may be problematic for some types of phylogenetic analysis and/or be indicative of problematic samples.

Once the marker genes are identified, the individual gene alignments of the amino acid sequences for each gene are generated with MAFFT [118]. Prior to alignment, redundant sequences of each gene are identified, and only the unique alleles are input to MAFFT. In particular, for each gene identified by CD-HIT, we first look for exact string matches within each gene cluster and select the representative unique alleles. The unique alleles are input to MAFFT, and the amino acid alignment is generated. The output gene alignment is then parsed, replicating the aligned sequence corresponding to each duplicated allele, generating the alignment for the entire set of alleles. Because MSA is computationally intensive, minimizing the number of sequences to align helps to reduce the associated computational overhead, leading to significantly reduced memory consumption and run-time.

Finally, the individual genes are concatenated into a single alignment to be used in downstream analysis. The alignment can then be input to commonly used programs for generating phylogenetic trees, such as RaXML or FastTree, to create a maximum likelihood Maximum Likelihood (ML) tree or approximate ML tree, respectively. We have included the ability to directly generate a neighbor-joining tree within the R package, to allow users to create a tree easily. *cognac* is well suited to generating alignments for extremely large datasets, and in these instances, the computational workload for ML based methods may be prohibitive, and therefore creating a neighbor-joining tree may present a good option. Neighbor-joining trees are a distance-based method that requires a much less computational overhead relative to ML based methods. While ML methods are likely to produce better results, the increased speed may be desirable for situations where a high degree of precision is not required.

Additionally, several optional outputs may be generated. We provide the functionality to generate a nucleotide alignment by mapping the corresponding codons to the amino acid alignment. We use gap placement in the amino acid alignment to position the corresponding codons from the nucleotide sequence of each gene, generating a codon-aware nucleotide alignment. This has the added benefit of increasing the number of variant positions in the alignment, which are a product of synonymous substitutions. This is potentially useful for applications where maximizing variation is key. We also provide functionality for parsing the alignments, including eliminating gap positions, removing non-variant positions, partitioning the alignment into the individual gene alignments, removing low-quality alignment positions, and creating distance matrices.

cognac was developed for R version 4.0.2. C++ code was integrated via the Rcpp package (version 1.0.3) and was written using the C++11 standard [119]. Multithreading is enabled in the C++ code via RcppParallel, which provides wrapper classes for R objects used by Intel Threading Building Blocks parallel computing library. Multithreading for R functions was enabled via the future.apply package (version 1.3.0). Functions for analysis of phylogenetic trees were enabled via the APE R package (version 5.3). [120].

2.4 Results

To demonstrate the utility of our tool, we created genus-level core gene alignments for 27,529 genomes from eight clinically relevant species of bacteria (Table 2.1). The number of genomes included from each genera had a wide range from 24 for Pluralibacter to 11,639 for Escherichia. *cognac* was run, requiring that at least 1000 genes that qualify as core genes included in the align-

Genus	Number of genomes	Total number of coding sequences	Number of core genes	Alignment length (amino acid residues)	Run time (min)	Memory usage (GB)
Citrobacter	262	1,356,975	1864	590,749	14.78	3.4
Enterobacter	1947	9,575,752	1671	551,522	105.81	37.39
Escherichia	11,639	61,042,774	1353	387,857	693.38	223.81
Klebsiella	9879	55,944,623	1957	631,196	980.95	184.86
Pluralibacter	24	131,798	1919	611,547	2.88	1.21
Proteus	207	806,518	1081	305,078	4.94	2.41
Pseudomonas	3051	19,509,251	1065	313,694	95.42	47.83
Serratia	520	2,673,835	1109	327,628	14.1	6.84

Table 2.1: Description of dataset and run statistics for the analysis in this study

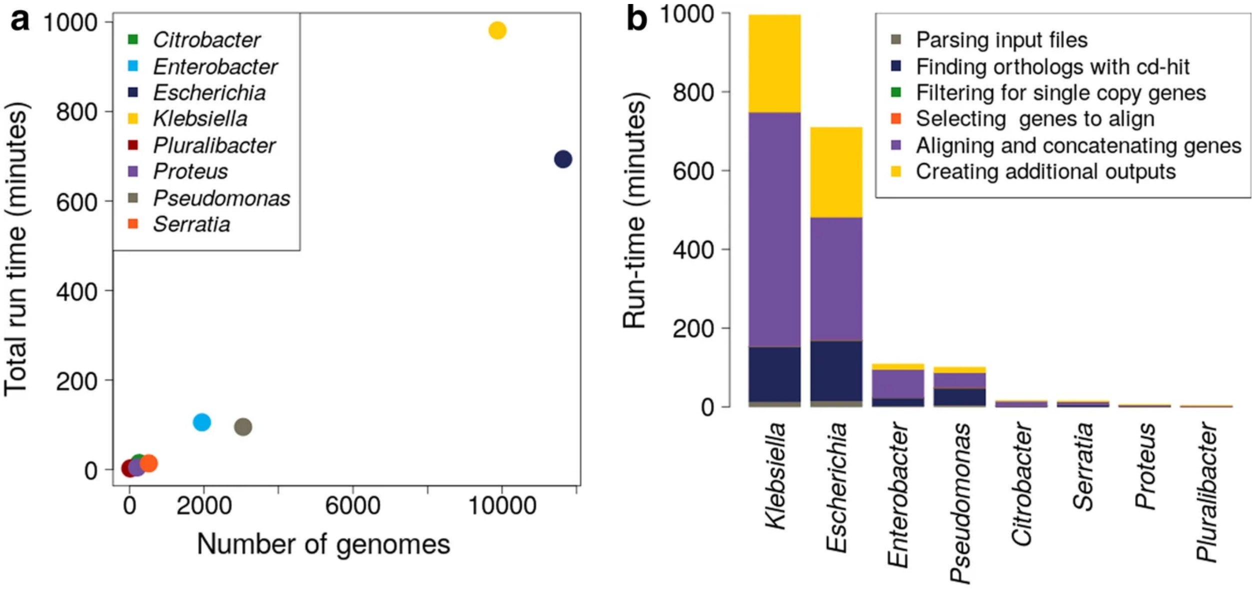


Figure 2.2: *cognac* is able to maintain reasonable run time even for very large datasets. All runs include: generating the amino acid alignment, mapping back to the nucleotide sequence, creating a distance matrix, and neighbor-joining tree. (A) For each genus, the run time was plotted against the number of genomes included in the analysis. (B) The composition of the run time by step

ment and genomes missing greater than 1% of core genes were removed. This was an extensive data set with the potential for inaccurate species assignment or assemblies to be of poor quality, ensuring that these genomes do not limit the number of core genes included. Additionally, for our test runs, we included the optional steps to generate the nucleotide alignment, create a pairwise single nucleotide variant distance matrix from the nucleotide alignment, and generate a neighbor-joining tree.

All runs finished in less than a day, and ranged from three minutes to 16 h and 21 min (Table 2.1). Run-time grew linearly as the number of genomes increased (Figure 2.2A). For all runs, with the exception of *Pseudomonas*, generating the MAFFT alignments was the largest portion of the total run-time (Figure 2.2B). The CD-HIT step was the highest fraction of runtime for *Pseudomonas* due to the larger genome size and the large degree of pan-genome diversity observed for this genus (Table 2.1).

To assess the magnitude of the reduction in the number of sequences that were aligned by selecting only the unique alleles of each gene, which is related to increased computational efficiency, we calculated the number of unique alleles per core gene as a fraction of the number of genomes (Table 2.1, Fig. 3a). We observed a strong inverse relationship between the number of genomes included and the number of unique alleles identified within the dataset (Fig. 2.3b). As a fraction of the number of genomes, *Klebsiella* had the lowest range of unique alleles with 0.02% (n=2) to 6.07% (n=600), with a median of 1.13% (n=112). *Pluralibacter* had the fewest genomes and had the highest proportion of unique alleles, with a maximum value of 79.9% unique alleles (n=19). This substantially decreases the number of sequences that need to be aligned, enabling *cognac* to scale to very large datasets. Because organisms are related genealogically, sequences in the genome are not independent, sharing a common ancestor. Therefore adding additional genomes does not necessarily expand the number of unique alleles for any genes, and all of the sequences may be represented by a substantially reduced subset of the number of samples.

We then wanted to analyze the effect of converting the amino acid alignments to nucleotide alignments with respect to amplifying the sequence diversity. The raw number of pairwise substitutions was calculated between all genomes from both the amino acid alignment and nucleotide normalized to the alignment length (Fig. 4). This greatly expanded the quantity of genetic variation contained in the alignment, although to different degrees for different datasets. This may reflect non-biological processes. For example, different data sets may have more diversity due to non-random sampling of the diversity within each genus. Additionally, the magnitude of the phylogenetic distances between isolates may not be uniform within different taxonomic assignments. However, biological factors may also play a role in the observed genetic distances. For example, the lowest amount of diversity was observed in *Pseudomonas*. The published mutation rate for E. coli is 2.5 times higher than that of *P. aeruginosa*, suggesting that the differences in diversity may be a function of the mutation rate in these organisms [121].

Because identifying core genes is central to *cognac* performing effectively, we tested the ability of our approach to rely on the input gene annotations and CD-HIT for clustering to identify a set of core genes accurately. We used the Panaroo pipeline to define gene content and benchmark the *cognac* algorithm. Panaroo has been proposed as a more rigorous approach for identifying gene families in prokaryotic genomes [122]. First, CD-HIT clusters genes by sequence identity, then a graph-based algorithm that identifies gene families by comparing the genomic context across the input genomes to more accurately identify genes related by vertical descent. This pipeline accounts for many common errors in genome annotation arising from multiple sources: errors in the assembly, fragmentation of the genomic sequence across multiple contigs, contamination, and the diversity within gene families. To compare the ability of *cognac's* default algorithm to identify core genes with that of the Panaroo pipeline, we used a real-world dataset of *Klebsiella*

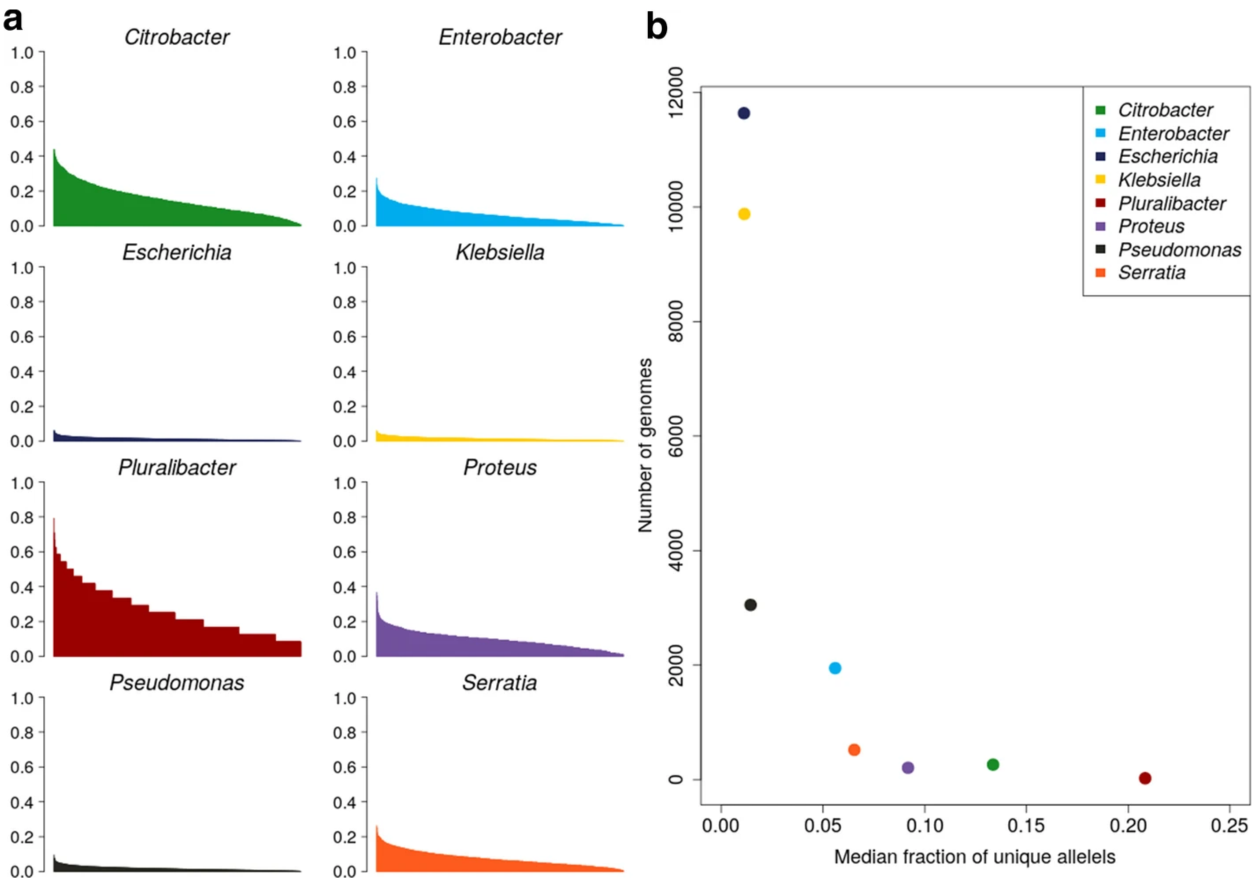


Figure 2.3: The fraction of unique alleles per gene is inversely proportional to the number of genomes in the dataset. (A) The distribution of unique alleles per core gene included in the alignment as a fraction of the number of genomes. (B) The relationship between the number of genomes and the median fraction of unique alleles for each gene

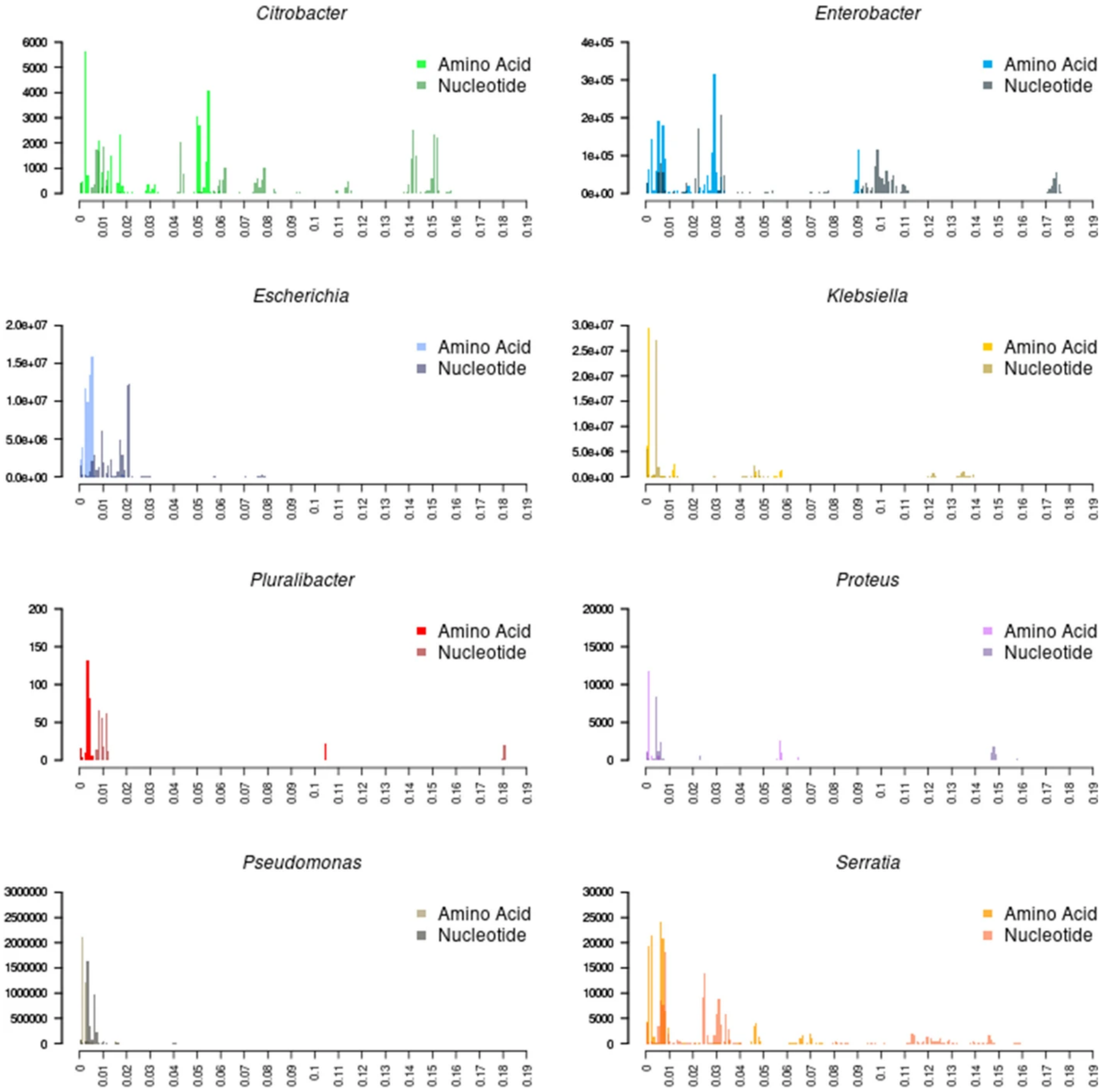


Figure 2.4: Mapping back to the nucleotide sequence of core gene alignments expands the number of variants for phylogenetic analysis. Pairwise distance matrices were constructed with the raw number of substitutions of the amino acid and nucleotide alignments. Histograms show the distribution of substitutions per position in the alignment. Lighter colors represent the amino acid distances, and darker colors represent the nucleotide distances

pneumoniae genomes collected from University Hospital in Ann Arbor, Michigan. This dataset constitutes a representative sampling of the species diversity in *Klebsiella pneumoniae*, containing 95 unique sequence types representing infection and colonization isolates.

We ran *cognac* and Panaroo with the default parameters. In total, *cognac's* CD-HIT command identified 3472 core genes. Of these core genes, 3332 exist as single copies in all input genomes and would be included in the output alignment. Panaroo identified 3808 core genes, and 3249 correspond to single-copy genes. There was a significant overlap between the two sets of core genes, with 2989 genes classified as core and single-copy genes by both methods. Two genes were classified as core by *cognac* and not by Panaroo. For each gene, the pairwise identity was calculated from the corresponding MAFFT alignment. Comparison of the genetic distances for these genes revealed that these genes were high identity: a minimum pairwise sequence identity of 97% and 96%, respectively, was observed. Panaroo defines orthologues by gene neighborhood, and these genes were transposed into a different locus in a subset of genomes and were classified into different gene clusters by Panaroo. However, given the high identity of these genes, it is likely that these genes are indeed a core orthologous group. Core genes classified by Panaroo and not *cognac* were primarily due to differences in sequence length. *cognac* uses a threshold of 80%; however, Panaroo uses no such restriction. In 282 cases (82.9%), Panaroo identified clusters with genes that differed in length by more than 80%. The median length differential between the longest and shortest sequence in these clusters was 30.7% as a fraction of the longer sequence, with a range of 4.32-79.8%. In the remaining 58 cases, gene clusters were erroneously split between multiple clusters by CD-HIT.

Three hundred thirty-eight genes were classified as multi-copy by Panaroo and single copy by *cognac*. In 336 cases, the differences in length above the 80% threshold could explain why at least one allele was not classified into the same cluster by CD-HIT. The percent difference as a fraction of the longer sequence ranged from 4.63-79.7%, with a median of 50.3%. In the remaining two cases, genes classified into a single cluster by Panaroo were divided into two separate clusters by CD-HIT separately, both of which were core genes.

The run-time for the *cognac's* marker gene detection, including running CD-HIT and parsing the results, was 2.65 minutes. The total run-time for Panaroo was 186.3 minutes. In this case, there are 336 additional core genes, corresponding to an increase of 9.68%. However, while Panaroo could identify a higher number of core genes, this came at the cost of a 70.3% increase in run-time. This analysis supports *cognac's* ability to rapidly identify a majority of core genes. Still, some sacrifice is made regarding speed in terms of missing a subset of core genes that more rigorous approaches may identify.

2.5 Discussion

We present a method to rapidly identify over 1000 marker genes and generate concatenated gene alignments that is capable of handling diverse bacterial genomes. Recently, we used this method to generate a core genome alignment and maximum likelihood tree for 52 genomes in the family *Bacteroidetes*, illustrating the utility of this tool to create gene trees over large phylogenetic distances [123]. Importantly, phylogenetically informative marker genes are selected using a data-driven approach, without any knowledge of the input genomes a priori, which allows for flexible selection of marker genes that are tailored to any input dataset.

Our approach relies fundamentally on amino acid sequence comparisons. Translation provides a natural compression algorithm, which has several advantages. First, the amino acid sequences have a third of the length of the corresponding nucleotide sequence. Because the length of the input sequences is a major contributor to the computational complexity of MSA, this reduction in length significantly improves performance and scalability [124, 125]. Additionally, amino acid sequences have a higher degree of conservation relative to nucleotide sequences [126]. This enables us to leverage redundancy in the codon code to identify orthologous genes and generate more accurate alignments. This enables a more robust and rapid approach for identifying and aligning orthologous genes, especially when applied to phylogenetically diverse datasets.

When performing computationally intensive procedures, amino acid sequences have many advantages; however, nucleotide alignments may be preferable for some applications. To address this, we provide the optional functionality to map the corresponding codons back to the amino acid alignment to return the nucleotide alignment. This can substantially increase the sequence variation contained in the alignment, which may be useful for applications where it is important to distinguish between closely related isolates. Additionally, we leverage the information contained in the amino acid sequences to produce a codon-aware alignment. This allows for greater accuracy in the placement of functional residues within the gene sequence and reduces the potential for misalignment of codons that may occur when aligning nucleotide sequences.

An important feature of our algorithm is that it relies only on annotated whole-genome assemblies, which provides several advantages over commonly used techniques of aligning raw sequencing reads to a reference genome. First, with respect to the size of the files, assemblies are a small fraction of the files containing the raw sequencing data. Second, *cognac* does not require selection of a reference genome. Different choices of reference genome have been shown to have large influences on the quality of the output alignment, potentially amplifying the frequency of mapping errors [127]. Additionally, the mapping accuracy is severely compromised when considering diverse datasets, even within a species. This limits the application of this method to diverse datasets. Finally, since our approach relies on assemblies, this enables us to analyze genomes sequenced on different platforms, allowing for increased sample size. For this study, we used only high-quality assemblies. Low-quality assemblies may be missing substantial numbers of genes, limiting the number of core genes identified for a given dataset, especially for small numbers of genomes.

Other assembly-based methods for estimating the genomic distance between genomes use dimensionality reduction techniques such as k-mers or the MinHash algorithm to estimate the distance between genomes [128, 129]. These methods have the advantage that they can leverage non-coding regions as a source of additional variation; however, the natural structure of the data is lost. Our method not only allows for an estimation of the genetic distances between isolates but also produces an alignment that can be used in downstream applications. This has the potential to leverage the alignment to identify recombinogenic genes and has the potential for use in gaining biological insights into molecular evolution.

Our algorithm was able to scale to extremely large datasets. For a data set of 11,639 *Escherichia* genomes, we were able to generate a neighbor-joining tree from a nucleotide concatenated gene alignment in less than 17 h. This is accomplished by reducing the computational overhead of MSA in two ways: (1) translating the sequences, effectively reducing their length, and (2) reducing the number of sequences by only aligning unique alleles. For extremely large datasets, this results in an approximately 99% reduction in the number of sequences that need to be aligned, allowing for great improvements in scalability, and allowing for application to extremely large datasets.

2.6 Conclusions

In summary, *cognac* is a robust, rapid method for generating concatenated gene alignments that scales to extremely large datasets. Our method uses a data-driven approach for the identification of phylogenetic markers, which are efficiently aligned and concatenated into a single alignment for downstream phylogenetic analysis. The pipeline is open source and freely available as an R package. We expect our tool will be generally useful for many different types of analysis and will enable evolutionary insights in a broad range of applications.

2.7 Availability of Data and Materials

Genomes for this study were downloaded from the Pathosystems Resource Integration Center (PATRIC) [130], and are available from https://www.patricbrc.org/. All available genomes from the genera of interest available as of 06/01/2020 that were isolated from humans and met the criteria for good quality were downloaded from the PATRIC FTP server. Quality was assessed for completeness, contamination, coarse consistency, and fine consistency via the CheckM algorithm within the PATRIC genome annotation service [131, 132]. Additional genomes used in this study were collected as part of a longitudinal study of carbapenem-resistant organisms and are available from RefSeq under BioProject PRJNA603790, and PRJNA690239 [133]. All genome annotations were generated with RAST [114]. *cognac* source code is available at https://github.com/rdcrawford/cognac. Scripts used in benchmarking are available at https://github.com/rdcrawford/cognac_paper. Additionally, a docker image is available at https://hub.docker.com/repository/docker/rdcrawford/cognac.

CHAPTER 3

A Comparative Analysis of Regional Horizontal Gene Transfer Across Three Distinct US States

3.1 Introduction

CRE represents a critical antibiotic resistance threat due to their high rates of multi-drug resistance, ability to transmit effectively in healthcare settings, and their association with high rates of morbidity and mortality in infected patients [5]. While there are epidemic strains of CRE that dominate in certain regions, overall CRE is extremely diverse with respect to both species and strains [134, 135, 136, 71]. This diversity has been driven by the frequent horizontal transfer of carbapenemase enzymes within and between *Enterobacterales* species, with commonly observed species of CRE including *Klebsiella pneumoniae, Escherichia coli, Citrobacter freundii, and Enterobacter cloacae* [63]. Moreover, for each of these species, many strains are commonly observed in healthcare settings, with new prominent strains constantly emerging [137, 36]. Thus, preventing CRE infections requires not just preventing the clonal spread of known CRE threats but also detecting the emergence and spread of new CRE strains with epidemic potential. Critical to containing new strains of CRE is an improved understanding of the pathways by which mobile elements harboring carbapenemase enzymes are spread, and the factors determining whether a strain acquiring a carbapenemase will go on to propagate within a healthcare facility and across a region.

Whole-genome sequencing has proved transformative for our ability to track the transmission pathways of antibiotic-resistant organisms [138]. However, progress in tracking the transfer of carbapenemases, and other ARGs, has lagged behind due to technical challenges associated with characterizing the mobile elements carrying them [35]. While ARGs are most often carried on plasmids, an individual ARG can be present in many different plasmid contexts, and plasmids can harbor a variable set of different ARGs [6]. Making things even more complex is that even within lineages stably associated with carbapenemases, the associated transposon can jump into different strain-associated plasmids [33]. This lack of stability of ARGs in the context of the plasmids that disseminate them can make it challenging to deduce their pathways of transfer between different strains and species. Thus, past studies tracking the specific pathways of ARGs transfer have been largely restricted to transfer events within patients or hospitals, where the transfer is caught in action, or to cases where the transfer is to a recipient lineage that is typically not associated with the ARGs [37]. However, most transfer events likely occur in the gut of asymptomatically colonized patients, thereby not being directly detected, and multiple transfer events into individual lineages are likely commonplace [139, 3]. Thus, effectively monitoring the emergence and spread of resistance threats more comprehensively will require new genomic analysis strategies.

In this study, we sought to trace the importation, transmission, and horizontal transfer of the KPC gene across regional healthcare networks in three US states by performing genomic analysis on comprehensive collections of regional KPC-harboring CRE strains. To enable this, we developed a novel marker gene-based approach that enabled us to track KPC plasmid transfer using short read sequencing data and identify transfer events occurring between circulating strains in the same region. Using this approach, we show that while the horizontal transfer of KPC frequently occurs in all three states, the strains and species involved, as well as the overall contribution to regional CRE burden, differ substantially across the three states.

3.2 Results

3.2.1 Using Phylogenetically Informed Marker Genes to Monitor the Presence of Plasmids

We developed a novel phylogenetic approach to track the spread of KPC among different CRE species and strains across three US states. While plasmids harboring KPC, and other ARGs, can be highly variable in gene content, we hypothesized that there may be defined sets of core plasmid genes that can serve as a reliable marker of plasmid presence and thereby be used to track plasmid transfer into different bacterial lineages. To identify these marker genes, we extracted genes from 328 complete KPC plasmids and identified clusters of genes that were gained and lost across the phylogeny of 1823 Enterobacterales genomes in a highly correlated manner. To test the accuracy of our approach to identifying genes that travel together, we used a set of publicly available complete genomes and evaluated whether genes grouped into marker gene clusters were found to co-occur on the same plasmids (Figure 3.1A). In total, the precision was very high, with 96.6% of cluster genes located on the same contig in complete genomes (Figure S3). Across all marker gene clusters predicted to be present in complete genomes, 84.0% of clusters had 100% precision (Figure S3A). Thus, the identified clusters of genes were reliable markers for the presence of a plasmid.

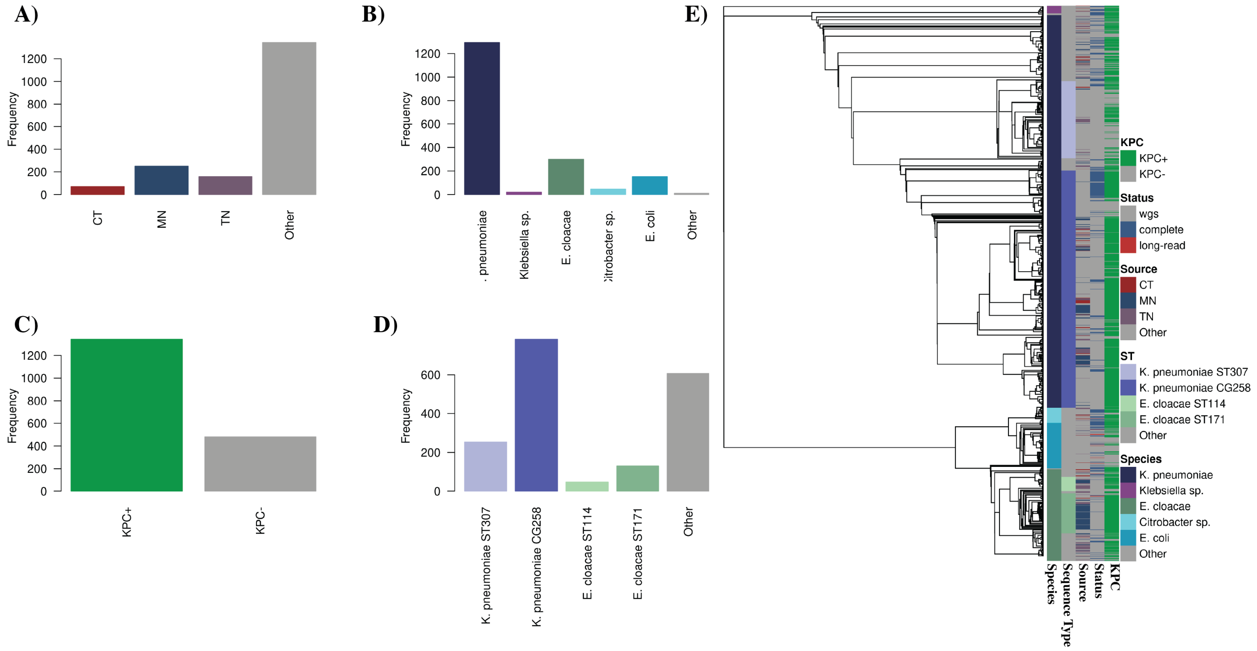


Figure 3.1: Description of genomic data used in this study. Counts of the isolates included with respect to: (A) source of collection from the three EIP collection states and public sources, (B) species, (C) KPC status, and (D) sequence type. (E) Approximate maximum likelihood tree of core genes shared across the isolates included in the analysis, rooted at the midpoint.

3.2.2 Identifying Marker Gene Sets Linked to the Gain and Loss of the KPC gene

While all of the genes in this analysis had been observed on a KPC-carrying plasmid previously, plasmids differ with respect to the strength of their association with KPC (Figure S3E). To focus our analysis on gene clusters which are reliable markers of KPC transfer, we only included clusters with enrichment in KPC co-transitions (i.e., KPC gain/loss occurring on the same branch of the phylogeny and in the same direction as cluster gain/loss (Figure S4)). This resulted in 35 marker gene clusters with a strong association with KPC; 92.28% of the KPC+ isolates from the current study were assigned at least one cluster, with a median of two clusters assigned per genome (Figure 3.2B). The presence of multiple clusters could reflect plasmid fusion or strains harboring multiple plasmids associated with KPC and transposon-mediated hopping between them.

Next, we examined the functions of genes assigned to the KPC-associated gene clusters to understand what makes them good markers (Figure S5). While most genes lacked annotation, most of the annotated genes assigned to clusters were assigned to essential plasmid functions, including conjugation and DNA processing enzymes. In contrast, these marker gene clusters rarely included plasmid cargo, such as AMR genes. This functional analysis supports our approach, having iden-

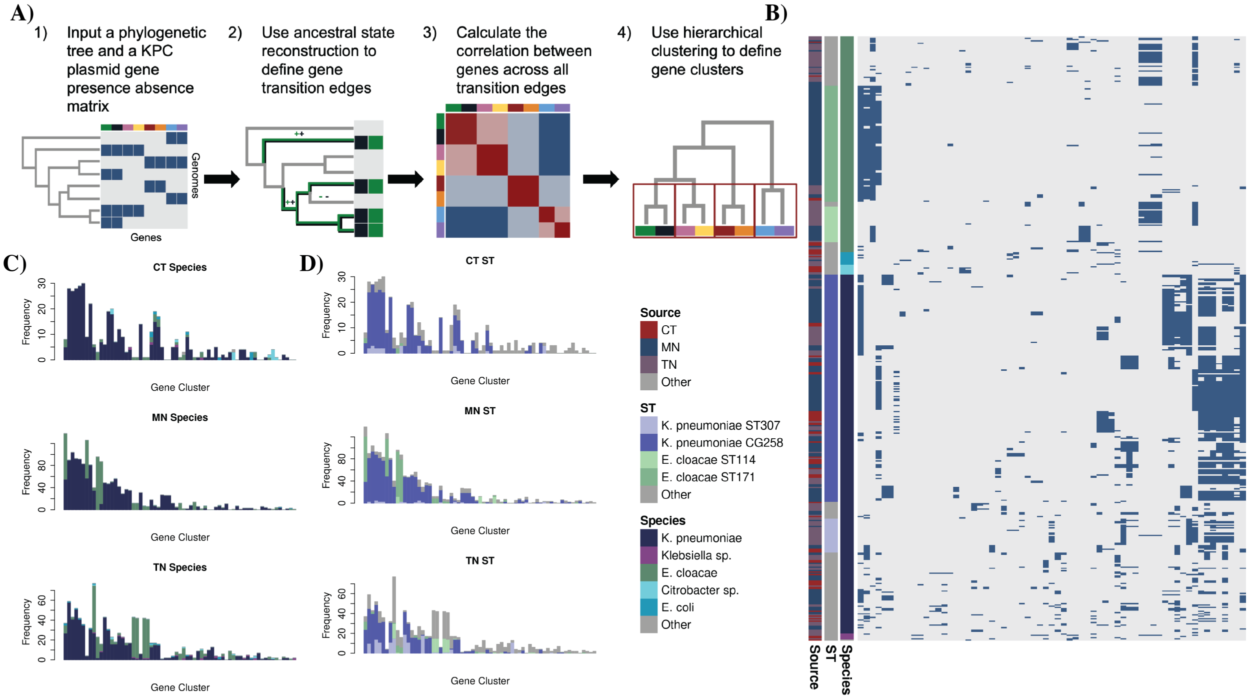


Figure 3.2: Distribution of KPC associated gene clusters in the EIP genomes. (A) Description of the algorithm to identify gene clusters to serve as markers of KPC plasmid presence. First, a gene presence-absence matrix and a phylogenetic tree are input. Ancestral state reconstruction is then used to define gene transition edges. Next, the correlation coefficient for each pair of genes is calculated by their presence and absence on each transition edge. Finally, hierarchical clustering is used to define gene clusters. (B) Heatmap indicates gene clusters' presence and absence across the entire EIP dataset. Rows are sorted by the midpoint-rooted approximate likelihood tree, and columns are sorted with hierarchical clustering. (C) Counts of each gene cluster are stratified by state, with colors indicating species assignment. (D) Counts of each gene cluster are stratified by state, with colors indicating sequence type assignment.

tified a set of conserved plasmid backbone genes whose presence can be used to monitor plasmid transfer.

3.2.3 KPC-associated Plasmid Marker Genes are Widely Disseminated Across Geography and Phylogeny

Having identified gene markers of KPC plasmids in our isolate collection, we next sought to describe their spread across species, strains, and states. A majority of marker gene clusters were identified in CRE genomes from all three states, illustrating the widespread transmissibility of these healthcare-associated lineages and the plasmids they carry. While gene clusters differed with respect to the strength of association between different species and strains, 92.9% of the KPC-associated gene clusters were observed in*K. pneumoniae* at least once (Figure 3.2C). This finding is consistent with the ability of*K. pneumoniae* to acquire and maintain diverse plasmids 1, and supports its hypothesized role as a key trafficker of antibiotic resistance 1. Within*K. pneumoniae*, we observed that certain clonal groups harbored much of the cluster diversity, with clonal group (CG) 258, in particular, is associated with 71.4% of marker gene clusters (Figure 3.2D). Thus, despite a largely stable relationship between CG258 and KPC, different plasmids are inconsistently associated with KPC, which is consistent with previous work which has shown frequent shuffling of plasmids harboring KPC in CG258 [36].

In contrast to the diversity of clusters associated with CG258, there are other examples where a set of KPC-associated gene clusters appear stably associated with a lineage. For example, *Enterobacter cloacae* ST171 was stably associated with three clusters, and in the few cases of cluster loss, there was no evidence of onward transmission. However, while these clusters were stably associated with ST171, these clusters do not appear restricted to a particular genetic background, as the individual gene clusters are observed across many other *Enterobacter* and *Klebsiella* STs (Figure 3.2).

3.2.4 Horizontal Transfer of KPC-associated Gene Clusters is Frequent Within Regions, but Imported Clusters Spread More Widely

We next sought to differentiate between KPC-associated gene cluster acquisitions which occurred within a state, representing recent HGT, and those that occurred outside of a given state, indicating importation from another location. To this end, we leveraged a large collection of contextual genomes from other regions and performed ancestral reconstruction to estimate where in the lineage importation events occurred. We then identified which of the KPC-associated gene cluster acquisitions occurred prior to importation into a given state (Figure 3.3A). Of all gene cluster ac-

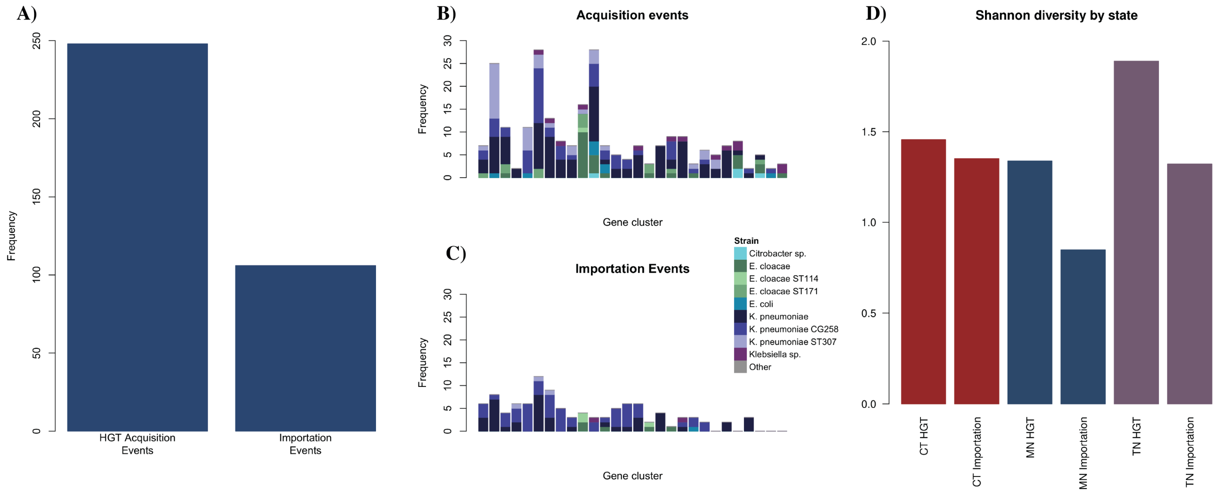


Figure 3.3: Characterization of gene cluster acquisition events across the input dataset. (A) Counts of the classification of cluster acquisition events for KPC-associated gene clusters: HGT events first observed in a state and importation events where a lineage was associated with a cluster before being observed in a given state. The prevalence of KPC-associated gene clusters across the strains observed in this study characterized as: (B) HGT and (C) importation events, respectively. (D) Shannon diversity index of the HGT acquisition edges stratified by state and HGT, importation status of the cluster acquisition event.

quisitions that were observed, most of them corresponded to HGT events within the state (n = 248), with fewer events occurring prior to importation (n = 106). However, despite intra-state HGT events being more common, historic events propagated more widely, as evidenced by the prevalence of clusters originating from importation (n = 807) being higher than the prevalence of clusters originating from intra-state HGT (n = 452) (Figure 3.3B). This observation is consistent with imported strains that are already stably associated with KPC being more apt to spread across regional healthcare networks. A Wilcoxon rank sum test confirmed that significantly more isolates arose from importation relative to HGT clusters (p = 6.291e-09).

Although overall imported KPC-associated clusters spread more widely than those acquired within states, we next explored whether there were differences in the extent of dissemination after importation vs. acquisition across states, species, and STs. Consistent with the overall trends, we found that in MN and CT, imported clusters spread more widely than those derived from local HGT events (Figure 3.4A, p = 2.52x10-7 and 3.46x10-4, respectively). In both states, this trend was driven by large clonal expansions traced back to imported CG258 isolates, as compared to singleton observations for many local HGT events in non-clonal lineages. In addition to the larger clonal expansions associated with imported clusters, they also tended to spread to multiple facilities more than locally acquired clusters, supporting both their stability and propensity for healthcare

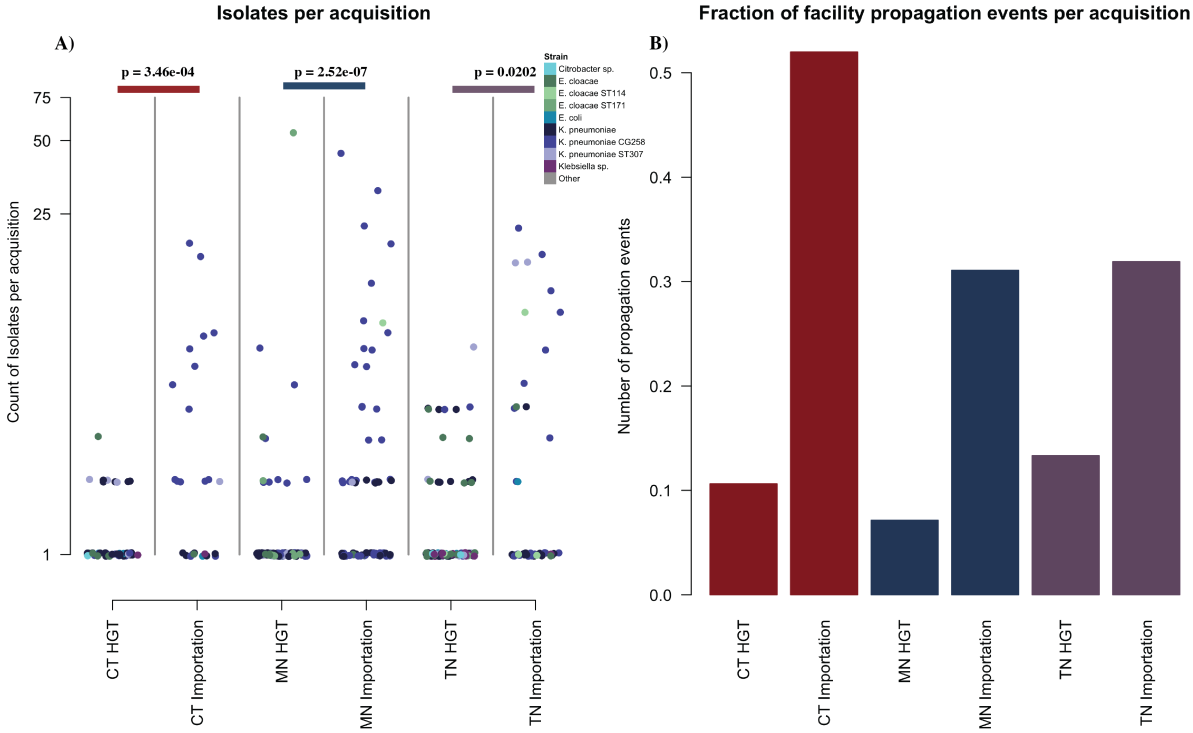


Figure 3.4: Analysis of the propagation of KPC plasmid genes. (A) Counts of isolates observed after HGT or acquisition events by state and importation status. Wilcoxon rank sum test p-values the significance between the distribution of counts with respect to HGT and importation isolates in each state. (B) Fraction of KPC-associated gene cluster acquisitions which are observed in more than one facility stratified by state and HGT, importation status of the cluster acquisition event.

transmission (Figure 3.4B).

3.2.5 Nucleotide Identity of KPC-associated Cluster Sequences is Correlated With the Site of Isolation

While most KPC plasmid-associated gene clusters were observed across all three states, we hypothesized that if we are truly detecting recent HGT occurring between isolates within each region that the sequence variation within these genes would be reflective of the proximity at which isolates were sampled. To test this hypothesis, we first considered all isolates, regardless of whether they represent local HGT, and asked whether cluster genes tend to be more similar within a state versus between states. Indeed, supporting the local proliferation of KPC plasmid-associated gene clusters via transmission and HGT, we observe significantly higher nucleotide identity among cluster genes within a state than between states (Figure S6A). Next, we focused on high-confidence cases of local HGT. We hypothesized that if HGT was indeed local and recent, HGT recipients should have a high cluster sequence identity to a putative donor strain within the same state. To this end, we subset down to HGT singletons, defined as isolates derived from HGT within a given state, but for which it was the only isolate observed descending from that cluster acquisition event (i.e., isolate count in Figure 3.4 of one). Supporting our detection of local HGT, we observed significantly greater nucleotide identity between gene clusters from HGT singletons and gene clusters in other isolates within a state than those collected outside the state (Figure 3.5A).

Having observed that local HGT events are discernable at the level of gene cluster nucleotide identity, we next sought to evaluate whether, within a state, we could detect signatures of HGT within individual healthcare facilities. Indeed, KPC-associated gene clusters from HGT singletons have significantly higher nucleotide identity than gene clusters from other isolates from the same healthcare facility as compared to other facilities from the same state (Figure 3.5B). Taken together, these results corroborate our hypothesis that our approach is detecting recent HGT events and that using marker gene clusters provides sufficient resolution to provide insight into the locations of putative HGT events.

3.2.6 Monitoring of KPC-associated Gene Clusters Enables Tracking Clonal Dissemination and HGT Within and Between Healthcare Facilities and Regions

Having found that genetic variation within KPC-associated gene clusters can enable tracking within and between states, we next examined in detail the transfer and clonal spread of the cluster with the highest number of HGT events. This cluster was observed in all three states and in a vari-

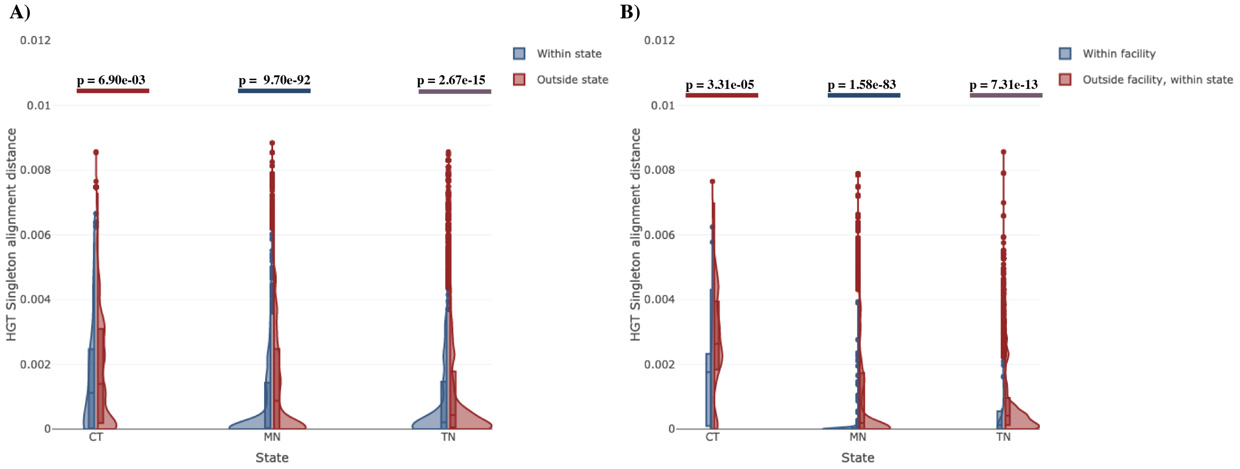


Figure 3.5: Alignment distance distribution of HGT isolates within and between sources of isolation. Concatenated gene alignments for all gene clusters were generated individually, and the pairwise alignment substitution distance was calculated for HGT singletons, isolates which were the only observed as the only isolate with an identified cluster acquisition as these were the most likely set of HGT recipient isolates. Distance distributions were then compared for isolates collected (A) within and outside each state and (B) within a facility and outside of the facility, within the same state. Wilcoxon rank sum test p-values are shown.

ety of species and ST contexts (Figure 3.6). To facilitate deeper insights into transfer patterns, we compared the core-genome phylogeny with the phylogeny constructed from plasmid-cluster genes (Figure 3.6). Examining the co-phylogeny plots for these two trees shows clear patterns of HGT in different strain backgrounds, as evidenced by the independent acquisition of the gene cluster numerous times (Figure 3.6A). Moreover, there is a clear geographic partitioning of HGT events, whereby isolates from MN and TN form distinct clusters on the gene cluster phylogeny, with these local plasmid gene clusters entering different species and strain backgrounds within the respective regions. In MN, this gene cluster was observed primarily in CG258, with multiple local transfer events, primarily into less prevalent *Klebsiella* and *Enterobacter* STs. In contrast, in TN, this plasmid gene cluster is far more widely dispersed, having been transferred to ST307 Klebsiella, ST110 Enterobacter, as well as several less prevalent STs of *Klebsiella*, *Enterobacter*, *E. coli* and *Citrobacter*.

Lastly, to provide more regional context to the transfer of this plasmid gene cluster, we examined its presence across healthcare facilities in MN and TN over time (Figure 3.6D). In MN, this cluster was first identified in CG258 in 2012, with many isolates appearing in multiple facilities after a single importation event. Subsequently, beginning in 2015, several HGT events were observed into*K. pneumoniae* and*E. cloacae* isolates observed across several facilities. In TN, importation events of this cluster were observed in *E. cloacae* ST171, *E. cloacae* ST114, and*K. pneumoniae*

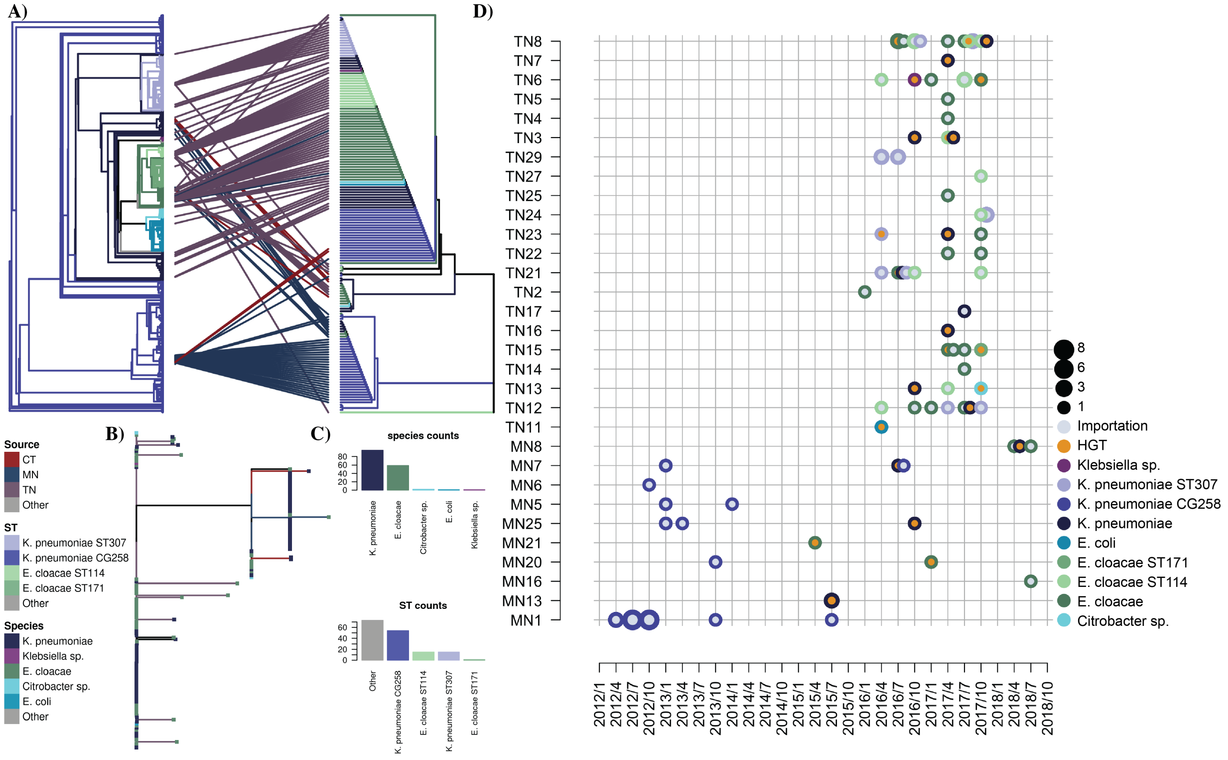


Figure 3.6: Analysis of a gene cluster propagation across MN and TN. (A) co-phylo plot showing the approximate maximum likelihood trees for the core gene alignment (left) and the plasmid gene tree (right). Edges are colored by strain in both trees, and the lines connecting tips of the trees are colored by the state of isolation (B) concatenated gene tree for the plasmid genes showing the edge lengths. (C) Counts of species (top) and sequence types (bottom) are predicted to have the gene cluster. (D) Timeline of acquisition events with time on the x-axis and facility on the y-axis. The size of the points is proportional to the number of isolates and colored by species. Inner circles indicate HGT or importation into facilities.

ST307, with dissemination across multiple facilities. In total, we observed 21 HGT events for this cluster occurring in TN facilities, with 71% of these HGT isolates having overlapping facility exposure with at least one importation isolate.

3.3 Disscussion

Public health departments are increasingly embracing genomic surveillance to track emerging threats. To effectively use these data to track not just clonal spread but also dissemination of mobile elements will require novel genomic analysis strategies to overcome the complex evolutionary trajectories of plasmids harboring cargo of interest. In this work, we defined a new method for identifying marker genes for tracking the spread of KPC plasmids by exploiting the correlated movement of plasmid genes with each other and with KPC. Applying this approach to comprehensive collections of KPC-carrying isolates from three US states, we were able to track KPC-carrying plasmids as they spread across large geographic distances, were transferred into local bacterial populations, and spread across regional healthcare facilities.

Several previous studies of HGT in hospital isolates characterized HGT events occurring in a single hospital [140, 37, 38, 141]. This study adds to previous work by quantifying the degree to which various strains propagate after an initial HGT event, thereby enabling comparisons of the dynamics of HGT across diverse geographic locations. Additionally, many studies of HGT in hospitalized patients have focused on HGT between different genera [37, 71, 31]. The method we have developed here has the advantage that it can be used to identify HGT events that occur between closely related isolates, potentially even within the same sequence type. This is especially beneficial because HGT has been shown to happen at increasing rates between isolates that are more closely related phylogenetically [71]. Therefore only focusing on HGT within the same genera would miss a substantial number of HGT events. Another advantage is that by relying on annotated assemblies, we can incorporate genome sequences generated by different sequencing technologies: incorporating long-read and short-read sequencing data in the same analysis. Therefore we can leverage more of the available data to address questions regarding the evolution and phylogeography of bacterial pathogens.

For gene clusters classified as HGT and importation, the magnitude of onward spread was significantly different. The total burden of KPC was largely attributable to the importation of strains already harboring KPC, which proliferated at higher rates relative to strains acquiring clusters within a region. Imported strains were mainly *K. pneumoniae*, but not exclusively. The fact that we were able to match these imported lineages with outside isolates of the same strain also carrying the same plasmid suggests a stable relationship between these strains and the plasmids they carry, evidenced by transmission across large geographic distances. This suggests that these strains are well adapted to the hospital environment and able to proliferate at higher rates relative to other strains. *Klebsiella pneumoniae* CG258, in particular, was observed with a notable diversity of plasmid gene clusters and carried them over large distances, which has been documented previously [39]. Recent HGT isolates may not have the same ability for transmission and are outcompeted, potentially due in part to the fitness costs associated with plasmid carriage. In contrast, laboratory studies have shown that *Klebsiella pneumoniae* can maintain large conjugative AMR plasmids with minimal fitness cost [19]. In contrast, *E. coli* isolates were shown to have significant fitness costs attributed to the conjugative machinery present on these plasmids, which was deleted as these strains were propagated in culture, alleviating the fitness cost of the plasmid. This may present a mechanism by which *Klebsiella pneumoniae* provides a vessel facilitating plasmid evolution via recombination and rearrangement, which can then be spread to isolates.

In TN, we noted a greater diversity of strains in which HGT was observed, as well as wider dissemination subsequent to HGT. In contrast, in CT and MN we observed wider dissemination of imported KPC harboring strains. The reasons for these differences are unclear. Variation in HGT and transmission may be due to differences in the underlying patient populations, prevalence of antibiotic usage, practices with respect to patient transfer, or differences in infection control practices. Unfortunately, data on these metrics are unavailable, so we are unable to access this directly. TN is notable for having higher antibiotic prescribing rates with 1046 prescriptions per 1000 residents in contrast to MN and CT, which have 447 and 778 prescriptions per 1000 residents, respectively [142]. Higher rates of antibiotic prescriptions may create a commensurate selective pressure to acquire ARGs within circulating bacterial populations creating a diverse pool of AMR isolates, which compete with the clonal lineages also circulating in the region. While only three states are represented in this analysis, this raises an intriguing possibility for future research to analyze the association between antibiotic prescribing and HGT of ARGs within bacterial populations.

This study has several limitations. First, while we have assembled a large, diverse dataset of isolates to characterize gene cluster importation and HGT events, our approach relying on previous collections of public isolates is far from a representative sample of the bacterial populations circulating in the healthcare system. Therefore we may overstate the number of HGT events observed within a given state because we can only identify the first observed instance of a plasmid and strain pair as HGT. Additionally, this dataset is biased in that it disproportionately contains infection isolates, as opposed to surveillance isolates. This may cause undersampling of lineages carrying KPC which are capable of colonization and disseminating KPC plasmids but have a low potential for virulence. Previous work has shown that infection isolates represent the tip of the iceberg, and therefore we are likely missing some transmission events [45]. Additionally, we do not attempt to differentiate between multiple HGT events in a given strain. This may overstate the number of propagation events, which are, in reality, numerous independent HGT events. Our method using hierarchical clustering assigns a gene to a single cluster. While this is useful for identifying specific markers of a plasmid, some gene cassettes may be components of multiple diverse plasmids and therefore do not appear strongly correlated to any given set of plasmid genes. This limits the fraction of total plasmid genes assigned to any given cluster, dividing a plasmid into multiple clusters. We only analyzed genes that were found on previously sequenced KPC plasmids. This limits our ability to identify new plasmids, which may function to shuttle KPC through bacterial populations. This method could be extended in the future to include all genes within a given dataset; however, the diversity of isolates, plasmids, and regions made this intractable for our analysis. Finally, the non-overlapping intervals of isolate collection do not provide a uniform snapshot of the prevalence of CRE during the study. Also, differences in the duration of isolate collection may limit the ability to compare across the different regions. The extended duration of collection and greater quantity of isolates in MN may provide a better representation of the circulating bacterial populations relative to TN and CT. These differences in the time of collection and quantity of genomic data available may limit the accuracy of statistics and limit the number of propagation events observed in the states with shorter collection periods.

In conclusion, this analysis represents the first comparative analysis of HGT across multiple statewide healthcare networks. We have shown that in a majority of cases *K. pneumoniae* CG258 are imported into a region and introduce the KPC plasmids they carry into diverse lineages of *Enterobacterales* not previously associated with KPC. This was observed with multiple strains within *K. pneumoniae* CG258 and multiple plasmids. We showed that while strains representing predicted recent KPC acquisitions within a state were less prevalent, these newly carbapenem-resistant strains have the potential to cause infections. In several cases, clonal isolates with the same KPC-plasmid genes suggest the possibility of a single HGT event and subsequent transmission, resulting in multiple antibiotic-resistant infections.

While *K. pneumoniae* was the first species with epidemic potential observed with KPC; this was not the only strain capable of disseminating KPC across geographic regions. *E. cloacae* ST114 and *E. cloacae* ST171 were both observed importing KPC plasmids across multiple states, illustrating the importance of genomic surveillance in characterizing emerging threats. Understanding the epidemiology and evolution of antibiotic resistance requires not only the challenging task of tracking transmission networks of bacterial pathogens in patient populations but also tracking the transmission of plasmids within bacterial populations, adding enormous complexity. This work provides a framework for leveraging large datasets of bacterial genome sequences to characterize KPC-carrying plasmids and track their introduction into diverse bacterial populations, which enhances the potential for virulence with severe consequences for public health.

3.4 Materials and Methods

Whole genome sequences were obtained for isolates collected as part of the Emerging Infections Program (EIP) as described previously: confirmed carbapenem-resistant KPC+ isolates were collected in three geographically diverse states: Connecticut (CT), Minnesota (MN), and Tennessee (TN) [143]. Briefly, CT isolates were collected across the entire state from 2017-2018. MN isolates were collected from 2021-2018. TN isolates were collected from 2016-2017 from all counties except for counties in the Memphis-Delta and Northeast Tennessee region due to high healthcare utilization rates by residents from states other than Tennessee.

Public *Enterobacerales* genomes used for this study were downloaded from PATRIC as of 04/23/2021 [144]. Additionally, assemblies from PRJNA603790, PRJNA690239, PRJNA401340, and PRJNA415194 were also included for a total of 74,367 assemblies. All genomes were annotated using RAST [114]. Redundant assemblies from previous CDC studies were excluded from the analysis (PRJNA292901, PRJNA292904, PRJNA288601, and PRJNA272863).

Concatenated gene alignments for the core genes of the full data-set of *Enterobacteriaceae* assemblies, and the EIP isolates and their matched nearest KPC+/KPC- neighbors, and complete KPC+ genomes were generated with cognac (v1.0) [145]. FastTree was used to generate approximate maximum likelihood trees for the core gene alignment of the EIP isolates, their matched neighbors, and the complete KPC+ genomes [146].

To define gene content, CD-HIT (v4.7) was used to cluster genes by amino acid sequence similarity [117]. Then for each gene cluster, the corresponding nucleotide sequences of each gene were extracted and aligned with MAFFT (v7.310) [118]. Nucleotide alignments were then parsed to identify gene clusters with at least 99% identity over the aligned sequence or fewer than two substitutions. If an allele met these criteria for more than one gene cluster, it was assigned to the cluster that minimized the number of substitutions.

To identify genetic markers of KPC plasmids, we implemented an approach to reconstruct the ancestral states of each gene across the phylogeny, and define HGT events by the acquisition of sets of genes that are correlated in their patterns of acquisition and deletion across multiple, independent branches of the phylogenetic tree. To this aim, we employed maximum parsimony ancestral state reconstruction to estimate the ancestral character states of each gene. This algorithm identifies the minimum number of character state changes necessary to explain the distribution of character states at the tips of the tree and has a long history of use in evolutionary biology [147, 148]. For each gene, we then identified all the transition edges: the edges on the tree where there was a change in absence or presence, representing the period in which a gene was acquired or deleted. Ancestral reconstruction was performed by inputting the fastTree of the concatenated, core-gene alignment and a binary gene presence in the absence of each plasmid gene observed more than once. For each gene, the binary vector representing gene presence and absence was input to the MPR function in APE to estimate the presence of a gene at each node [120].

The MPR results were parsed to identify the sets of genes that were acquired or deleted on each edge, respectively. A gene transition event by gene matrix was created for each set of transitions with at least two genes. Pearson's correlation coefficient for the genes across multiple transition edges was calculated using the stats package [149]. The gene correlations were then converted to the true euclidean distance, and agglomerative hierarchical clustering was performed using the "agnes" function from the cluster package (v2.1.2). To define the cluster transition edges, we identified all edges where at least 50% of the genes were present for each cluster. Cluster acquisition edges were then defined as the last edge, proximal to the root with at least 50% of the cluster genes. Cluster deletion edges were defined as any edge descending from an acquisition edge, with fewer than 50% of cluster genes. Genomes that maintained at least 50% of genes from the acquisition edge to the corresponding tip were classified as positive for a cluster. Many KPC-associated gene clusters exhibited highly correlated patterns of presence and absence across the EIP genomes. Because these clusters likely represented components of the same plasmid, we then merged clusters that were 90% similar with respect to the genomes they were present in, ensuring that all merged clusters were always acquired on the same edge for a final total of 65 gene clusters most relevant to our analysis.

For each set of genes corresponding to a KPC-associated gene cluster, the corresponding nucleotide alignments from cluster-positive genomes were concatenated, and FastTree was used to generate approximate maximum likelihood trees [146]. MPR was also performed by location and strain to classify each edge of the tree.

All code generated for this study is available at https://github.com/rdcrawford. All analyses were performed using R version 4.1.1 20. Plots and manipulation of phylogenetic trees was performed using the Ape (v5.5) and phytools (v1.0-1) packages [120, 150]. Multithreading was enabled via the future.apply package (v1.0). Heatmaps were generated with pheatmap (v1.0.12).

CHAPTER 4

Phenotypic and Genomic Diversification in Complex Carbohydrate-Degrading Human Gut Bacteria

4.1 Preamble

This work represents a collaboration exploring the breadth and complexity of carbohydrate metabolism in the microbiome and the genomic events that shape these behaviors. I performed the genomic analysis of horizontal gene transfer: I developed and applied the method for identification of HGT loci involving core genes, made the phylogenetic tree, performed BLAST analysis, and I presented these results in figures 4.6, 4.7, and supplemental figure B7. My coauthors conducted the remainder of the analysis. We published this work in mSystems: Pudlo, N. A., Urs, K., Crawford, R., Pirani, A., Atherly, T., Jimenez, R., ... & Martens, E. C. (2022). Phenotypic and Genomic Diversification in Complex Carbohydrate-Degrading Human Gut Bacteria. mSystems, 7(1), e00947-21.

4.2 Introduction

Microbial communities in the distal intestines of humans and other mammals play critical roles in the digestion of dietary polysaccharides [151, 152, 153]. Unlike proteins, lipids, and simple sugars, which can be assimilated in the small intestine, the vast majority of nonstarch polysaccharides (fibers) transit undegraded to the distal gut due to a lack of requisite enzymes encoded in the human genome [74]. Microbial transformation of dietary fiber polysaccharides into host-absorbable organic and short-chain fatty acids is a beneficial process that unlocks otherwise unusable calories from our diet [154], shapes the composition and behavior of the gut microbial community [155, 156, 157], provides preferred nutrients directly to the colonic epithelium [158, 159, 160], and shapes the development of immune cell populations [161, 162].

The abundance of dietary fiber in the mammalian diet and the substantial chemical diversity within this class of molecules provide a prominent selective pressure that drives genome evolution and diversification within symbiotic bacterial populations. The genomes of individual human gut bacteria frequently encode dozens to hundreds more polysaccharide-degrading enzymes than humans secrete into the gastrointestinal tract, reflecting gut microbial adaptations to degrade dietary fibers [153, 74]. As examples, the genomes of a few well-studied Gram-negative *Bacteroides* (*Bacteroides thetaiotaomicron, Bacteroides ovatus, and Bacteroides cellulosilyticus*) encode between 250 and over 400 CAZymes that collectively equip them to target nearly all commonly available dietary polysaccharides [163, 164, 165]. However, none of these three species is by itself capable of degrading all available polysaccharides, a conclusion that was supported by early phenotypic surveys of cultured human gut bacteria that encompassed species from other phyla [166]. These findings suggest that individual microbes fill multiple, specific carbohydrate degradation niches and that a diverse community is required to ensure degradation of the entire repertoire of dietary fibers. Given that hundreds of different microbial species typically coexist in an individual over long time periods [167], it is important to understand how many different polysaccharide metabolism pathways are present within the individual microbial species that compose a community and how these traits are represented across strains and species. If some species possess very similar phenotypic abilities, they may be functional surrogates or compete for similar niches and therefore seldom cooccur. Members of the *Bacteroidetes* phylum are often among the most numerous bacteria in the human colonic microbiota, with members of the genus *Bacteroides* often prominent in individuals from industrialized countries [167, 168, 169]. These bacteria are well appreciated for their abilities to degrade a broad range of polysaccharides [163, 164, 165, 170, 171] and modify disease states in a bacterial species-specific fashion [172, 173, 174]. In this study, we empirically measured the abilities of members of 29 different *Bacteroidales* species to grow on a custom panel of carbohydrates that span the diversity of plant, animal, and microbial polysaccharides. Our results reveal a wide range of metabolic breadth between different species, indicating that some have evolved to be carbohydrate generalists, while others have become metabolically specialized to target just one or a few nutrients. A pangenome analysis of several related strains provides insight into the evolutionary events that shape carbohydrate utilization among these important symbionts and reveals a dizzying mosaic of heterogeneity at the level of discrete gene clusters mediating polysaccharide metabolism. Based on the analysis of several variable loci, we provide evidence to support a mechanism of lateral gene transfer that may account for this mosaic architecture. Our results provide a glimpse into the metabolic breadth and diversity of an important group of human gut bacteria toward polysaccharide metabolism. Given the large amount of genomic and metagenomic sequence information that has been generated from the human microbiome, phenotypic studies such as the one presented here represent important next steps in deciphering the functionality of these organisms in their native gut habitat.

4.3 Results

Phenotypes are the ultimate measures of biological function. However, large-scale phenotypic analyses are still uncommon in surveys of the human gut microbiome, which have instead relied on sequence-based approaches to infer function, often with substantial uncertainty. This lack of phenotypic information is due partly to a lack of high-density (e.g., strain level) culture representation for the dominant taxa combined with a lack of defined growth conditions to measure the behavior of these organisms. With the resurgence of gut microbial culturing, both of these gaps have begun to close [175, 176, 177, 178], revealing an urgent need for scalable platforms to define the actual behavior of these organisms. To address this gap, we assembled a collection of human and animal gut Bacteroidetes and constructed a custom anaerobic phenotyping platform centered around carbohydrate metabolism, a key function that symbiotic gut microorganisms contribute to mammalian digestion [74]. This array consists of 45 different carbohydrates (30 polysaccharides and 15 monosaccharides) that span the repertoire of common sugars and linkages present in dietary plants and meat, as well as host mucosal secretions and some rare nutrients consumed in regional populations or as food additives (see Figure B1 in the supplemental material for a summary of polysaccharide structures).

The carbohydrate utilization abilities of 354 different human and animal *Bacteroidetes* strains were measured by individually inoculating each into this custom growth array and automatically monitoring anaerobic growth every 10 to 20min for 4 days (see Materials and Methods). Based on the 16S rRNA gene sequence for each strain, this collection encompasses 29 different species based on the requirement that each strain possesses at least 98% 16S rRNA gene identity to a named type strain in a given species (Table S1a) (note that all but three strains, which were all related to each other and to *Bacteroides uniformis*, met this criterion). The resulting 31,860 individual growth curves were first inspected manually and then subjected to automated analysis to quantify total growth and growth rate parameters for each substrate (see Materials and Methods). A normalization scheme was employed to compensate for general growth differences in the two differently defined medium formulations employed (see Table S1a for a full list of strains assayed and all raw and normalized growth measurements; see Figure B2 in the supplemental material for an analysis of replicates).

4.3.1 Members of the Same Species Possess Similar Carbohydrate Utilization Profiles

Growth results are summarized in Figure 4.1 and 4.2 and Figure B3 in the supplemental material. Whether considered from the perspective of how many species degrade a particular polysaccharide

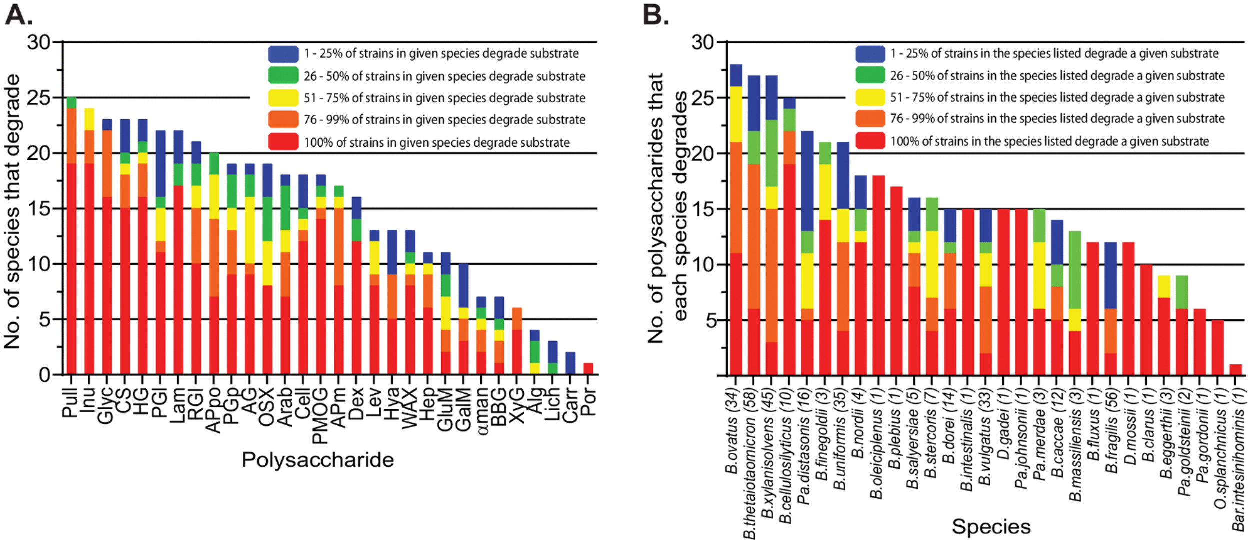


Figure 4.1: (A) The number of species out of 29 tested that degrade each polysaccharide is listed in order of decreasing degradation frequency from left to right. Since not all strains within a given species necessarily have the metabolic potential to utilize each polysaccharide, colors illustrate the percentage of strains within each degrading species that possess the indicated ability. (B) The number of polysaccharides that a given species degrades is shown in decreasing order. The number of strains tested for each species is listed in parentheses, and colors represent the percentage of strains in each indicated species that degrade each glycan counted toward the total.

(Figure 4.1A) or how many individual polysaccharides are targeted by members of a particular species (Figure 4.1B), there was substantial variability in carbohydrate utilization among the organisms surveyed (range, 1 to 28 polysaccharides degraded per strain; mean, 15.6). Some polysaccharides like soluble starch/glycogen were degraded by a majority of the species tested, and yet others like the edible seaweed polysaccharides carrageenan and porphyran were used by just one or two strains.

Given the diversity in observed carbohydrate utilization phenotypes, we wished to address if closely related strains display similar abilities or instead if strains of the same species have diverged from one another. To assist in visualizing the overall trends in carbohydrate utilization across this phylum, we performed unsupervised clustering of the strains based on their carbohydrate utilization profiles. While many species are not deeply represented by multiple strains, clustering based on a combination of normalized growth and rate measurements largely grouped strains of the same species together (Figure 4.2), and as expected, this clustering was driven mostly by polysaccharide utilization abilities (see Figure B4 in the supplemental material).

Our data reveal that strains belonging to several individual species possess more similar polysaccharide-degrading abilities to each other than their more distant relatives, a finding that has importance for interpreting or predicting function based on community sequencing data. As

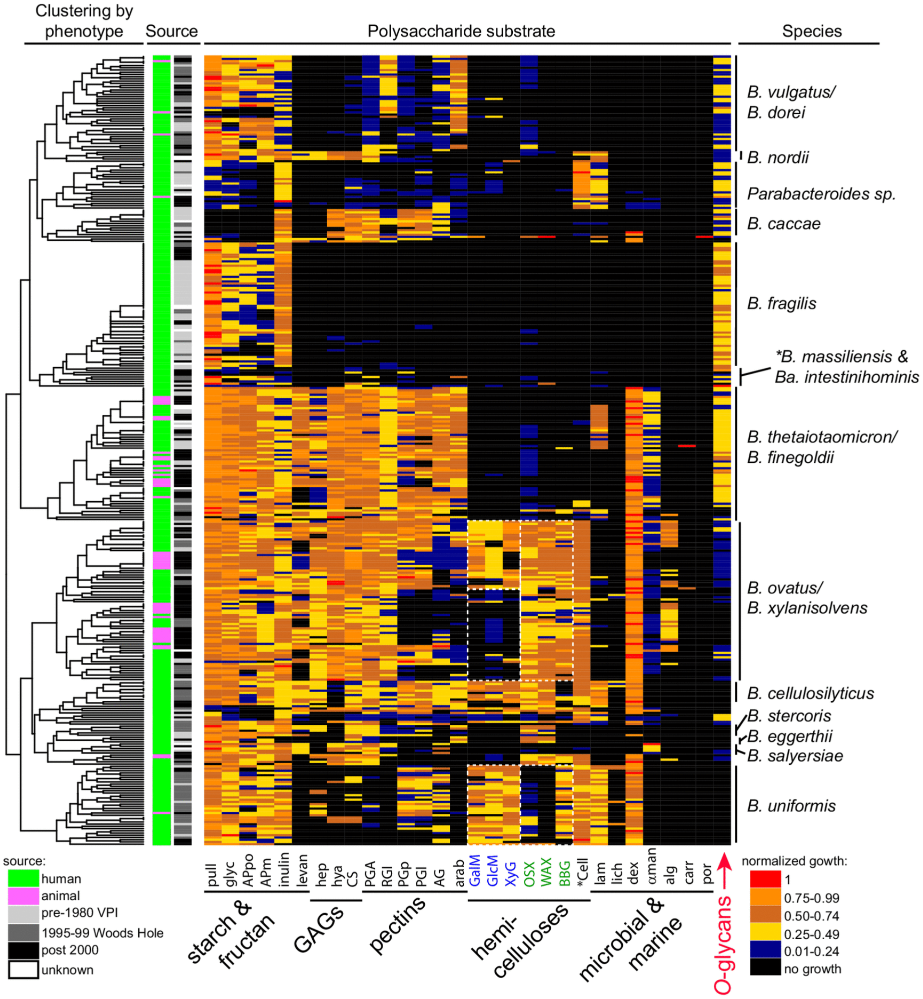


Figure 4.2: Species are clustered by glycan utilization phenotype based on normalized total growth level (Figure B4B). The magnitude of growth is indicated by the heatmap scale at the bottom right. Columns at the left indicate the source (human or animal) and time period of isolation. The cladogram at the far left shows the results of unsupervised clustering of the data based on the normalized growth data shown. The species designations at the right are the results of 16S rRNA gene sequencing (>98% identity to the species type strain was used to assign species). The region containing mucin specialists *B. massiliensis* and *B. intestinihominis* is indicated but marked with an asterisk because the 4 strains in these 2 species are not clustered perfectly in this region. All raw and normalized growth and rate data for individual strains may be found in Table S1. See Figure B3 for an expanded heatmap with monosaccharide data and individual strain names labeled. All processed growth curves are available as source data.

examples, all 56 strains of *B. fragilis* clustered together, reflecting their generally restricted abilities to utilize forms of soluble starch/glycogen, inulin, and mucus O-glycans. Likewise, all 36 strains of B. uniformis, a species with a broader metabolic capacity that includes digestion of plant cell wall hemicelluloses, were also grouped together into a single branch. The inclusivity of these groupings was generally independent of the time period when strains were isolated or whether they were isolated from humans or other mammals (Figure 4.2). Another important feature of the observed species clustering is that the grouping does not mirror the overall phylogeny of the gut *Bacteroidetes*. Rather, phylogenetically separated species often group adjacent to one another based on similarities in carbohydrate metabolism (e.g., *B. ovatus/B. xylanisolvens* and *B. cellulosilyticus*, and *B. vulgatus/B. dorei and B. fragilis*) (see Figure 4.3A for a phylogenetic tree based on conserved housekeeping genes) [179, 180]. It is interesting to directly compare *B. fragilis* and *B. vulgatus/B. dorei*, which are two groups with deep strain representation (Figure 4.2). Despite being phylogenetically more distant, members of these two species possess similar abilities to degrade starch and related molecules (glycogen and pullulan), inulin, and mucin O-glycans. The major distinguishing feature between these groups is the presence of some pectin utilization, which is often weak, among strains of *B. vulgatus/B. dorei*. Indeed, acquisition of growth abilities that are unique with respect to species with an otherwise similar potential may be one way that species avoid direct competition for the same niches.

Some polysaccharides, especially those present in the cell walls of dietary plants, occur in the same physical context and presumably traverse the gut together, potentially exerting selective pressure for bacteria to use them simultaneously. To test for the co-occurrence of different polysaccharide utilization abilities within the 354 individual strains, we calculated the pairwise correlations between the utilization of any two polysaccharides by the same strain (see Figure B5 in the supplemental material). This test might reveal tendencies to coutilize different polysaccharides that are chemically different (positive correlation) or avoid using substrates from incompatible niches (negative correlation), if they exist. The presence of two different soluble starches (potato and maize amylopectin) and two starch-like glycans (glycogen and pullulan) provides an internal control since they are essentially identical in their sugar and linkage chemistry but vary in the proportion and placement of branches as well as polymer length, crystallinity, and solubility (Figure B1). These four molecules are utilized through a single degradation/transport system in the type strain of *B. thetaiotaomicron*, which was included in our study [181]. As expected, the abilities to use these four polysaccharides were among the strongest positive correlations (between 44% and 75%); although, there was not a perfect correlation suggesting that some finer adaptation may exist even for different structural forms of a chemically similar molecule.

We also observed positive correlations in the ability of bacteria to simultaneously utilize polysaccharides within two different groups of plant cell wall polysaccharides (pectins and hemi-

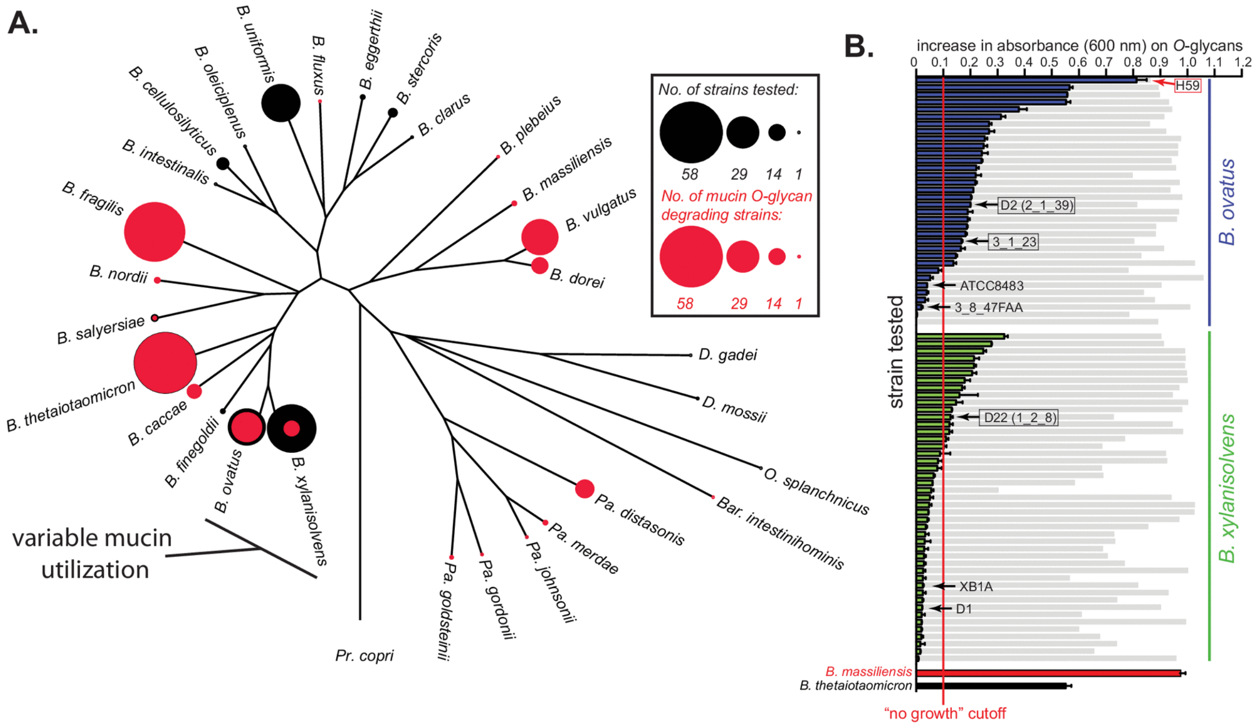


Figure 4.3: Host mucin O-glycan metabolism within the Bacteroides. (A) A phylogenetic tree based on housekeeping genes that compares mucin O-glycan utilization across species. The diameter of the black circles represents the number of strains tested within each species (sample depth), whereas the size of the overlaid red circle corresponds to the number of strains exhibiting O-glycan metabolism. Note that some species have either full or no penetrance of this phenotypic trait and yet others like *B. ovatus*/*B. xylanisolvens* have more extensive variability among strains. (B) Strains of *B. ovatus* (blue) and *B. xylanisolvens* (green) that show variable growth abilities on mucin O-glycan (n=2 growth assays per bar, error bars are range between values). Gray histogram bars are total growth controls on an aggregate of the monosaccharides that all strains of these two species grow on (Table S1) and are provided as a reference for overall growth ability on a non O-glycan substrate. Data from two established O-glycan degraders, namely, *B. massiliensis* and *B. thetaiotaomicron*, are also shown for reference. Species with black arrows were used for pangenome analyses to compare genetic traits associated with mucin O-glycan metabolism. We performed RNA-seq on three strains included in this pangenome analysis (black boxes) that were positive for O-glycan utilization and an additional strain, namely, *B. ovatus* NLAE-zl-H59 (red arrow, box), to see if there were unique genes/PULs present in strains that have the ability to grow on mucin O-glycans.

celluloses), as well as animal tissue glycosaminoglycans (Figure B5, green boxes highlight the 3 separate groups containing substrates with positive correlations within that group, although a weaker correlation can be observed across groups). These correlations occurred despite the fact that the polysaccharides within each of these groups often possess different structures but might co-occur in plant material or digested animal tissue. In the case of the hemicelluloses, there was even some apparent separation based on dicotyledonous versus monocotyledonous sources. The predominantly dicot hemicelluloses (Figure 4.2, blue labels) and monocot hemicelluloses (Figure 4.2, green labels) show some exclusivity with respect to the bacteria that utilize them. Many *B. ovatus*/*B. xylanisolvens* strains lack the ability to utilize the three dicot hemicelluloses (GalM, GlcM, and XyG), whereas the ability to degrade those from monocots (OSX, WAX, and BBG) is distributed more evenly. B. uniformis has a partially opposite pattern, preferring substrates from dicots, while only degrading one of the two major monocot structures (BBG) and poorly degrading the two xylans tested (OSX and WAX). Similar observations were also made for pectins and GAGs and could reflect adaptations to simultaneously harvest different nutrients from digesta particles derived from dicot plant cell walls or animal tissue ingested in a carnivorous diet. Finally, there was a positive correlation between the utilization of α-mannan and dextran, two microbial polysaccharides that are not known to occur together in foods or other sources of these polysaccharides (Figure B5).

4.3.2 Specialization for Mucus O-linked Glycans

The most noteworthy correlation between polysaccharide utilization traits was observed between the utilization of host-produced mucin O-glycans and many of the other polysaccharides tested. Growth on a total of 19/30 polysaccharides showed negative correlations with the ability to utilize O-glycans, with the strongest negative correlations being between O-glycans and the seven different hemicelluloses (Figure B5). This negative correlation is observed easily by comparing the rightmost column in Figure 4.2 (O-glycan utilization) with the respective columns for hemicellulose degradation. Because this trend was observed across several species, it suggests that there could be a more general exclusive relationship between the two niches associated with foraging on mucus and hemicellulose. This idea is further supported by experiments described below, which suggest that isolates of *B. ovatus* and *B. xylanisolvens*, both adept hemicellulose consumers, are in the process of losing the ability to degrade O-glycans, relative to an ancestor that contained multiple gene clusters involved in the metabolism of these structures.

Interestingly, the mucin O-glycan mixture was the only substrate for which we observed absolute metabolic specialization among the substrates tested. A single, and only available strain of Barnesiella intestinihominis exhibited the ability to exclusively utilize mucin O-glycans, along with a subset of the sugars that are contained in these structures (Figure 4.2; Table S1a). Three strains of *Bacteroides massiliensis* exhibited similar behavior with very strong growth on mucin O-glycans and only weak growth on soluble starches and a few other polysaccharides (Figure 4.2; Table S1a). These three *B. massiliensis* strains were also restricted in the repertoire of simple sugars with which they could metabolize; this list is limited to those found in mucin and other host glycans (galactose, N-acetylgalactosamine, N-acetylglucosamine, N-acetylneuraminic acid, and L-fucose; weak fructose utilization by one strain was the only exception). Members of these two species are represented poorly in culture collections and remain lightly studied. However, their specific adaptations for host mucin glycans may render them important members of the microbiota, potentially thriving at the interface between the gut lumen and host tissue and relying exclusively on the host to be sustained. The continuous supply of mucin in vivo could explain why some species have become specialized for it as a nutrient, whereas dietary fiber degraders may need to be more generalist since the substrates available to them change with the host's meals.

4.3.3 Pangenome Reconstruction Reveals Extensive Genetic Diversification Among Related *Bacteroides* Members

With a view of the carbohydrate utilization traits present in our gut *Bacteroidetes* collection, we next sought to determine if certain variable traits were being gained or lost within strains of certain species and if available genomes provide insight into the mechanisms driving genomic adaptations to particular nutrients. Connections between polysaccharide utilization phenotypes and the underlying genes involved have been explored systematically for a few *Bacteroides* species (*B. thetaiotaomicron*,*B. ovatus*, and *B. cellulosilyticus*) with partial analyses in others [155, 165, 170, 171, 182, 183, 184, 185]. These studies have revealed that, in essentially all cases, the ability to degrade a particular polysaccharide is conferred by one or more clusters of coexpressed genes termed Polysaccharide-Utilization Locus (PUL)s) [186]. PULs share defining features, such as genes encoding homologs of outer membrane TonB-dependent transporters (SusC-like), surface glycan-binding proteins (SGBPs; or SusD- and SusE/F-like), usually an associated sensor/transcriptional regulator, and one or more degradative CAZymes (glycoside hydrolase [GH], polysaccharide lyase [PL], and carbohydrate esterase [CE]), as well as other enzymes like sulfatases or proteases. Since the presence of one or more cognate PULs is required to utilize a given polysaccharide and these genes typically exhibit large increases in gene expression in response to their growth substrate, we rationalized that we could focus on traits that were variable in closely related strains and locate the associated PULs by transcriptomic analysis to gain insight into the basis of their acquisition or loss.

To test this hypothesis, we focused on members of two closely related species, *B. ovatus* and

B. xylanisolvens, for which there is noticeable interstrain variation in their ability to use mucin O-glycans (Figure 4.2 and 3). The investigation of these two species also benefits from substantial culture depth and many strains with available sequences. The O-glycans attached to mucins represent a diverse family of over one hundred different structures [187], albeit with common linkage patterns (Figure B1). Correspondingly, the ability to utilize these glycans is a complex trait, involving the simultaneous expression of at least 6 to 13 different O-glycan-inducible PULs in *B. thetaiotaomicron*, *B. massiliensis*, *B. fragilis*, and *Bacteroides caccae* [155, 170, 183]. Among the *B. ovatus* and *B. xylanisolvens* strains that surpassed the threshold for growth on O-glycans, there was a continuous gradient of growth abilities, which could be attributed to variations in PUL content and therefore gradations in the ability of the strains to access the many different structures in the complex O-glycan mixture (Figure 4.3B). One hypothesis to explain this observation is that some *B. ovatus* and *B. xylanisolvens* strains have gained the ability to utilize O-glycans relative to an ancestor that lacked this phenotype. If so, the PULs they express during O-glycan degradation might be unique to their genomes and may indicate HGT, as has been the case for the acquisition of phenotypes such as porphyran, agarose, and $gamma$-carrageenan utilization in gut *Bacteroides*, which are all components of integrative conjugative elements or mobilizable plasmids [179, 188]. An alternative hypothesis is that some *B. ovatus* and *B. xylanisolvens* strains are in the process of losing this ability from a common ancestor. If so, the genomes of nondegraders may still contain some PULs that are homologous to those present in more proficient O-glycan-degrading strains, but these strains may have lost a key step(s) that has eroded their ability to express this phenotype.

To distinguish these hypotheses, we selected seven strains (black arrows in Figure 4.3B) that vary in their ability to degrade O-glycans and for which genome sequences exist. Note that three strains that degrade O-glycans were chosen initially because they were among the strongest degraders in our data set with sequenced genomes when we initiated these experiments. We later identified strains with better O-glycan growth abilities and address one of these (strain H59) separately below. Four of the selected strains were *B. ovatus* (two positive and two negative for O-glycan degradation); three strains were *B. xylanisolvens* (one weakly positive and two negative for O-glycan degradation). One of these strains (*B. xylanisolvens* XB1A) has a finished circular genome and was used as a scaffold to align the remaining six draft genome sequences, with manual curation (see Materials and Methods), resulting in a nearly contiguous pangenome sequence that captures the spatial arrangement of homologous and variable genes that are present in these seven strains (see Table S2a in the supplemental material) (see https://www.ericmartenslab.org/ for downloadable physical maps of the pangenome).

An analysis of the *B. ovatus*/*B. xylanisolvens* pangenome revealed remarkable variability in gene content among just the seven strains used. A total of 12,960 different genes were delineated based on at least 90% identity in their translated amino acid sequence (Table S2a). Remarkably, only 2,264 (17.5%) of these genes were shared among all 7 strains. The largest proportion of genes (7,244, 55.9%) was present only in 1 of the 7 strains. Separating two major classes of core PUL functions, SusC/D homologs and degradative CAZymes (GH, PL, and CE), revealed that these key components of *Bacteroidetes* polysaccharide metabolism were also represented heavily in the "accessory gene" pool that is not common to all strains (Figure 4.4A).

Through informatics-based and manual annotation of gene clusters containing typical PUL functions, we delineated between 180 and 236 different PULs in the reconstructed pangenome (ambiguity is caused by many PULs occurring adjacent to each other; although in many cases separation of adjacent PULs according to individual genomes allowed us to make more precise delineations) (Table S2b). A direct comparison of the O-glycan-degrading and nondegrading strains revealed that there was a substantial number of genes (3,351) that were unique to the 3 O-glycan degrading strains, including genes belonging to 51 PULs (Figure 4.4B). However, such a distribution in gene content might be expected given the overall large proportion of noncore genes in these seven strains, and there was correspondingly no indication that all three O-glycan-degrading strains shared overlapping PULs with each other; no PULs were common to all three O-glycan degraders, and only five PULs were shared by any two strains (Figure 4.4C). Considering that there are 51 total PULs that are unique to the mucin-degrading strains, if these strains have gained the ability to degrade O-glycans from an ancestral lineage that lacked this ability, it likely occurred by the acquisition of separate gene clusters. To more directly distinguish between the two hypotheses given above, we performed transcriptional profiling on all three O-glycan-degrading strains to determine if the PUL genes that they express during O-glycan degradation are indeed unique to these strains.

Compared with reference growth in minimal medium containing glucose (MM-glucose), the *B. xylanisolvens* D22, *B. ovatus* 3-1-23, and *B. ovatus* D2 strains activated the expression of 196, 227, and 359 total genes more than 10-fold, and these gene lists included components of 14, 19, and 42 different PULs, respectively (see Table S3a to c in the supplemental material). As expected from studies in other Bacteroides, these PULs were scattered throughout the genome (see Figure B6 in the supplemental material), suggesting that they are regulated autonomously in response to glycan cues present in the O-glycan mixture. Strikingly, the majority of PULs that contained O-glycan-activated genes (63/75, 84%) were not unique to the O-glycan-degrading strains (Table S3a to c; Figure B6). Moreover, in each of the three strains analyzed, the most highly upregulated PULs were also often shared with non-mucin-degrading strains. These observations lend support to the hypothesis that strains of *B. ovatus* and *B. xylanisolvens* are in the process of losing the ability to utilize O-glycans relative to a common ancestor that possessed a more expansive gene repertoire to successfully access these nutrients. However, we cannot rule out that individual nondegrading strains are separately acquiring PULs that are associated with mucin degradation and retaining them without the full benefit that presumably occurs with the ability to fully execute this growth phenotype. This latter idea is consistent with interspecies PUL exchange observations elaborated below.

Finally, because we subsequently identified a *B. ovatus* strain (NLAE-zl-H59, red arrow in Figure 4.3B) with a substantially higher ability to use O-glycans relative to the strains used for pangenome construction, we performed an additional transcriptome sequencing (RNA-seq) analysis on this strain. Compared with a glucose reference, this strain activated 373 total genes in response to O-glycans, including genes from 30 different PULs (Table S3d). Among these PULS, 26 activated PULs were also present in 1 of the 7 strains in our pangenome and 24 were homologous to PULs in strains that did not degrade O-glycans. However, this strain did activate the expression of genes within four PULs that were completely unique to its genome compared with the seven strains used for pangenome reconstruction, suggesting that it could possess additional genes that augment its ability to grow on mucin O-glycans. This increased PUL expression could be responsible for the enhanced growth of the H59 strain on O-glycans, especially if genes included within these unique PULs are responsible for key metabolic steps required for efficient O-glycan utilization.

4.3.4 Evidence That Intergenomic Recombination Has Driven *Bacteroides* Pangenome Evolution

Similar to other bacteria, we observed that many accessory genes in the *B. ovatus* and *B. xylanisolvens* pangenome are located in contiguous clusters or "islands," often involving PULs or capsular polysaccharide synthesis gene cluster (Table S2a). In contrast to previously identified *Bacteroides* PULs that have more obviously been subjects of lateral transfer [179, 188, 189] and are associated with Integrative and conjugative element (ICE)s), most of the variable genomic regions that we identified were not associated with functions indicative of mobile DNA. Instead, these regions are often located precisely in between one or more core genes (i.e., those common to all seven strains; herein referred to as "genomic nodes") that flank each side of the variable gene segment (Figure 4.5A and B).

Several intergenomic transfer mechanisms might account for the observed mosaic structure of the *B. ovatus*-*B. xylanisolvens* pangenome. The first is the movement of genes into a recipient genome by conjugation of mobile ICEs. While such events would be expected to leave behind residual genes involved in mobilization and transfer, which were not observed, these DNA vehicles are known to target a subset of core genes, such as tRNAs [189], and may have undergone subsequent genomic deletion events that eliminated the mobile DNA. Two other known mechanisms of bacterial HGT are natural competence and phage transduction, of which neither has been

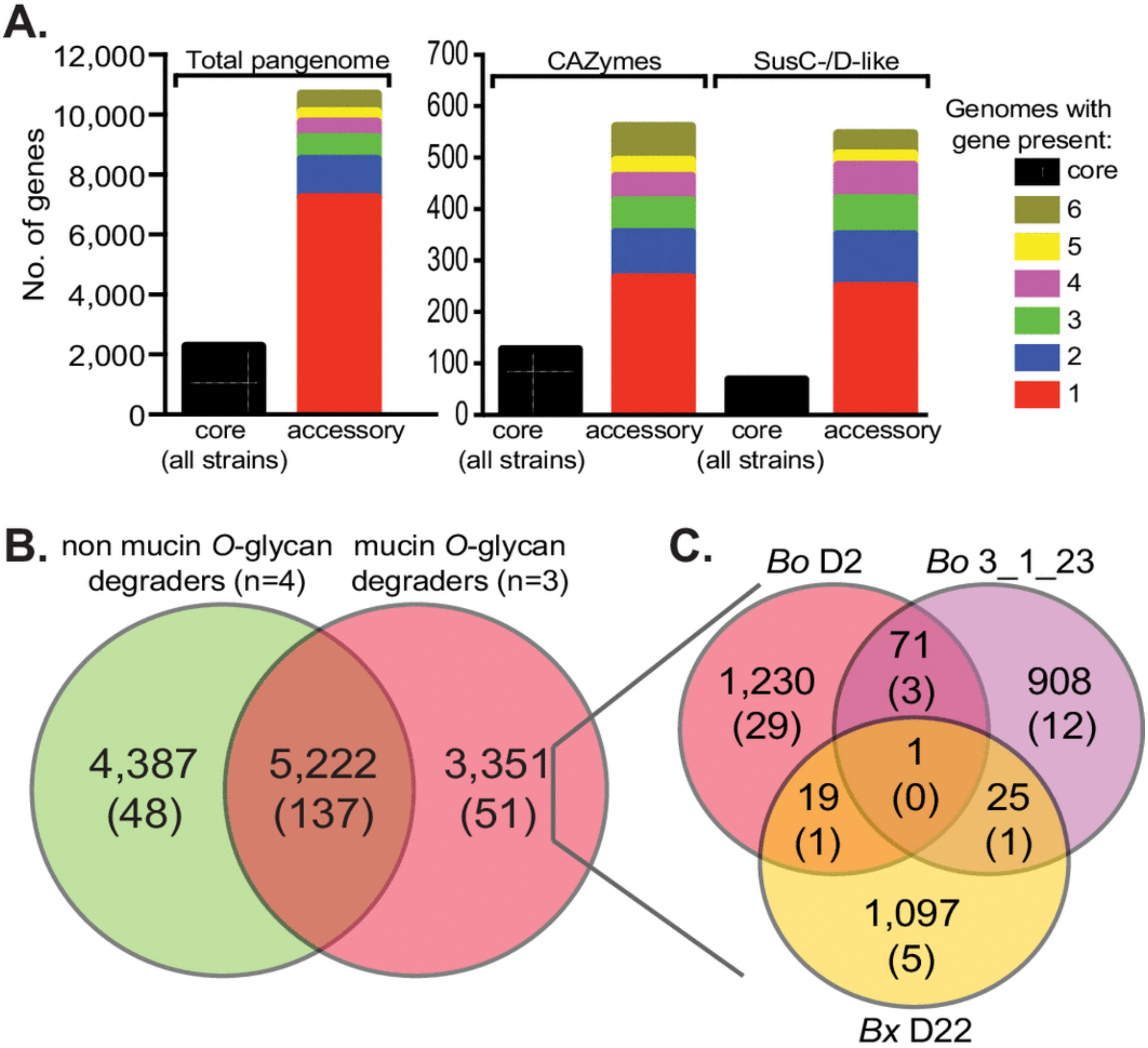


Figure 4.4: Distribution of all genes as well as core polysaccharide utilization functions in the *B. ovatus*/*B. xylanisolvens* pangenome. (A) Left, shows the number of core genes (i.e., those present in all 7 strains used for pangenome construction) compared with genes present in 2 to 7 of the individual strains. Right, shows the same distribution of genes assigned to PULs or particular degradative CAZyme families (GH, PL, and CE) (see Tables S2 and S3 for more detailed assignments). (B) The distribution of genes between mucin-degrading (n=3) and nondegrading (n=4) strains used to construct the pangenome. Top numbers indicate total genes, while numbers in parentheses indicate the number of PULs (not individual PUL genes) in each category. (C) Distribution of the genes that are unique to the three mucin-degrading strains within each genome. Genes/PULs are numbered as described for panel B. Note that no PULs are shared by all three strains.

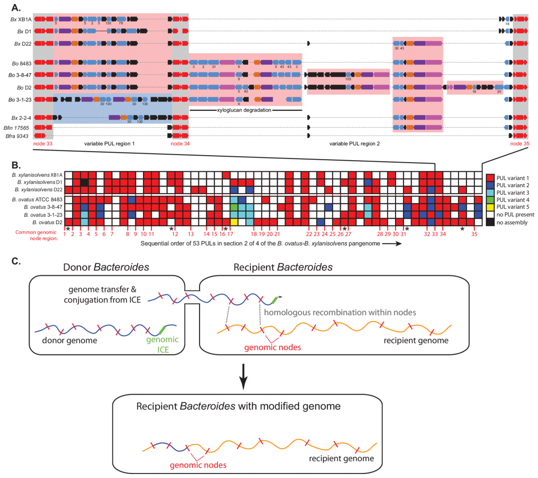


Figure 4.5: Pangenome diversification in *B. ovatus* and *B. xylanisolvens*. (A) A higher-resolution view of a region of the *B. ovatus* and *B. xylanisolvens* pangenome shows the variable presence of at least 6 different PULs occurring between 3 genomic nodes (nodes 33 to 35 in this quarter of the total pangenome). Segment 2 of the physical pangenome map was selected because the first segment was initiated with numerous small contigs and this segment contained previously validated genes for xyloglucan metabolism [190]. Node genes are colored red; while susC-like and susD-like genes are colored purple and orange, respectively; and glycoside hydrolase genes in light blue. GH family numbers are given below select PULs starting from the top to indicate potential specificity, and new numbers are only added going down the schematic if the family assignments are different, indicating a different PUL. A well-studied *B. ovatus* PUL for xyloglucan degradation [190] is shown in the center and occurs variably between two nodes and also has variable gene content. The two bottom genomes are from different species, namely, *Bacteroides finegoldii* (Bfin) and *Bacteroides fragilis* (Bfra) and show less complex genome architecture with the *Bacteroides fragilis* region possessing no PULs. (B) A broader view of the genome region in panel A, showing that the same mosaic pattern is common across the pangenome. Only PULs are illustrated, although many other genes were also variable in these regions. The numbers at the bottom delineate the presence of 35 different core gene nodes (as in panel A, some nodes contain multiple core genes) in this section of the genome, and the presence of homologous or unique PULs is illustrated according to the color code at right (see Figure B6 for high-resolution physical maps of the pangenome with PUL annotations). Note that in some cases up to five different PULs were located at one location.(C) A schematic showing the proposed mechanism of genome exchange based on previous studies ([191, 192, 193]) and observations presented here. Genomic ICEs that are either partially active (excision deficient but capable of initiating DNA strand breakage and conjugation) or activated in trans by the presence of an exogenous conjugative transposon initiate genome mobilization from a donor into a recipient. If sufficient homology between node genes exists in the recipient, homologous recombination between two nodes can replace a section of the recipient with a segment from the donor. Note that genomic regions are shown as linear fragments for simplicity but would be circular.

observed in members of *Bacteroidetes*.

A final potential mechanism is the direct conjugation of the chromosome from a donor bacterium into a related recipient, followed by subsequent homologous recombination between flanking nodes to add or delete intervening DNA in the recipient genome (Figure 4.5C). This mechanism is conceptually similar to high-frequency recombination (Hfr) transfer in Escherichia coli and has already been described for *B. thetaiotaomicron* and *B. fragilis*. The mechanism involves chromosomal ICE that may have lost their ability to circularize from the genome and instead act as transfer initiation points to conjugate a donor genome into a recipient, sometimes in response to the activity of other ICE or conjugative transposons sometimes in response to the activity [191, 192, 193]. If such a mechanism was active more broadly in HGT between Bacteroides, we would expect that some of the core/node genes involved would reflect sequence identities that were more similar to the donor bacterium from which they originated and this difference would be detectable more easily if the transfer was between members of different species like *B. ovatus* and *B. xylanisolvens*. Moreover, such transfer events could result either in the introduction of new genes into the recipient or elimination of genes depending on the genetic content in between recombination nodes from the donor chromosome.

To test this hypothesis, we took a bioinformatics approach aimed at first identifying high-confidence examples of interspecies recombination involving core genes and then assessed whether those genes were associated with the cotransfer of adjacent or intervening accessory genes (Figure 4.6A). We collected a data set of 33 *B. ovatus* and *B. xylanisolvens* genomes, which represent a subsample of the isolates for which we generated phenotypic data. We identified a set of 1,384 core genes—expectedly smaller than the core genome of the 7 strains used above due to additional strains being added—that are present as a single copy in all members of both species. To identify cases of putative interspecies HGT via homologous recombination at core genes, we searched for instances in which a core gene sequence from either species was more similar to the corresponding gene in the other species. We calculated the median distance of each strain-specific core gene to all other alleles of that core gene in strains belonging to both species (Figure 4.6B, blue and red boxes indicate the core genes that are more similar to alleles in the other species). Among these candidate HGT genes/loci, we then investigated if any of these putative transfer events have resulted in pangenome diversification by searching for the presence of any accessory gene(s) that was observed only adjacent to a core gene with evidence of HGT.

In total, we identified 29 different loci at which the exchange of core genes appeared to have occurred and adjacent accessory genes were identified, including 7 that appeared to involve the transfer of PULs (Figure 4.7A, see Figure B8 in the supplemental material). Similar numbers of potentially transferred loci were identified for each species (16 loci in *B. xylanisolvens* and 13 loci in *B. ovatus*). Among the candidate HGT events, variable numbers of accessory genes were

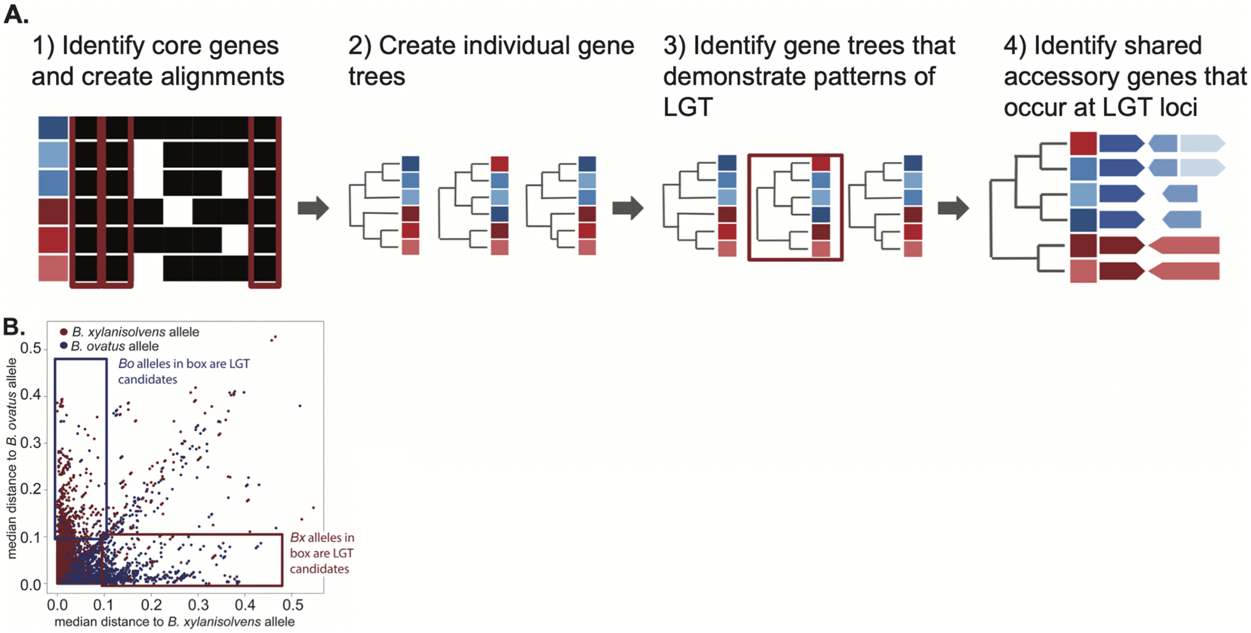

A.
1) Identify core genes and create alignments
2) Create individual gene trees
3) Identify gene trees that demonstrate patterns of LGT
4) Identify shared accessory genes that occur at LGT loci
B.
● B. xylanisolvens allele
● B. ovatus allele
Bo alleles in box are LGT candidates
Bx alleles in box are LGT candidates
median distance to B. ovatus allele
0.0
0.1
0.2
0.3
0.4
0.5
median distance to B. xylanisolvens allele
0.0
0.1
0.2
0.3
0.4
0.5

Figure 4.6: (A) Schematic of the workflow to identify putative HGT core genes, which is described as follows: align genes and build corresponding trees for each core gene, determine the median substitution distances for each allele of a core gene in a given strain to both species, and identify loci with an identical conserved structure between isolates of opposite species. (B) Plot of median distances for all core genes identified in the 33 genomes analyzed. The boxes show the regions containing genes for which the median distance was >0.1 to the assigned species for a given strain and <=0.1 for the opposite species to which a strain is assigned. These genes were determined to be high-confidence examples of core/node genes that had been replaced by an allele from the other species.

found within the loci ranging from 1 to 13 genes (Figure 4.7A, Figure B8). More genes (57 total) appeared to be transferred into *B. ovatus* than into *B. xylanisolvens* (36 total).

Finally, we determined if any of the identified HGT events could explain differential phenotypes measured by our high-throughput growth assay by modifyingdddd the complement of PULs in individual genomes. As a specific example, we focused on a PUL that was associated previously with β-mannan degradation [170, 194] that was among our candidate loci with evidence of transfer from a *B. xylanisolvens* ancestor into two *B. ovatus* strains. The presence of this PUL (PUL-A in Figure 4.7A and B) was observed in all strains with the ability to grow on the β-mannan galactomannan (GalM), including two strains of *B. ovatus* (ATCC 8483 and CL02T12C04) for which the flanking node regions were more similar to *B. xylanisolvens*. We showed previously that the deletion of this PUL from *B. ovatus* ATCC 8483 eliminated growth on GalM and glucomannan (GluM) [194], suggesting that it was both acquired from a *B. xylanisolvens* strain and conferred growth on these two β-mannans. However, the presence of this PUL was not correlated perfectly with growth on GalM, and several strains that lacked PUL-A still exhibited robust growth. Thus, we searched for other PULs that harbor GH26 family enzymes and determined that all of the other strains that grow on GalM, but lack PUL-A, harbor another candidate GalM PUL (PUL-B, Figure 4.7B) at a different genomic location and some strains possess both (Figure 4.7A). Gene expression analysis by quantitative PCR (qPCR) revealed that PUL-B was expressed highly in strains that lacked PUL-A during growth in GalM (Figure 4.7C) and every strain that grew robustly on GalM had at least one of these two PULs. While we had previously shown that PUL-A was required for GlcM growth in *B. ovatus* ATCC 8483, there were a number of other strains (red "+" symbols in Figure 4.7A) that displayed a weaker ability to grow only on GlcM, while lacking both of the GalM-associated PULs, suggesting the presence of additional PULs that confer the ability to grow on variant β-mannans. Such a presence of multiple nonorthologous PULs that confer the same or similar functions, and some which may be moving between genomes of related species by the putative HGT mechanisms noted above, complicates the process of understanding the genotype-phenotype relationships in human gut *Bacteroidetes* but will need to be resolved to make better functional predictions from sequence-based data.

4.4 Discussion

In this study, we leveraged a scalable, high-throughput quantitative growth platform to characterize the phenotypic abilities that are present in a sample of hundreds of *Bacteroidetes* strains from the human and animal gut. Our anaerobic screening technique is directly applicable to other bacterial phyla from the human gut and other environments. Moreover, it can be adapted to include new polysaccharides or to focus on different nutrient utilization or chemical resistance phenotypes.

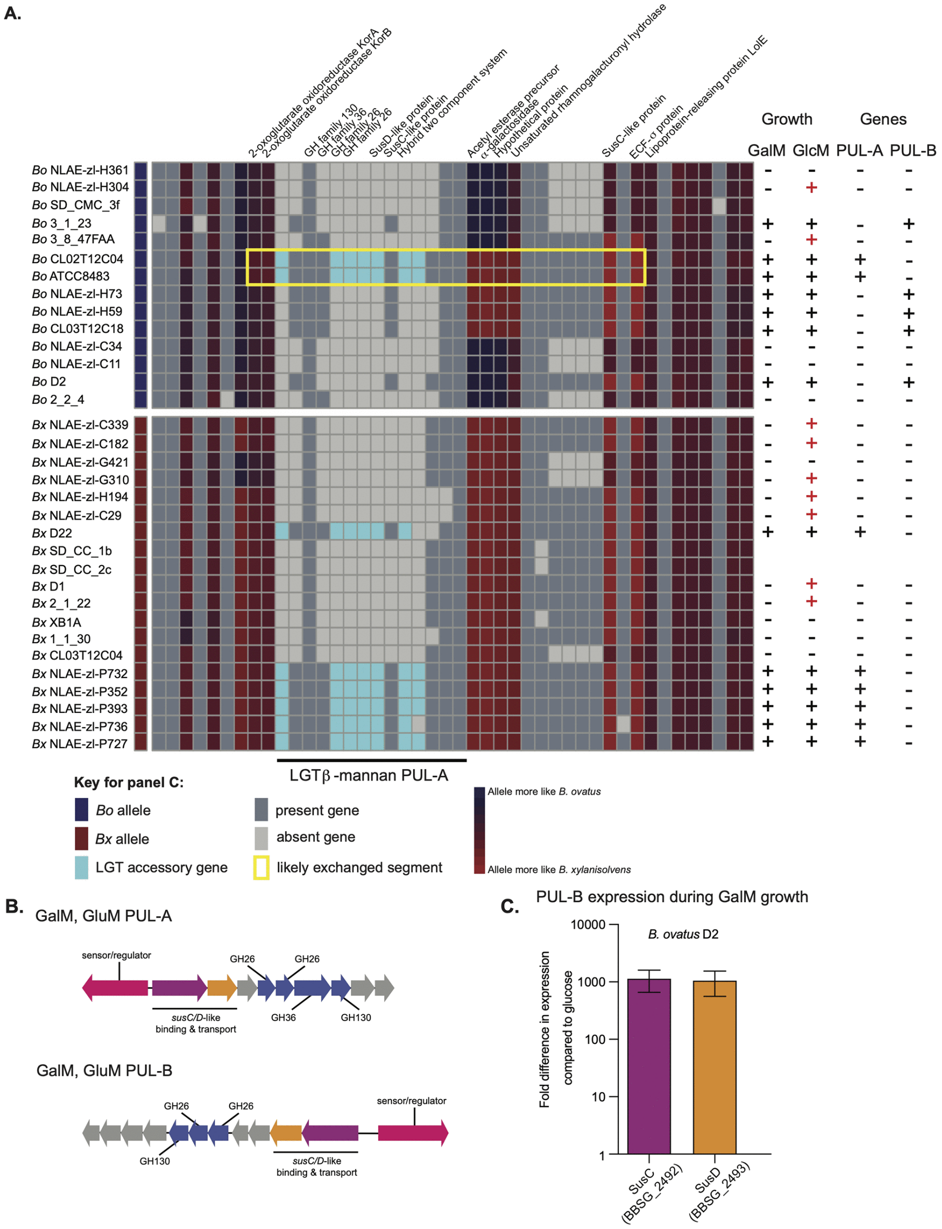


Figure 4.7: Evidence that a PUL for β-mannan metabolism has been laterally transferred into *B. ovatus*. (A) A region of the *B. ovatus*/*B. xylanisolvens* pangenome that contains a PUL involved in galactomannan (GalM) and glucomannan (GluM) degradation. This PUL is present in six strains of *B. xylanisolvens* and two strains of *B. ovatus*, and in the latter cases, flanking node genes exhibit signatures of being derived from HGT with a *B. xylanisolvens* donor (the yellow box highlights a potential recombination region). The columns at the left indicate the growth of each strain on GalM or GluM. The ability to grow on GalM is correlated fully with the presence of one of two different PULs, or both, that are transcriptionally activated during growth on this substrate (23). Notably, some strains (red "+") are able to grow weakly on GluM but do not possess either of the identified PULs, suggesting that additional, partially orthologous PULs exist that confer the ability to use only GluM. (B) Schematics of PUL-A and PUL-B associated with GalM and GlcM utilization. In *B. ovatus* ATCC 8384, elimination of PUL-A eliminates both of these growth abilities. (C) Expression analysis by qPCR of two sentinel genes from PUL-B in*B. ovatus* strain D2 that lacks PUL-A but still exhibits robust growth on GalM.

The current study, in concert with future applications of phenotypic screening, will help close the gap between our largely sequence-based view of the human gut microbiota and the functions that its members provide. However, instances like the ones investigated here for mucin glycan and β-mannan utilization by *Bacteroidetes*serve as a warning that the presence or absence of genes that are associated experimentally with a particular function do not always indicate that the phenotype is expressed or not. Pangenome reconstruction for *B. ovatus* and *B. xylanisolvens* revealed extensive variability between strains of these closely related species, which is not unexpected for bacteria that engage in HGT. However, the lack of mobile DNA signatures for the majority of accessory genes and evidence of intergenomic recombination between species at core genes provide new insight into what may be a prominent mechanism of genome diversification in members of this phylum. The previously described intergenomic transfer mechanisms in *B. thetaiotaomicron* and *B. fragilis* required the presence of active or inactive ICEs, highlighting the potential roles for these mobile elements in not just shaping genomes directly but also indirectly through their ability to catalyze the exchange of broader genomic segments. In *B. thetaiotaomicron*, genome transfer was determined to initiate at genomically integrated ICEs of which there are four in the type strain of *B. thetaiotaomicron* (VPI-5482). They have not been shown to be fully functional for circularization and mobilization. However, the introduction and activation of an additional, excision-proficient conjugative transposon (either cTnDOT or cTnERL) [191], which shares common features with the genomic ICEs, catalyzed the expression of genes in the genomic ICEs and transfer of parts of the genome in a manner that requires recA and homologous DNA to be present in the recipient [191]. An additional study in *B. fragilis* showed that conjugation from a strain with multiple genomic ICEs, with one or more presumably retaining transfer activity, results in the transfer of up to 435 Kb of chromosome into a recipient that initiates near genomic ICEs, with individual transfer events being of variable size. The latter observation suggests that intergenomic recombination could then occur at different homologous regions (i.e., the core gene nodes observed in the pangenome), which could depend on the amount of genomic DNA transferred and the length/homology of available recombination sites. Given that the number of ICEs in individual genomes is variable and their ability to be activated by functional conjugative transposons that are circulating in the ecosystem may also vary, it will be interesting to determine in future work if there are hot spots for genome transfer or if certain strains/species are dominant genome donors that could play a disproportionate role.

The phenotypic similarity between members of the same species (e.g., *B. ovatus* and *B. xylanisolvens*) and the large amount of gene diversity, including genes involved in carbohydrate metabolism, present a paradox and raise the question of why the genome diversification observed in strains of *B. ovatus* and *B. xylanisolvens* has not pushed members of these species to behave more differently and cluster based on phenotype with members of other species. One answer may be the apparent exclusion of some traits, such as mucin O-glycan/hemicellulose metabolism, which may limit the fitness advantage associated with acquiring new phenotypes. A second emerges from the proposed genome-exchange mechanism for which we offer new bioinformatics support. Since this intergenomic exchange relies on homologous recombination, its frequency should decrease between genomes that are more divergent. Thus, this strategy may be one mechanism through which only closely related bacteria can share traits that are advantageous with other close relatives. The presence of nonorthologous PULs that confer the same function (e.g., GluM and GalM utilization), of which some appear to be subjected to HGT, further complicates interpretations of genotype-to-phenotype relationships in these bacteria. Based on the prevalence data, it seems that PUL-A is a GalMan utilization system that is more prevalent in, and perhaps also originated in, *B. xylanisolvens*, and it is also capable of transfer to *B. ovatus*. PUL-B is more prevalent in *B. ovatus* and may have origins in that species, at least with respect to *B. xylanisolvens* where it has so far not been observed. Notably, the genome transfer mechanism proposed here does not account for how new genes can be incorporated between conserved nodes. Rather, this variability must pre-exist among different strains and therefore be created by different inter- and intragenomic diversification mechanisms. Nevertheless, the data that we report here underscore the notion that individual gut symbiont genomes are not just highly variable but also dynamically so.

4.5 Materials and Methods

4.5.1 Bacterial Strains and Growth Conditions.

A total of 354 human and animal gut *Bacteroidetes* were included in this study. A complete list is provided in Table S1b, along with species designation based on 16S rRNA gene sequencing and associated metadata. Abigail Salyers (University of Illinois, Urbana-Champagne) kindly provided many of the strains, and 2 large portions of this collection were isolated over several decades, as follows: 99 strains with “WH” designations were collected from fecal samples of healthy human volunteers as part of the Woods Hole Summer Course on Microbial Diversity in the late 1990s, and 95 additional strains with “VPI” designations were collected from human samples at the Virginia Polytechnic Institute in the 1960s to 1970s. Species classifications were made based on alignment of a minimum of 734 bp of 16S rRNA gene sequence to a database containing the type strains of >29 named human gut *Bacteroidetes* species using the classify.seqs command with Bayesian settings in the program mothur [195]; assignment for each strain was also checked manually by BLAST [196]. Isolates with at least 98% 16 rRNA gene sequence identity to the type strain of a named species were labeled with that species designation. This classification strategy included all except for 3 of the 354 strains examined, which ranged between 96.6% and 96.7% sequence identity to the B. uniformis ATCC type strains, and based on sequential isolate numbers might be clones from the same individual (see WH15, WH16, and WH17 entries in Table S1a). Because of the small number of strains that did not satisfy our 98 % cutoff, we grouped these unclassified strains with their nearest relative and labeled them as more divergent in Table S1a; although, in most cases, the carbohydrate phenotypes of these strains were very similar to other members of the B. uniformis group. All strains were grown routinely in an anaerobic chamber (Coy Lab Products, Grass Lake, MI) at 37°C under an atmosphere of 5% H2, 5% CO2, and 90% N2 on brain heart infusion (BHI; Beckton Dickinson) agar that included 10% defibrinated horse blood (Colorado Serum Co.) and gentamicin (200 $micro$g/mL). A single colony was picked into either tryptone-yeast extract-glucose (TYG) media [197] or modified chopped-meat carbohydrate broth (Table S1b) and then subcultured into a minimal medium (MM) formulation that contained a mixture of monosaccharides, vitamins, nucleotides, amino acids, and trace minerals (Table S1b provides components and a complete recipe).

4.5.2 Carbohydrate Growth Array Setup and Data Collection

Two different minimal medium formulations were used in the carbohydrate growth arrays (Table S1a lists the formulation used for each isolate). The simpler of the two formulations (medium 1) was identical to the above MM, except that no carbohydrates were included and the medium was prepared at a 2× concentration. The second minimal medium formulation (medium 2) was identical to medium 1 but included beef extract (0.5% [wt/vol] final concentration) as an additional supplement. We initially attempted to cultivate all of the species tested using only medium 1 but determined that beef extract was specifically required to allow the growth of some species, especially *Parabacteroides spp., Barnesiella intestinihominis, Odoribacter splanchnicus*, and the branch of *Bacteroides* that includes *Bacteroides plebeius* and *B. massiliensis*. Growth in the absence of an added carbohydrate source was generally not observed or very low, except with *Parabacteroides* that were often able to grow to a low level on the added 0.5% beef extract. The corresponding negative-control wells for each strain assayed were averaged, and this value was subtracted from the total growth calculation of the corresponding to strain on other carbohydrates tested (all raw growth curves are provided as source data). Despite several attempts to supplement minimal media with different components or employ more stringent anaerobic methods, we were unable to cultivate several common *Bacteroidetes* genera/species (*Prevotella spp., Paraprevotella spp., Alistipes spp.*, and *Bacteroides coprocola and Bacteroides coprophilus*) in these two MM formulations and therefore did not include them in this study. All of these isolates grew readily in rich medium, suggesting that they have specific nutritional requirements that were not met in the MM formulations used.

Carbohydrate growth arrays were run as described previously [170] using a list of carbohydrates (see reference 23 for a complete list with supplier information) that were present in duplicate, nonadjacent wells of a 96-well plate; 2 additional wells contained no carbohydrate and served as negative controls. Each MM was prepared as a 2× concentrated stock without carbohydrates (MM-no carb). An aliquot of each strain was taken from a MM-monosaccharides culture (grown for 16 to 20 h) and was centrifuged to pellet cells. Bacteria were resuspended in the same volume of 2× MM-no carb and then centrifuged again prior to suspension in a volume of 2× MM-no carb that was equal to the original volume. These washed bacterial cells were then inoculated at a 1:50 ratio into 2× MM-no carb, and the suspension was added in equal volume ($100 micro$L/well) to the 96 wells of the carbohydrate growth array. Each well of the carbohydrate growth array contained $100 micro$L of 2× carbohydrate stock (10 to 20 mg/mL); thus, when diluted 2-fold, it resulted in 1× MM containing a unique carbohydrate and a bacterial inoculum that was identical to other wells. Growth arrays were monitored at kinetic intervals of 10 to 20 minutes using a microplate stacking device and coupled absorbance reader (Biotek Instruments, Winooski, VT), and data were recorded for 4 d (variable kinetic interval times reflect variations in the number of microtiter plates present in a given batch).

4.5.3 Carbohydrate Growth Array Data Processing.

Growth data were processed according to the following workflow: (i) data for each strain were exported from Gen5 software (Biotek Instruments, Winooski, VT) into Microsoft Excel and a previously described automated script was employed to call the points at which growth began (min) and ended (max) [170], (ii) each file was checked manually to validate that appropriate calls were made and the min and max values edited if needed (generally, only due to obvious baselining artifacts or erroneously high calls caused by temporary bubbles or precipitation); (iii) "total growth" (A600 max - A600 min) and "growth rate" [(A600 max - A600 min)/(t max - t min)] were calculated for each strain on each substrate (A600 is the absorbance value at 600 nm that corresponds to each min and max point; t is the corresponding time values in minutes; when necessary, the growth level associated with the average negative-control growth was subtracted from the total growth value), and (iv) individual cultures in which total growth was $<= 0.1$ were scored as "no growth" and their A600 values converted to 0. Only assays in which both replicates showed an increase in A600 of $>=$0.1 were considered growth; if the 2 replicate assays were discordant (one positive, one negative), then both values were converted to 0.

To normalize the results for each strain, the substrate(s) that provided maximum total growth and growth rate values were determined, and they were set to 1.0. All other growth values for a given strain were normalized to this maximum value, providing a range of values between 0 and

1.0. We next normalized growth ability across individual substrates using the previously normalized values for each individual strain; the strain with the maximum total growth and growth rate values were identified (many of these were already set to 1.0). Then, the corresponding values for each other species on that particular substrate were calculated as a fraction of the maximum value for that substrate, yielding a range of values between 0 and 1.0 for each substrate. These values were used to create the heat map shown in Figure 4.2 and Figure B3, and all raw and normalized values are provided in Table S1a.

4.5.4 Data Clustering and Statistics.

Heatmaps and corresponding dendrograms were generated using the "heatmap" function in the "stats" package of R (version 3.4.0) which employs unsupervised hierarchical clustering (complete linkage method) to group similar carbohydrate growth profiles. Pearson correlation was used to calculate the co-occurrence of the ability to grow on each pair of different substrates. The normalized growth value for each substrate was compared with the corresponding growth values on all other substrates using the Pearson correlation test in R, and these values are displayed in the Pearson correlation plot in Figure B5.

4.5.5 Pangenome Reconstruction for *B. ovatus* and *B. xylanisolvens* Strains.

Since one of the seven strains used for pangenome reconstruction (*B. xylanisolvens* XB1A) was assembled into a single circular chromosome, we used this genome as a scaffold for the contigs representing the remaining six strains. Contigs from the six unfinished strains were aligned against the XB1A genome using a combination of Mauve (49), to align and orient larger contigs, and reciprocal best BLAST-hit analysis using $\geq$ 90% amino acid identity to identify likely homologs, to provide finer resolution. Contigs from draft genome assemblies or *B. xylanisolvens* XB1A were broken as needed to accommodate the inclusion of unique accessory genes but only in circumstances where genes on both sides of the break could be aligned to homologs in one or more genomes with a contig that spanned that breakpoint. After constructing a preliminary assembly, we analyzed the size distribution of putative homologous open reading frames (ORFs) as a measure of assembly accuracy and to identify variations in genetic organization that might be attributable to real genetic differences such as frame shifts, which would result in two homologous gene calls of smaller size in the genome containing the frameshift. Any variation in $>$50% of homologous ORF size was inspected manually using the "orthologous neighborhood viewer, by best BLAST hit" function in the U.S. Dept. of Energy Integrated Microbial Genomes (IMG) website. Introduced contig breaks are documented in Table S2a. GenVision software (DNAstar, Madison, WI) was used to visualize and label selected functions in the pangenome assembly and also display RNA-seq data as a function of shared and unique PULs. Downloadable physical maps of the reconstructed pangenome are provided online at https://www.ericmartenslab.org/people.

4.5.6 RNA-seq analysis

For RNA-seq, *B. xylanisolvens* and *B. ovatus* cells were grown to mid-exponential phase on either purified mucin O-linked glycans (purified in-house from Sigma type III porcine gastric mucin) or glucose as a reference as previously described[170]. Total RNA was extracted using an RNeasy kit (Qiagen) and treated with Turbo DNase I (Ambion), and mRNA was enriched using the bacterial Ribo-Zero rRNA removal kit (Epicentre). Residual mRNA was converted to sequencing libraries using TruSeq barcoded adaptors (Illumina) and sequenced at the University of Michigan Sequencing Core in an Illumina HiSeq instrument with 24 samples multiplexed per lane. Barcoded data were demultiplexed and analyzed using the Arraystar software package with Qseq (DNAstar). All RNA-seq data are available publicly from the National Institutes of Health Gene Expression Omnibus Database under accession numbers GSM4714867 to GSM4714890.

4.5.7 Core Gene Determination and Detection of HGT Events Between *B. Ovatus* and *B. Xylanisolvens* Strains

The core gene alignment was generated with cognac [145]. The alignment was then partitioned into the individual component genes, and approximate maximum likelihood gene trees were generated with FastTree [128]. Cophylogenetic distances were calculated with ape [198]. A distance threshold of greater than 0.1 to the same species and less than 0.1 to the opposite species was used to identify alleles bearing signatures of HGT. All analyses were performed in R (version 3.6.3) [149]. All code developed for this project are available online at https://github.com/rdcrawford/bacteroides_hgt.

CHAPTER 5

Conclusions

5.1 Concluding Remarks

The vast quantity of bacterial whole-genome sequencing data presents enormous opportunities to address basic and translational research questions. The number of microbial genome sequences is growing exponentially; more than 1.2 million bacterial genomes are publicly available in GenBank for the scientific community to access [199]. These data collectively represent an enormous potential to understand the dynamics of microbial genome evolution, provide a better understanding of the genetic determinants of virulence, the transmission of pathogenic isolates across large geographic distances, and genome structure and function in commensal microbes and pathogenic isolates alike. Seizing the opportunity presented by these data to answer these important scientific questions will require bioinformatics tools capable of analyzing massive datasets to extract valuable insights.

This dissertation provides a methodological framework to leverage large bacterial whole-genome sequencing datasets to facilitate the study of bacterial genome evolution. We also illustrate the utility of these computational tools with applications to the analysis of HGT in two important contexts: dissemination of antibiotic resistance genes across regional healthcare networks and HGT resulting in expansion of the pan-genome in commensal members of the human microbiome. By developing these methods and providing these proof-of-principal analyses, this dissertation provides a framework facilitating future research in bacterial genomics and the genomic epidemiology of infectious disease.

5.1.1 Large-scale Phylogenetic Analysis

Methods

Chapter Two describes our R package cognac, a tool for generating concatenated gene alignments for phylogenetic analysis of bacterial whole-genome sequencing data. We illustrate that cognac can efficiently generate alignments for extremely large genomic datasets. Unlike other tools capable of handling large numbers of genomes, cognac has the benefit of generating an alignment that can be used in downstream phylogenetic analysis to gain insights into molecular evolution, and not only an estimate of genetic distance, as in alternative approaches [200, 128].

Results and Implications

We illustrate the utility of cognac on several datasets of varying compositions representing diverse genre of Enterobacteriaceae. Despite the large numbers and diversity of genomes in the input datasets, cognac was able to identify shared phylogenetic marker genes for these data and efficiently generate an alignment that can be used for downstream phylogenetic analysis. Generating the individual gene alignments with only the representative unique alleles in the input dataset was highly effective for reducing the computational overhead associated with multiple sequence alignment of large numbers of sequences. Additionally, we incorporated multithreading at multiple steps, further enhancing speed. These design features enable the application of this algorithm to large genomic datasets while maintaining a high degree of efficiency.

This dissertation also describes use cases for this software illustrating its utility for genomic analysis in two contexts. Chapter Four describes a method leveraging cognac alignments to identify recombinogenic core genes. These shared genes may facilitate recombination of these HGT sequences, enabling pan-genome expansion and phenotypic evolution in Bacteroides. We show that these chromosomal HGT events result in the mosaic architecture observed in these species, identifying many resulting from HGT between two closely related species. In Chapter Three, we further illustrate that this tool can generate alignments for enormous datasets: applying this software to a dataset of over 72,000 microbial whole-genome sequences from the family *Enterobactiacae*. Additionally, in our previous work we have illustrated the utility of cognac for diverse applications, including: highly clonal datasets of *Klebsiella pneumoniae* carrying the NDM carbapenemase and highly diverse isolates from the order *Bacteroidales* [123, 201].

In summary, we present a broadly applicable software package with an easy-to-use interface in R, which is useful for a wide range of applications in computational genomics.

5.1.2 HGT of KPC in Clinical Isolates

Methods

In Chapter Three, we present an analysis of the transmission of KPC plasmids across regional healthcare networks in three US states. This study represents the first large-scale analysis of HGT across different geographic regions. To make this analysis possible, we implemented a novel phylogenetic method to facilitate the identification of core plasmid genes, effectively serving as markers of KPC plasmid presence. Previous work has shown that stable marker genes of plasmid presence can be used to characterize KPC plasmids; however, identification of plasmid marker genes is nontrivial [202]. This method enables a generalized, data-driven approach to identify plasmid marker genes, which we illustrate can be used to track the spread of diverse KPC plasmids as they are transmitted through bacterial populations. These plasmids facilitate the introduction of KPC into novel genetic contexts, including lineages that are not strongly associated with KPC but maintain KPC plasmids, express carbapenem-resistant phenotype, and are capable of causing multiple independent infections.

Results and Implications

Our phylogenetic approach for identifying plasmid transmission enabled us to estimate the point in time where plasmid transmission occurred in different lineages across the phylogenetic tree. We use this information to classify isolates: whether there is evidence of importation, a stable association of a linage and plasmid before introduction into a state, and recent HGT events likely occurring within a given region. While most KPC carrying isolates within a given state represent imported lineages with prior association with KPC, we identified a substantial proportion of infection isolates that likely recently acquired KPC. Occasionally, these recent HGT lineages exhibit evidence of transmission to other patients. For example, *Enterobacter cloacae* ST171 in MN revealed evidence of HGT of KPC plasmids within the region. These strains maintained a long-term, stable relationship with the plasmids they carry. These KPC-carrying clonal lineages then spread broadly through the healthcare network to many different facilities in the state. These results highlight the importance of genomic surveillance to monitor the prevalence of antibiotic-resistant clonal lineages as they spread throughout the healthcare network. Furthermore, this highlights the potential for plasmids to spread from high-risk clonal lineages into diverse bacterial populations, which can cause antibiotic-resistant infections.

5.1.3 Phenotypic Adaptaion via HGT in the Microbiome

Methods

Chapter Four presents an analysis of phenotypic and genotypic diversification in *Bacteroides* species with microbiological and genomic data. Phenotypic data revealed substantial diversity in carbohydrate utilization phenotypes. Frequently carbohydrate utilization profiles were conserved within specific species or closely related species indicating that these species have evolved to occupy a specific metabolic niche. However, this was not always the case, whereby closely related species exhibited highly varied carbohydrate utilization profiles. In particular, we examined two closely related species, *Bacteroides ovatus* and *Bacteroides xylanisolvens*. By comparing the gene distance distributions between alleles belonging to each species respectfully, we could identify core genes, which were highly divergent from the alleles of the same species, and had a high degree of similarity to alleles isolated from the opposite species. Detailed examination of the loci containing these putative HGT genes yielded several instances where there were multiple core genes with similar patterns of variation at the same locus. Additionally, we identified accessory genes with conserved synteny and colinearity at these HGT loci representing pan-genome expansion events. This analysis revealed many loci that exhibit evidence of HGT between these two species and identified instances of these events that enabled new metabolic phenotypes.

Results and Implications

These data provide an exploration of the phenotypic capacity of microbes and the genomic events underlying polysaccharide utilization by members of the human microbiota. Our simple method for identifying HGT loci by comparing distance distributions in core genes identified many loci which had been exchanged by two closely related species of *Bacteroides*. Specifically, we identified a polysaccharide utilization locus for β-mannan utilization, which was common in *B. xylanisolvens* and present in a subset of *B. ovatus*. In *B. ovatus* these accessory genes were only present in instances where there was evidence of HGT in the core genes flanking these genes. Molecular characterization of growth on *Bacteroides*.Specifically, we identified a polysaccharide utilization locus for β-mannans and revealed that these genes mediated robust growth on this substrate. *Bacteroides*, including the species represented in this study, represent fundamental members of the microbiota. Frequently, these strains can have acute beneficial effects on human physiology and fill fundamental roles in the metabolism of a broad range of substrates, including the polysaccharides discussed here. Understanding the underlying genomic events that shape the structure and function of the microbiota will further enhance our understanding of the functional composition of healthy microbial communities and how this interfaces with the host to improve human health.

5.2 Future Directions

This dissertation led to the development of three open-source methods for the study of bacterial genomics. The *cognac* package is a resource with diverse applications in bacterial genomics and has been used in various contexts in computational genomics and the genomic epidemiology of infectious disease. Next, the method for the identification of genes that serve as specific markers of plasmid presence could be applied to diverse questions in the area of plasmid biology and bacterial genome evolution in many different contexts. Finally, the method for identifying core genes has broad application to many species and can be further used to study the composite nature of bacterial genomes. Wholistically, these methods represent tools that can be used directly by the genomics community to address many important questions and represent stepping-stones that can be further built upon to extend their utility further. In the following sections, I describe further improvements that could be made to these methods and potential future applications.

5.2.1 Chapter Two: Future Directions For *cognac*

As demonstrated in chapter 2, *cognac* can generate alignments for large sets of bacterial genomes, with customizable parameters for extensions in diverse applications. Herein, we describe future applications for cognac and potential expansions of this package to enhance its utility.

Just as *cognac* was able to be applied to gain insights into large datasets of bacterial whole-genome sequences, this could further be applied to other domains of the tree of life. While there are many methods available for generating alignments, *cognac* represents a rapid method for generating alignments with demonstrated ability to handle enormous datasets. This can be applied in many applications better to understand the structure of datasets of whole-genome sequences before subsequent analyses. For example, this could be used to divide an input dataset into appropriate subsets for traditional read-mapping-based genome alignment. Additionally, complete genomes could be included in the cognac alignment step to identify an optimal reference genome for a given set of genomes. Additionally, finding an optimal out-group for phylogenetic analysis is non-trivial, and *cognac* could be used to identify an optimal out-group based on distance metrics in the core genes.

While *cognac* was only tested on bacterial genomes, future applications of these methods could be in the phylogenetic analysis of other types of microbial genome sequences. The benefits of cognac for studying phylogenetically diverse bacteria could apply to studying fungi, viruses, and organelles. For example, fungi are prominent members of the human microbiome; however, they remain relatively understudied relative to their bacterial counterparts [203]. This method could potentially be applied to sets of diverse fungal genomes, especially in instances where there is an absence of a known reference genome appropriate for the analysis. Additionally, this could be used to study diverse human viruses or bacteriophages, which share a common ancestor but differ in their gene content, as has been applied previously [204, 205]. Additionally, genomic sequences of organelles, such as plastids or mitochondria, are appropriate for analysis with cognac. Concatenated gene alignments are common in the plastid literature, and cognac could be applied to sets of plastid genomes as well [206]. Future work could illustrate the utility of cognac for other types of microbial genomes to study microbial evolution and genomic epidemiology.

A limitation of *cognac* is that there is currently no built-in functionality for masking recombinogenic genes, which are readily available within the package. Recombination has been shown to influence the topology of phylogenetic trees, and therefore masking recombination would have the potential to produce trees with greater accuracy [207]. Statistical approaches for identifying recombination, such as those implemented in Gubbins, or, machine learning methods for outlier detection, such as Isolation Forest, are promising strategies for identifying recombination in future work [208].

5.2.2 Chapter Three: Future Directions for Tracing the Spread of Antibiotic Resistance in Bacterial Populations

The methods outlined in Chapter Three demonstrated the ability to identify modules of plasmid genes with stable relationships that can be used to track strains and their ARG-carrying plasmids as they are transmitted vertically and horizontally. We demonstrated the ability of the phylogenetic methods implemented here to capture plasmid acquisitions within closely related isolates, even of the same sequence type with sufficient context.

Future work using this method could delve deeper into the mechanisms governing the stability of plasmids and the bacterial lineages that carry them. Plasmids have associated fitness costs that are strain-dependent [6]. These methods could be applied to studying plasmid evolution and potentially identifying the specific features of plasmids that enhance transmissibility. In turn, a stable association with the plasmids requires an amenable host, and future work could address factors that predict a plasmid's stability in bacterial lineages. Future research could employ bacterial genome-wide association studies between strains and plasmids to identify host factors associated with plasmid presence and genetic factors representing co-evolution between chromosomes and plasmids that foster a stable relationship. An understanding of these factors would enhance the capacity to survey for plasmids and strain/plasmid associations that pose a serious risk to public health.

Public health departments increasingly use whole-genome sequencing to identify emerging threats and inform infection control practices [138]. Tools for gaining insights from this inflow of data are urgently needed for use by clinicians and public health professionals. While most epidemiological investigations focus on the transmission of clonal lineages in healthcare networks, this work illustrates the substantial burden of infections caused by isolates that have only recently acquired KPC. This work highlights the threat of highly transmissible plasmids. Dissemination of these plasmids has the potential to enhance pathogenicity and antibiotic resistance resulting in untreatable infections. Lessons learned from this dissertation can be applied to inform strategies to track the transmission of isolates and identify the dissemination of antibiotic resistance or virulence-enhancing genes as they are disseminated through bacterial populations.

Another application of this method is better understanding the distribution of mobile elements across an input dataset. Frequently, researchers will select a subset of isolates that were sequenced with short-read sequencing technologies for long-read sequencing, which provides insights into the genomic structure [39, 37]. This approach could be used to provide a first pass at characterizing plasmid sequences from short-read sequencing data and then performing long-read sequencing to confirm these results and gain a better understanding of plasmid structure.

5.2.3 Chapter Four: Future Directions for Understanding Phenotypic and Genomic Diversification in the Microbiome

The results presented in Chapter Four present methodologies to understand the phenotypic function of diverse microorganisms. A better understanding of how microbes collectively function to interface with human physiology and modify disease states are of great clinical interest. A fundamental understanding of the metabolic capacity of these microbes and their role in microbial communities can better inform strategies to manipulate microbial communities for the benefit of patients.

The method for detecting HGT loci in core genes could also be expanded further. In this work, we used a distance threshold, which was only capable of detecting HGT core genes that were greatly divergent from the other members of that species. Future work could implement more sophisticated statistical approaches to more accurately model the observed vs. expected variation within genes and provide a more accurate estimate of which genes result from HGT. Another addition to this method would be incorporating multi-species comparisons within the gene distance distributions. In our analysis, we only included two closely related species, an oversimplification of the HGT network that shapes the genome evolution in these isolates. The method outlined here could enable large-scale analysis investigating the natural history of genomes of interest to facilitate the identification of these events in a systematic way. Previous research into the genomic composition of pathogenic isolates has shown that *K. pneumonaie* ST258 is a hybrid generated by HGT of 1.1 megabase pair stretch of the ST442 genome into an ST11 genetic background [16]. The application of our method for analysis of genetic variation on the chromosome could address these questions about the genomic landscape shaping the evolution in these instances. Further work could use this framework to compare the distribution of gene distances across multiple species or strains. These analyses could identify which genes are commonly transferred, which genes are well maintained, which strains exchange genes frequently, and infer the functions within a shared niche within this HGT network. Large-scale analysis of this type could be used to determine the origins of HGT sequences in the genomes and provide insight into the genomic events underlying the evolution of bacterial lineages in diverse contexts.

5.3 Conclusions

The work presented in this dissertation was motivated by the goal of understanding the dynamics of bacterial genome evolution via HGT. Understanding the HGT in the context of antibiotic resistance is especially critical because of the tremendous associated costs for affected patient populations. Each chapter implements a new method for studying bacterial genomes and applying these methods to study bacterial genome evolution in multiple contexts: evolution of antibiotic resistance in pathogenic microbes and metabolic phenotypes in commensal microbes. Future development and application of the work presented here will facilitate future research into bacterial genome evolution and an understanding of the consequences for human health.

APPENDIX A

Supporting Information for Chapter 3

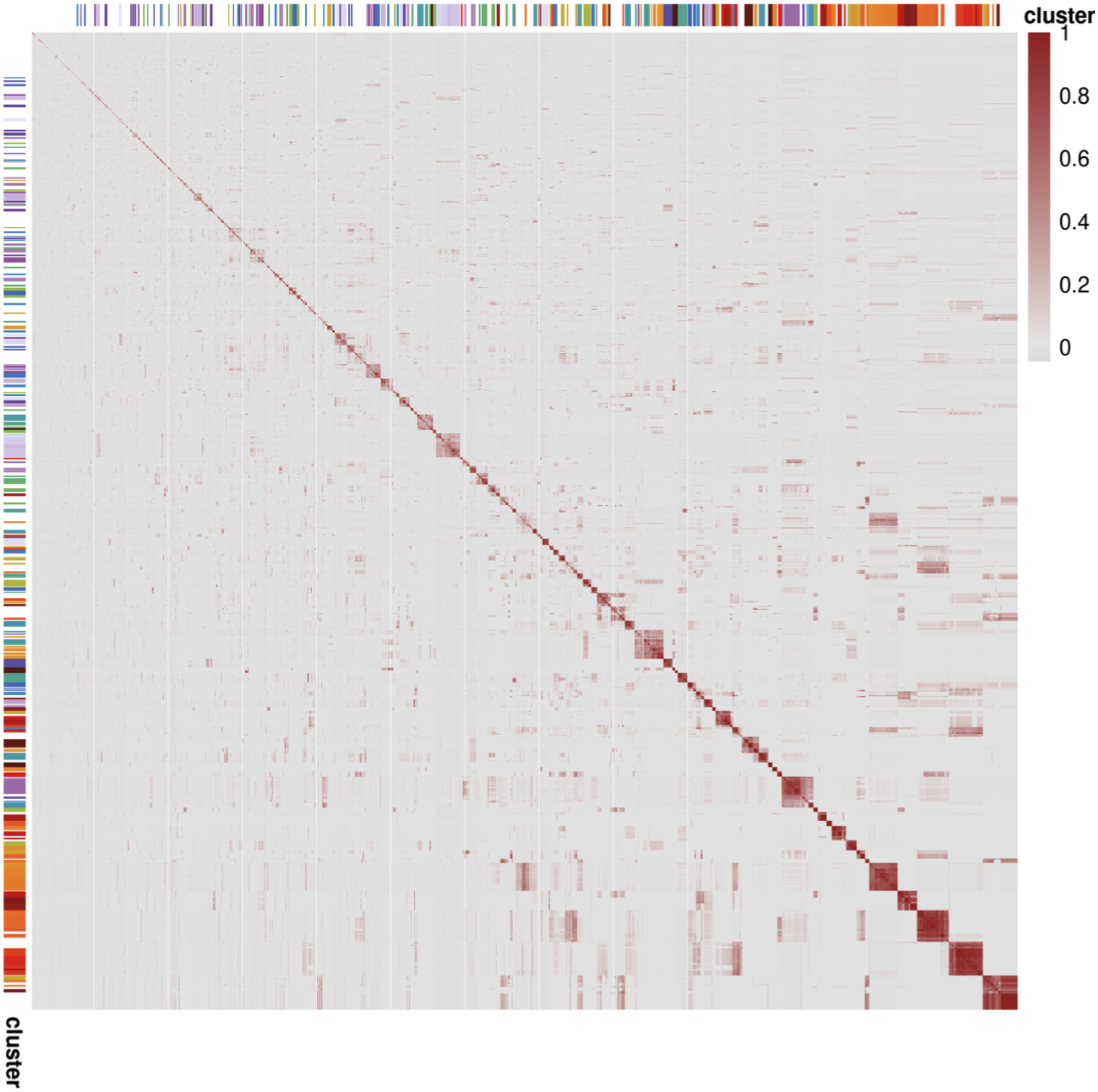


Figure A.1: Correlations between KPC plasmid genes. Gene correlation matrix for the 7596 KPC plasmid genes. Matrix is sorted by the hierarchical clustering results. Row and column annotation show the resulting gene clusters.

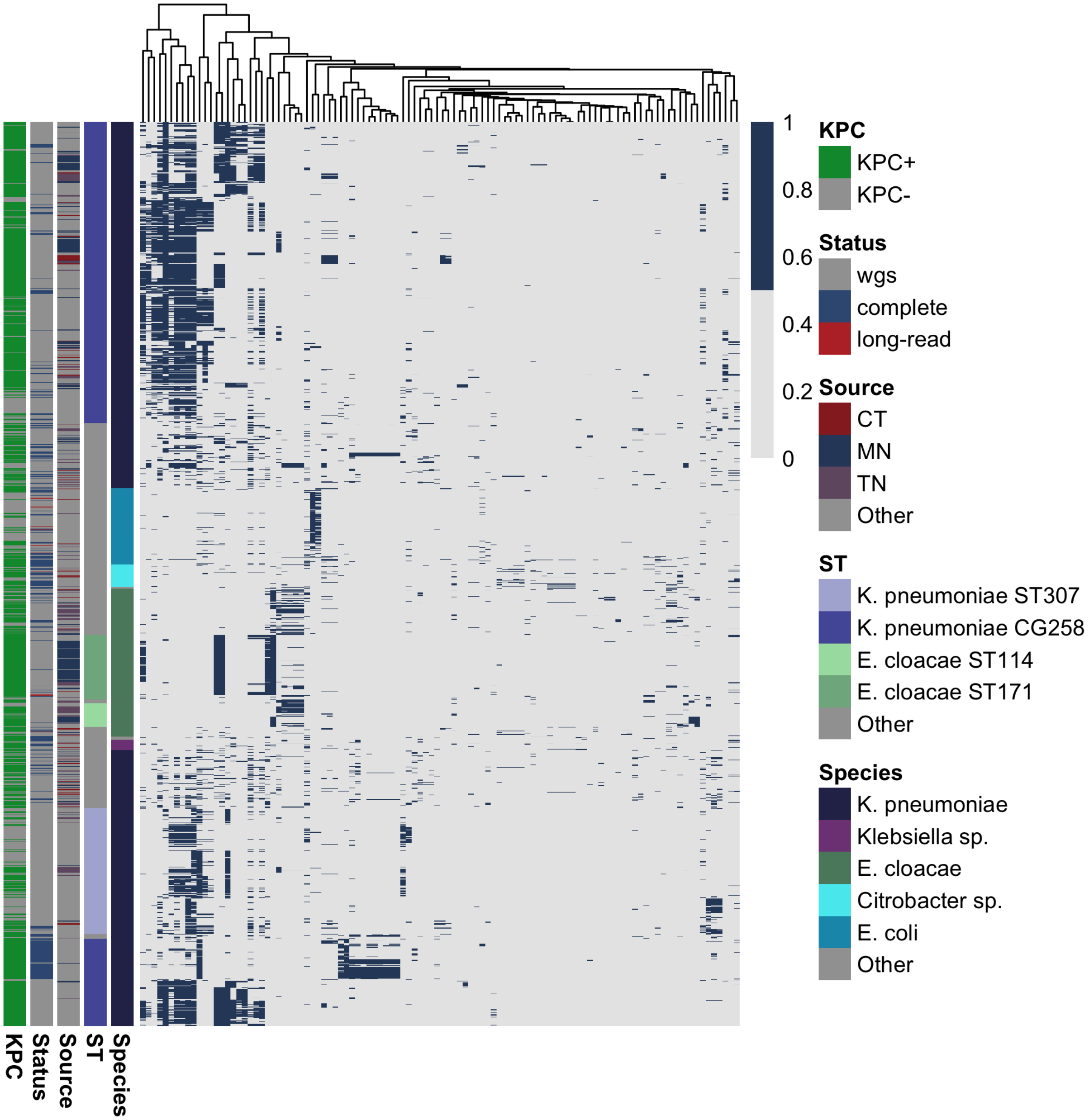


Figure A.2: Distribution of gene clusters present in EIP genomes across the entire dataset. Rows are sorted by the midpoint-rooted approximate maximum likelihood tree base off of the concatenated, core gene alignment. Columns are sorted by hierarchical clustering.

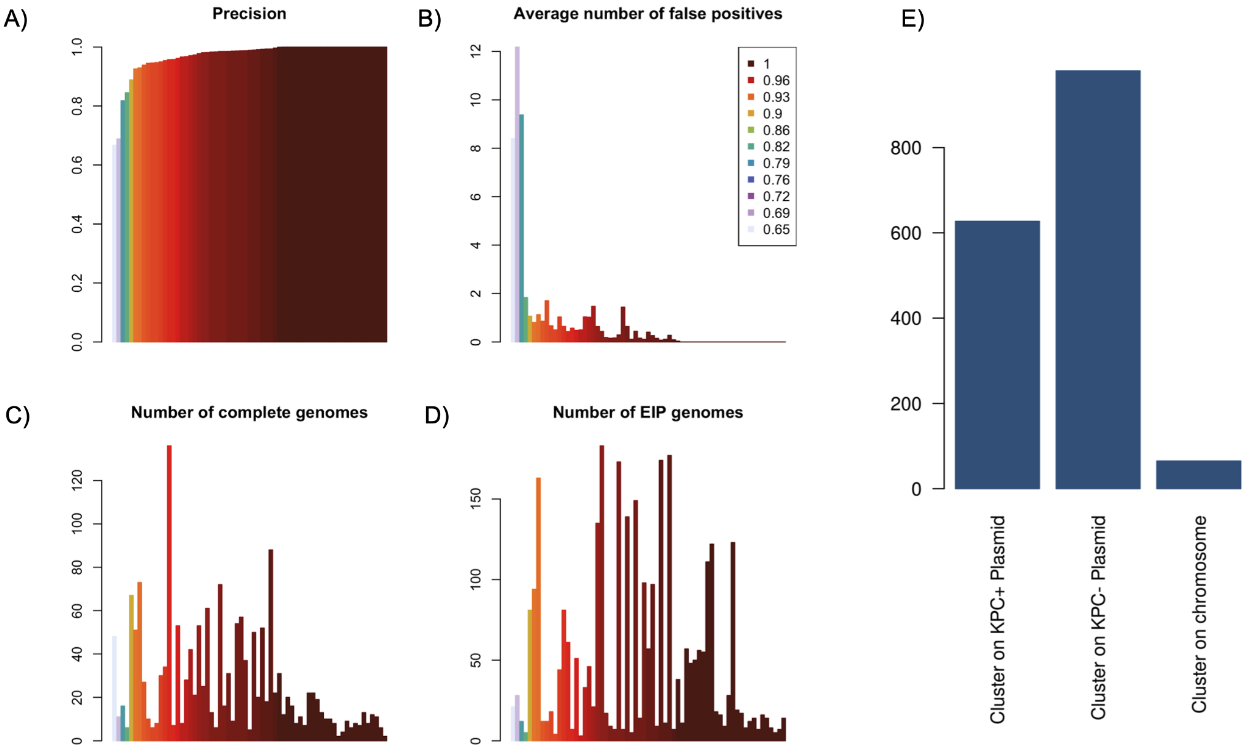


Figure A.3: Assessment of gene cluster distribution in the complete genomes. (A) Precision of each gene cluster calculated from using the complete genomes true positives are those present on the molecule containing the most cluster genes and false positives are genes present on another contig. (B) Average number of false positives per cluster. (C) The number of complete genomes each cluster is present in to evaluate precision. (D) number of EIP genomes each cluster is present in. (E) Counts of molecule types on which the identified clusters are present.

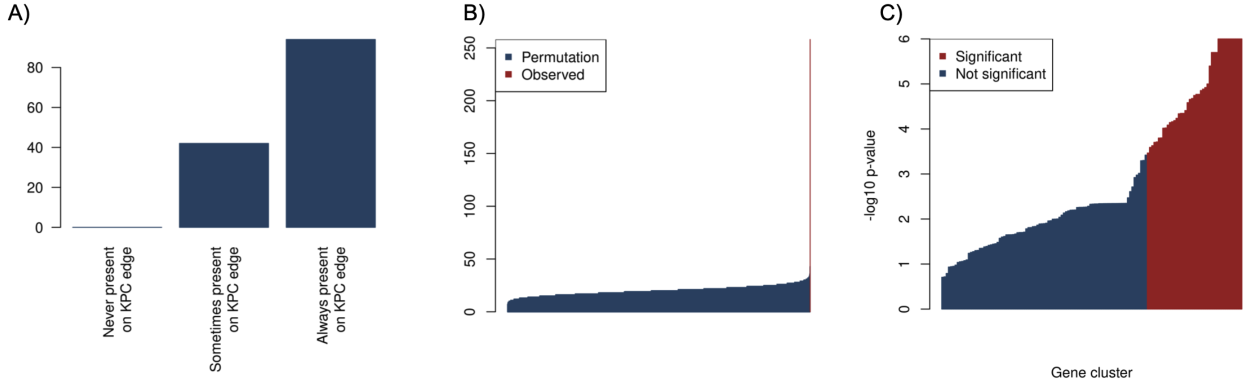


Figure A.4: Identification of gene clusters significantly associated with KPC status. (A) Count of clusters predicted to appear which appear never, variably, or always on edges where KPC was also predicted to be present. (B) Histogram showing the distribution of the total counts of the total number of gene cluster, KPC co-transitions generated from permuting cluster acquisition edges across one million permutations. The observed number of KPC cluster co-transitions shown in red. (C) P-values from individual clusters. Clusters which are significant after bonferroni correction are highlighted in red.

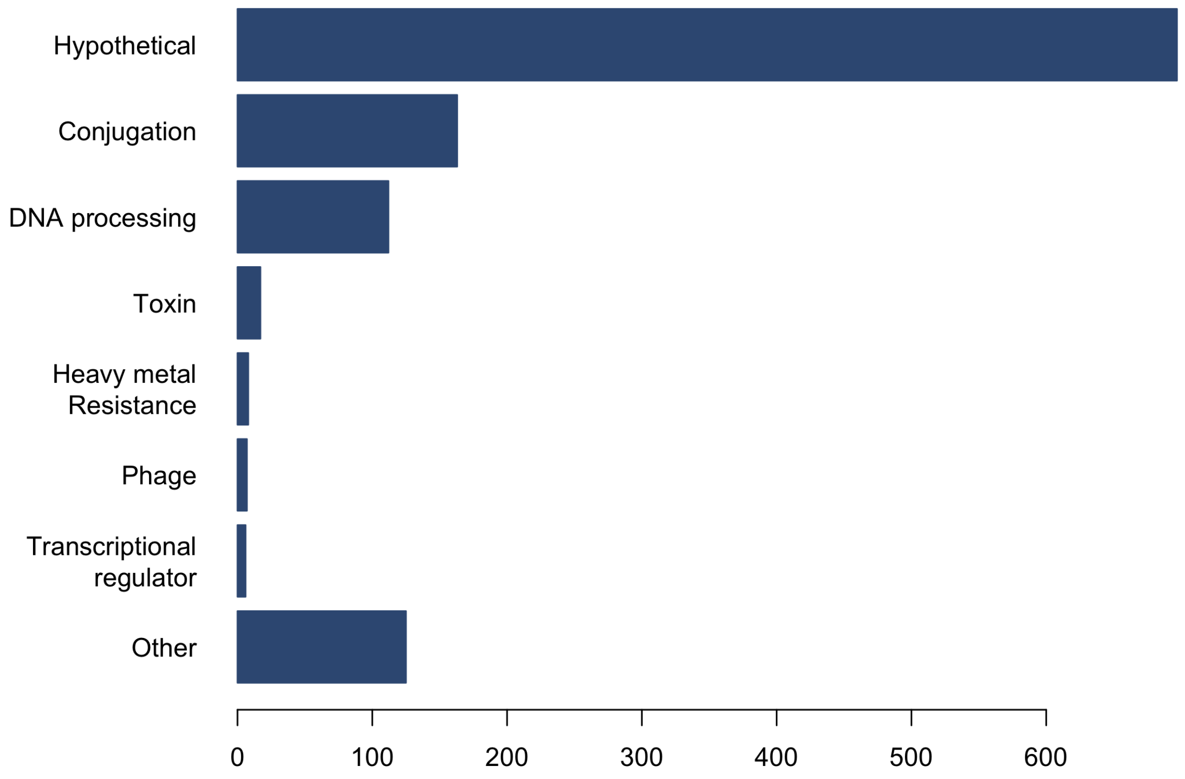


Figure A.5: Annotations assigned to KPC associated cluster genes. Counts of the annotations assigned by RAST with functions assigned to the selected classes of plasmid genes.

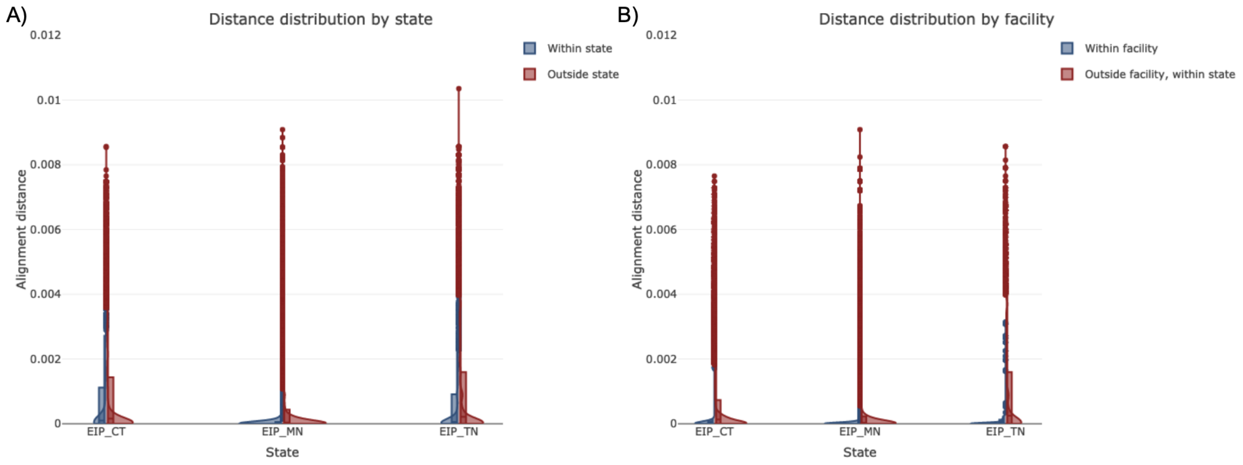


Figure A.6: Plasmid gene distance distributions for concatenated gene alignments for all isolates by (A) state and (B) facility

APPENDIX B

Supporting Information for Chapter 4

Figure S1

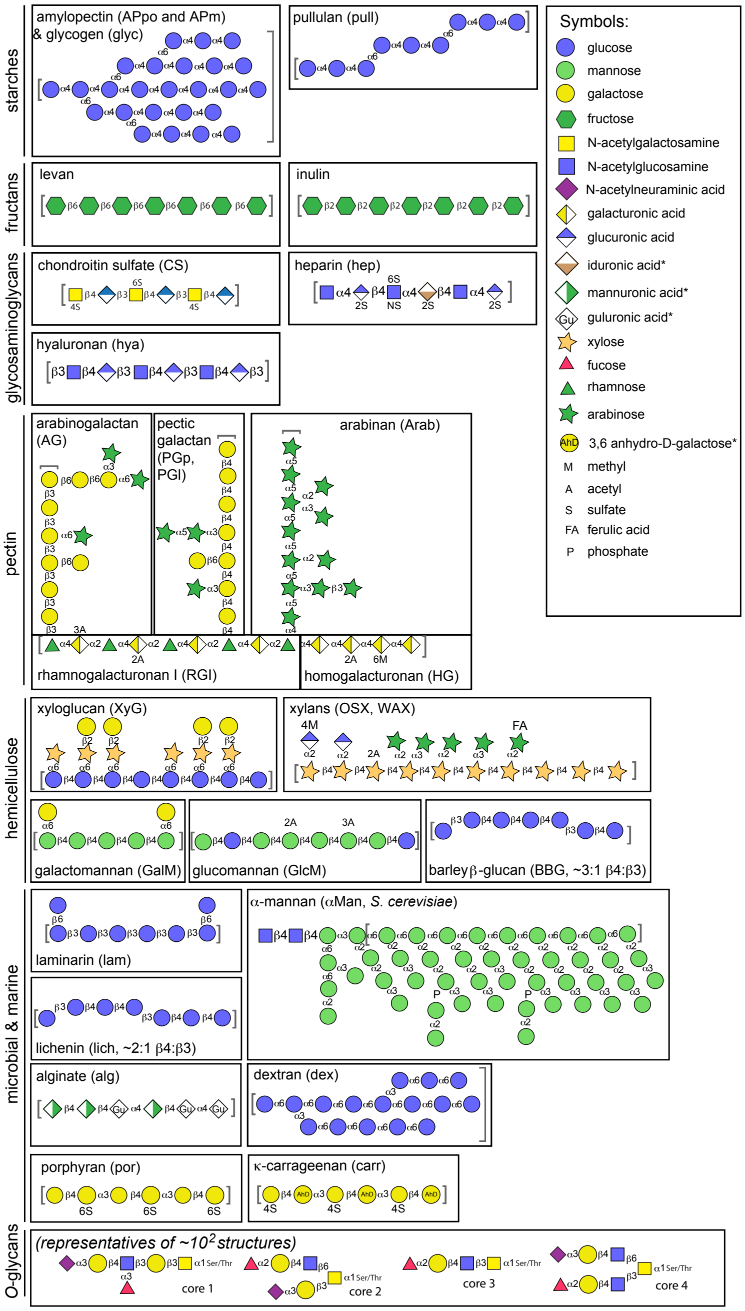
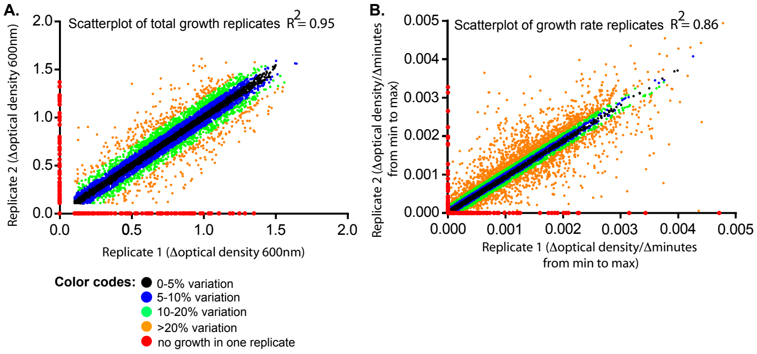


Figure B.1: Schematics of the polysaccharides used in this study with sugar composition and linkages schematized according to the "Symbol nomenclature for glycans" standard format and based on the symbol key provided at the right. Linkages are labeled as α or β, and the number provided represents the carbon position in the recipient sugar. The carbon in the donor sugar is carbon-1 in all cases except N-acetyl neuraminic acid and is not shown. Note that pectic galactan (potato and lupin), xylan (oat spelt and wheat arabinoxylan), and amylopectin (potato and maize) can have variable structures based on the plant source. Abbreviations for several polysaccharides are provided in parentheses and used throughout the text and figures.

Figure S2



C.

Polysaccharides:

Substrate	R^2 between growth values	R^2 between rate values
AG	0.99	0.89
alg	0.96	0.96
α-mann	0.93	0.76
APm	0.94	0.92
APpo	0.96	0.91
arab	0.98	0.96
BBG	0.95	0.70
carr	0.93	0.96
Cell	0.96	0.81
CS	0.96	0.91
dex	0.96	0.87
GalM	0.96	0.98
GlcM	0.93	0.85
glyc	0.96	0.94
hep	0.96	0.85
hya	0.91	0.88
inulin	0.92	0.89
lam	0.96	0.96
levan	0.96	0.88
lich	0.80	0.45
MOG	0.98	0.97
OSX	0.93	0.83
PGA	0.97	0.89
PGl	0.95	0.96
PGp	0.92	0.92
por	0.85	0.85
pull	0.84	0.78
RGI	0.96	0.98
WAX	0.97	0.42
XyG	0.92	0.72

Monosaccharides:

Substrate	R^2 between growth values
Ara	0.90
Fru	0.90
Fuc	0.93
Gal	0.60
GalA	0.86
GalNAc	0.93
Glc	0.69
GlcA	0.87
GlcNAc	0.72
GlcNH3	0.93
Man	0.88
NeuNAc	0.86
Rha	0.96
Rib	0.94
Xyl	0.85

Figure B.2: Correlation of replicate growth and rate measurements. Two replicate measurements were made for each of the two parameters recorded, namely, total growth (A) and growth rate (B), for each species on each carbohydrate substrate. Data points are color coded based on whether the two replicates exhibited variation between 0% and 5% (black), 5% and 10% (blue), 10% and 20% (green), ¿20% (orange), or growth in one assay and no growth in the other (red). (C) A linear function was fitted (with red points omitted) to calculate an r2 value for the data set associated with the utilization of each individual substrate. Measurements on some substrates were more variable than on others due, at least in part, to the tendency of these substrates to partially precipitate or retrograde during growth, which yielded variable levels of background absorbance.

Figure S3

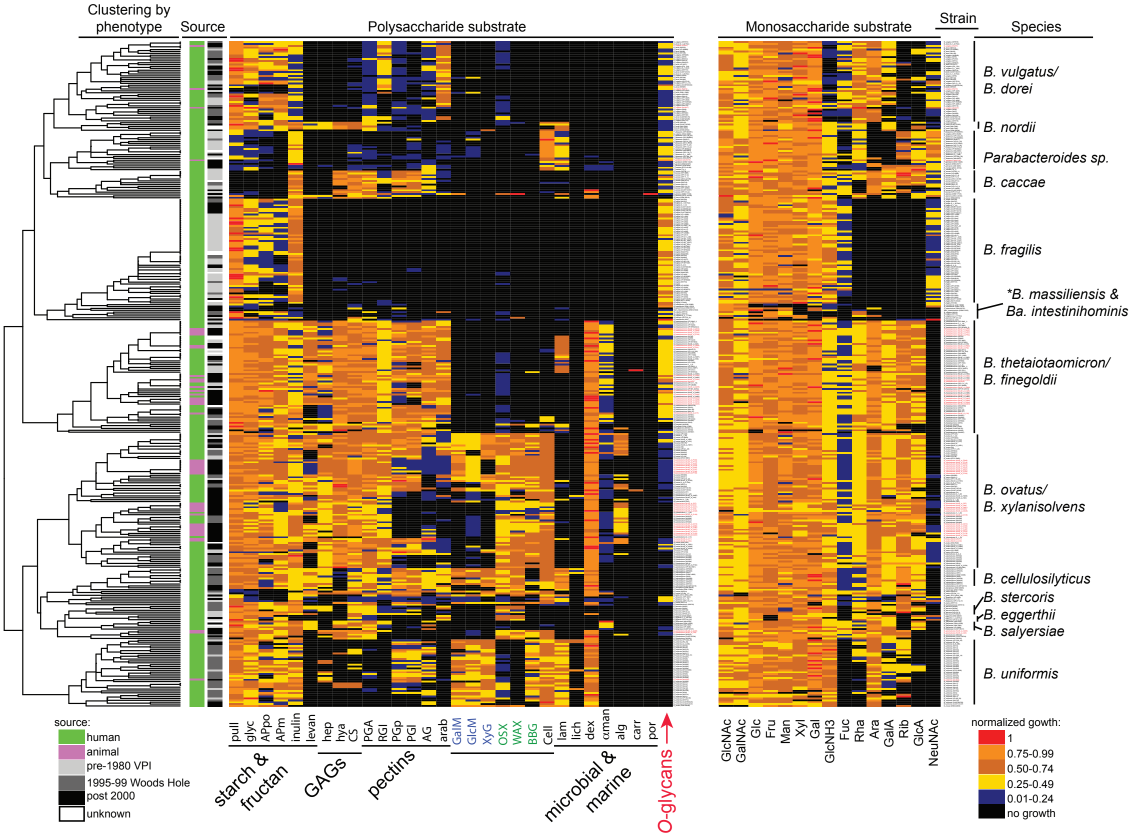


Figure B.3: Heatmap identical to the one shown in Fig. 2 main text, except that monosaccharide growth data are included. Strain names are also noted at the far right side of each panel (best viewed in electronic PDF form with magnification), and animal strains are labeled in red font.

Figure S4

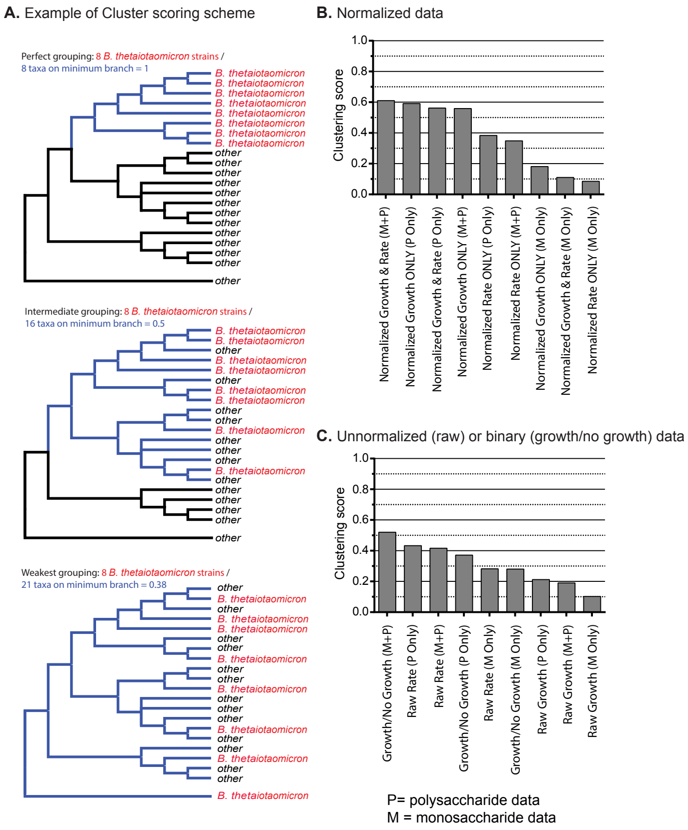
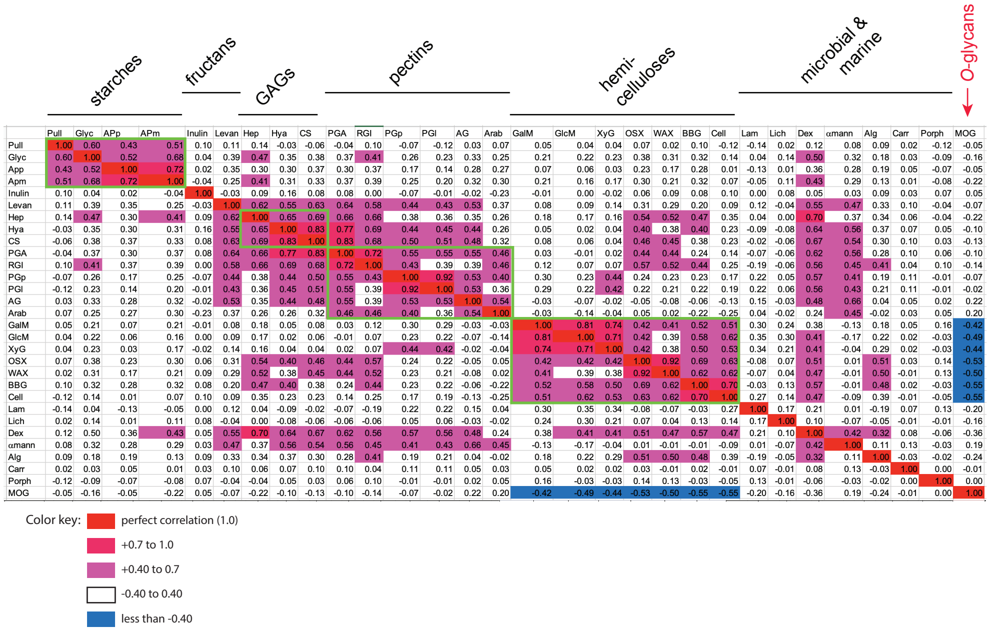


Figure B.4: Scheme for evaluating which aspects of growth phenotype data are most influential. A) Clustering strains that belong to the same species using hypothetical *B. thetaiotaomicron* data as an illustrative example. A quantitative index was used in which the number of strains tested is divided by the minimum number of branches needed to encompass all of the strains for that species, with a perfect score being “1” (e.g., eight *B. theta* strains divided by the minimum of eight branches needed to encompass all strains in the top example). (B) Actual clustering index data for the raw and normalized growth and rate data gathered for 354 different Bacteroidetes strains. M and P stand for “monosaccharide” and “polysaccharide” growth, respectively. One of the two most optimal conditions, which incorporates normalized growth data on polysaccharides only, was used to construct Figure 4.2 and Figure B3

Figure S5



	starches				fructans		GAGs			pectins					hemi-celluloses							microbial & marine							O-glycans
	Pull	Glyc	APp	APm	Inulin	Levan	Hep	Hya	CS	PGA	RGI	PGp	PGI	AG	Arab	GalM	GlcM	XyG	OSX	WAX	BBG	Cell	Lam	Lich	Dex	αmann	Alg	Carr	Porph	MOG
Pull	1.00	0.60	0.43	0.51	0.10	0.11	0.14	-0.03	-0.06	-0.04	0.10	-0.07	-0.12	0.03	0.07	0.05	0.04	0.04	0.07	0.02	0.10	-0.12	-0.14	0.02	0.12	0.08	0.09	0.02	-0.12	-0.05
Glyc	0.60	1.00	0.52	0.68	0.04	0.39	0.47	0.35	0.38	0.37	0.41	0.26	0.23	0.33	0.25	0.21	0.22	0.23	0.38	0.31	0.32	0.14	0.04	0.14	0.50	0.32	0.18	0.03	-0.09	-0.16
App	0.43	0.52	1.00	0.72	0.02	0.35	0.30	0.30	0.37	0.30	0.37	0.17	0.14	0.28	0.27	0.07	0.06	0.03	0.23	0.17	0.28	0.01	-0.13	0.01	0.36	0.28	0.19	0.05	-0.07	-0.05
Apm	0.51	0.68	0.72	1.00	-0.04	0.25	0.41	0.31	0.33	0.37	0.39	0.25	0.20	0.32	0.30	0.21	0.16	0.17	0.30	0.21	0.32	0.07	-0.05	0.11	0.43	0.29	0.13	0.01	-0.08	-0.22
Inulin	0.10	0.04	0.02	-0.04	1.00	-0.03	0.09	0.16	0.08	0.08	0.00	-0.07	-0.01	-0.02	-0.23	-0.01	0.00	-0.02	0.06	0.09	0.08	0.10	0.00	0.08	0.05	0.03	0.09	0.03	0.07	0.05
Levan	0.11	0.39	0.35	0.25	-0.03	1.00	0.62	0.55	0.63	0.64	0.58	0.44	0.43	0.53	0.37	0.08	0.09	0.14	0.31	0.29	0.20	0.09	0.12	-0.04	0.55	0.47	0.33	0.10	-0.04	-0.07
Hep	0.14	0.47	0.30	0.41	0.09	0.62	1.00	0.65	0.69	0.66	0.66	0.38	0.36	0.35	0.26	0.18	0.17	0.16	0.54	0.52	0.47	0.35	0.04	0.00	0.70	0.37	0.34	0.06	-0.04	-0.22
Hya	-0.03	0.35	0.30	0.31	0.16	0.55	0.65	1.00	0.83	0.77	0.69	0.44	0.45	0.44	0.26	0.05	0.02	0.04	0.40	0.38	0.40	0.23	-0.09	-0.08	0.64	0.56	0.37	0.07	0.05	-0.10
CS	-0.06	0.38	0.37	0.33	0.08	0.63	0.69	0.83	1.00	0.83	0.68	0.50	0.51	0.48	0.32	0.08	0.06	0.04	0.46	0.45	0.38	0.23	-0.02	-0.06	0.67	0.54	0.30	0.10	0.03	-0.13
PGA	-0.04	0.37	0.30	0.37	0.08	0.64	0.66	0.77	0.83	1.00	0.72	0.55	0.55	0.55	0.46	0.03	-0.01	0.02	0.44	0.44	0.24	0.14	-0.07	-0.06	0.62	0.56	0.28	0.10	0.06	-0.10
RGI	0.10	0.41	0.37	0.39	0.00	0.58	0.66	0.69	0.68	0.72	1.00	0.43	0.39	0.39	0.46	0.12	0.07	0.07	0.57	0.52	0.44	0.25	-0.19	-0.06	0.56	0.45	0.41	0.04	0.10	-0.14
PGp	-0.07	0.26	0.17	0.25	-0.07	0.44	0.38	0.44	0.50	0.55	0.43	1.00	0.92	0.53	0.40	0.30	0.23	0.44	0.24	0.23	0.23	0.17	0.22	0.05	0.57	0.41	0.19	0.11	-0.01	-0.07
PGl	-0.12	0.23	0.14	0.20	-0.01	0.43	0.36	0.45	0.51	0.55	0.39	0.92	1.00	0.53	0.36	0.29	0.22	0.42	0.22	0.21	0.22	0.19	0.22	0.06	0.56	0.43	0.21	0.11	-0.01	-0.02
AG	0.03	0.33	0.28	0.32	-0.02	0.53	0.35	0.44	0.48	0.55	0.39	0.53	0.53	1.00	0.54	-0.03	-0.07	-0.02	-0.05	-0.08	-0.06	-0.13	0.15	-0.03	0.48	0.66	0.04	0.05	0.02	0.22
Arab	0.07	0.25	0.27	0.30	-0.23	0.37	0.26	0.26	0.32	0.46	0.46	0.40	0.36	0.54	1.00	-0.03	-0.14	-0.04	0.05	0.02	-0.22	-0.25	0.04	-0.02	0.24	0.45	-0.02	0.03	0.05	0.20
GalM	0.05	0.21	0.07	0.21	-0.01	0.08	0.18	0.05	0.08	0.03	0.12	0.30	0.29	-0.03	-0.03	1.00	0.81	0.74	0.42	0.41	0.52	0.51	0.30	0.24	0.38	-0.13	0.18	0.05	0.16	-0.42
GlcM	0.04	0.22	0.06	0.16	0.00	0.09	0.17	0.02	0.06	-0.01	0.07	0.23	0.22	-0.07	-0.14	0.81	1.00	0.71	0.42	0.39	0.58	0.62	0.35	0.30	0.41	-0.17	0.22	0.02	-0.03	-0.49
XyG	0.04	0.23	0.03	0.17	-0.02	0.14	0.16	0.04	0.04	0.02	0.07	0.44	0.42	-0.02	-0.04	0.74	0.71	1.00	0.42	0.38	0.50	0.53	0.34	0.21	0.41	-0.04	0.29	0.02	-0.03	-0.44
OSX	0.07	0.38	0.23	0.30	0.06	0.31	0.54	0.40	0.46	0.44	0.57	0.24	0.22	-0.05	0.05	0.42	0.42	0.42	1.00	0.92	0.69	0.63	-0.08	0.07	0.51	0.01	0.51	0.03	0.14	-0.53
WAX	0.02	0.31	0.17	0.21	0.09	0.29	0.52	0.38	0.45	0.44	0.52	0.23	0.21	-0.08	0.02	0.41	0.39	0.38	0.92	1.00	0.62	0.62	-0.07	0.04	0.47	-0.01	0.50	-0.01	0.13	-0.50
BBG	0.10	0.32	0.28	0.32	0.08	0.20	0.47	0.40	0.38	0.24	0.44	0.23	0.22	-0.06	-0.22	0.52	0.58	0.50	0.69	0.62	1.00	0.70	-0.03	0.13	0.57	-0.01	0.48	0.02	-0.03	-0.55
Cell	-0.12	0.14	0.01	0.07	0.10	0.09	0.35	0.23	0.23	0.14	0.25	0.17	0.19	-0.13	-0.25	0.51	0.62	0.53	0.63	0.62	0.70	1.00	0.27	0.14	0.47	-0.09	0.39	-0.01	-0.05	-0.55
Lam	-0.14	0.04	-0.13	-0.05	0.00	0.12	0.04	-0.09	-0.02	-0.07	-0.19	0.22	0.22	0.15	0.04	0.30	0.35	0.34	-0.08	-0.07	-0.03	0.27	1.00	0.17	0.21	0.01	-0.19	0.07	0.13	-0.20
Lich	0.02	0.14	0.01	0.11	0.08	-0.04	0.00	-0.08	-0.06	-0.06	-0.06	0.05	0.06	-0.03	-0.02	0.24	0.30	0.21	0.07	0.04	0.13	0.14	0.17	1.00	0.10	-0.07	-0.05	-0.01	-0.01	-0.16
Dex	0.12	0.50	0.36	0.43	0.05	0.55	0.70	0.64	0.67	0.62	0.56	0.57	0.56	0.48	0.24	0.38	0.41	0.41	0.51	0.47	0.57	0.47	0.21	0.10	1.00	0.42	0.32	0.08	-0.06	-0.36
αmann	0.08	0.32	0.28	0.29	0.03	0.47	0.37	0.56	0.54	0.56	0.45	0.41	0.43	0.66	0.45	-0.13	-0.17	-0.04	0.01	-0.01	-0.01	-0.09	0.01	-0.07	0.42	1.00	0.11	0.13	-0.03	0.19
Alg	0.09	0.18	0.19	0.13	0.09	0.33	0.34	0.37	0.30	0.28	0.41	0.19	0.21	0.04	-0.02	0.18	0.22	0.29	0.51	0.50	0.48	0.39	-0.19	-0.05	0.32	0.11	1.00	-0.03	-0.02	-0.24
Carr	0.02	0.03	0.05	0.01	0.03	0.10	0.06	0.07	0.10	0.10	0.04	0.11	0.11	0.05	0.03	0.05	0.02	0.02	0.03	-0.01	0.02	-0.01	0.07	-0.01	0.08	0.13	-0.03	1.00	0.00	-0.01
Porph	-0.12	-0.09	-0.07	-0.08	0.07	-0.04	-0.04	0.05	0.03	0.06	0.10	-0.01	-0.01	0.02	0.05	0.16	-0.03	-0.03	0.14	0.13	-0.03	-0.05	0.13	-0.01	-0.06	-0.03	-0.02	0.00	1.00	0.00
MOG	-0.05	-0.16	-0.05	-0.22	0.05	-0.07	-0.22	-0.10	-0.13	-0.10	-0.14	-0.07	-0.02	0.22	0.20	-0.42	-0.49	-0.44	-0.53	-0.50	-0.55	-0.55	-0.20	-0.16	-0.36	0.19	-0.24	-0.01	0.00	1.00

Color key:
🟥 perfect correlation (1.0)
■ +0.7 to 1.0
🟪 +0.40 to 0.7
□ -0.40 to 0.40
🟦 less than -0.40

Figure B.5: Pearson correlation plot to determine if individual growth abilities co-occur in the same strains. For each substrate pair, the values shown indicate the positive or negative correlation value that both substrates will be used by any of the strains among the 354 surveyed. Positive or negative correlations that are ≥ 0.40 are shown in the colors indicated.

Figure S6

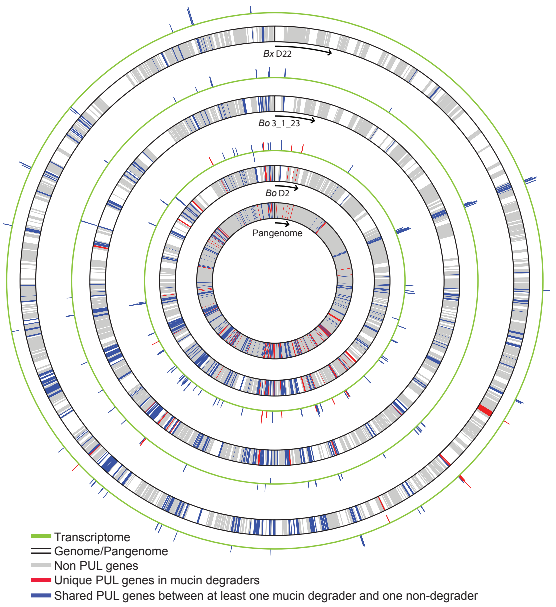

Bx D22
Bo 3_1_23
Bo D2
Pangenome
Transcriptome
Genome/Pangenome
Non PUL genes
Unique PUL genes in mucin degraders
Shared PUL genes between at least one mucin degrader and one non-degrader

Figure B.6: Circular pangenome illustration and corresponding transcriptomic-based analysis of PULs that are activated during growth on O-glycans in strains *B. xylanisolvens* D22, *B. ovatus* 3_1_23, and *B. ovatus* D2.

Supplemental Figure 7. Bacteroides LGT Loci

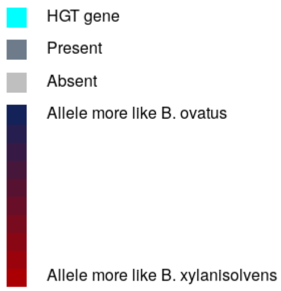


Figure B.7: Individual maps of high-confidence intergenomic exchange events between *B. ovatus* and *B. xylanisolvens* strains. The method for the identification of these loci was the same as the method described in the text for the example in Figgure 4.6C. Examples involving PULs are shown first, and examples showing non-PUL genes are shown second.

BIBLIOGRAPHY

[1] Eugene V Koonin. Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research*, 5, 2016.

[2] Brian J Arnold, I Huang, William P Hanage, et al. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, pages 1–13, 2021.

[3] Ilana Lauren Brito. Examining horizontal gene transfer in microbial communities. *Nature Reviews Microbiology*, 19(7):442–453, 2021.

[4] Luis Boto, Manuel Pineda, and Rafael Pineda. Potential impacts of horizontal gene transfer on human health and physiology and how anthropogenic activity can affect it. *The FEBS Journal*, 286(20):3959–3967, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/febs.15054.

[5] Centers for Disease Control and Prevention. Antibiotic resistance threats in the united states, 2019. 2019.

[6] Alvaro San Millan. Evolution of plasmid-mediated antibiotic resistance in the clinical context. *Trends in Microbiology*, 26(12):978–985, 2018.

[7] Amy J. Mathers, Gisele Peirano, and Johann D. D. Pitout. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant enterobacteriaceae. *Clinical Microbiology Reviews*, 28(3):565–591, 2015. Publisher: American Society for Microbiology.

[8] Ravi Jain, Maria C. Rivera, and James A. Lake. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7):3801–3806, 1999. Publisher: Proceedings of the National Academy of Sciences.

[9] J. Peter Gogarten and Jeffrey P. Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687, 2005. Number: 9 Publisher: Nature Publishing Group.

[10] Christopher M. Thomas and Kaare M. Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, 3(9):711–721, 2005. Number: 9 Publisher: Nature Publishing Group.

[11] M G Lorenz and W Wackernagel. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological Reviews*, 58(3):563–602, 1994. Publisher: American Society for Microbiology.

[12] Melanie Blokesch. In and out—contribution of natural transformation to the shuffling of large genomic regions. *Current Opinion in Microbiology*, 38:22–29, 2017. Mobile genetic elements and HGT in prokaryotes * Microbiota.

[13] Ross S McInnes, Gregory E McCallum, Lisa E Lamberte, and Willem van Schaik. Horizontal transfer of antibiotic resistance genes in the human gut microbiome. *Current Opinion in Microbiology*, 53:35–43, 2020.

[14] Christine L." "Schneider, Stephen T. Abedon, Benjamin H. Burrowes, and Malcolm L. McConville. *Bacteriophage-Mediated Horizontal Gene Transfer: Transduction*. Springer International Publishing, Cham, 2021.

[15] Nelson Frazão, Ana Sousa, Michael Lässig, and Isabel Gordo. Horizontal gene transfer overrides mutation in escherichia coli colonizing the mammalian gut. *Proceedings of the National Academy of Sciences*, 116(36):17906–17915, 2019.

[16] Liang Chen, Barun Mathema, Johann D. D. Pitout, Frank R. DeLeo, Barry N. Kreiswirth, and George Jacoby. Epidemic klebsiella pneumoniae st258 is a hybrid strain. *mBio*, 5(3):e01355–14, 2014.

[17] Chris Smillie, M. Pilar Garcillán-Barcia, M. Victoria Francia, Eduardo P. C. Rocha, and Fernando de la Cruz. Mobility of plasmids. *Microbiology and Molecular Biology Reviews*, 74(3):434–452, 2010.

[18] Eric Cascales and Peter J. Christie. The versatile bacterial type IV secretion systems. *Nature Reviews Microbiology*, 1(2):137–149, 2003. Number: 2 Publisher: Nature Publishing Group.

[19] Andreas Porse, Kristian Schønning, Christian Munck, and Morten O.A. Sommer. Survival and evolution of a large multidrug resistance plasmid in new clinical bacterial hosts. *Molecular Biology and Evolution*, 33(11):2860–2873, 2016.

[20] R. Craig MacLean and Alvaro San Millan. Microbial evolution: Towards resolving the plasmid paradox. *Current Biology*, 25(17):R764–R767, 2015.

[21] Vincent Burrus and Matthew K Waldor. Shaping bacterial genomes with integrative and conjugative elements. *Research in Microbiology*, 155(5):376–386, 2004. Genome plasticity and the evolution of microbial genomes.

[22] François Delavat, Ryo Miyazaki, Nicolas Carraro, Nicolas Pradervand, and Jan Roelof van der Meer. The hidden life of integrative and conjugative elements. *FEMS Microbiology Reviews*, 41(4):512–537, 03 2017.

[23] Amanda C. Carroll and Alex Wong. Plasmid persistence: costs, benefits, and the plasmid paradox. *Canadian Journal of Microbiology*, 64(5):293–304, 2018. PMID: 29562144.

[24] Yuri I. Wolf, Igor B. Rogozin, Nick V. Grishin, and Eugene V. Koonin. Genome trees and the tree of life. *Trends in Genetics*, 18(9):472–479, 2002.

[25] W. Ford Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128, 1999. Publisher: American Association for the Advancement of Science.

[26] Gur Sevillya, Orit Adato, and Sagi Snir. Detecting horizontal gene transfer: a probabilistic approach. *BMC Genomics*, 21(1):106, 2020.

[27] Robert G. Beiko, Timothy J. Harlow, and Mark A. Ragan. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*, 102(40):14332–14337, 2005. Publisher: Proceedings of the National Academy of Sciences.

[28] Yoji Nakamura, Takeshi Itoh, Hideo Matsuda, and Takashi Gojobori. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics*, 36(7):760–766, 2004. Number: 7 Publisher: Nature Publishing Group.

[29] Santiago Garcia-Vallvé, Anton Romeu, and Jaume Palau. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Research*, 10(11):1719–1725, 2000. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[30] Howard Ochman, Jeffrey G. Lawrence, and Eduardo A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000. Number: 6784 Publisher: Nature Publishing Group.

[31] Hao Zhou, Juan Felipe Beltrán, and Ilana Lauren Brito. Functions predict horizontal gene transfer and the emergence of antibiotic resistance. *Science Advances*, 7(43):eabj5056, 2021. Publisher: American Association for the Advancement of Science.

[32] Mathieu Groussin, Mathilde Poyet, Ainara Sistiaga, Sean M. Kearney, Katya Moniz, Mary Noel, Jeff Hooker, Sean M. Gibbons, Laure Segurel, Alain Froment, Rihlat Said Mohamed, Alain Fezeu, Vanessa A. Juimo, Sophie Lafosse, Francis E. Tabe, Catherine Girard, Deborah Iqaluk, Le Thanh Tu Nguyen, B. Jesse Shapiro, Jenni Lehtimäki, Lasse Ruokolainen, Pinja P. Kettunen, Tommi Vatanen, Shani Sigwazi, Audax Mabulla, Manuel Domínguez-Rodrigo, Yvonne A. Nartey, Adwoa Agyei-Nkansah, Amoako Duah, Yaw A. Awuku, Kenneth A. Valles, Shadrack O. Asibey, Mary Y. Afihene, Lewis R. Roberts, Amelie Plymoth, Charles A. Onyekwere, Roger E. Summons, Ramnik J. Xavier, and Eric J. Alm. Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell*, 184(8):2053–2067.e18, 2021.

[33] Christian Brandt, Adrian Viehweger, Abhijeet Singh, Mathias W. Pletz, Daniel Wibberg, Jörn Kalinowski, Sandrina Lerch, Bettina Müller, and Oliwia Makarewicz. Assessing genetic diversity and similarity of 435 KPC-carrying plasmids. *Scientific Reports*, 9(1):11223, 2019. Number: 1 Publisher: Nature Publishing Group.

[34] Leho Tedersoo, Mads Albertsen, Sten Anslan, and Benjamin Callahan. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Applied and environmental microbiology*, 87(17):e00626–21, 2021.

[35] Sergio Arredondo-Alonso, Rob J. Willems, Willem van Schaik, and Anita C. Schürch. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics*, 3(10):e000128, 2017.

[36] Sophia David, Sandra Reuter, Simon R. Harris, Corinna Glasner, Theresa Feltwell, Silvia Argimon, Khalil Abudahab, Richard Goater, Tommaso Giani, Giulia Errico, Marianne Aspbury, Sara Sjunnebo, Edward J. Feil, Gian Maria Rossolini, David M. Aanensen, and Hajo Grundmann. Epidemic of carbapenem-resistant klebsiella pneumoniae in europe is driven by nosocomial spread. *Nature Microbiology*, 4(11):1919–1929, 2019. Number: 11 Publisher: Nature Publishing Group.

[37] Daniel R Evans, Marissa P Griffith, Alexander J Sundermann, Kathleen A Shutt, Melissa I Saul, Mustapha M Mustapha, Jane W Marsh, Vaughn S Cooper, Lee H Harrison, and Daria Van Tyne. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. *eLife*, 9:e53886, 2020. Publisher: eLife Sciences Publications, Ltd.

[38] Anna E. Sheppard, Nicole Stoesser, Daniel J. Wilson, Robert Sebra, Andrew Kasarskis, Luke W. Anson, Adam Giess, Louise J. Pankhurst, Alison Vaughan, Christopher J. Grim, Heather L. Cox, Anthony J. Yeh, the Modernising Medical Microbiology (MMM) Informatics Group, Costi D. Sifri, A. Sarah Walker, Tim E. Peto, Derrick W. Crook, and Amy J. Mathers. Nested russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene blaKPC. *Antimicrobial Agents and Chemotherapy*, 60(6):3767–3778, 2016. Publisher: American Society for Microbiology.

[39] Sophia David, Victoria Cohen, Sandra Reuter, Anna E. Sheppard, Tommaso Giani, Julian Parkhill, the European Survey of Carbapenemase-Producing Enterobacteriaceae (EuSCAPE) Working Group, the ESCMID Study Group for Epidemiological Markers (ESGEM), Gian Maria Rossolini, Edward J. Feil, Hajo Grundmann, and David M. Aanensen. Integrated chromosomal and plasmid sequence analyses reveal diverse modes of carbapenemase gene spread among klebsiella pneumoniae. *Proceedings of the National Academy of Sciences*, 117(40):25043–25054, 2020. Publisher: Proceedings of the National Academy of Sciences.

[40] Rauf Salamzade, Abigail L. Manson, Bruce J. Walker, Thea Brennan-Krohn, Colin J. Worby, Peijun Ma, Lorrie L. He, Terrance P. Shea, James Qu, Sinéad B. Chapman, Whitney Howe, Sarah K. Young, Jenna I. Wurster, Mary L. Delaney, Sanjat Kanjilal, Andrew B. Onderdonk, Cassiana E. Bittencourt, Gabrielle M. Gussin, Diane Kim, Ellena M. Peterson, Mary Jane Ferraro, David C. Hooper, Erica S. Shenoy, Christina A. Cuomo, Lisa A. Cosimi, Susan S. Huang, James E. Kirby, Virginia M. Pierce, Roby P. Bhattacharyya, and Ashlee M. Earl. Inter-species geographic signatures for tracing horizontal gene transfer and long-term persistence of carbapenem resistance. *Genome Medicine*, 14(1):37, 2020.

[41] Sergio Arredondo-Alonso, Rob J. Willems, Willem van Schaik, and Anita C. Schürch. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics*, 3(10):e000128, 2017.

[42] Chau-Chyun Sheu, Ya-Ting Chang, Shang-Yi Lin, Yen-Hsu Chen, and Po-Ren Hsueh. Infections caused by carbapenem-resistant enterobacteriaceae: an update on therapeutic options. *Frontiers in microbiology*, 10:80, 2019.

[43] Mark S Wilke, Andrew L Lovering, and Natalie CJ Strynadka. beta-lactam antibiotic resistance: a current structural perspective. *Current Opinion in Microbiology*, 8(5):525–533, 2005. Antimicrobials / Edited by Malcolm Page and Christopher T Walsh · Genomics / Edited by Stephan C Schuster and Gerhard Gottschalk.

[44] Keith Poole. Resistance to β-lactam antibiotics. *Cellular and Molecular Life Sciences CMLS*, 61(17):2200–2223, 2004.

[45] Johann D. D. Pitout, Patrice Nordmann, and Laurent Poirel. Carbapenemase-producing klebsiella pneumoniae, a key pathogen set for global nosocomial dominance. *Antimicrobial Agents and Chemotherapy*, 59(10):5873–5884, 2015.

[46] Latania K Logan and Robert A Weinstein. The epidemiology of carbapenem-resistant enterobacteriaceae: the impact and evolution of a global menace. *The Journal of infectious diseases*, 215(suppl_1):S28–S36, 2017.

[47] Edward P Abraham and Ernst Chain. An enzyme from bacteria able to destroy penicillin. *Nature*, 146(3713):837–837, 1940.

[48] Patricia A Bradford, Carl Urban, Noriel Mariano, Steven J Projan, James J Rahal, and Karen Bush. Imipenem resistance in klebsiella pneumoniae is associated with the combination of act-1, a plasmid-mediated ampc beta-lactamase, and the foss of an outer membrane protein. *Antimicrobial agents and chemotherapy*, 41(3):563–569, 1997.

[49] FM MacKenzie, Ken J Forbes, T Dorai-John, SGB Amyes, and Ian M Gould. Emergence of a carbapenem-resistant kiebsiella pneumoniae. *The Lancet*, 350(9080):783, 1997.

[50] George A. Jacoby. Ampc &#x3b2;-lactamases. *Clinical Microbiology Reviews*, 22(1):161–182, 2009.

[51] Min-Jeong Park, Taek-Kyung Kim, Wonkeun Song, Jae-Seok Kim, Han-Sung Kim, and Jacob Lee. An increase in the clinical isolation of acquired ampc β-lactamase-producing klebsiella pneumoniae in korea from 2007 to 2010. *Annals of laboratory medicine*, 33(5):353, 2013.

[52] Guillaume Arlet and George A Jacoby. Plasmid-determined ampc-typeβ-lactamases. *Antimicrob. Agents Chemother*, 2002.

[53] George A Jacoby, Debra M Mills, and Nancy Chow. Role of β-lactamases and porins in resistance to ertapenem and other β-lactams in klebsiella pneumoniae. *Antimicrobial agents and chemotherapy*, 48(8):3203–3206, 2004.

[54] Mark E Rupp and Paul D Fey. Extended spectrum β-lactamase (esbl)-producing enterobacteriaceae. *Drugs*, 63(4):353–365, 2003.

[55] Johann D. D. Pitout, Patrice Nordmann, Kevin B. Laupland, and Laurent Poirel. Emergence of Enterobacteriaceae producing extended-spectrum beta-lactamases (ESBLs) in the community. *Journal of Antimicrobial Chemotherapy*, 56(1):52–59, 05 2005.

[56] Patricia A Bradford. Extended-spectrum β-lactamases in the 21st century: characterization, epidemiology, and detection of this important resistance threat. *Clinical microbiology reviews*, 14(4):933–951, 2001.

[57] Mariana Castanheira, Patricia J Simner, and Patricia A Bradford. Extended-spectrum β-lactamases: An update on their characteristics, epidemiology and detection. *JAC-antimicrobial resistance*, 3(3):dlab092, 2021.

[58] H Knothe, P Shah, V Krcmery, M Antal, and S Mitsuhashi. Transferable resistance to cefotaxime, cefoxitin, cefamandole and cefuroxime in clinical isolates of klebsiella pneumoniae and serratia marcescens. *Infection*, 11(6):315–317, 1983.

[59] GA Jacoby, AA Medeiros, TF O'Brien, ME Pinto, and H Jiang. Broad-spectrum, transmissible beta-lactamases. *The New England journal of medicine*, 319(11):723–724, 1988.

[60] JOHN P Quinn, D Miyashiro, D Sahm, R Flamm, and K Bush. Novel plasmid-mediated beta-lactamase (tem-10) conferring selective resistance to ceftazidime and aztreonam in clinical isolates of klebsiella pneumoniae. *Antimicrobial agents and chemotherapy*, 33(9):1451–1456, 1989.

[61] KE Goodman, PJ Simner, PD Tamma, and AM Milstone. Infection control implications of heterogeneous resistance mechanisms in carbapenem-resistant enterobacteriaceae (cre). *Expert review of anti-infective therapy*, 14(1):95–108, 2016.

[62] Anne Marie Queenan and Karen Bush. Carbapenemases: the versatile &#x3b2;-lactamases. *Clinical Microbiology Reviews*, 20(3):440–458, 2007.

[63] Johann DD Pitout and Kevin B Laupland. Extended-spectrum $beta$-lactamase-producing enterobacteriaceae: an emerging public-health concern. *The Lancet Infectious Diseases*, 8(3):159–166, 2008.

[64] Li Liu, Xiaowei Chen, Geir Skogerbø, Peng Zhang, Runsheng Chen, Shunmin He, and Da-Wei Huang. The human microbiome: a hot spot of microbial horizontal gene transfer. *Genomics*, 100(5):265–270, 2012.

[65] N. Shterzer and I. Mizrahi. The animal gut as a melting pot for horizontal gene transfer. *Canadian Journal of Microbiology*, 61(9):603–605, 2015.

[66] Lara Kern, Suhaib K Abdeen, Aleksandra A Kolodziejczyk, and Eran Elinav. Commensal inter-bacterial interactions shaping the microbiota. *Current Opinion in Microbiology*, 63:158–171, 2021.

[67] Aditi Kanhere and Martin Vingron. Horizontal gene transfers in prokaryotes show differential preferences for metabolic and translational genes. *BMC evolutionary biology*, 9(1):1–13, 2009.

[68] Yongfei Hu, Xi Yang, Jing Li, Na Lv, Fei Liu, Jun Wu, Ivan YC Lin, Na Wu, Bart C Weimer, George F Gao, et al. The bacterial mobile resistome transfer network connecting the animal and human microbiomes. *Applied and environmental microbiology*, 82(22):6672–6681, 2016.

[69] Kimberly A Bliven and Anthony T Maurelli. Evolution of bacterial pathogens within the human host. *Microbiology spectrum*, 4(1):4–1, 2016.

[70] Jan-Hendrik Hehemann, Gaëlle Correc, Tristan Barbeyron, William Helbert, Mirjam Czjzek, and Gurvan Michel. Transfer of carbohydrate-active enzymes from marine bacteria to japanese gut microbiota. *Nature*, 464(7290):908–912, 2010.

[71] Chris S. Smillie, Mark B. Smith, Jonathan Friedman, Otto X. Cordero, Lawrence A. David, and Eric J. Alm. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376):241–244, 2011. Number: 7376 Publisher: Nature Publishing Group.

[72] Mislav Acman, Lucy van Dorp, Joanne M Santini, and Francois Balloux. Large-scale network analysis captures biological features of bacterial plasmids. *Nature communications*, 11(1):1–11, 2020.

[73] Mor N Lurie-Weinberger, Michael Peeri, and Uri Gophna. Contribution of lateral gene transfer to the gene repertoire of a gut-adapted methanogen. *Genomics*, 99(1):52–58, 2012.

[74] Abdessamad El Kaoutari, Fabrice Armougom, Jeffrey I Gordon, Didier Raoult, and Bernard Henrissat. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, 11(7):497–504, 2013.

[75] Francesco Riva, Valentina Riva, Ester M. Eckert, Noemi Colinas, Andrea Di Cesare, Sara Borin, Francesca Mapelli, and Elena Crotti. An environmental escherichia coli strain is naturally competent to acquire exogenous DNA. *Frontiers in Microbiology*, 11, 2020.

[76] Didier Debroas and Cléa Siguret. Viruses as key reservoirs of antibiotic resistance genes in the environment. *The ISME Journal*, 13(11):2856–2867, 2019. Number: 11 Publisher: Nature Publishing Group.

[77] Dietmar Fernández-Orth, Elisenda Miró, Maryury Brown-Jaque, Lorena Rodríguez-Rubio, Paula Espinal, Judith Rodriguez-Navarro, Juan José González-López, Maite Muniesa, and Ferran Navarro. Faecal phageome of healthy individuals: presence of antibiotic resistance genes and variations caused by ciprofloxacin treatment. *Journal of Antimicrobial Chemotherapy*, 74(4):854–864, 2019.

[78] Dongchang Sun, Katy Jeannot, Yonghong Xiao, and Charles W. Knapp. Editorial: Horizontal gene transfer mediated bacterial antibiotic resistance. *Frontiers in Microbiology*, 10, 2019.

[79] Neil Woodford, Jane F. Turton, and David M. Livermore. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiology Reviews*, 35(5):736–755, 09 2011.

[80] Srujana Mohanty, Ritu Singhal, Seema Sood, Benu Dhawan, Arti Kapil, and Bimal K Das. Citrobacter infections in a tertiary care hospital in northern india. *Journal of Infection*, 54(1):58–64, 2007.

[81] Melissa L Hargreaves, Kristin M Shaw, Ginette Dobbins, Paula M Snippes Vagnone, Jane E Harper, Dave Boxrud, Ruth Lynfield, Maliha Aziz, Lance B Price, Kevin AT Silverstein, et al. Clonal dissemination of enterobacter cloacae harboring bla kpc-3 in the upper midwestern united states. *Antimicrobial agents and chemotherapy*, 59(12):7723–7734, 2015.

[82] MJD Dautzenberg, MR Haverkate, MJM Bonten, and MCJ Bootsma. Epidemic potential of escherichia coli st131 and klebsiella pneumoniae st258: a systematic review and meta-analysis. *BMJ open*, 6(3):e009971, 2016.

[83] Christine Martineau, Xuejing Li, Cindy Lalancette, Thérèse Perreault, Eric Fournier, Julien Tremblay, Milagros Gonzales, Étienne Yergeau, and Caroline Quach. Serratia marcescens outbreak in a neonatal intensive care unit: new insights from next-generation sequencing applications. *Journal of clinical microbiology*, 56(9):e00235–18, 2018.

[84] Edwin C Pereira, Melissa Anacker, Jeana Houseman, Mary E Horn, Timothy J Johnson, Ruth Lynfield, Paula Snippes Vagnone, Medora Witwer, and Susan Kline. A cluster of carbapenemase-producing enterobacter cloacae complex st171 at a tertiary care center demonstrating an ongoing regional threat. *American journal of infection control*, 47(7):767–772, 2019.

[85] L.W. Riley. Pandemic lineages of extraintestinal pathogenic escherichia coli. *Clinical Microbiology and Infection*, 20(5):380–390, 2014.

[86] Sheila Adams-Sapper, Binh An Diep, Francoise Perdreau-Remington, and Lee W Riley. Clonal composition and community clustering of drug-susceptible and-resistant escherichia coli isolates from bloodstream infections. *Antimicrobial agents and chemotherapy*, 57(1):490–497, 2013.

[87] Giancarlo Ripabelli, Michela Lucia Sammarco, Massimiliano Scutellà, Valentina Felice, and Manuela Tamburro. Carbapenem-resistant kpc-and tem-producing escherichia coli st131 isolated from a hospitalized patient with urinary tract infection: first isolation in molise region, central italy, july 2018. *Microbial Drug Resistance*, 26(1):38–45, 2020.

[88] Lin Gong, Na Tang, Dongke Chen, Kaiwen Sun, Ruiting Lan, Wen Zhang, Haijian Zhou, Min Yuan, Xia Chen, Xiaofei Zhao, et al. A nosocomial respiratory infection outbreak of carbapenem-resistant escherichia coli st131 with multiple transmissible bla kpc–2 carrying plasmids. *Frontiers in microbiology*, 11:2068, 2020.

[89] Brian D Johnston, Paul Thuras, Stephen B Porter, Melissa Anacker, Brittany VonBank, Paula Snippes Vagnone, Medora Witwer, Mariana Castanheira, and James R Johnson.

Global molecular epidemiology of carbapenem-resistant escherichia coli (2002–2017). *European Journal of Clinical Microbiology & Infectious Diseases*, pages 1–13, 2021.

[90] Miran Tang, Xin Kong, Jingchen Hao, and Jinbo Liu. Epidemiological characteristics and formation mechanisms of multidrug-resistant hypervirulent klebsiella pneumoniae. *Frontiers in Microbiology*, 11:581543, 2020.

[91] Kelly L .Wyres and Kathryn E Holt. Klebsiella pneumoniae as a key trafficker of drug resistance genes from environmental to clinically important bacteria. *Current Opinion in Microbiology*, 45:131–139, 2018. Antimicrobials * Microbial systems biology.

[92] Jorge A. Ramos-Castañeda, Alberto Ruano-Ravina, Raquel Barbosa-Lorenzo, Jaime E. Paillier-Gonzalez, Javier C. Saldaña-Campos, Diego F. Salinas, and Elkin V. Lemos-Luengas. Mortality due to kpc carbapenemase-producing klebsiella pneumoniae infections: Systematic review and meta-analysis: Mortality due to kpc klebsiella pneumoniae infections. *Journal of Infection*, 76(5):438–448, 2018.

[93] Melissa L. Hargreaves, Kristin M. Shaw, Ginette Dobbins, Paula M. Snippes Vagnone, Jane E. Harper, Dave Boxrud, Ruth Lynfield, Maliha Aziz, Lance B. Price, Kevin A. T. Silverstein, Jessica L. Danzeisen, Bonnie Youmans, Kyle Case, Srinand Sreevatsan, and Timothy J. Johnson. Clonal dissemination of enterobacter cloacae harboring blaKPC-3 in the upper midwestern united states. *Antimicrobial Agents and Chemotherapy*, 59(12):7723–7734, 2015. Publisher: American Society for Microbiology.

[94] Angela Gomez-Simmonds, Yue Hu, Sean B. Sullivan, Zheng Wang, Susan Whittier, and Anne-Catrin Uhlemann. Evidence from a new york city hospital of rising incidence of genetically diverse carbapenem-resistant enterobacter cloacae and dominance of ST171, 2007–14. *Journal of Antimicrobial Chemotherapy*, 71(8):2351–2353, 2016.

[95] Hajime Kanamori, Christian M. Parobek, Jonathan J. Juliano, David van Duin, Bruce A. Cairns, David J. Weber, and William A. Rutala. A prolonged outbreak of KPC-3-producing enterobacter cloacae and klebsiella pneumoniae driven by multiple mechanisms of resistance transmission at a large academic burn center. *Antimicrobial Agents and Chemotherapy*, 61(2):e01516–16, 2017. Publisher: American Society for Microbiology.

[96] Shawn E Hawken, Laraine L Washer, Christopher L Williams, Duane W Newton, and Evan S Snitkin. Genomic investigation of a putative endoscope-associated carbapenem-resistant enterobacter cloacae outbreak reveals a wide diversity of circulating strains and resistance mutations. *Clinical Infectious Diseases*, 66(3):460–463, 2018.

[97] Chulsoo Ahn, Alveena Syed, Fupin Hu, Jessica A. O'Hara, Jesabel I. Rivera, and Yohei Doi. Microbiological features of KPC-producing enterobacter isolates identified in a u.s. hospital system. *Diagnostic Microbiology and Infectious Disease*, 80(2):154–158, 2014.

[98] Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature reviews genetics*, 13(5):303–314, 2012.

[99] Martin Kreitman. Nucleotide polymorphism at the alcohol dehydrogenase locus of drosophila melanogaster. *Nature*, 304(5925):412–417, 1983.

[100] Scott V Edwards. Is a new and general theory of molecular systematics emerging? *Evolution: International Journal of Organic Evolution*, 63(1):1–19, 2009.

[101] Nicolas Galtier and Vincent Daubin. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):4023–4029, 2008.

[102] Antonis Rokas, Dirk Kruger, and Sean B Carroll. Animal evolution and the molecular signature of radiations compressed in time. *science*, 310(5756):1933–1938, 2005.

[103] Francesca D Ciccarelli, Tobias Doerks, Christian Von Mering, Christopher J Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *science*, 311(5765):1283–1287, 2006.

[104] Hervé Philippe, Nicolas Lartillot, and Henner Brinkmann. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Molecular biology and evolution*, 22(5):1246–1253, 2005.

[105] Qiyun Zhu, Uyen Mai, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G Sanders, Pedro Belda-Ferre, Gabriel A Al-Ghalith, Evguenia Kopylova, Daniel McDonald, et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nature communications*, 10(1):1–14, 2019.

[106] Richard G Olmstead and Jennifer A Sweere. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the solanaceae. *Systematic Biology*, 43(4):467–481, 1994.

[107] Jessica W Leigh, Edward Susko, Manuela Baumgartner, and Andrew J Roger. Testing congruence in phylogenomic analysis. *Systematic Biology*, 57(1):104–115, 2008.

[108] João Tonini, Andrew Moore, David Stern, Maryia Shcheglovitova, and Guillermo Ortí. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. *PLoS currents*, 7, 2015.

[109] Sudhindra R Gadagkar, Michael S Rosenberg, and Sudhir Kumar. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 304(1):64–74, 2005.

[110] J Rajendhran and P Gunasekaran. Microbial phylogeny and diversity: small subunit ribosomal rna sequence analysis and beyond. *Microbiological research*, 166(2):99–110, 2011.

[111] Anand Patwardhan, Samit Ray, and Amit Roy. Molecular markers in phylogenetic studies-a review. *Journal of Phylogenetics & Evolutionary Biology*, 2014, 2014.

[112] Martin CJ Maiden, Jane A Bygraves, Edward Feil, Giovanna Morelli, Joanne E Russell, Rachel Urwin, Qing Zhang, Jiaji Zhou, Kerstin Zurth, Dominique A Caugant, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145, 1998.

[113] Nicola Segata, Daniela Börnigen, Xochitl C Morgan, and Curtis Huttenhower. Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, 4(1):1–11, 2013.

[114] Ramy K Aziz, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, et al. The rast server: rapid annotations using subsystems technology. *BMC genomics*, 9(1):1–15, 2008.

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

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

[117] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[118] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[119] Dirk Eddelbuettel and Romain Francois. Rcpp: Seamless r and c++ integration. *Journal of statistical software*, 40:1–18, 2011.

[120] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.

[121] Jeremy R Dettman, Jacqueline L Sztepanacz, and Rees Kassen. The properties of spontaneous mutations in the opportunistic pathogen pseudomonas aeruginosa. *BMC genomics*, 17(1):1–14, 2016.

[122] Gerry Tonkin-Hill, Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh, John A Lees, Rebecca A Gladstone, Stephanie Lo, Christopher Beaudoin, R Andres Floto, et al. Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome biology*, 21(1):1–21, 2020.

[123] Nathan T Porter, Andrew J Hryckowian, Bryan D Merrill, Jaime J Fuentes, Jackson O Gardner, Robert WP Glowacki, Shaleni Singh, Ryan D Crawford, Evan S Snitkin, Justin L Sonnenburg, et al. Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in bacteroides thetaiotaomicron. *Nature microbiology*, 5(9):1170–1181, 2020.

[124] Kazutaka Katoh, John Rozewicki, and Kazunori D Yamada. Mafft online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*, 20(4):1160–1166, 2019.

[125] Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. *Journal of computational biology*, 1(4):337–348, 1994.

[126] Eugene Koonin and Michael Y Galperin. *Sequence—evolution—function: computational approaches in comparative genomics*. Springer Science & Business Media, 2002.

[127] Stephen J Bush, Dona Foster, David W Eyre, Emily L Clark, Nicola De Maio, Liam P Shaw, Nicole Stoesser, Tim EA Peto, Derrick W Crook, and A Sarah Walker. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. *GigaScience*, 9(2):giaa007, 2020.

[128] John A Lees, Simon R Harris, Gerry Tonkin-Hill, Rebecca A Gladstone, Stephanie W Lo, Jeffrey N Weiser, Jukka Corander, Stephen D Bentley, and Nicholas J Croucher. Fast and flexible bacterial genomic epidemiology with poppunk. *Genome research*, 29(2):304–316, 2019.

[129] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):1–14, 2016.

[130] Joseph J Gillespie, Alice R Wattam, Stephen A Cammer, Joseph L Gabbard, Maulik P Shukla, Oral Dalay, Timothy Driscoll, Deborah Hix, Shrinivasrao P Mane, Chunhong Mao, et al. Patric: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and immunity*, 79(11):4286–4298, 2011.

[131] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[132] Bruce Parrello, Rory Butler, Philippe Chlenski, Robert Olson, Jamie Overbeek, Gordon D Pusch, Veronika Vonstein, and Ross Overbeek. A machine learning-based service for estimating quality of genomes using patric. *BMC bioinformatics*, 20(1):1–9, 2019.

[133] Mary K Hayden, Michael Y Lin, Karen Lolans, Shayna Weiner, Donald Blom, Nicholas M Moore, Louis Fogg, David Henry, Rosie Lyles, Caroline Thurlow, et al. Prevention of colonization and infection by klebsiella pneumoniae carbapenemase–producing enterobacteriaceae in long-term acute-care hospitals. *Clinical Infectious Diseases*, 60(8):1153–1161, 2015.

[134] F. Arena, F. Vannetti, V. Di Pilato, L. Fabbri, O. L. Colavecchio, T. Giani, C. Marraccini, R. Pupillo, C. Macchi, F. Converti, and G. M. Rossolini. Diversity of the epidemiology of carbapenemase-producing enterobacteriaceae in long-term acute care rehabilitation settings from an area of hyperendemicity, and evaluation of an intervention bundle. *Journal of Hospital Infection*, 100(1):29–34, 2018.

[135] Minhui Miao, Huiyan Wen, Ping Xu, Siqiang Niu, Jingnan Lv, Xiaofang Xie, José R. Mediavilla, Yi-Wei Tang, Barry N. Kreiswirth, Xia Zhang, Haifang Zhang, Hong Du, and Liang Chen. Genetic diversity of carbapenem-resistant enterobacteriaceae (CRE) clinical isolates from a tertiary hospital in eastern china. *Frontiers in Microbiology*, 9, 2019.

[136] Beiwen Zheng, Hao Xu, Lihua Guo, Xiao Yu, Jinru Ji, Chaoqun Ying, Yunbo Chen, Ping Shen, Huiming Han, Chen Huang, Shuntian Zhang, Tao Lv, and Yonghong Xiao. Genomic and phenotypic diversity of carbapenemase-producing enterobacteriaceae isolates from bacteremia in china: A multicenter epidemiological, microbiological, and genetic study. *Engineering*, 2020.

[137] Liang Chen, Barun Mathema, Kalyan D. Chavda, Frank R. DeLeo, Robert A. Bonomo, and Barry N. Kreiswirth. Carbapenemase-producing klebsiella pneumoniae: molecular and genetic decoding. *Trends in Microbiology*, 22(12):686–696, 2014.

[138] Anita C. Schürch and Willem van Schaik. Challenges and opportunities for whole-genome sequencing–based surveillance of antibiotic resistance. *Annals of the New York Academy of Sciences*, 1388(1):108–120, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nyas.13310.

[139] Bruce Y. Lee, Sarah M. Bartsch, Kim F. Wong, Diane S. Kim, Chenghua Cao, Leslie E. Mueller, Gabrielle M. Gussin, James A. McKinnell, Loren G. Miller, and Susan S. Huang. Tracking the spread of carbapenem-resistant enterobacteriaceae (CRE) through clinical cultures alone underestimates the spread of CRE even more than anticipated. *Infection Control & Hospital Epidemiology*, 40(6):731–734, 2019. Publisher: Cambridge University Press.

[140] Amos Adler, Efrat Khabra, Svetlana Paikin, and Yehuda Carmeli. Dissemination of the blaKPC gene by clonal spread and horizontal gene transfer: comparative study of incidence and molecular mechanisms. *The Journal of Antimicrobial Chemotherapy*, 71(8):2143–2146, 2016.

[141] S. Breurec, N. Guessennd, M. Timinouni, T. T. H. Le, V. Cao, A. Ngandjio, F. Randrianirina, J. M. Thiberge, A. Kinana, A. Dufougeray, J. D. Perrier-Gros-Claude, P. Boisier, B. Garin, and S. Brisse. Klebsiella pneumoniae resistant to third-generation cephalosporins in five african and two vietnamese major towns: multiclonal population structure with two major international clonal groups, CG15 and CG258. *Clinical Microbiology and Infection*, 19(4):349–355, 2013. Publisher: Elsevier.

[142] Centers for Diesase Control. Current report | antibiotic use | CDC.

[143] Shelley S. Magill, Ghinwa Dumyati, Susan M. Ray, and Scott K. Fridkin. Evaluating epidemiology and improving surveillance of infections associated with health care, united states. *Emerging Infectious Diseases*, 21(9):1537–1542, 2015.

[144] Alice R. Wattam, David Abraham, Oral Dalay, Terry L. Disz, Timothy Driscoll, Joseph L. Gabbard, Joseph J. Gillespie, Roger Gough, Deborah Hix, Ronald Kenyon, Dustin Machi, Chunhong Mao, Eric K. Nordberg, Robert Olson, Ross Overbeek, Gordon D. Pusch, Maulik Shukla, Julie Schulman, Rick L. Stevens, Daniel E. Sullivan, Veronika Vonstein, Andrew

Warren, Rebecca Will, Meredith J. C. Wilson, Hyun Seung Yoo, Chengdong Zhang, Yan Zhang, and Bruno W. Sobral. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*, page gkt1099, 2013. 00000.

[145] Ryan D. Crawford and Evan S. Snitkin. cognac: rapid generation of concatenated gene alignments for phylogenetic inference from large, bacterial whole genome sequencing datasets. *BMC Bioinformatics*, 22(1):70, 2021.

[146] M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7):1641–1650, 2009.

[147] Jeffrey B. Joy, Richard H. Liang, Rosemary M. McCloskey, T. Nguyen, and Art F. Y. Poon. Ancestral reconstruction. *PLOS Computational Biology*, 12(7):e1004763, 2016. Publisher: Public Library of Science.

[148] David L. Swofford and Wayne P. Maddison. Reconstructing ancestral character states under wagner parsimony. *Mathematical Biosciences*, 87(2):199–229, 1987.

[149] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2021.

[150] Liam J. Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012.

[151] Nicole M Koropatkin, Elizabeth A Cameron, and Eric C Martens. How glycan metabolism shapes the human gut microbiota. *Nature Reviews Microbiology*, 10(5):323–335, 2012.

[152] Harry J Flint, Karen P Scott, Sylvia H Duncan, Petra Louis, and Evelyne Forano. Microbial degradation of complex carbohydrates in the gut. *Gut microbes*, 3(4):289–306, 2012.

[153] Nathan T Porter and Eric C Martens. The critical roles of polysaccharides in gut microbial ecology and physiology. *Annual review of microbiology*, 71:349–369, 2017.

[154] NI McNeil. The contribution of the large intestine to energy supplies in man. *The American journal of clinical nutrition*, 39(2):338–342, 1984.

[155] Mahesh S Desai, Anna M Seekatz, Nicole M Koropatkin, Nobuhiko Kamada, Christina A Hickey, Mathis Wolter, Nicholas A Pudlo, Sho Kitamoto, Nicolas Terrapon, Arnaud Muller, et al. A dietary fiber-deprived gut microbiota degrades the colonic mucus barrier and enhances pathogen susceptibility. *Cell*, 167(5):1339–1353, 2016.

[156] Erica D Sonnenburg, Samuel A Smits, Mikhail Tikhonov, Steven K Higginbottom, Ned S Wingreen, and Justin L Sonnenburg. Diet-induced extinctions in the gut microbiota compound over generations. *Nature*, 529(7585):212–215, 2016.

[157] Erica D Sonnenburg and Justin L Sonnenburg. Starving our microbial self: the deleterious consequences of a diet deficient in microbiota-accessible carbohydrates. *Cell metabolism*, 20(5):779–786, 2014.

[158] Laura Wrzosek, Sylvie Miquel, Marie-Louise Noordine, Stephan Bouet, Marie Joncquel Chevalier-Curt, Véronique Robert, Catherine Philippe, Chantal Bridonneau, Claire Cherbuy, Catherine Robbe-Masselot, et al. Bacteroides thetaiotaomicron and faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. *BMC biology*, 11(1):1–13, 2013.

[159] Vadivel Ganapathy, Muthusamy Thangaraju, Puttur D Prasad, Pamela M Martin, and Nagendra Singh. Transporters and receptors for short-chain fatty acids as the molecular link between colonic bacteria and the host. *Current opinion in pharmacology*, 13(6):869–874, 2013.

[160] SI Cook and JH Sellin. Short chain fatty acids in health and disease. *Alimentary pharmacology & therapeutics*, 12(6):499–507, 1998.

[161] Patrick M Smith, Michael R Howitt, Nicolai Panikov, Monia Michaud, Carey Ann Gallini, Mohammad Bohlooly-y, Jonathan N Glickman, and Wendy S Garrett. The microbial metabolites, short-chain fatty acids, regulate colonic treg cell homeostasis. *Science*, 341(6145):569–573, 2013.

[162] Myunghoo Kim, Yaqing Qie, Jeongho Park, and Chang H Kim. Gut microbial metabolites fuel host antibody responses. *Cell host & microbe*, 20(2):202–214, 2016.

[163] Jian Xu, Michael A Mahowald, Ruth E Ley, Catherine A Lozupone, Micah Hamady, Eric C Martens, Bernard Henrissat, Pedro M Coutinho, Patrick Minx, Philippe Latreille, et al. Evolution of symbiotic bacteria in the distal human intestine. *PLoS biology*, 5(7):e156, 2007.

[164] Brandi L Cantarel, Pedro M Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. The carbohydrate-active enzymes database (cazy): an expert resource for glycogenomics. *Nucleic acids research*, 37(suppl_1):D233–D238, 2009.

[165] Nathan P McNulty, Meng Wu, Alison R Erickson, Chongle Pan, Brian K Erickson, Eric C Martens, Nicholas A Pudlo, Brian D Muegge, Bernard Henrissat, Robert L Hettich, et al. Effects of diet on resource utilization by a model human gut microbiota containing bacteroides cellulosilyticus wh2, a symbiont with an extensive glycobiome. *PLoS biology*, 11(8):e1001637, 2013.

[166] AA Salyers, JR Vercellotti, SE West, and TD Wilkins. Fermentation of mucin and plant polysaccharides by strains of bacteroides from the human colon. *Applied and environmental microbiology*, 33(2):319–322, 1977.

[167] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *nature*, 464(7285):59–65, 2010.

[168] Ruth E Ley, Micah Hamady, Catherine Lozupone, Peter J Turnbaugh, Rob Roy Ramey, J Stephen Bircher, Michael L Schlegel, Tammy A Tucker, Mark D Schrenzel, Rob Knight, et al. Evolution of mammals and their gut microbes. *science*, 320(5883):1647–1651, 2008.

[169] Paul B Eckburg, Elisabeth M Bik, Charles N Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R Gill, Karen E Nelson, and David A Relman. Diversity of the human intestinal microbial flora. *science*, 308(5728):1635–1638, 2005.

[170] Eric C Martens, Herbert C Chiang, and Jeffrey I Gordon. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell host & microbe*, 4(5):447–457, 2008.

[171] Eric C Martens, Elisabeth C Lowe, Herbert Chiang, Nicholas A Pudlo, Meng Wu, Nathan P McNulty, D Wade Abbott, Bernard Henrissat, Harry J Gilbert, David N Bolam, et al. Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS biology*, 9(12):e1001221, 2011.

[172] Seth M Bloom, Vinieth N Bijanki, Gerardo M Nava, Lulu Sun, Nicole P Malvin, David L Donermeyer, W Michael Dunne Jr, Paul M Allen, and Thaddeus S Stappenbeck. Commensal bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. *Cell host & microbe*, 9(5):390–403, 2011.

[173] Christina A Hickey, Kristine A Kuhn, David L Donermeyer, Nathan T Porter, Chunsheng Jin, Elizabeth A Cameron, Haerin Jung, Gerard E Kaiko, Marta Wegorzewska, Nicole P Malvin, et al. Colitogenic bacteroides thetaiotaomicron antigens access host immune cells in a sulfatase-dependent manner via outer membrane vesicles. *Cell host & microbe*, 17(5):672–680, 2015.

[174] Sarkis K Mazmanian, June L Round, and Dennis L Kasper. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature*, 453(7195):620–625, 2008.

[175] Ilias Lagkouvardos, Rüdiger Pukall, Birte Abt, Bärbel U Foesel, Jan P Meier-Kolthoff, Neeraj Kumar, Anne Bresciani, Inés Martínez, Sarah Just, Caroline Ziegler, et al. The mouse intestinal bacterial collection (mibc) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nature microbiology*, 1(10):1–15, 2016.

[176] Hilary P Browne, Samuel C Forster, Blessing O Anonye, Nitin Kumar, B Anne Neville, Mark D Stares, David Goulding, and Trevor D Lawley. Culturing of 'unculturable'human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604):543–546, 2016.

[177] Jean-Christophe Lagier, Saber Khelaifia, Maryam Tidjani Alou, Sokhna Ndongo, Niokhor Dione, Perrine Hugon, Aurelia Caputo, Frédéric Cadoret, Sory Ibrahima Traore, El Hadji Seck, et al. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature microbiology*, 1(12):1–8, 2016.

[178] Melanie Tramontano, Sergej Andrejev, Mihaela Pruteanu, Martina Klünemann, Michael Kuhn, Marco Galardini, Paula Jouhten, Aleksej Zelezniak, Georg Zeller, Peer Bork, et al. Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. *Nature microbiology*, 3(4):514–522, 2018.

[179] Jan-Hendrik Hehemann, Amelia G Kelly, Nicholas A Pudlo, Eric C Martens, and Alisdair B Boraston. Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. *Proceedings of the National Academy of Sciences*, 109(48):19786–19791, 2012.

[180] Kazune Tamura, Glyn R Hemsworth, Guillaume Déjean, Theresa E Rogers, Nicholas A Pudlo, Karthik Urs, Namrata Jain, Gideon J Davies, Eric C Martens, and Harry Brumer. Molecular mechanism by which prominent human gut bacteroidetes utilize mixed-linkage beta-glucans, major health-promoting cereal polysaccharides. *Cell reports*, 21(2):417–430, 2017.

[181] Joseph A Shipman, James E Berleman, and Abigail A Salyers. Characterization of four outer membrane proteins involved in binding starch to the cell surface of bacteroides thetaiotaomicron. *Journal of bacteriology*, 182(19):5365–5372, 2000.

[182] Guillaume Déjean, Kazune Tamura, Adriana Cabrera, Namrata Jain, Nicholas A Pudlo, Gabriel Pereira, Alexander Holm Viborg, Filip Van Petegem, Eric C Martens, and Harry Brumer. Synergy between cell surface glycosidases and glycan-binding proteins dictates the utilization of specific beta (1, 3)-glucans by human gut bacteroides. *Mbio*, 11(2):e00095–20, 2020.

[183] Nicholas A Pudlo, Karthik Urs, Supriya Suresh Kumar, J Bruce German, David A Mills, and Eric C Martens. Symbiotic human gut bacteria with variable metabolic priorities for host mucosal glycans. *MBio*, 6(6):e01282–15, 2015.

[184] Jordane Despres, Evelyne Forano, Pascale Lepercq, Sophie Comtet-Marre, Gregory Jubelin, Christophe Chambon, Carl J Yeoman, Margaret E Berg Miller, Christopher J Fields, Eric Martens, et al. Xylan degradation by the human gut bacteroides xylanisolvens xb1a t involves two distinct gene clusters that are linked at the transcriptional level. *BMC genomics*, 17(1):1–14, 2016.

[185] Erica D Sonnenburg, Hongjun Zheng, Payal Joglekar, Steven K Higginbottom, Susan J Firbank, David N Bolam, and Justin L Sonnenburg. Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell*, 141(7):1241–1252, 2010.

[186] Nicolas Terrapon, Vincent Lombard, Elodie Drula, Pascal Lapébie, Saad Al-Masaudi, Harry J Gilbert, and Bernard Henrissat. Puldb: the expanded database of polysaccharide utilization loci. *Nucleic Acids Research*, 46(D1):D677–D683, 2018.

[187] Malin EV Johansson, Jessica M Holmén Larsson, and Gunnar C Hansson. The two mucus layers of colon are organized by the muc2 mucin, whereas the outer layer is a legislator of host–microbial interactions. *Proceedings of the national academy of sciences*, 108(Supplement 1):4659–4665, 2011.

[188] Nicholas A Pudlo, Gabriel Vasconcelos Pereira, Jaagni Parnami, Melissa Cid, Stephanie Markert, Jeffrey P Tingley, Frank Unfried, Ahmed Ali, Austin Campbell, Karthik Urs, et al.

Extensive transfer of genes for edible seaweed digestion from marine to human gut bacteria. *bioRxiv*, 2020.

[189] Eric C Martens, Amelia G Kelly, Alexandra S Tauzin, and Harry Brumer. The devil lies in the details: how variations in polysaccharide fine-structure impact the physiology and evolution of gut microbes. *Journal of molecular biology*, 426(23):3851–3865, 2014.

[190] Johan Larsbrink, Theresa E. Rogers, Glyn R. Hemsworth, Lauren S. McKee, Alexandra S. Tauzin, Oliver Spadiut, Stefan Klinter, Nicholas A. Pudlo, Karthik Urs, Nicole M. Koropatkin, A. Louise Creagh, Charles A. Haynes, Amelia G. Kelly, Stefan Nilsson Cederholm, Gideon J. Davies, Eric C. Martens, and Harry Brumer. A discrete genetic locus confers xyloglucan metabolism in select human gut bacteroidetes. *Nature*, 506(7489):498–502, 2014. Number: 7489 Publisher: Nature Publishing Group.

[191] Kyung Moon, Justin Sonnenburg, and Abigail A Salyers. Unexpected effect of a bacteroides conjugative transposon, ctndot, on chromosomal gene expression in its bacterial host. *Molecular microbiology*, 64(6):1562–1571, 2007.

[192] Fasahath Husain, Kevin Tang, Yaligara Veeranagouda, Renata Boente, Sheila Patrick, Garry Blakely, and Hannah M.YR 2017 Wexler. Novel large-scale chromosomal transfer in bacteroides fragilis contributes to its pan-genome and rapid environmental adaptation. *Microbial Genomics*, 3(11):e000136, 2017. Publisher: Microbiology Society,.

[193] Gabrielle Whittle, Nathan Hamburger, Nadja B. Shoemaker, and Abigail A. Salyers. A bacteroides conjugative transposon, CTnERL, can transfer a portion of itself by conjugation without excising from the chromosome. *Journal of Bacteriology*, 188(3):1169–1174, 2006. Publisher: American Society for Microbiology.

[194] Sumitha K. Reddy, Viktoria Bågenholm, Nicholas A. Pudlo, Hanene Bouraoui, Nicole M. Koropatkin, Eric C. Martens, and Henrik Stålbrand. A $beta$-mannan utilization locus in bacteroides ovatus involves a GH36 $alpha$-galactosidase active on galactomannans. *FEBS Letters*, 590(14):2106–2118, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1873-3468.12250.

[195] Patrick D. Schloss, Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, Brian B. Oakley, Donovan H. Parks, Courtney J. Robinson, Jason W. Sahl, Blaz Stres, Gerhard G. Thallinger, David J. Van Horn, and Carolyn F. Weber. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009. Publisher: American Society for Microbiology.

[196] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[197] Lillian V. Holdeman, 1929, Walter Edward C. Moore, and Elizabeth P. Cato. *Anaerobe laboratory manual*. 1977.

[198] Andrei-Alin Popescu, Katharina T. Huber, and Emmanuel Paradis. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in r. *Bioinformatics*, 28(11):1536–1537, 2012.

[199] Eugene V Koonin, Kira S Makarova, and Yuri I Wolf. Evolution of microbial genomics: conceptual shifts over a quarter century. *Trends in Microbiology*, 29(7):582–592, 2021.

[200] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, 2016.

[201] Zena Lapp, Ryan Crawford, Arianna Miles-Jay, Ali Pirani, William E Trick, Robert A Weinstein, Mary K Hayden, Evan S Snitkin, Michael Y Lin, and CDC Prevention Epicenters Program. Regional spread of blaNDM-1-containing klebsiella pneumoniae ST147 in post-acute care facilities. *Clinical Infectious Diseases*, 73(8):1431–1439, 2021.

[202] Angela Gomez-Simmonds, Medini K. Annavajhala, Nina Tang, Felix D. Rozenberg, Mehrose Ahmad, Heekuk Park, Allison J. Lopatkin, and Anne-Catrin Uhlemann. Population structure of blaKPC-harbouring IncN plasmids at a new york city medical centre and evidence for multi-species horizontal transmission. *Journal of Antimicrobial Chemotherapy*, page dkac114, 2022.

[203] J. Christian Pérez. Fungi of the human gut microbiota: Roles and significance. *International Journal of Medical Microbiology*, 311(3):151490, 2021.

[204] Soo Jen Low, Mária Džunková, Pierre-Alain Chaumeil, Donovan H. Parks, and Philip Hugenholtz. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order caudovirales. *Nature Microbiology*, 4(8):1306–1315, 2019. Number: 8 Publisher: Nature Publishing Group.

[205] Etienne P. de Villiers, Carmina Gallardo, Marisa Arias, Melissa da Silva, Chris Upton, Raquel Martin, and Richard P. Bishop. Phylogenomic analysis of 11 complete african swine fever virus genome sequences. *Virology*, 400(1):128–136, 2010.

[206] Matthew A. Gitzendanner, Pamela S. Soltis, Gane K.-S. Wong, Brad R. Ruhfel, and Douglas E. Soltis. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany*, 105(3):291–301, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajb2.1048.

[207] Mikkel H Schierup and Jotun Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2):879–891, 2000.

[208] Nicholas J. Croucher, Andrew J. Page, Thomas R. Connor, Aidan J. Delaney, Jacqueline A. Keane, Stephen D. Bentley, Julian Parkhill, and Simon R. Harris. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic Acids Research*, 43(3):e15, 2015.