THE UNIVERSITY OF CHICAGO

REVEALING THE ECOLOGICAL INTERACTIONS, EVOLUTIONARY HISTORIES AND NICHE BOUNDARIES OF PREVALENT HUMAN GUT PLASMIDS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MICROBIOLOGY

BY
EMILY CLARE FOGARTY

CHICAGO, ILLINOIS
AUGUST 2023

Copyright © 2023 by Emily Clare Fogarty

All Rights Reserved

This thesis is dedicated to my grandma, who was saving for my education before I even knew what school was.

"Even when not fully attained, we become better by striving for a higher goal." - *Victor Frankl, Man's Search for Meaning*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

All supplementary tables for Chapter 2 were submitted as a separate supplemental file in ProQuest.

All supplementary tables for Chapter 3 are available at https://doi.org/10.6084/m9.figshare.22336666.

ACKNOWLEDGMENTS

I am eternally grateful for all of the people who have helped me get to where I am today. I feel incredibly fortunate that grad school has been some of the best years of my life.

Committee on Microbiology, and really the University of Chicago in general. I feel so fortunate to have come here and I have so much gratitude for the university that gave me a home for the last 6 years and changed the trajectory of my entire life.

Meren, you had total faith in my ability to learn to code, despite me showing up with no idea what even the command line was. You led by example, pushed me to be the best possible scientist I could be, and showed us all that standing up for what you believe in scientifically, ethically, or both encourages others to do the same, even when it's unconventional. You taught me that work environment is critical, and promoted an ecosystem of learning, sharing and camaraderie. Those pre-covid lab memories will hold a special place in my heart forever.

Laurie, I can't begin to describe how grateful I am to you for taking me into your lab and giving me a second scientific home. I've lost track of how many times I've laughed and cried in your office, and feel so fortunate to have you as both a role model, and many ways, a friend.

Mike, you were like a 3rd unofficial advisor to me. Mike and I spent >300 hours together on zoom writing and editing the plasmid prediction work, and I learned more about science, writing and life that I would have ever imagined. Although thanks to Mike, maybe I could now develop a machine learning model to predict that.

My thesis committee, thank you for all your support over the years. I don't know how many students can say they have all of their committee members on their thesis paper, and I enjoyed working with each one of you.

To the Meren Lab, you know who you are. My favorite example of how incredibly supportive our lab is, was Evan spending 6 hours one day to give me a crash course in R. But each and every one of you made coming to work an absolute joy.

To my grad school friends, Meike, Matt, Steven, Kourtney, Evan, Cat, Devon, Jimmy, Vaughn, Fernando, Liz and everyone else that I haven't named: Sunday night dinners are forever <3. It wouldn't have been the same experience without you.

To my physical therapists and chiropractors, and especially Dusten, thank you for putting me back together, tboth mentally and physically. Dusten, I needed the tough love and I can't thank you enough for convincing me that I can get back to where I was pre-injury.

To my family, and most of all my Mom, there are no words- your love and support means the world over to me.

And last, I want to thank myself. Grad school is hard, but chronic pain starting at 25 is harder. For the last half of grad school I dealt with a severe back injury from weightlifting. Despite 3 years of chronic pain and inability to do many of the things I love, this experience has taught me to be grateful for simple things in my life, has forced me to let go of my ego in many situations, and has changed my perspective on what success means to me. It took an incredible amount of perserverence and self-determination to make it through those years, but it's changed me into a more reflective, grateful, and self-sufficient person.

ABSTRACT

Microbes exist in virtually all environments on Earth's surface. As asexually reproducing organisms, one strategy microbes employ to adapt to these environments is horizontal gene transfer, the movement of genetic material from one cell to another rather than parent to offspring. Plasmids, small circular DNA molecules that often carry beneficial traits, can facilitate this transfer of genetic material. In this dissertation, I discuss the identification of plasmid sequences, and many different aspects of their subsequent characterization. In the first half of this thesis, I describe the development of a machine-learning based model that I used to predict 68,350 plasmid sequences from human gut metagenomes. Downstream characterization of the genetic content of these plasmids reveals evolutionary patterns called 'plasmid systems' resulting from plasmid recombination. Plasmid systems are comprised of backbone genes, encoding basic functions for plasmid replication, and cargo genes, encoding fitness determining genes. I then present an example where the environmental variable of chloramphenicol usage correlates with the acquisition of of chloramphenicol resistance as cargo genes in plasmid systems. In the second half of this thesis, I focus on a particularly prevalent and abundant plasmid called pBI143 that is present in up to 92% of individuals across 4,513 metagenomes. I show that the host range of pBI143 is broad, spanning Bacteroides, Parabacteroides and Phocaeicola, and that pBI143 can transfer between these genera. pBI143 is specific humans, appearing in only human- and sewage-associated metagenomes, and lacking in any other environmental samples. pBI143 only encodes 2 genes, repA for plasmid replication, and mobA for plasmid transfer, and exists in 3 predominant versions that differ in their repA sequences. Across our metagenomes, I show that pBI143 is under strong purifying selection, and that it is monoclonal in most individuals. pBI143 is transferred from mothers to infants, and I suggest that the monoclonal nature of pBI143 may be due to priority effects of the first pBI143 version to inhabit the gut after birth. To address how pBI143 impacts the bacterial hosts, I construct isogenic strain sets of cells

with and without the naive version of pBI143 and compete these strains in gnotobiotic mice, which shows no clear fitness benefit or detriment to host cells carrying this plasmid. However, pBI143 is able to take up additional cargo genes in nature, suggesting it acts as a "discretionary parasite", where it transitions between a state of benefiting or parasitizing the host cell. The plastic nature of this plasmid makes it a good candidate for gene delivery to human gut microbiomes. Similar to other mobile genetic elements, I showed that pBI143 increases its copy number *in vitro* when the host cell is stressed, and tested this same phenomenon in naturally occurring stressful environments to demonstrate that pBI143 also increase its copy number during inflammatory bowel disease and has future potential as a diagnostic biomarker. Finally, given the widespread, abundant, and human specific nature of pBI143, we showed that this plasmid can be used as an amplifiable biomarker of human fecal contamination in water samples.

CHAPTER 1

INTRODUCTION

1.1 The human gut microbiome

Microbial life can be found in nearly every environment on Earth's surface. From thermal vents to the human gut, there are microbes that eke out an existence on the most minimal nutrients and those that grow with an abundance of resources. Regardless of environment, microbes tend to grow in intradependent communities with complex symbiotic relationships between various members. These communities often grow and develop on or within a larger host. In humans, extensive microbial colonization begins at birth and much of the initial seeding comes from the mother. Although skin and vaginal strains colonize more transiently, maternal strains appear to robustly colonize the infant gut [Ferretti et al., 2018]. Over the next year of life, the infant's gut community will develop, maintaining some strains from the mother, and acquiring others from other environments, which will eventually coalesce to form a more stable, adult-like community [Bäckhed et al., 2015, Ferretti et al., 2018].

In adults, many efforts have been made to define a 'core' set of microbial taxa or functions that are present in most individuals and which may constitute a 'healthy' gut community [Fan and Pedersen, 2020]. However, there are challenges in identifying a core community. First and foremost, core could be defined as 1) the organisms present in the most individuals, 2) the most abundant in individuals, the most stable across time, 3) the most ecologically important, 4) the most functionally important, 5) the best adapted to the host, or 6) some combination of these metrics [Neu et al., 2021, Risely, 2020]. The most common approach is to define core as the organisms present in some large fraction of the population. The second difficulty in identifying a core is that the level of taxonomic resolution will determine which organisms are considered core. For example, if we examine phylum-level resolution, we may find taxa present in most individuals, while if we look at population-level resolution, that

number will drop dramatically. The final issue in determining a core microbiome is that the components that we consider core have traditionally been defined as living organisms, however we don't take into account other biological entities like mobile genetic elements.

Despite the difficulties in determining which factors constitute a "healthy" microbiome, much of microbiome research has focused on finding specific microbes or alterations to the gut communities that may influence disease status or progression. Until the sequencing revolution in the late 2000s [Kris A. Wetterstrand, 2019], this work primarily focused on organisms that had been cultivated from the gut. This approach is still in use today, where cultivation and the downstream experiments it makes possible remains one of the primary methods for mechanistic understandings of how microbes interact with each other and the human host. 16S rRNA sequencing revolutionized our understanding of the diversity of microbes across environments. The 16S gene is conserved across all bacteria, and can be used to identify the taxa present in a sample. However, 16S sequencing is primer-based leading to biases in amplification, and is primarily limited to taxonomic assignments. Metagenomes, the data resulting from sequencing all DNA present in an environmental sample, offers a more comprehensive understanding of the genetic content of organisms and their unique functional capabilities, and is the primary data type used throughout this work. Through a combination of the approaches described above, scientists have implicated the human gut microbiome in a plethora of human ailments including cancer [Chattopadhyay et al., 2021], inflammatory diseases [Schirmer et al., 2019, Henke et al., 2019], and metabolic disorders [Sharma et al., 2020].

1.2 Horizontal gene transfer in the human gut

The microbial community in the gut is extremely dense; the average human gut, for example, is thought to carry 3 x 10^{13} bacterial cells alone - a 1:1 ratio with human cells [Sender et al., 2016]. The close cell-to-cell contact and diversity of cells of this environment

promotes the exchange of DNA from one cell to another - a process known as "horizontal gene transfer" (HGT). Although large segments of the host chromosome can be moved via HGT, the process is usually facilitated by mobile genetic elements (MGE). MGEs can be categorized into three primary groups: 1) Viruses, double or single stranded DNA or RNA that hijacks the host replication machinery to package its genome into particles that can infect new hosts 2) transposons, genomic islands that use recombinases to integrate into existing DNA and 3) plasmids, often circular, double-stranded DNA moieties capable of independent replication that rarely carry essential genes [Frost et al., 2005]. Nature, however, is prone to ignoring these discrete categories and many mobile genetic elements exist with lifestyles that blur these distinctions. Serendipitous discoveries have uncovered phages carrying plasmid segregation proteins [Oliva et al., 2012], plasmids encoding phage capsids [Chen et al., 2012], and phagemids, which can integrate into the genome like a prophage or replicate in the cytoplasm like a plasmid [Dokland, 2019]. Despite the gray areas between many MGEs, categorizing them into groups with shared features can be useful to study their ecological roles.

To exist as the drivers of HGT, MGE must move between individual host cells. The well-studied methods of MGE transfer include transformation, transduction, and conjugation [Frost et al., 2005]. Transformation occurs when DNA is taken up from the environment. Transduction is gene transfer mediated by phage infection. Conjugation is the movement of DNA via a rod-like structure called a pilus that forms a bridge between two cells in close physical proximity.

The density and diversity of microorganisms in the human gut facilitates all forms of HGT, which drives adaptation to changing environmental conditions more rapidly than conventional evolution by introducing 'ready-to-use' genetic material to a host cell. MGEs can carry a plethora of genes, such as those that allow the host to enhance metabolism, increase virulence and resist antibiotics [Al-Shayeb et al., 2022, Johnson and Nolan, 2009,

Jacob and Hobbs, 1974]. MGEs can alter ecosystem dynamics, yet these alternations differ radically in their outcomes depending on the lifestyle and transfer method of the element in question. The acquisition of a beneficial trait from a conjugative plasmid may push a bacterial population to dominate an environment. Conversely, an especially virulent lytic phage infection may decimate a once thriving population. A deeper understanding of the interplay between MGEs and their hosts is critical for broader insights into microbial ecology.

1.3 The plasmid paradox - why are plasmids maintained in cells?

Of all MGEs, plasmids have been well-studied for their applications to biotechnology and their ability to rapidly disseminate important fitness determinants throughout a population [Li et al., 2018, Smith, 1985]. A gene for antibiotic resistance, for example, may radically alter the ability of a bacterial cell to survive a harsh environment. A simple interpretation of natural selection would assume that plasmids which provide clear benefits for their bacterial hosts should be maintained in the population. However, this does not take into account the ability of genes to transfer from plasmids to their host chromosome over time, resulting in a phenomenon known as the 'plasmid paradox'.

The plasmid paradox assumes that because plasmids are independently replicating DNA in the cytoplasm, their maintenance imposes some fitness burden on the cell. The plasmid can temporarily overcome that burden by proving a beneficial function to the host cell, but if beneficial plasmids persist in cells for long enough, the fitness-enhancing genes will eventually migrate to the chromosome rendering the plasmid redundant [MacLean and Millan, 2015]. Through this logic, plasmids should eventually be purified from the population, however we do not observe that in nature.

Perhaps even more surprising is the maintenance of 'cryptic plasmids', those which do not carry a known beneficial function for the host cell and are often thought of as parasites. Although one argument for the maintenance of cryptic plasmids is that they must provide an

unknown benefit to the host cell, this would not allow them to escape the plasmid paradox in the long term. Indeed, a parasitic plasmid would also be selected against at both the individual level, where cells may lose unnecessary or redundant plasmids, and at the population level, where cells with costly plasmids will be less fit compared with their plasmidless competitors [Iranzo et al., 2016]. It is likely that for parasitic plasmids to be maintained in a population, they must be horizontally transferred between cells at a rate high enough to combat both of these levels of negative selection [Iranzo et al., 2016, Novozhilov et al., 2005]. Due to the impossibility of experimentally testing the transfer rate for all plasmids (although see [Gordon, 1992, Wan et al., 2011]) most studies rely on modeling and the field is divided over the question of whether plasmid transfer rate is sufficient to explain their maintenance in cells [Carroll and Wong, 2018, Svara and Rankin, 2011, Bergstrom et al., 2000].

Assuming that the rate of plasmid transfer is high enough, Iranzo et al argue that the existence of parasitic plasmids is inevitable by demonstrating that if a cell prevented all HGT, clonally replicating populations will have no means to restore or acquire new genetic material and will eventually succumb to Muller's ratchet – the accumulation of deleterious mutations to the point of cell death [Iranzo et al., 2016]. Thus, the occurrence of parasitic plasmids may be an inescapable by-product of genetic mechanisms that are in place to avoid Muller's ratchet.

1.4 Identifying plasmids from bacterial communities

Since the discovery of plasmids by J. Lederberg and W. Hayes in the 1950s, people primarily studied plasmids through cultivation of the host organism [Helinski, 2022]. Circular DNA can be isolated from cells using an approach that takes advantage of circular DNA renaturing and remaining soluble more effectively than linear DNA post denaturation [Lorsch, 2013]. To study the functions of plasmid proteins, researchers developed approaches like using restriction enzymes to recombine sections of plasmids and observing the result-

ing phenotypes [Smith, 1985]. The work of countless scientists resulted in our fundamental understandings of plasmids, such as their ability to transfer between cells and the types of genes that are typically encoded. For example, most, if not all plasmids encode some form of replication protein, which typically interacts with the host replication machinery to control plasmid replication [Lu et al., 1998].

As sequencing became more affordable, access to bacterial genomes and plasmids became more commonplace. Not only could we sequence thousands of whole genomes, but also metagenomes, the combined DNA of two or more organisms living in an environment like water, soil or feces. The short DNA sequences called 'reads' that are generated from a genome or metagenome can be assembled together to form continuous segments of DNA called 'contigs'. With the unprecedented access to naturally occurring environments that are offered by (meta)genomic data, the field improvised multiple sequence-based identification methods for plasmid contigs.

A relatively straightforward approach for identifying plasmids from genomic data is to determine which contigs are assembled as circular molecules [Antipov et al., 2016]. This approach ensures that you are identifying complete plasmids, but will inevitably include other circular MGEs like integrative conjugative elements or phages while missing linear or integrated plasmids [Hinnebusch and Tilly, 1993].

Another method to identify plasmids from assembled sequence data is to identify contigs that contain common plasmid genes, such as those encoding for replication or mobilization proteins [Robertson and Nash, 2018]. While this is an effective approach to identify some potential plasmids, there are multiple caveats. First, the gene in question must already have a known function in databases. Second, many genes are shared between plasmids and other types of mobile genetic elements and may result in an incorrect classification. Third, this approach will only capture the plasmids that possess that particular gene, likely a small fraction of all plasmids present in a sample.

The application of machine learning to plasmid identification has resulted in the development of multiple plasmid classifiers [Andreopoulos et al., 2022, Krawczyk et al., 2018, Pellow et al., 2020, 2021, Yu et al., 2020, Camargo et al., 2023]. The pipeline to develop a machine learning model for predicting plasmids typically involves taking a database of known plasmids and known chromosomes, designating 'features' of plasmids versus chromosomes, and asking the model to 'learn' the features that are more commonly present in plasmids versus chromosomes. How the features are chosen drastically impacts the model performance. Classifiers, (for example PlasFlow [Krawczyk et al., 2018]) that are based on kmers, nucleotide substrings of a given length that are common within the same genome, will result in models that are very effective at identifying plasmids that are similar to those it was trained on, but will fail to predict truly novel plasmids. Recently, a new generation of gene-based models have emerged. Models trained with genes as features appear to be the most effective tools to date for predicting novel plasmids [Camargo et al., 2023, Andreopoulos et al., 2022, Yu et al., 2020]. As the quality of the tools for identifying plasmids from complex data increases, we can expand our knowledge of plasmid ecology across environments, and better understand their evolutionary histories and implications for different ecosystems.

1.5 Thesis topics

Overall, this thesis showcases the premise that I aimed to build my PhD on: learning to use data-driven insights to direct wet lab experiments (and vice versa). I combine model development, genomic exploration, and experimental biology to address fundamental questions regarding the ecological distribution, evolutionary history, and functional significance of plasmids in the human gut. While generating this body of work, I developed rigorous quantitative skills to reproducibly analyze terabyte-sized genomic datasets, along with the ability to develop hypotheses based on insights from data and to critically evaluate other genomics-based research. However, the aspect of this work of which I am most proud is

the genetic engineering component. With little direct guidance I spent a year planning and executing the non-trivial task of constructing isogenic bacterial strains that differed only in a small, markerless plasmid. These details are buried deep in methods of Chapter 3, but taught me two valuable lessons: 1) with enough investment of time and energy you can learn to do pretty much anything and 2) to save yourself some of that time and energy, ask for guidance from people who know what they're doing.

This thesis contains two main sections: the development and deployment of a new method to identify plasmids from metagenomic data (Chapter 2), and the identification and characterization of an incredibly prevalent cryptic plasmid (Chapter 3). In Chapter 2, I discuss in detail our development of a machine learning-based classifier, PlasX, to predict plasmid sequences from assembled metagenomes. At the time of development, the field lacked the ability to predict truly novel plasmids from metagenomic data. We used PlasX to predict 68,000 unique human gut plasmids - increasing our reference sequences by an order of magnitude. We used secondary validation methods to confirm that these sequences are plasmids by comparing to databases, determining the circularity of assembled contigs, and experimentally demonstrating the ability of a predicted plasmids to transfer between cells. With our large collection of human gut plasmids, we defined the evolutionary concept of 'plasmid systems', in which a smaller 'backbone plasmid' is confirmed to carry the genes necessary for replication, and larger 'cargo plasmids' carry the backbone plus fluctuating accessory gene content that varies depending on environmental pressures.

In Chapter 3, I dive deep into the characterization of pBI143, the most prevalent, experimentally confirmed plasmid of our collection of 68,350. I use hundreds of bacterial isolate genomes and thousands of publicly available metagenomes to make ecological observations about the prevalence, abundance and host range of pBI143. Due to the wide distribution of pBI143 and its specificity to humans, I show it is a more sensitive marker of human fecal contamination in water than the current bacterial markers. Next, I take a population

genetics approach to characterize the nucleotide-level diversity present across human populations and show that pBI143 is highly conserved and monoclonal in most individuals. I further investigate its monoclonal nature and show that pBI143 is transferred from mothers to infants and due to priority effects, the initial version usually colonizes the infant long term. Throughout these ecological characterizations, the outstanding question was whether pBI143 impacts the bacterial host fitness. I experimentally construct the bacterial strains to test this, and show through competition experiments that pBI143 does not have a clear negative or positive impact on the fitness of the host. Instead, I propose that it likely acts as a 'discretionary parasite', by transiently uptaking additional DNA that may benefit the host cell, then losing it to regain a parasitic form. Finally, I demonstrate that similar to other mobile genetic elements, pBI143 has mechanisms in place to respond to bacterial host stress. I show that both *in vitro* and *in silico* the copy number of pBI143 is significantly higher when exposed to oxidative stress, and propose its potential as an effective proxy for measuring microbial stress response in the gut.

CHAPTER 2

THE GENETIC AND ECOLOGICAL LANDSCAPE OF PLASMIDS IN THE HUMAN GUT

This chapter is derived from the following publication:

Michael K.Yu*, **Emily C. Fogarty***, and A. Murat Eren. 2022. "The genetic and ecological landscape of plasmids in the human gut." bioRxiv. https://doi.org/10.1101/2020.11.01.361691.

*** Co-first authors**

2.1 Author contributions

Conceptualization: MKY, AME Methodology: MKY, ECF, AME Investigation: MKY, ECF, AME Visualization: MKY, ECF, AME Funding acquisition: MKY, AME Project administration: MKY, AME Supervision: MKY, AME Writing – original draft: MKY, ECF Writing – review and editing: MKY, ECF, AME Data curation: MKY, ECF Formal Analysis: MKY, ECF Resources: ECF, MKY Software: MKY Validation: ECF

2.2 Abstract

Plasmids are mobile genetic elements found across all domains of life. As plasmids often encode determinants of fitness, their evolution is intertwined with their hosts. However, naturally occurring plasmids remain far less understood than their hosts due to the lack of frameworks to recognize plasmids and to classify them into evolutionary groups. Here we trained a machine learning model that recognizes plasmids based on genetic architecture with state-of-the-art accuracy. We applied this model to a global collection of human gut metagenomes to identify 68,350 unique plasmids, 13,280 of which had a very high model

confidence and represent more than an order of magnitude increase over the number of known plasmids that we detected in this environment. To understand the evolution of these plasmids, we developed a generalizable approach that enabled us to define 1,169 'plasmid systems'. Each system consists of plasmids that share a backbone sequence containing core plasmid functions, such as replication and conjugation, but vary in cargo genes that are often critical to the host, such as antibiotic resistance, amino acid biosynthesis, and tRNA modification. Members of the same system are often found in geographically distinct human populations, revealing cargo genes that likely respond to environmental selection. The ecological patterns of plasmids we observed could not be explained by microbial taxonomy. This work uncovers the tremendous diversity of plasmids and demonstrates the need to characterize them as a separate component of microbiomes distinct from their hosts.

2.3 Introduction

Plasmids are a type of mobile genetic element [Frost et al., 2005] that occur in all domains of life [Kumar et al., 2021, Kazlauskas et al., 2019]. They typically exist as extrachromosomal and circular DNA, replicate semi-independently of their hosts, and often transfer between cells as a mechanism of horizontal gene transfer [del Solar et al., 1998, Khan, 1997, Lilly and Camps, 2015, Thomas, 2003, Summers, 1993]. A hallmark of plasmids is their remarkably diverse capacity to impact their microbial hosts, by carrying fitness-determining functions [del Solar et al., 1998, Khan, 1997, Lilly and Camps, 2015] such as antibiotic resistance genes [Jacob and Hobbs, 1974, Poyart-Salmeron et al., 1990] and virulence factors [Lan et al., 2003, Meletzus et al., 1993]. Plasmids also exhibit many interesting genetic properties, such as frequent recombination, which can result in recurrent "backbone" sequences that are shared by multiple plasmids [Sen et al., 2011, Holt et al., 2007, Fernandez-Lopez et al., 2017]. These backbone sequences often encode for core replication and transfer machinery [Sen et al., 2011, Oliva et al., 2020, Holt et al., 2007, Orlek et al., 2017, Norberg et al., 2011] that can determine

their host range [Norberg et al., 2011, Heuer and Smalla, 2012] as well as copy number in a specific host [Sota et al., 2010]. Experiments in model systems and cultured organisms have revealed the critical impact of plasmids in microbial phenotypes and survival especially for pathogens with medical significance. Yet, our understanding of the diversity, ecology, and genetic architecture of naturally occurring plasmids are far from complete.

Recent advances in metagenomics offer unprecedented access to the entire DNA content of an environment without the need for cultivation [Handelsman, 2004]. In particular, metagenomic assembly and binning strategies have enabled the reconstruction and characterization of microbial genomes de novo [Chen et al., 2020b], including those in the human gut [The Human Microbiome Project Consortium, 2012] where microbes have been associated with health and disease states [Manor et al., 2020, de Vos et al., 2022]. Metagenomic approaches have also been applied to study plasmid content [Smalla et al., 2015], but this application has been limited to enriching for plasmid DNA through library preparation techniques or to surveying only a small handful of metagenomes at a time [Jones and Marchesi, 2007, Delaney et al., 2018, Brown Kav et al., 2012, Krawczyk et al., 2018, Pellow et al., 2020, Antipov et al., 2019]. Over the past decade, the number of publicly available metagenomes has rapidly increased, numbering in the tens of thousands, creating an opportunity to study plasmids at an unprecedented scale in complex ecosystems.

Comprehensive insights into plasmid ecology and evolution require effective computational strategies for de novo identification of plasmids, which remains a challenge [Hou et al., 2021]. Several computational strategies have been developed to identify plasmids in sequence collections [Arredondo-Alonso et al., 2017, Orlek et al., 2017, Andreopoulos et al., 2022]. Many of these approaches rely on k-mer patterns learned from reference plasmid sequences [Zhou and Xu, 2010, Krawczyk et al., 2018, Pellow et al., 2020], exploit known functions such as replication or conjugation genes [Carattoli et al., 2014b, Robertson and Nash, 2018, Garcillán-Barcia et al., 2009], or use a combination of these features [Andreopoulos et al.,

2022]. While these features can help identify plasmids similar to those in public databases, they are of limited utility to recognize novel plasmids. Other approaches focus on circularity of sequences during (meta)genomic assembly [Antipov et al., 2019, Rozov et al., 2017, Pellow et al., 2021]; however, this strategy overlooks plasmids that are linear, integrated, or found as assembly fragments, and may confuse other types of circular mobile elements for plasmids.

Here, we present (1) a machine learning approach to identify plasmids in complex microbial ecosystems, (2) and a novel algorithm to gain insights into plasmid evolution at scale. Specifically, we identify a collection of 68,350 non-redundant plasmids in the human gut microbiome that were more genetically diverse than reference plasmids and substantially more prevalent across global human populations. Using a novel network partitioning algorithm, we organize this large-scale sequence collection into 'plasmid systems' based on shared backbone sequences, and demonstrate that plasmid systems provide a framework for studying the selection of plasmids by environmental pressures.

2.4 Results

2.4.1 A plasmid classification system based on de novo gene families

To enable a systematic study of plasmid sequences for machine learning, we compiled a reference set of 16,827 plasmids and 14,367 chromosomal sequences from public databases (Figure 2.1A, Table 2.1). In these sequences, we identified 51.2 million open reading frames and annotated them with functions defined in the Cluster of Orthologous Genes (COG) [Galperin et al., 2015] and Pfam [El-Gebali et al., 2019] databases. We also used MMseqs2 [Steinegger and Söding, 2017] to organize genes de novo into 2,322,750 gene families and removed those that contained only one gene. The remaining 1,090,132 de novo families enabled a more comprehensive analysis by accounting for 95% of all plasmid genes (Figures 2.1B and 2.2A). Using de novo gene families also substantially increased our ability to identify

genes that were enriched in plasmids, independently of available gene function databases (Figures 2.2B and 2.2C).

We used this reference database to train a machine learning model, PlasX, that distinguishes between plasmids and chromosomes based on genetic architecture (Figure 2.1C). PlasX is a logistic regression, which assigns a positive or negative coefficient to gene families that are likely to originate from sequences that are of plasmid or non-plasmid origin. The coefficients of gene families within a sequence are summed to calculate a prediction score, ranging from 0 to 1, where a score of >0.5 designates that a sequence is more likely to be a plasmid than not. To improve performance, PlasX uses a technique called elastic net regularization, which identifies gene families with redundant or noisy signals and then minimizes the usage of these families by setting their coefficients equal or close to zero. Consequently, only a non-redundant and informative set of gene families can impact predictions by having coefficients far from zero (Figure 2.2D). For training and evaluating PlasX, we used 10kb slices of the reference sequences, to normalize for the fact that chromosomes are generally much longer than plasmids and to improve downstream application of PlasX on sequence collections that may contain a large number of fragmented sequences, such as assembled metagenomes.

Benchmarking the efficacy of a plasmid prediction algorithm is a non-trivial task. Evaluating an algorithm's performance on sequences that are similar to those used during training, or comparing it to older approaches that were trained on a much smaller number of sequences are common pitfalls that inflate accuracy estimates. Here we implemented a more realistic evaluation framework to compare PlasX to three state-of-the-art algorithms, PlasClass [Pellow et al., 2020], PPR-Meta [Fang et al., 2019], and Platon [Schwengers et al., 2020]. We first evaluated performance in 4-fold cross-validation, using a 'naive' randomized splitting of sequences into training and test data (Figure 2.2E). PlasX achieved nearly perfect accuracy, with the highest area under the precision-recall curve (AUCPR=0.99) compared to all other

methods (Figure 2.2F). While naive splitting is a common evaluation technique, it is not a fair strategy as it can separate very similar sequences into training and test data, especially given the redundancy of sequences in public databases, and thus inflate the accuracy of classification. As a more accurate benchmark, we (1) designed an 'informed' split by first clustering plasmid and chromosomal sequences into subtypes and then keeping all sequences in the same subtype together in either the training or test data to better evaluate the ability of recognizing novel sequences and (2) assigned normalized weights to sequences to prevent well-studied plasmids from influencing the prediction ability disproportionately (see Methods). This advanced benchmark revealed a greater performance divide between PlasX (weighted AUCPR=0.70) and all other methods, with the next best method performing substantially worse (Platon, weighted AUCPR=0.23) (Figure 2.1D).

Plasmids can be difficult to distinguish from other mobile or integrated genetic elements because they share common features, including being extrachromosomal [Wozniak and Waldor, 2010, Antipov et al., 2019], facilitating horizontal gene transfer [Wang et al., 2016, Cuecas et al., 2017], or encoding traditional core functions like replication and mobilization [Robertson and Nash, 2018, Garcillán-Barcia et al., 2011]. To determine PlasX's ability to distinguish plasmids from other mobile genetic elements, we ran PlasX on all ICE sequences from the ICEberg database [Liu et al., 2019] (n=552) and all prophage sequences from the NCBI viral database (n=445). PlasX correctly classified 92.2% of ICEs as not plasmids, and 93.2% of NCBI viral database as not plasmids (Table 2.9 and 2.10). Platon could also distinguish prophages from plasmids (99.6% accuracy), but its classification accuracy was much lower compared than PlasX's for ICEs, as Platon classified 37.1% of ICEs as plasmids. Next, we ran PlasX on 21,012 plasmids that were added to PLSDB after we had already trained the model. PlasX performed very well, identifying 81.5% (17,128) of these sequences as plasmids (Table 2.11). Finally, we evaluated the performance of PlasX and other methods on a novel and recently characterized plasmid of Wolbachia, pWCP [Reveillaud et al.,

2019]. Since pWCP was not present in the training data for any of the plasmid prediction tools, it provided a unique opportunity to investigate whether this plasmid, which remained elusive until recently, could have been discovered through a de novo plasmid survey. PlasX was able to predict pWCP as a plasmid (score = 0.73), while all other methods, PlasClass [Pellow et al., 2020], PPR-Meta [Fang et al., 2019], Platon [Schwengers et al., 2020] and Deeplasmid [Andreopoulos et al., 2022], were unable to classify it as a plasmid, either labeling it incorrectly as a chromosome or reporting high uncertainty in their prediction (Table 2.8). Overall, these results suggest that PlasX, with its reliance on gene families rather than strictly defined sequence features, is unique in its ability to predict novel plasmids that are not present in existing databases with high accuracy.

PlasX's accuracy suggests it has learned insights into defining a "plasmid". To broaden our understanding of a plasmid, we used PlasX's coefficients to rank gene families by their importance in de novo identification of plasmids (Table 2.7). Among the 200 most important gene families, 19 were COGs and Pfams whose functional descriptions can be immediately recognized as being plasmid-associated because they contain keywords such as "plasmid", "replication", or "conjugation" (Figure 2.1E). However, another 9 COGs and Pfams did not have such a recognizable description. For example, a family of lipoproteins (PF05714) has been studied for conferring virulence in a few plasmids [Sukupolvi and O'Connor, 1990, Norris et al., 1992], but it is not generally thought of as a common plasmid function. Nonetheless, this family had the 17th highest coefficient of 1.678, consistent with its enrichment in 168 plasmids (36 plasmid subtypes) but only 2 chromosomes. While these results show that coefficients provide an approximate guide to understanding PlasX's logic, we caution that interpreting each coefficient by itself can be complicated as PlasX often sums up the coefficients of several families in a sequence to make a prediction. Further curation is necessary to understand which high-coefficient families are truly characteristic of plasmids.

De novo families also provided two types of novel insights about plasmids. One insight is

that 56.1% of de novo families can be thought as 'subfamilies' that group together a subset of genes within a COG or Pfam. As many COGs and Pfams contain both plasmid and chromosomal genes, these subfamilies can provide a deeper resolution of the bacterial gene pool by delineating plasmid- or chromosome-evolved lineages. For example, the Pfam PF10609 is a broad family of genes related to parA, a gene that drives the partitioning of chromosomes [Jalal and Le, 2020] and plasmids [Bouet and Funnell, 2019] during cell division. As this family is found on 35% of plasmids and 95% of chromosomes, it alone is not informative for identifying plasmids and thus has a coefficient close to zero (-0.023). However, PF10609 can be further dissected into plasmid-specific subfamilies, such as mmseqs_5_1535552 (coefficient +0.455), and chromosome-specific subfamilies, such as mmseqs_70_40217271 (coefficient -0.198), which become informative for PlasX to distinguish plasmids from chromosomes. Indeed, the maximum likelihood phylogenetic tree that relates the genes in these two subfamilies show a divergence of plasmids and chromosomes into monophyletic groups (Figure 2.1F), which is also reflected in their sequence alignment (Figure 2.1G). The second insight is that 35.5% (398,174) of de novo families have no overlap with any COG or Pfam. Many of these families have highly positive coefficients (e.g. 12,076 families have coefficients >0.1) that make a sequence appear substantially more like a plasmid and thus could represent fundamental but unexplored plasmid functions.

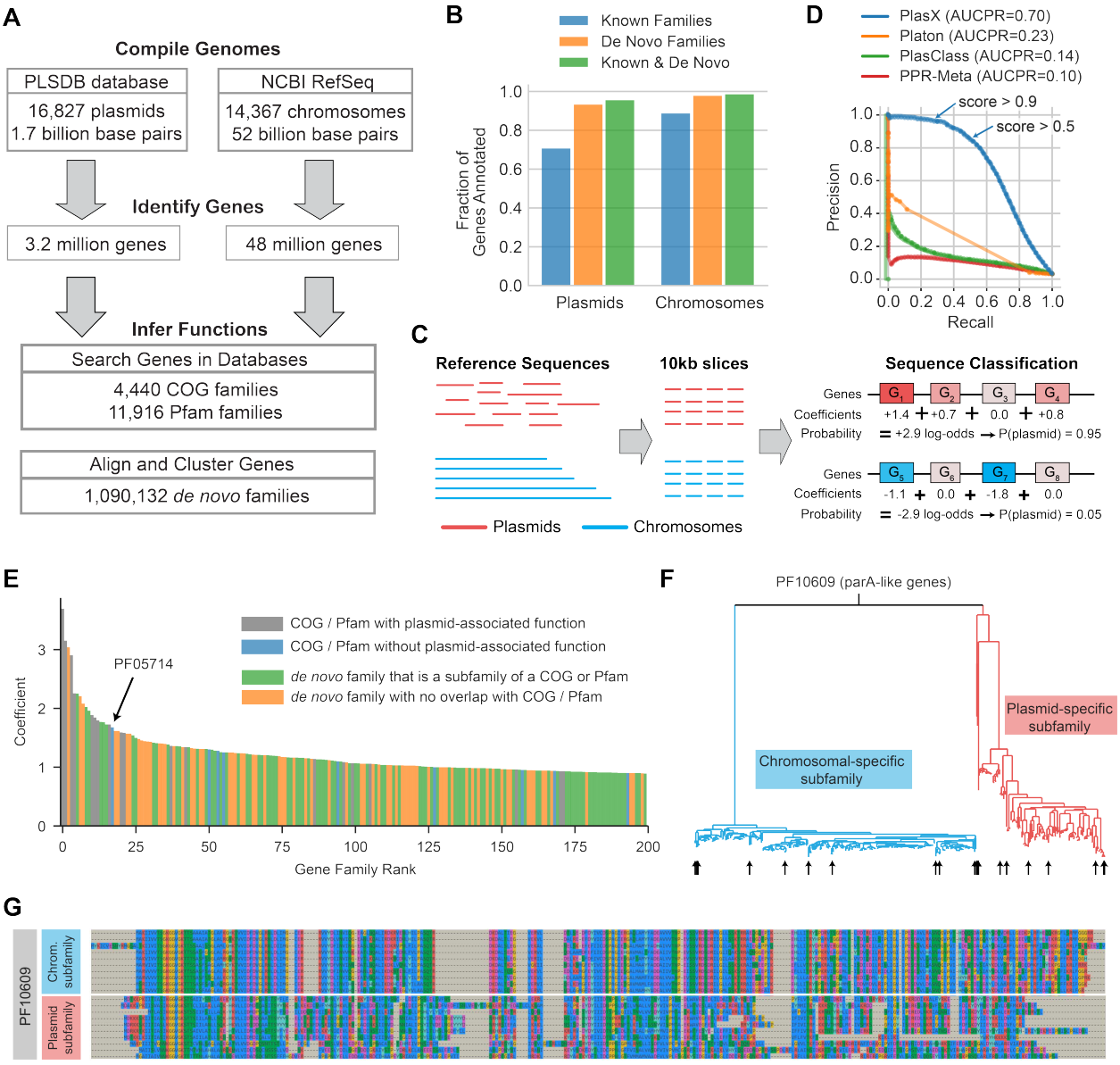

Figure 2.1: A machine learning model for classifying plasmids. (A) Our pangenomics workflow to characterize gene functions in a reference set of plasmids and chromosomes. (B) The fraction of all plasmids or all chromosomal genes that are annotated by using known families (blue), de novo families (orange), or a combination of both (green). (C) Training of PlasX. Reference sequences are sliced into 10kb windows and then prediction scores are made by a logistic regression that sums the contributions of gene families within a sequence. (D) Precision-recall curves comparing PlasX, Platon, PlasClass, and PPR-Meta. Except for PPR-Meta, every method was trained and evaluated using 4-fold cross-validation and an informed split. AUCPR was calculated using sequence weights for normalization. The arrows indicate the performance of PlasX using a score threshold of either >0.5 or >0.9.

Figure 2.1 continued: (E) The 200 gene families with the highest PlasX coefficients and thus most important for identifying plasmids. Gene families are ranked by their coefficient. (F) Maximum-likelihood phylogenetic tree of genes that are in PF10609 and also in either the plasmid-specific de novo subfamily mmseqs_5_1535552 (red) or chromosome-specific de novo subfamily mmseqs_70_40217271 (blue). (G) Sequence alignment of 10 representative genes from each subfamily (arrows in F)

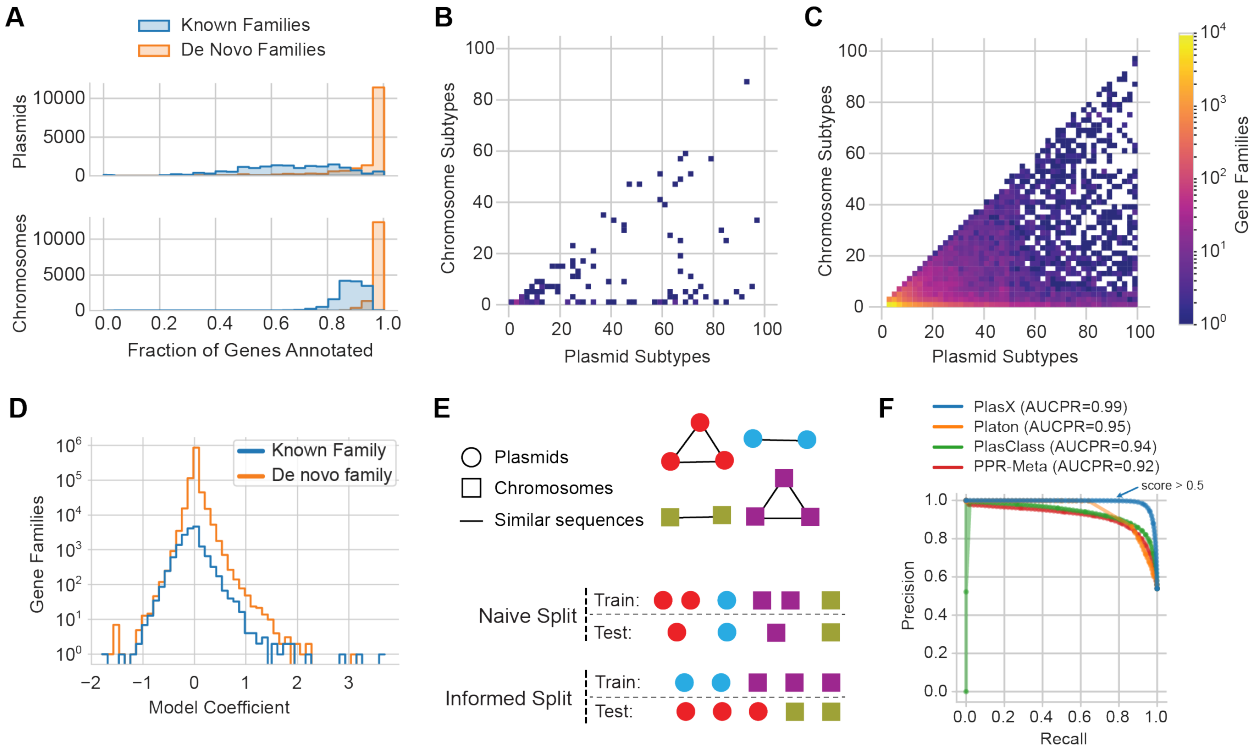

Figure 2.2: Additional analysis of PlasX. (A) Histograms of reference sequences, based on the fraction of genes that have known or de novo family annotations. (B-C) Two-dimensional histograms of known (B) and de novo (C) gene families, based on the number of plasmid and chromosomal subtypes that each family is found in. The number of gene families is log-scaled. Only the gene families that are enriched in plasmid subtypes (i.e. bottom-right triangle) are shown. (D) Histograms of the coefficients learned by PlasX, showing that the vast majority of coefficients are close to zero. (E) Diagrams of different training-test split configurations for cross-validation. A random 'naive' split of plasmids and chromosomal sequences results in training and test sets that have similar sequences, due to the existence of plasmid and chromosomal subtypes that contain highly similar sequences. An 'informed' split assigns all sequences of the same subtype to either training or test, creating a more representative evaluation of a model's ability to generalize to unseen sequences. Colors and edges represent sequences that are in the same subtype. (F) Precision-recall curves using 4-fold cross-validation and a naive split.

2.4.2 PlasX unveils a large database of new plasmids from the human gut microbiome

Having verified PlasX's ability to identify plasmid sequences, we applied it to survey naturally occurring plasmids in the human gut microbiome, an environment which harbors a diverse range of microbes and mobile genetic elements [Carr et al., 2021]. We assembled 36 million contigs from 1,782 human gut metagenomes, spanning culturally and geographically distinct human populations (Table 2.2). Running PlasX on these data resulted in a total of 226,194 predicted plasmids with a model score above 0.5 (Figures 2.3A and 2.4A, Table 2.3). Our predictions spanned a wide range of lengths, including 135 sequences that were longer than 100 kbp, but they were generally shorter than reference plasmids with a median length of 2.6 kbp versus 53.3 kbp, respectively (Figure 2.4B). This discrepancy can be partly explained by fragmentation during metagenomic assembly, as the median length of the entire set of contigs was 2.1 kbp, and only 50,310 (0.14%) contigs were longer than 100 kbp. To minimize this issue, we removed predictions that were likely assembly fragments because they were subsequences of other predictions in our collection and also did not appear to be circular elements themselves. This filter retained 100,719 predictions for downstream analyses (see Methods, Figure 2.5). While PlasX identifies contigs that are likely plasmids or plasmid fragments, throughout this manuscript we refer to these predicted sequences shortly as 'plasmids' for practical reasons.

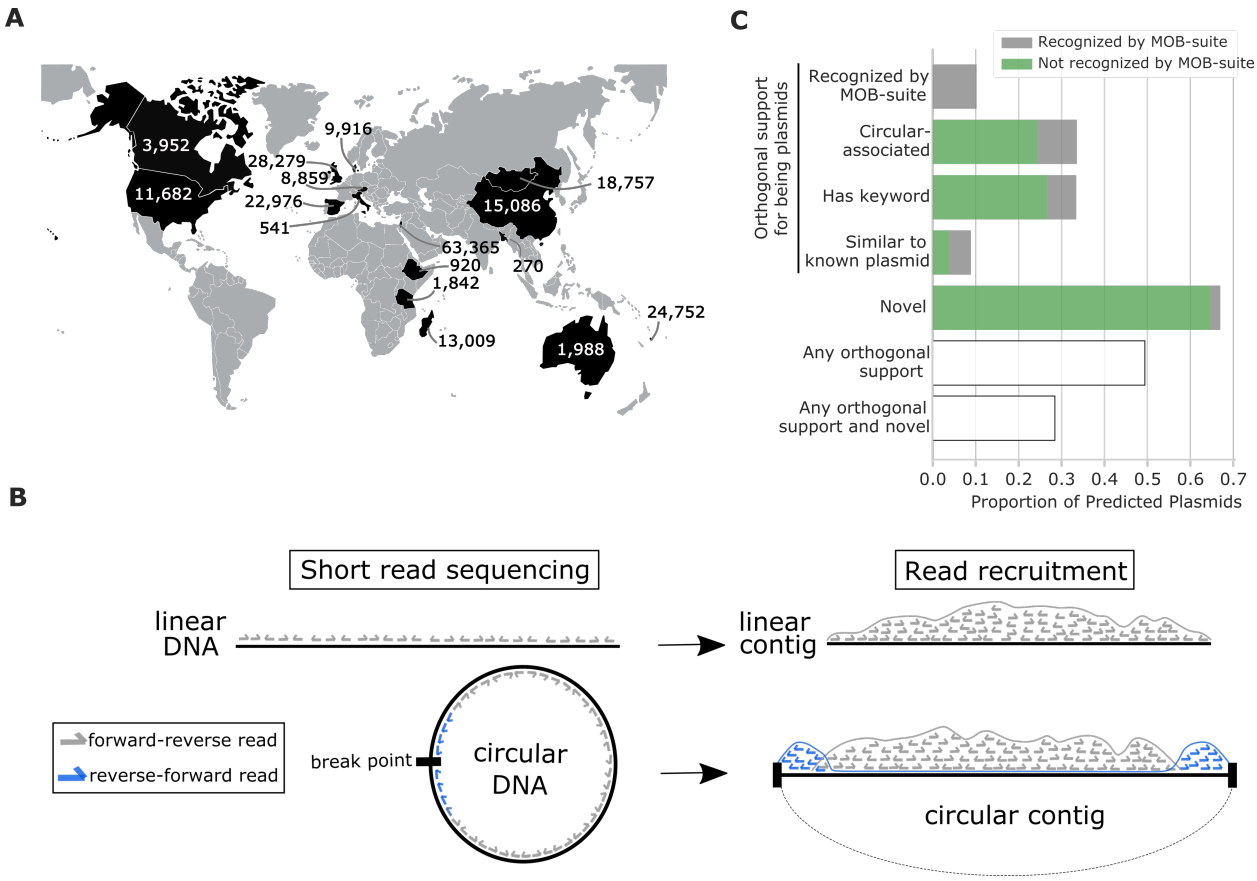

Figure 2.3: Plasmid prediction from metagenomes. (A) Number of plasmids predicted from each country. (B) Diagram of paired-end reads mapping to a linear versus a circular contig. Linear contigs have forward-reverse reads only, while circular contigs also have reverse-forward reads concentrated on the ends due to an artifact in contig assembly. (C) Orthogonal support for and novelty of the 100,719 non-fragment predictions.

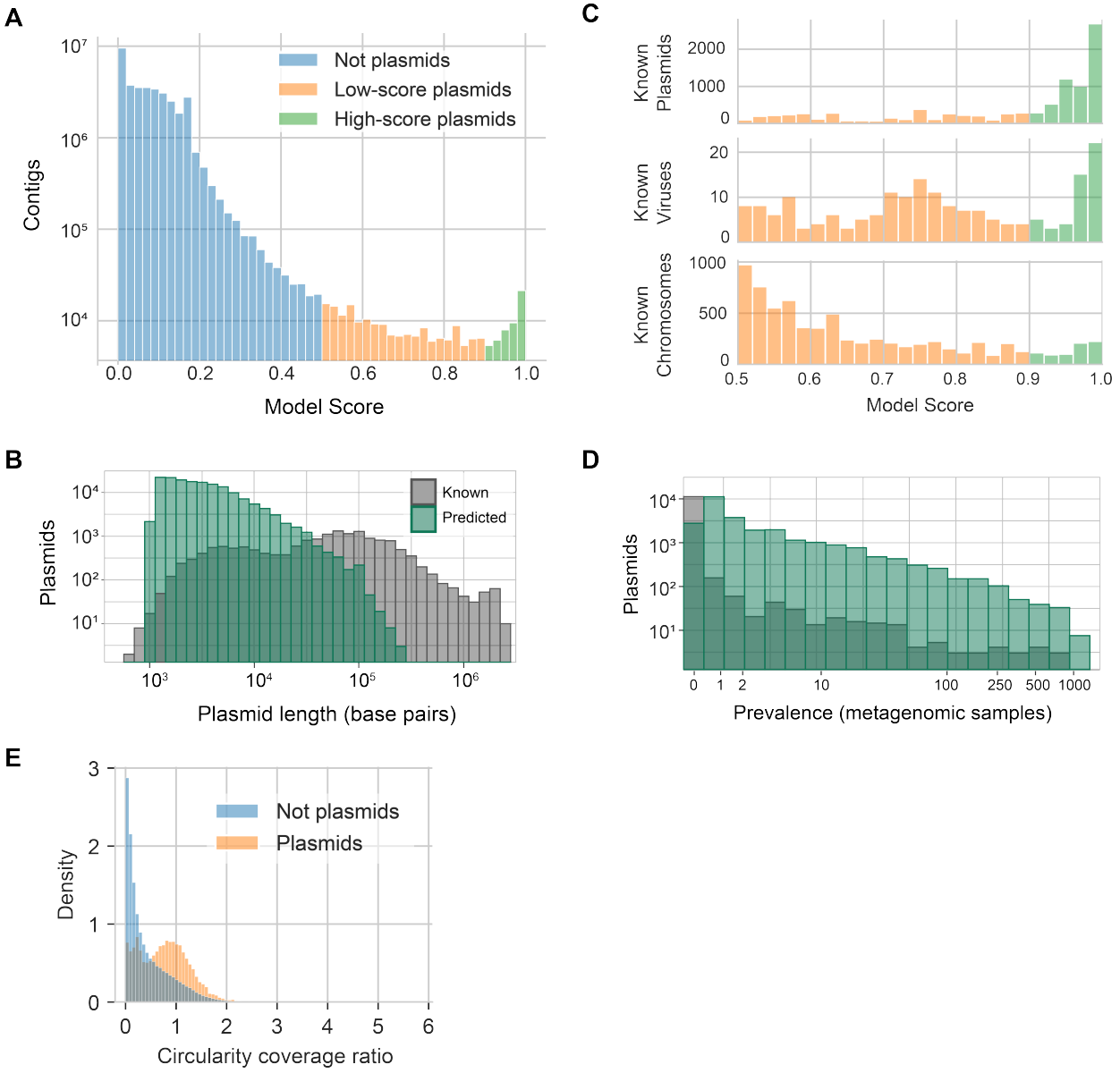

Figure 2.4: Additional analysis of predicted plasmids. (A) Model scores of all contigs assembled from all 1,782 metagenomes. 226,194 plasmids were predicted by applying a score threshold of >0.5. Of these, 50,163 plasmids were high-scoring (greater than 0.9 score). (B) The sequence length of known and predicted plasmids. (C) Model scores of predicted plasmids that matched a sequence in NCBI (greater than 90% alignment identity and greater than 90% coverage of the predicted plasmid). Predictions are labeled as a known 'plasmid', 'virus', or 'chromosome' based on the presence of these words in the description of the matching NCBI sequence. We searched NCBI for only the filtered set of 100,719 non-fragment predictions. (D) The prevalence of reference and predicted plasmids across all metagenomes. (E) We calculated a "circularity coverage ratio" as the number of supporting reverse-forward reads divided by the average coverage of a contig. All circular contigs are shown, and they are colored if they were predicted by PlasX as plasmids (orange) or not plasmids (blue).

To determine the circularity of predicted plasmids, we analyzed the orientation of paired-end reads recruited from metagenomes. This is a powerful strategy because if a contig occurred as a circular element in the environment, then matching reads from the same pair would be recruited to opposite ends of the contig in a 'reverse-forward' orientation, instead of the typical 'forward-reverse' orientation (Figure 2.3B). With this approach we found that 19,652 plasmid sequences were circular, and we designated them as high-confidence plasmids for downstream analyses. These circular plasmids spanned a range of sizes, with a median length of 4.4 kbp and 854/378/47 plasmids longer than 25/50/100 kbp. An additional 14,151 sequences were not circular themselves but were highly similar to a circular sequence. Together, these two types of sequences defined a set of 'circular-associated' sequences representing 33.6% (33,803/100,719) of predictions. Multiple factors can explain the lack of signal for circularity for the remaining plasmids, including insufficient sequencing depth to observe reverse-forward pairs, fragmented contigs, or the non-circular nature of some plasmids that occur linearly [Meinhardt et al., 1997] or are integrated in a chromosome [Kazlauskas et al., 2019]. There were 154,680 contigs that were not predicted to be plasmids but still appeared circular; however, these contigs tended to have a smaller number of supporting reverse-forward reads relative to their coverage (Figure 2.4E), which may indicate that they are other types of mobile elements such as viruses or ICEs that temporarily circularize.

Beyond circularity, confirming *in silico* whether a novel sequence represents a plasmid is a significant challenge. In the absence of single-copy core genes that have been vital to assess the completeness of non-plasmid and non-viral genomes assembled from metagenomes [Chen et al., 2020b], our understanding of the canonical features of plasmids is limited to a relatively small set of well-studied genes that are primarily derived from plasmids of model organisms in culture [Carattoli et al., 2014c, Robertson and Nash, 2018]. For instance, MOB-suite [Robertson and Nash, 2018] identified canonical features for plasmid replication

and conjugation in only 16.3% of the 16,827 PLSDB reference plasmid sequences used to train PlasX. This relatively small percentage reveals the limits of conventional approaches to identify plasmid features and thus foreshadows their limited utility to survey novel plasmids. Indeed, MOB-suite identified canonical features in only 10.1% of our predictions. Given this narrow sensitivity, we developed orthogonal data-driven strategies to increase confidence in our predictions.

For the remaining 89.8% (90,446/100,719) of predicted plasmid sequences in which MOB-suite did not find any canonical plasmid features, we performed several types of analyses to assess how many are true plasmids or novel sequences (Figure 2.3C). We found that 24.5% (24,689) of predictions were circular-associated sequences. 26.7% (26,921) were 'keyword-recognizable', as they contained a COG or Pfam function with the words 'plasmid' or 'conjugation'. And finally, 4.0% (3,996) were highly similar to a known plasmid sequence in NCBI, while 64.7% (65,117) were novel sequences with no hits to any sequence in NCBI. As these different subsets of plasmids are partially overlapping, we took their union to find that 49.4% (49,739) of predictions had some orthogonal support for being a plasmid, by MOB-suite or any of the first three types of analyses, and 28.5% (28,658) had such support and were novel (Table 2.3). Overall, these findings suggest that our collection of predicted plasmids include not only sequences that match known plasmids, facilitating the study of their diversity and gene pool in natural habitats, but also novel sequences that can further advance plasmid biology.

We further investigated the subset of predictions that were highly similar to a sequence in NCBI and categorized matches as either known plasmids (26.9%), chromosomes (21.3%), viruses (0.6%), or an unclear type of sequence (51.2%). A total of 189 predictions matched a known virus. Of these, 110 were recognized as plasmids by MOB-suite or keywords but also contained virus-related COG or Pfam functions, as indicated by the keywords 'virus', 'viral', and 'phage'. These predictions carry both plasmid and viral features, a phenomenon that has

previously been reported [Chen et al., 2012, Oliva et al., 2012, Dokland, 2019, Pfeifer et al., 2021]. Surprisingly, 808 predictions that matched a known chromosome were also circular-associated and recognized by MOB-suite or plasmid keywords. One explanation of these data is that these plasmids can switch between an extrachromosomal or a chromosome-integrated state.

While we have identified plasmids based on a score of >0.5, a stricter threshold could be used to filter for more confident predictions. For example, we identified a subset of 24,614 predictions with a score of >0.9. These high-scoring predictions were more likely to match known plasmids and less likely to match known chromosomes in NCBI, compared with predictions with a lower score between 0.5 and 0.9 (Figure 2.4C). High-scoring predictions also tended to be longer and enriched for circular sequences (Figure 2.6), suggesting that they are less likely to be assembly fragments. Nonetheless, a stricter threshold comes with an inevitable cost of not only removing noise but also bona fide plasmids. This tradeoff is most visible in cross-validation, where a threshold of >0.5 lies at an inflection point in the precision-recall curve of Figure 2.1D (with a precision of 0.850 and recall of 0.500). While applying a stricter threshold of >0.9 would provide a modest increase of 13% in precision (to 0.920), it would substantially decrease recall by 44% (to 0.280). As our understanding of plasmid diversity in metagenomes is greatly underdeveloped, we decided that a threshold of >0.5 provides a reasonable balance between precision and recall, such that the resulting predictions still contain potentially many novel plasmids to advance the field. A good example for this is the long-missed Wolbachia plasmid [Reveillaud et al., 2019], which has a score of 0.73. Furthermore, we found that 31.6% (31,847/100,719) of plasmids with lower prediction scores (between 0.5 and 0.9) had orthogonal support for being plasmids (Table 2.3).

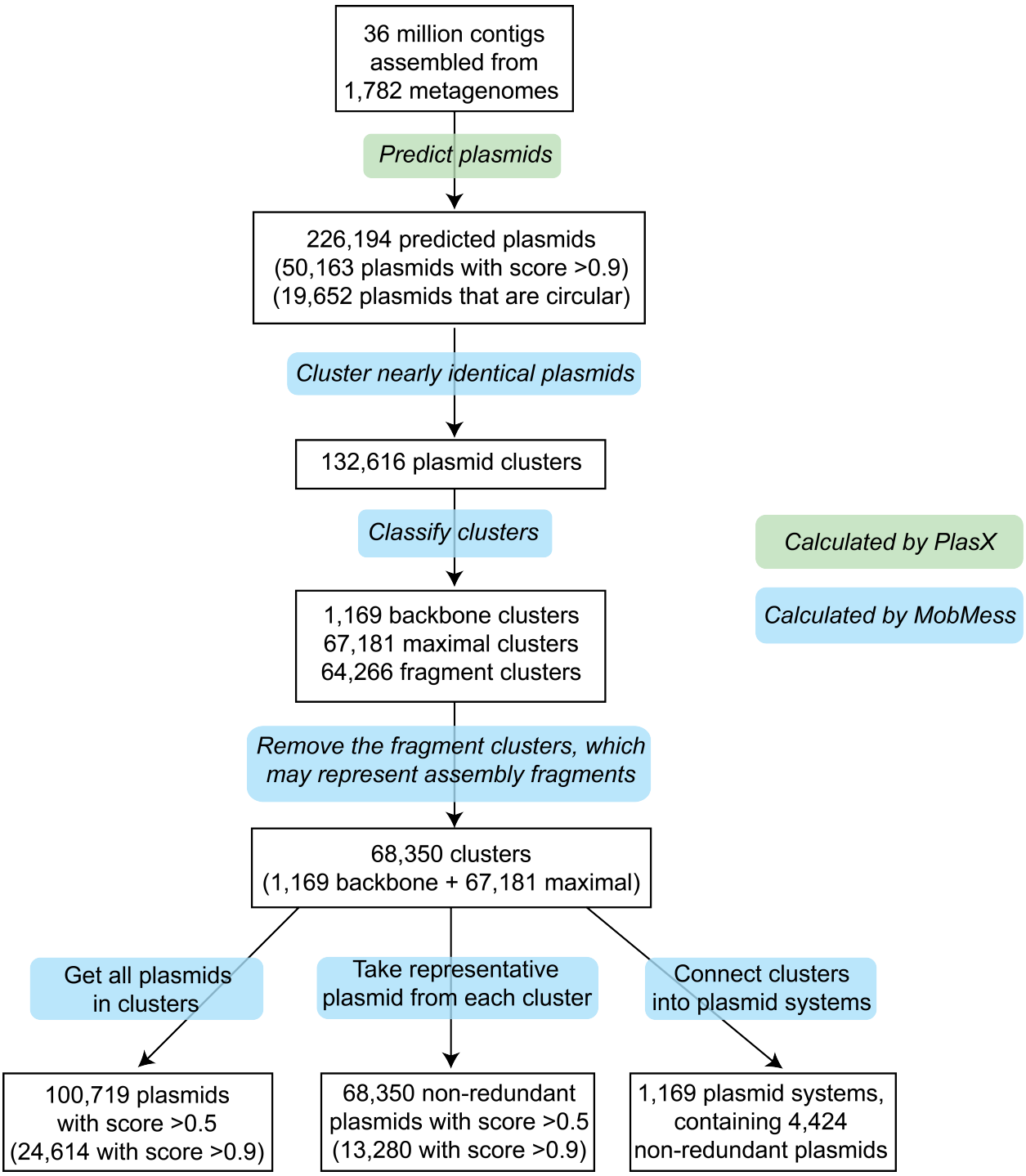

Figure 2.5: Workflow of predicting plasmids with PlasX and organizing them with MobMess.

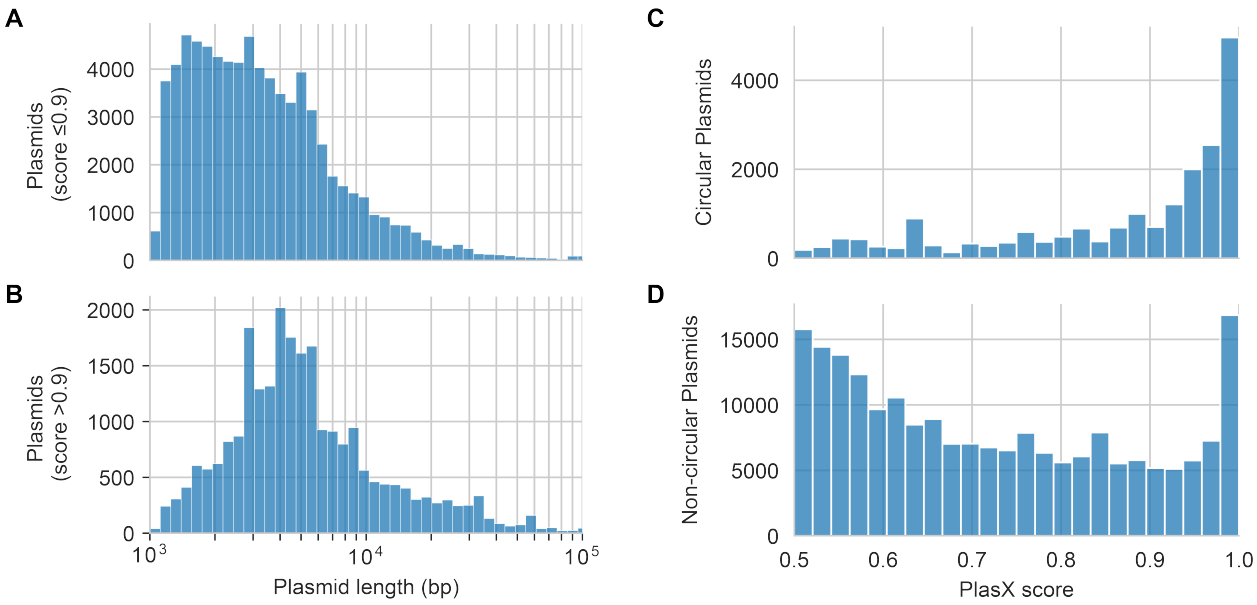

Figure 2.6: Relation between PlasX score and the length and circularity of predicted plasmids. (A-B) The distribution of plasmid lengths, for plasmids with a score of less than 0.9 (A) or >0.9 (B). (C-D) The distribution of PlasX scores for circular (C) or non-circular plasmids (D).

2.4.3 Plasmids predicted from metagenomes are found in isolate genomes and can transfer between microbial populations

To experimentally validate our metagenome-derived predictions as true plasmids of the human gut, we developed a pipeline for identifying predictions that are (1) present in human gut microbial isolates, (2) are circular in those isolates, and (3) can be naturally transferred to other microbes. First, we detected 127 of our predicted plasmids in 14 *Bacteroides* isolate genomes that we sequenced in a previous study [Vineis et al., 2016] (Figure 2.7A). Short-read sequencing of two of these isolates suggested that the predicted plasmids pFIJ0137_1 and pENG0187_1 were circular based on paired-end orientation (Figure 2.3C). We further confirmed their circularity using long-read sequencing. Following a previously described approach [Reveillaud et al., 2019], we identified and manually confirmed 500 long reads that align completely to a plasmid but not to the host chromosome (Figures 2.7A and 2.8). Some

of these long reads align across the artificial contig breakpoint, indicating these plasmids are extrachromosomal and circular (see Methods).

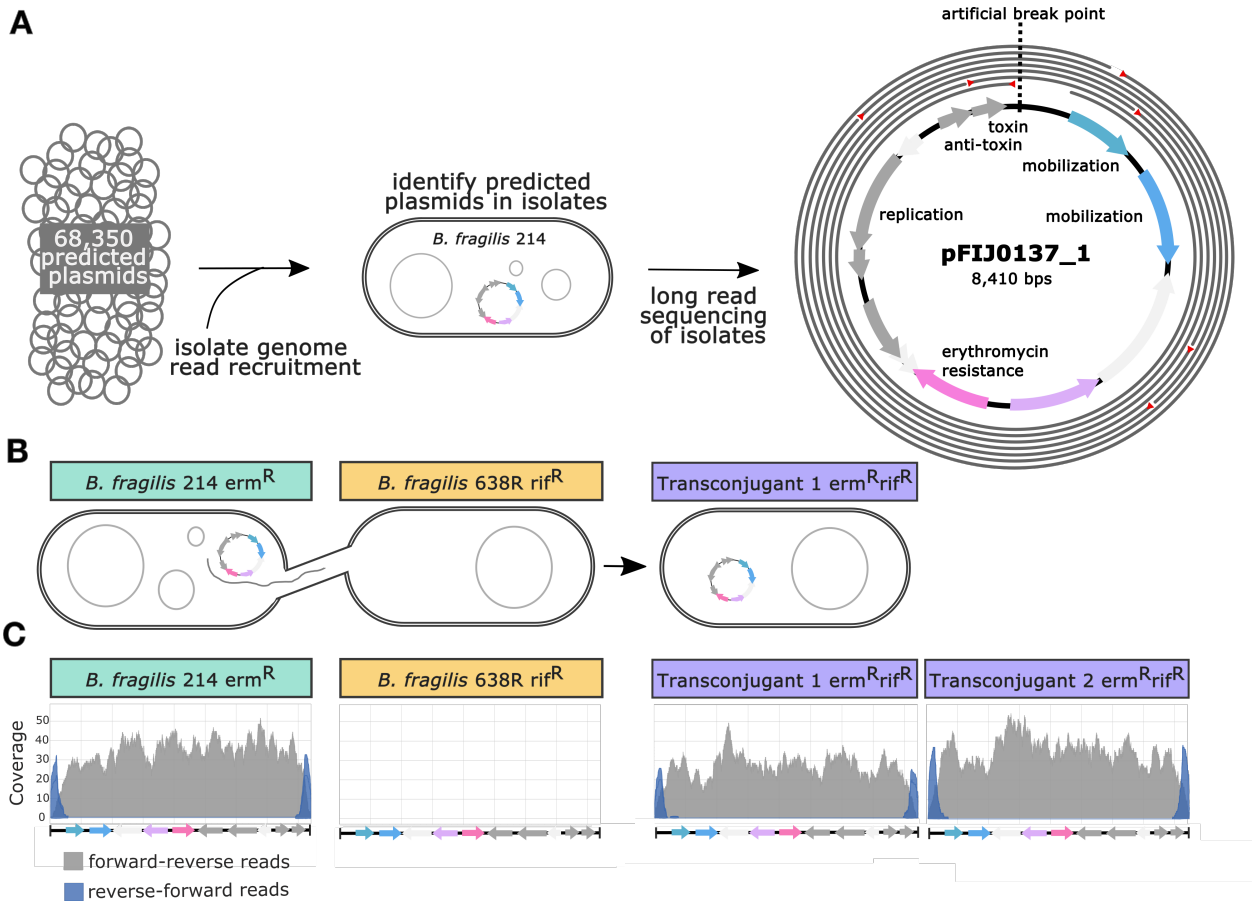

Figure 2.7: Experimental validation of plasmid predictions. (A) We recruited reads from the sequenced genomes of 14 *Bacteroides* isolates to determine which isolates contain our predicted plasmids. We further confirmed the presence and circularity of a predicted plasmid (pFIJ1037_1) in the isolate *B. fragilis* 214 by long read sequencing. Gray circles represent 7 (of 500) long reads that align to pFIJ0137_1. Red triangles designate the beginning of a long read. (B) Transfer of pFIJ1037_1 from *B. fragilis* 214 to *B. fragilis* 638R via conjugation and selection on erythromycin- and rifampicin-containing media. (C) Coverage plots showing read recruitment of *B. fragilis* whole-genome sequencing reads to the pFIJ1037_1 reference sequence, confirming transfer of pFIJ1037_1. Gray are forward-reverse reads, while blue are reverse-forward reads that indicate the circularity of pFIJ1037_1.

Finally, we tested the ability of pFIJ0137_1 to transfer between its host, *B. fragilis* 214 (one of 14 isolates from [Vineis et al., 2016]) to a well-known laboratory strain, *B. fragilis* 638R. We designed an experimental setup that takes advantage of the naturally encoded erythromycin resistance (ermR) on pFIJ0137_1 and the rifampicin resistance (rifR) of *B. fragilis* 638R. Specifically, we first mated isolates in the absence of antibiotics, and then selected for transconjugants on media containing both antibiotics (Figure 2.7B). While this plasmid lacks conjugation machinery, it contains two relaxases (blue genes in Figure 2.7A) and thus could be mobilized by different conjugative apparatus in the host cell. Through short-read sequencing of the donor, recipient, and resulting transconjugants, and by employing a read recruitment analysis, we confirmed that pFIJ0137_1 transferred from *B. fragilis* 214 to *B. fragilis* 638R (Figure 2.7C). This analysis also confirmed the circularity of pFIJ0137_1 in *B. fragilis* 214 and both *B. fragilis* 638R transconjugants. Besides a 68bp deletion, pFIJ0137_1 in *B. fragilis* 214 (isolated in Chicago, USA) was identical to the pFIJ0137_1 version assembled from a Fijian metagenome, suggesting a relatively recent transfer of this plasmid between unrelated human populations. These experimental results show that while PlasX identifies plasmids solely based on genetic architecture, it is capable of predicting plasmids that have canonical features of being extrachromosomal, circular, or transmissible between cells.

2.4.4 Novel plasmids are highly prevalent, reflect human biogeography, and unexplained by microbial taxonomy

Next, we sought to characterize the ecology of plasmids across human populations through metagenomic read recruitment. For this task, we first dereplicated the entire collection of reference and predicted plasmid sequences, where we assumed that any pair of plasmid sequences was redundant if at least 90% of either sequence aligned to the other with over 90% sequence identity. This analysis found 68,350 and 11,121 non-redundant sequences in

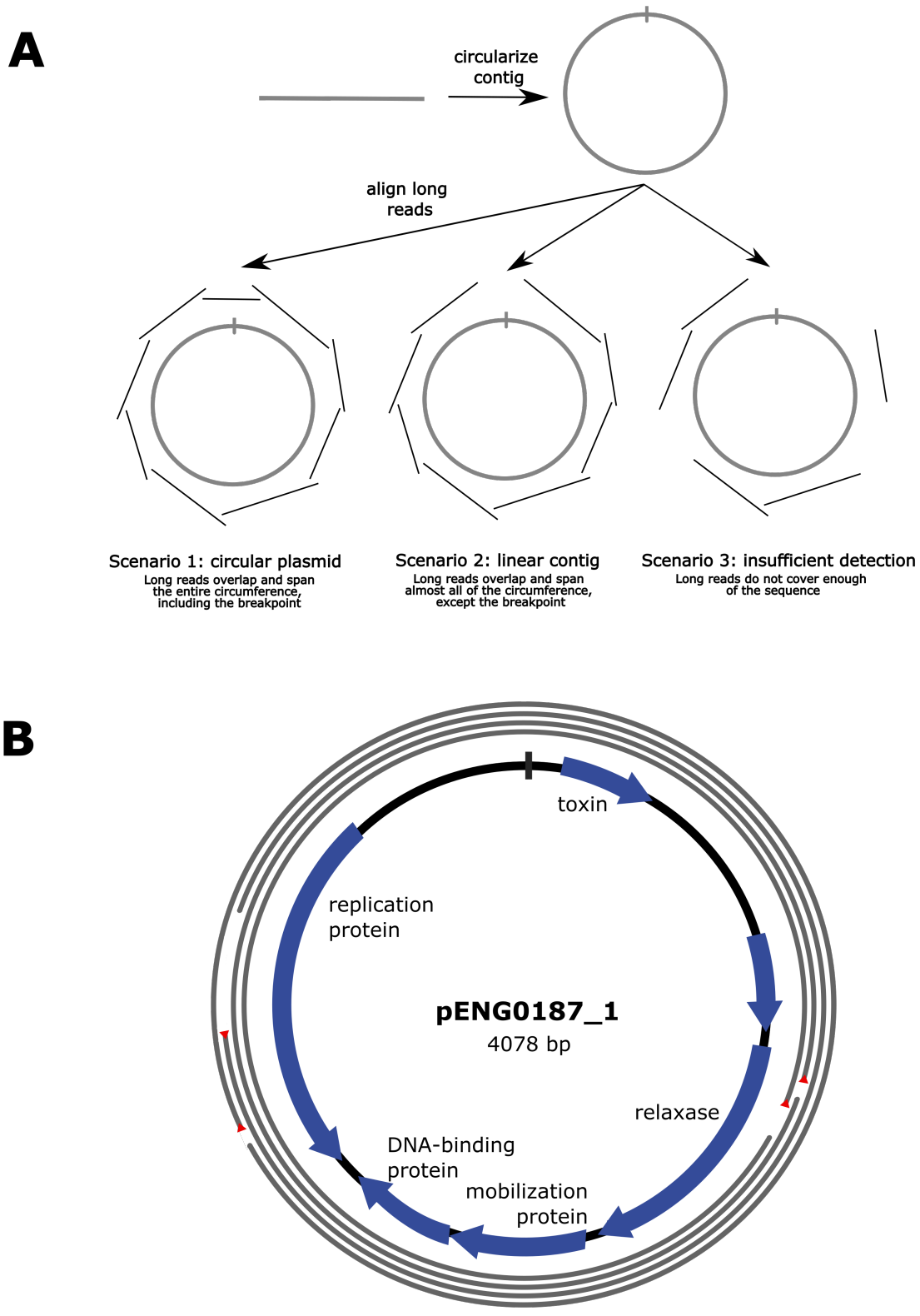

Figure 2.8: Long read circularity. (A) The process to identify circular plasmids using long

Figure 2.8 continued: read sequences. Contigs are always assembled as linear sequences even when originally circular in the environment. We can determine their original configuration by aligning long reads around the entire sequence. (B) *B. fragilis* 216 long reads aligned to pENG0187_1, demonstrating circularity. 4 of 500 reads are shown for simplicity. Red triangles designate the beginning of a long read.

the set of 226,194 predicted and 16,827 reference plasmids, respectively. Then, we used the non-redundant sets of plasmids to recruit reads from the 1,782 globally distributed human gut metagenomes. We labeled a plasmid as present in a metagenome if its 'detection' was greater than 0.95, where detection is the fraction of the sequence covered by at least one read (see Methods).

Our read recruitment analysis revealed that predicted plasmids were much more prevalent across human populations than reference plasmids. For instance, only 1.9% (211) of reference plasmids were present in at least two individuals in our dataset, suggesting the limited ecological relevance of reference plasmids to naturally occurring gut microbial communities. Indeed, many reference plasmids were isolated from a relatively small number of pathogens, such as *Escherichia coli, Salmonella enterica, Pseudomonas aeruginosa, Klebsiella pneumoniae*, and *Vibrio cholerae*, which are unlikely to be abundant in healthy humans. In contrast, 63.1% (43,114) of the predicted plasmids were present in at least two individuals (Figure 2.4D). Moreover, of the most highly prevalent plasmids found in greater than 100 individuals, 99.7% (5,400/5,414) were predicted plasmids while only 0.3% (14/5,414) were reference plasmids.

The prevalence of predicted plasmids suggests that they capture the biogeography and lifestyles of human populations more effectively than reference plasmids. To confirm this, we performed agglomerative clustering to construct a dendrogram that organizes metagenomes based on their plasmid content. Using reference plasmids for this clustering, we found that only 50.2% of metagenomes were arranged next to another metagenome from the same country (Figure 2.9A). In contrast, using predicted plasmids (Figure 2.9B) resulted in 74.0% of metagenomes arranged that way. We also organized metagenomes using a dimension-

ality reduction of predicted plasmids. This analysis shows that industrialized versus non-industrialized metagenomes can be distinguished solely by their plasmid content (Figure 2.9C). Dimensionality reduction also showed country-specific clustering (Figure 2.10).

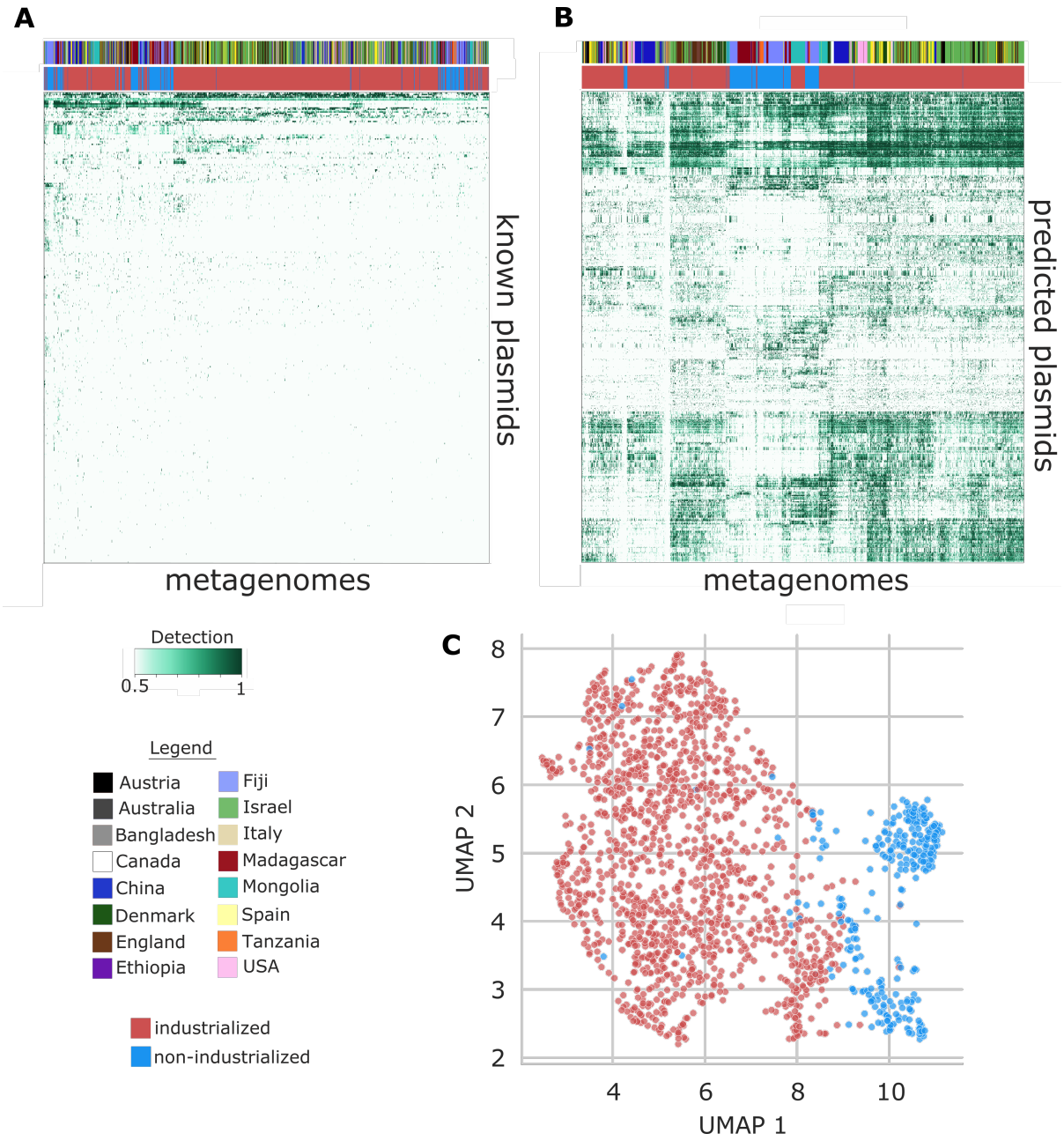

Figure 2.9: Global plasmid ecology. (A) Read recruitment of human gut metagenomes to 11,121 non-redundant reference plasmids. The heatmap shows the 338 plasmids that are present in at least one metagenome (greater than 0.95 detection). (B) Read recruitment to 68,350 non-redundant predicted plasmids. The heatmap shows the 1,000 most prevalent plasmids that are present in at least one metagenome and have PlasX score greater than 0.75. In A and B, column colors indicate country of origin and lifestyle (industrialized or non-industrialized). (C) Clustering of metagenomes based on the predicted plasmids that are present, using the UMAP dimensionality reduction method(McInnes, Healy, and Melville 2018). Metagenomes from industrialized or non-industrialized populations are colored red or blue, respectively.

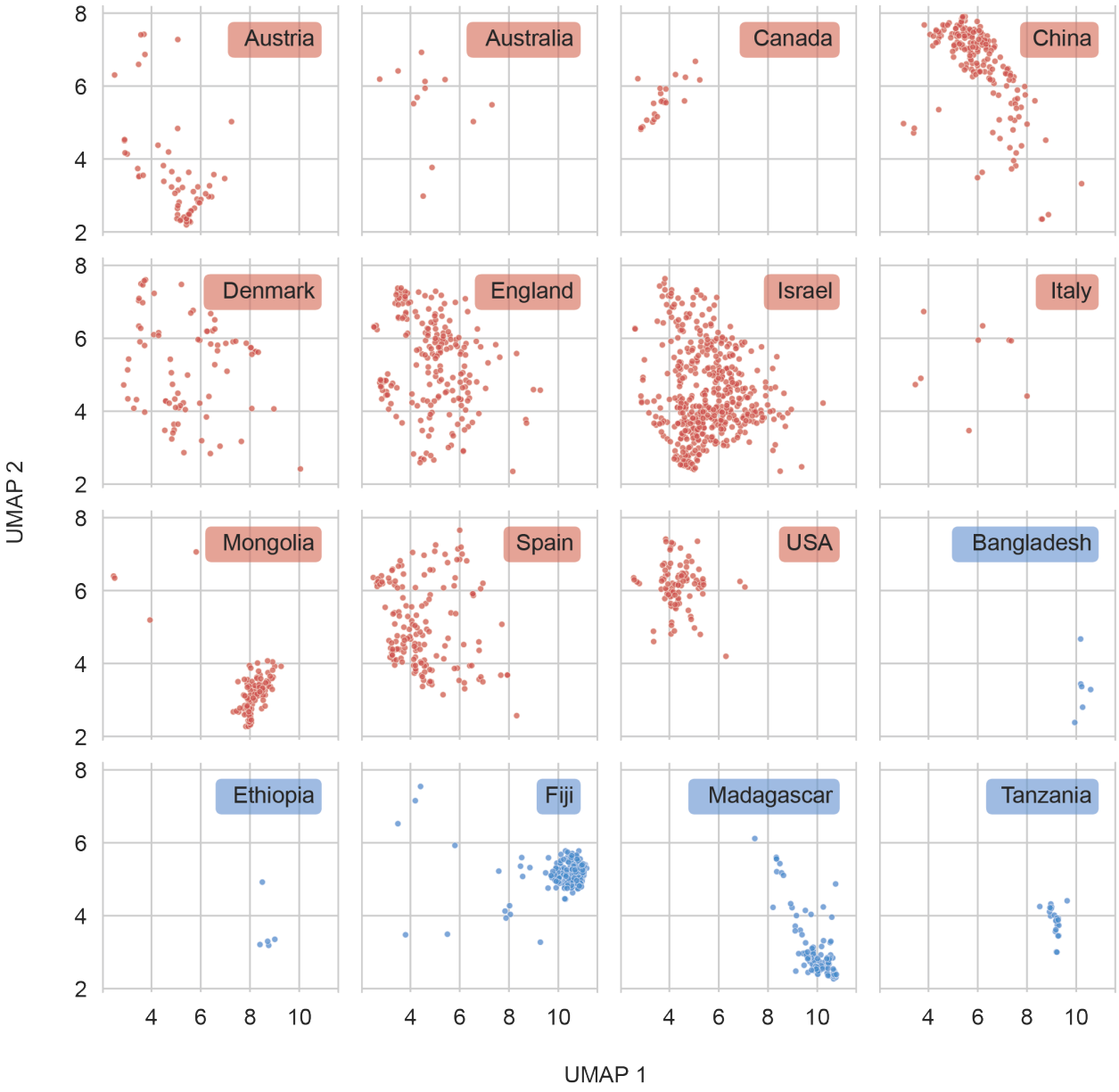

Figure 2.10: UMAP plots separated by country. Metagenomes have been partitioned to show clustering within each country

These results have parallels with previous studies that found associations between gut microbiota, as characterized by microbial taxonomy, and the geography and lifestyles of human populations [Monaghan et al., 2020, Mancabelli et al., 2017, Pasolli et al., 2019, Yatsunenko et al., 2012]. As geography is correlated to both plasmids and taxonomy, we wondered how many of our 68,350 plasmids are ecologically associated with and therefore explained by taxonomy. On one hand, such associations may be strong because plasmids are symbionts that rely on host machinery for replication and can have a narrow host range. On the other hand, such associations may be weak or nonexistent for two reasons. Some plasmids are known to have a range of multiple hosts [Klümper et al., 2015, Kohler et al., 2018, Bishé et al., 2019], which might not be neatly defined by a single species or even higher taxonomic category such as genus or phylum. Additionally, plasmids are often nonessential elements that can be gained or lost, such that nearly identical microbes can differ by the presence or absence of a plasmid or in the number of plasmid copies. Here, we systematically examined the ecological associations between plasmids and taxonomy to determine if plasmids comprise an independent component of microbial systems.

For every plasmid, we inferred its most likely host as the taxonomic group that had the most similar ecological distribution (see Methods). We surveyed taxonomic groups across all levels, from subspecies and species to class and phyla. We used two different formulas to calculate the ecological similarity between a plasmid and potential host: (1) the correlation in the abundance levels of the plasmid and host across metagenomes, and (2) how often the plasmid and host are found together in the same metagenome. Although some predicted plasmids had a high ecological similarity with their best matching taxonomic group, the vast majority of predicted plasmids had low similarity scores (median correlation = 0.04, median Jaccard = 0.21) (Figures 2.11A and 2.11B). We also observed low similarity scores even for reference plasmids that are isolated from a defined microbial host (Figures 2.11C and 2.11D). For example, the plasmid pDOJH10S and its cognate host, *Bifidobacterium longum*, were

present together in 10 metagenomes; however, 27 and 69 metagenomes contained only the plasmid or only the host, respectively (Figure 2.11E).

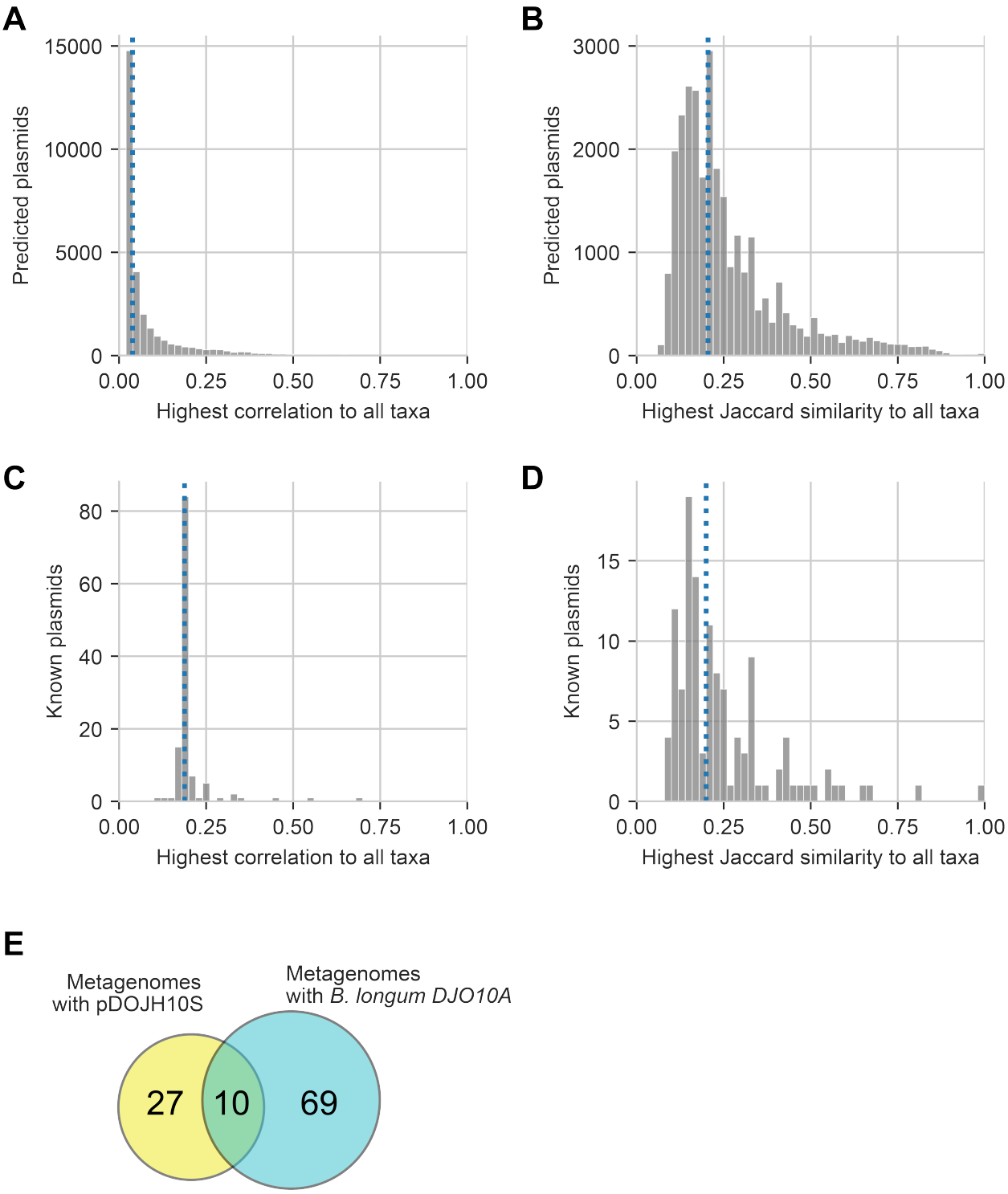

Figure 2.11: Comparison of the ecological distributions of plasmids and microbial taxonomy. We measured the association between every plasmid and taxon by calculating the correlation between their abundance levels across metagenomes, using the SparCC technique(Friedman and Alm 2012). As another association measure, we applied thresholds to the abundance

Figure 2.11 continued: levels and then calculated the Jaccard similarity between the metagenomes containing the plasmid versus those containing the taxon. We restricted analyses to plasmids that were present in at least 5 metagenomes. (A-B) For every predicted plasmid, we identified the taxon with the highest correlation (A) or Jaccard similarity (B). (C-D) We did the same to identify the best matching taxa of reference plasmids. Blue lines indicate the median of each distribution. (E) Venn diagram showing the discordance between the metagenomes containing a plasmid pDOJH10S and those containing its cognate host, a B. longum strain.

Overall, our findings suggest that plasmids are a highly complex and prevalent feature of microbiomes (Figures 2.9A, 2.9B, and 2.4D), forming an ecological dimension that can stratify human populations (Figures 2.9C and 2.10). With current methods of analysis, this stratification cannot be explained by microbial taxonomy alone (Figure 2.11). While high-throughput analyses of human gut microbiomes often focus on taxonomic features, it has been challenging to find significant or reproducible taxonomic associations that distinguish health and disease states(Lloyd-Price et al. 2019; Schloss 2018). As plasmids often carry key determinants for survival in an environment, we propose that systematic analysis of plasmid ecology is necessary to develop a complete understanding of the human microbiome.

2.4.5 Plasmid systems organize evolutionarily related plasmids by distinguishing backbone versus cargo content

Our large collection of predicted plasmids provides an unprecedented opportunity to study evolutionary patterns in plasmids and the extent to which they occur ecologically. Due to frequent genetic rearrangements, a hallmark of plasmid evolution is the reuse of a backbone and emergence of varying cargo/accessory genes [Sen et al., 2011, Orlek et al., 2017, Norberg et al., 2011, Fernandez-Lopez et al., 2017]. The backbone typically encodes machinery necessary for plasmid maintenance, while the cargo represents additional genetic content, such as antibiotic resistance or other fitness-determining functions. While backbones can be examined experimentally, most studies have identified them computationally.

Nonetheless, there are four major challenges to this computational task. First, there is a lack of consensus across studies on how to identify a plasmid backbone, with varying definitions based on nucleotide identity [Fernández-López et al., 2006, Sen et al., 2011, Holt et al., 2007], gene similarity [Norberg et al., 2011], or gene annotations [Garcillán-Barcia et al., 2015, Carattoli et al., 2005, 2014c]. Second, these methods do not verify that an identified backbone encodes a sufficient set of functions for plasmid replication. Third, these methods are typically designed to analyze a small set of plasmids in a single study or dataset. Finally, scaling methods to identify backbones in metagenomic data introduces extra complications related to plasmid redundancy and assembly fragments that could inflate the number of predicted backbones.

We designed a scalable algorithm called MobMess to study backbone structure in our collection of plasmids. Compared to previous methods, MobMess has the advantage of being able to simultaneously compare sequences without relying on gene annotations and to handle metagenomic issues of redundancy and fragmentation (see Methods). First, MobMess calculates pairwise alignments across all plasmids to build a sequence similarity network, in which a directed edge represents the containment of one plasmid within another (defined by greater than 90% sequence identity and greater than 90% coverage of the smaller plasmid) (Figures 2.12A and 2.13). Next, MobMess recognizes and collapses redundancy between plasmids. Finally, MobMess analyzes patterns of connectivity in the network to define and identify 'backbone plasmids' that satisfy two criteria. First, the backbone plasmid must be a circular element, inferred here by paired end orientation (Figure 2.3B), to ensure that it is not an assembly fragment and, importantly, that the genes present are sufficient for plasmid replication. Second, a backbone plasmid must be found as a subsequence within one or more 'compound plasmids'. These compound plasmids are composed of the backbone and additional cargo, indicating the ability to acquire or lose genes. We define a backbone and its compound plasmids as an evolutionary unit called a 'plasmid system' (Figure 2.12A).

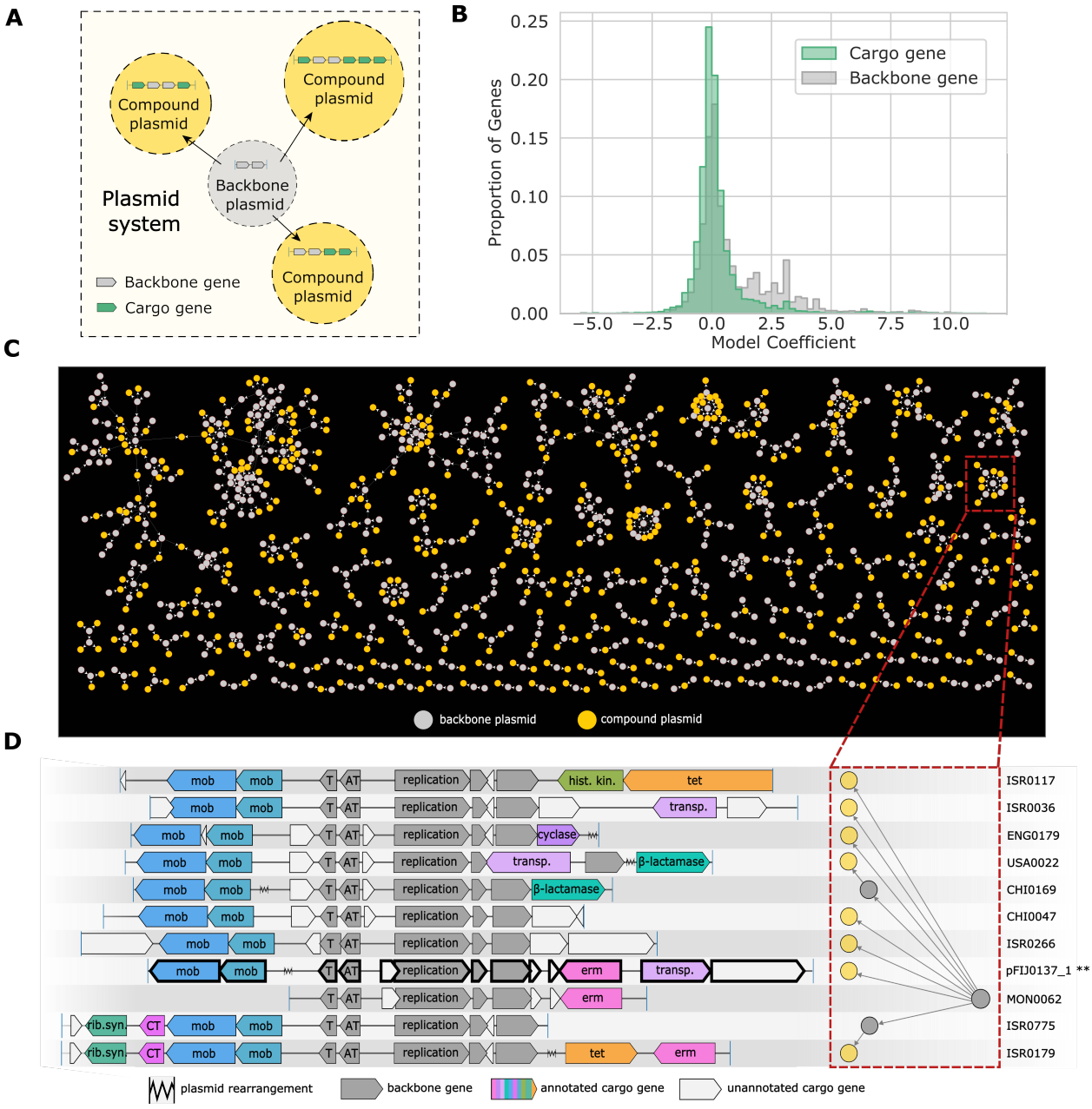

Figure 2.12: Identification of plasmid systems. (A) Network diagram of a plasmid system. (B) Distribution of model coefficients for backbone vs. cargo genes in the non-redundant set of 68,350 predicted plasmids. We excluded genes that lacked gene family annotations and thus have a coefficient of zero by default. We also excluded genes that were labeled as backbone with respect to some systems but cargo in others. (C) Network of all plasmid systems that contain greater than 3 non-redundant and high-confidence plasmids. Only these types of plasmids are shown. (D) Genetic architecture of plasmids in PS486, encased by a red box

Figure 2.12 continued: in C. Two plasmids in C are excluded. The system's backbone (assembled from metagenome MON0062) encodes 5 backbone genes (colored gray). Rib.syn.=riboflavin biosynthesis, CT=conjugative transfer, mob=mobilization, T=toxin, AT=anti-toxin, tet=tetracycline resistance, erm=erythromycin resistance, transp.=transposon, hist. kin.=histidine kinase.

This definition of plasmid systems enables a formal categorization of plasmids into evolutionarily cohesive groups and facilitates analyses of backbone versus cargo content and their ecology, much in the same way that pangenomes enable studies of core versus accessory gene content in microbial genomes. However, plasmid systems are a specific case of pangenomics, as it is unlikely to find a naturally occurring microbial genome composed only of core genes. In contrast, backbone plasmids represent a minimal entity that can propagate using only backbone genes. MobMess provides an automated framework and standardized vocabulary to study this concept across different studies and datasets.

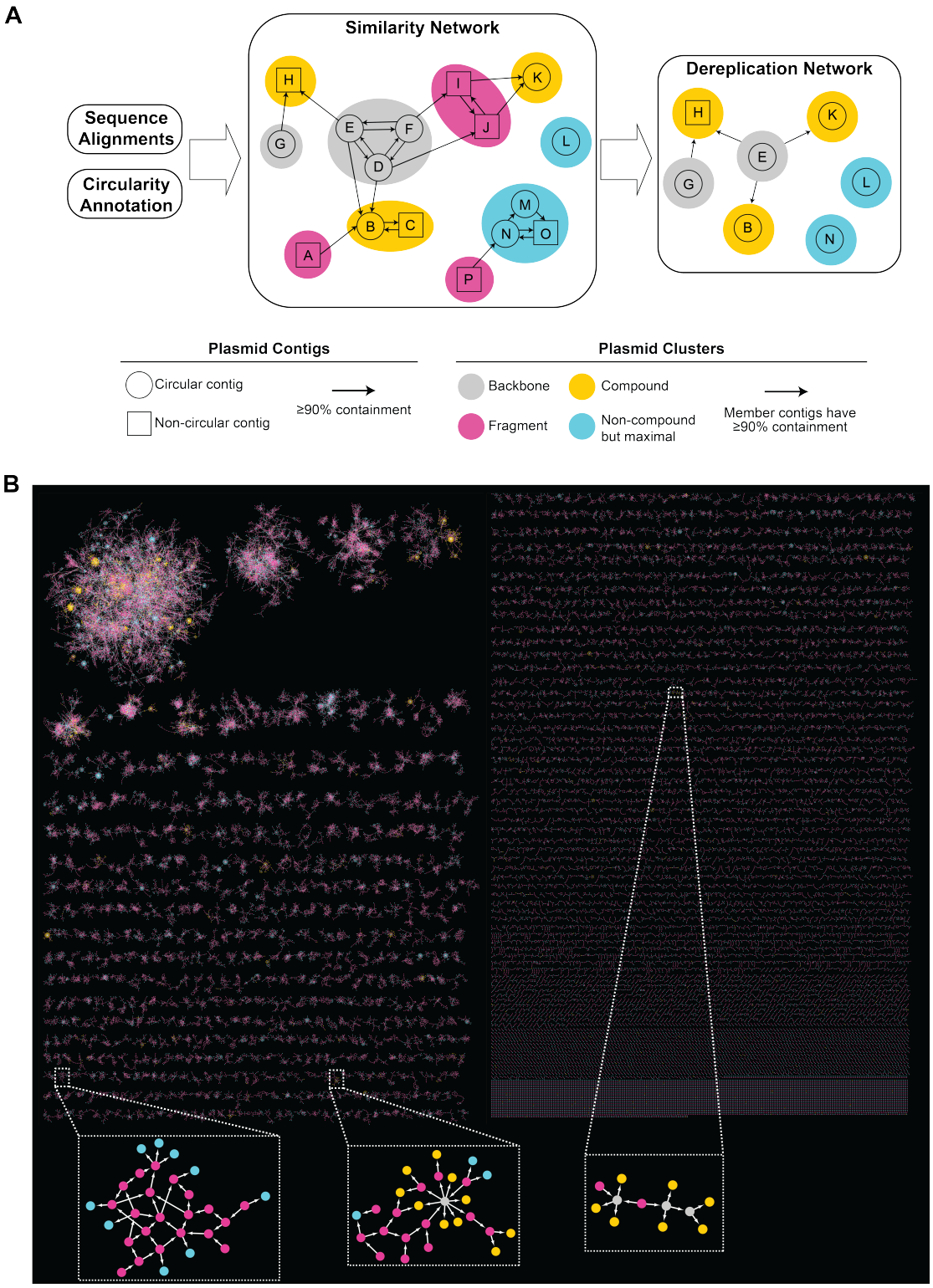

Figure 2.13: The MobMess algorithm and application to predicted plasmids. (A) Diagram of

Figure 2.13 continued: the MobMess algorithm for dereplicating plasmids and discovering plasmid systems. All-vs-all sequence alignments and circularity information are used to construct a similarity network of plasmid contigs. Similar contigs are clustered, and every cluster is labeled as either a backbone, fragment, compound, or non-compound maximal. A plasmid system consists of a backbone cluster and the compound clusters connected to the backbone. This example shows two systems: one system has G as the backbone (H is the compound plasmid), and another system has D, E, and F as the backbone (B, C, H, and K are the compound plasmids). To dereplicate, fragment clusters are discarded and a representative sequence is chosen for every non-fragment cluster. (B) Network of clusters of predicted plasmids. All clusters are shown except those that are not connected to any other cluster.

To define containment of plasmids within each other, we found that greater than 90% alignment identity and coverage was a natural threshold for two reasons. First, we examined the histogram of similarities between all pairs of predicted plasmids, revealing an average nucleotide identity (ANI) "valley" at around 85-90% identity (Figure 2.14A), although, similar to viruses [Bobay and Ochman, 2018], this drop was not as emphasized as those observed in the ANI between distinct bacterial taxa [Jain et al., 2018]. Second, we re-ran MobMess using varying thresholds. As the threshold is made stricter, plasmids gradually separated into distinct clusters, and consequently the number of non-redundant plasmids increased (Figure 2.14B-C). This growth in non-redundant plasmids occurred at a mostly constant rate from a threshold of 10% to 90%, but it suddenly accelerated from 90% to 100%. These results suggest that a threshold stricter than greater than 90% (e.g. greater than 95% or greater than 99%) would split highly similar plasmids into separate clusters.

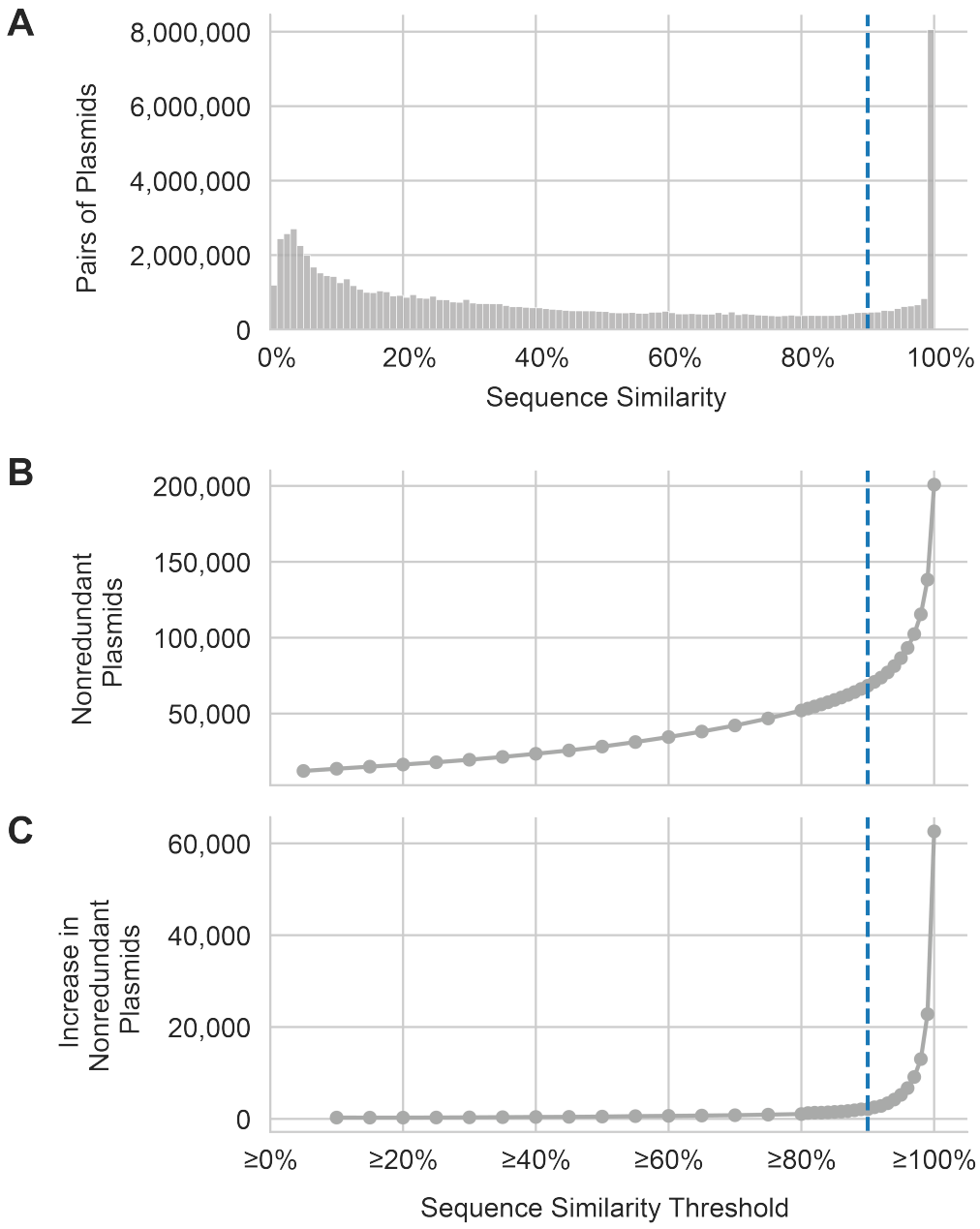

Figure 2.14: Choosing a similarity threshold for MobMess. (A) Histogram of similarities between every pair of the 226,194 predicted plasmid contigs. (B) We ran MobMess using different thresholds on the similarity, and then we calculated the number of non-redundant plasmids generated. (C) The derivative of the curve in B. The blue dashed lines represent our current greater than 90% similarity threshold.

Other methods have recently been developed to cluster thousands of plasmids [Redondo-Salvo et al., 2020, Acman et al., 2020], but unlike MobMess, they are not designed to identify plasmid systems or analyze metagenomic data. To compare methods, we ran MobMess on the same set of 9,894 reference plasmids analyzed by Redondo-Salvo et al [Redondo-Salvo et al., 2020] 2020) (Figure 2.15). In their study, Redondo-Salvo et al. constructed a plasmid similarity network with 79,727 edges. However, these edges span a wide range of similarity levels, where 66.5% of edges represent an alignment that covers <90% of either sequence (greater than 10% is not aligned) and 19.0% of edges have <70% alignment coverage (greater than 30% is not aligned). In contrast, MobMess applies a stricter threshold of greater than 90% coverage to construct a smaller but more refined set of 39,680 edges (connecting 25,270 unique pairs of plasmids). Moreover, Redondo-Salvo et al.'s edges are undirected, while MobMess's edges are directed to track smaller versus larger sequences. Retaining this extra information allowed MobMess to distinguish between the 10,860 pairs (43.0%) with unidirectional connections, representing a backbone contained in a compound plasmid, versus the 14,410 pairs (57.0%) with bidirectional connections, representing nearly identical plasmids.

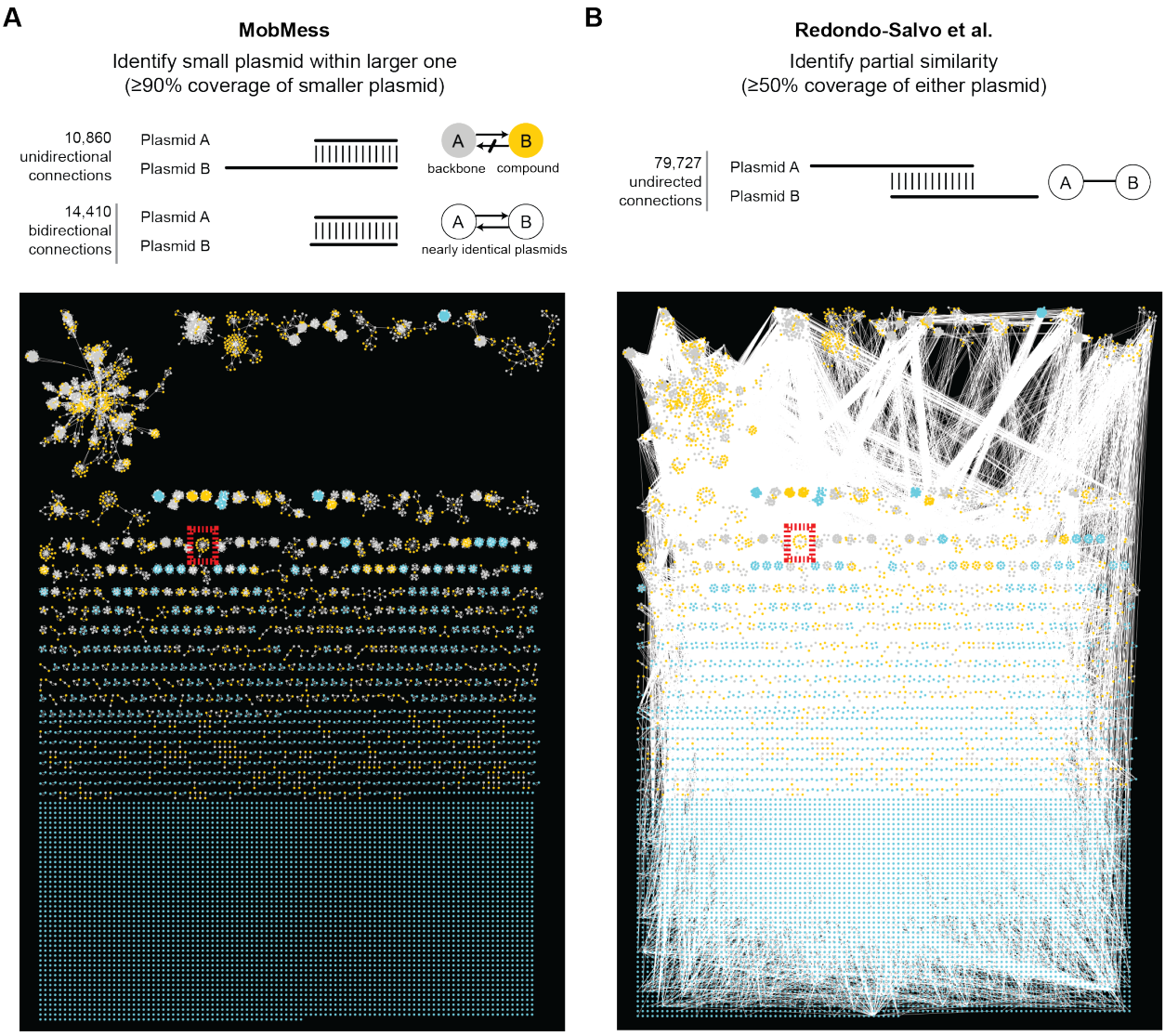

Figure 2.15: Conceptual differences in constructing plasmid similarity networks. We ran MobMess on the set of 9,894 reference plasmids analyzed by Redondo-Salvo et al.(Redondo-Salvo et al. 2020). MobMess constructs a network with directed edges, by aligning plasmids and determining if one plasmid is found as a subsequence within another. Redondo-Salvo et al. constructs a network with undirected edges, by determining whether two plasmids contain partial homology. (A-B) Visualization of the similarity networks. We used Cytoscape(Shannon et al. 2003) and the Prefuse directed layout algorithm(Heer, Card, and Landay 2005) to lay out the nodes in the MobMess network (A), and then we applied the same layout to the Redondo-Salvo et al. network (B). The red boxes represent the example shown in Figure 2.16.

Besides network construction, these methods also diverge in how they conceptually organize plasmids. MobMess dereplicates the 9,894 plasmids into 7,132 non-redundant sequences and then organizes them into 1,044 plasmid systems. In contrast, Redondo-Salvo et al. identified 641 clusters, or 'PTUs' [Redondo-Salvo et al., 2020]. We found that 135 PTUs did correspond one-to-one to a plasmid system in MobMess, but the other PTUs spanned a wide range of evolutionary relations. At one extreme, 251 PTUs were simple sets of nearly identical plasmids, representing recent and strong relations. At the other extreme, 45 PTUs were complex mixtures of distinct plasmid systems, representing distant and weak relations. For example, the largest PTU contained 2,460 plasmids, which MobMess further dissected into 1,481 non-redundant plasmids and 461 plasmid systems. Figure 2.16 demonstrates one such plasmid system, where MobMess precisely connects the system's backbone to its compound plasmids in a "star"-like topology, while the approach by Redondo-Salvo et al. connects almost every pair of these plasmids to each other, which obfuscates the internal organization of the plasmid system. Perhaps this is in part because the method by Redondo-Salvo et al. and another related method by Acman et al. [Acman et al., 2020] have only been tested on reference plasmids that have been completely assembled, while MobMess is designed to handle metagenomic data by distinguishing between fragmented versus complete (circular) plasmids.

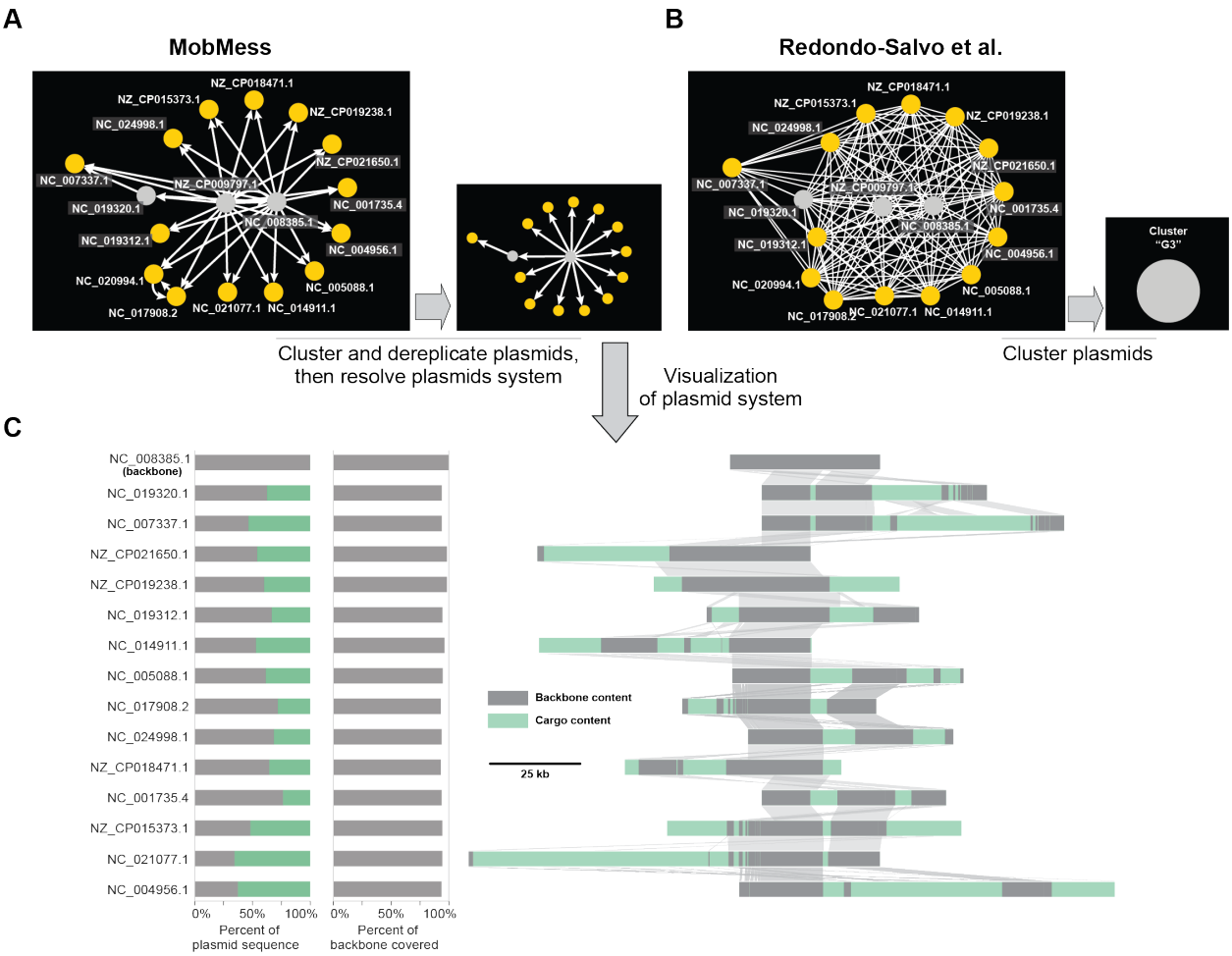

Figure 2.16: Comparison of MobMess versus Redondo-Salvo et al. for studying a plasmid system. (A-B) An example from the similarity networks in Figure 2.15, showing the connections between 17 plasmids from the same plasmid system. MobMess further collapses its network to dereplicate plasmids and reveal the plasmid systems's "star"-like topology, where a backbone connects to its compound plasmids. Redondo-Salvo et al. did recognize that these plasmids are related (represented by a cluster called "G3"), but they connected almost every pair of these plasmids in a "hairball" topology, obfuscating the system's internal organization. (C) Alignments of plasmids in the MobMess system. Subregions in every sequence are colored gray or green to represent backbone or cargo content, respectively. Ribbons between sequences represent the alignment of subregions. The barcharts show the total breakdown of each plasmid into backbone versus cargo, as well as the fraction of the backbone sequence ('NC_008385.1') that is found within the plasmid.

2.4.6 MobMess identifies 1,169 plasmid systems with conserved backbones and a wide repertoire of cargo functions

We ran MobMess on our predicted plasmids and identified a total of 1,169 plasmid systems, naming them PS1 (plasmid system #1) to PS1169. While plasmid systems captured a small fraction of the genetic diversity among non-redundant plasmids (6.5%, or 4,424/68,350), they captured a large fraction of all circular plasmid contigs (72.7%, or 14,285/19,652) (see Methods, Table 2.4). Plasmids that were part of a system tended to be longer and were more likely to be circular than plasmids that were not part of any system (Table 2.12). The requirement to be included in a plasmid system is that the sequence must not only be predicted as a plasmid (with score >0.5), but that there must also be at least one other predicted plasmid that shares the same backbone. Thus, while we previously applied a loose score threshold of >0.5 instead of >0.9 to identify plasmids, MobMess provides an independent de novo filter for plasmids with higher confidence. Indeed, we found that 16,663 plasmids with scores between 0.5 and 0.9 are part of a system.

Plasmid systems were highly heterogeneous in their genetic complexity. 37 plasmid systems contained sequences that could be classified among 7 different plasmid incompatibility types (Inc11, Inc18, IncFIB, IncFIC, IncI-gamma/K1, IncK2/Z, IncW) (Table 2.4). 602 plasmid systems contained at least 2 non-redundant compound plasmids, with the largest system containing 168 non-redundant compound plasmids (Figure 2.12C). For example, pFIJ1037_1, the plasmid we isolated and transferred between *B. fragilis* organisms, was part of PS486, a system containing 24 non-redundant plasmids and found across a total of 127 metagenomes. PS486's backbone consists of a replication protein and a toxin-antitoxin system, and the cargo genes include beta-lactamases, erythromycin resistance, tetracycline resistance and riboflavin biosynthesis (Figure 2.12D, Table 2.5). To understand how much genetic content is typically conserved or variable in a plasmid system, we calculated the percentage of genes on compound plasmids that were backbone genes versus cargo genes

(see Methods). Plasmid systems spanned a wide range of cargo gene percentages between 0% and 100%, with a median value of 40% (Figure 2.17). Conversely, the median backbone percentage was 60%. PlasX often assigned higher model coefficients to backbone genes in the non-redundant set of predicted plasmids, suggesting these genes define the 'essence' of a plasmid by encoding essential functions that promote the ability of a plasmid to exist as a distinct element from the chromosome, such as the genes for plasmid replication, *repA* (PF01051), and mobilization, *mobA* (PF03432) (Figure 2.12B). In contrast, PlasX assigned lower coefficients to cargo genes, suggesting they encode functions that are not universally essential but important for specific niches, such as nitrogen reductase, nifH (PF00142), and membrane transport, ompA (PF00691). Indeed, 24.1% (2,169/8,995) of backbone genes versus 13.4% (3,229/24,168) of cargo genes encoded COG and Pfam functions with descriptions related to plasmid replication, transfer, and maintenance (see Methods).

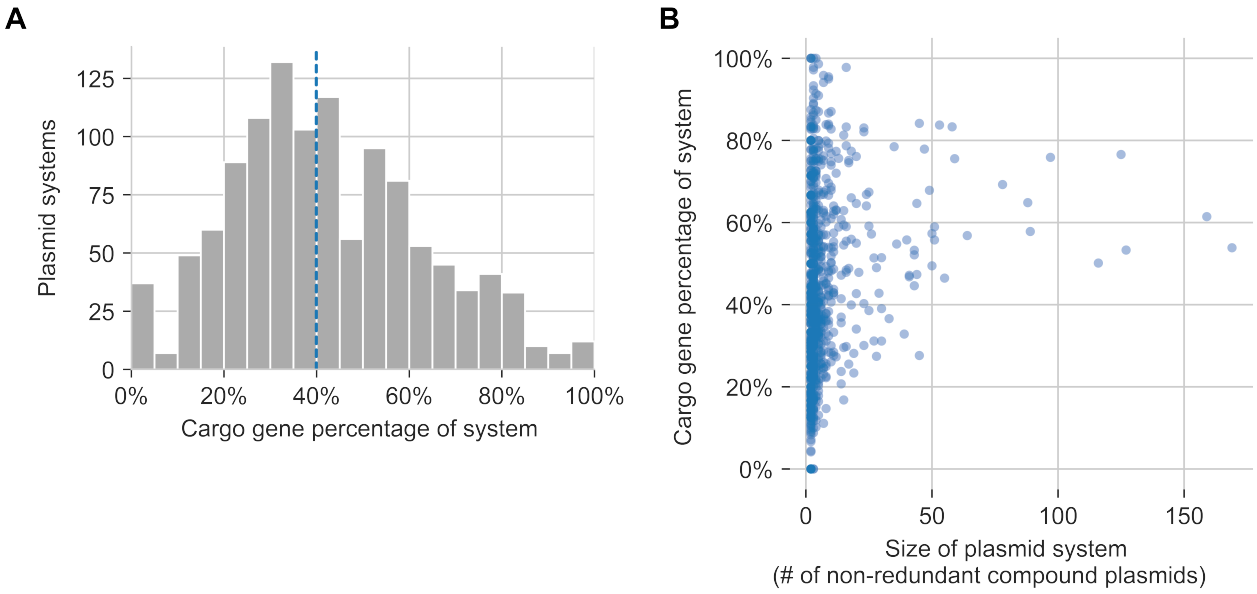

Figure 2.17: Backbone and cargo composition of plasmid systems. (A) For every plasmid system and compound plasmid in the system, we calculated the percentage of genes on the compound plasmid that were classified as cargo versus backbone genes (see Methods). We then averaged the cargo gene percentages across all compound plasmids in the system (x-axis). The vertical blue line shows the median at 40%. (B) Scatterplot of the cargo gene percentage versus the size of a plasmid system, showing a lack of correlation (R2 = 0.03). We defined the size as the number of non-redundant compound plasmids.

The most frequent type of function encoded on cargo genes was antibiotic resistance, including efflux pumps, which can provide general resistance to multiple antibiotics, and genes targeting specific classes of antibiotics, such as glycopeptides and beta-lactams (Figure 2.18A). This large-scale observation is consistent with numerous examples of known plasmids encoding resistance and further illustrates how the widespread presence of these plasmids pose a public health threat [Vrancianu et al., 2020, MacLean and San Millan, 2019, World Health Organization, 2018, Centers for Disease Control and Prevention, 2021].

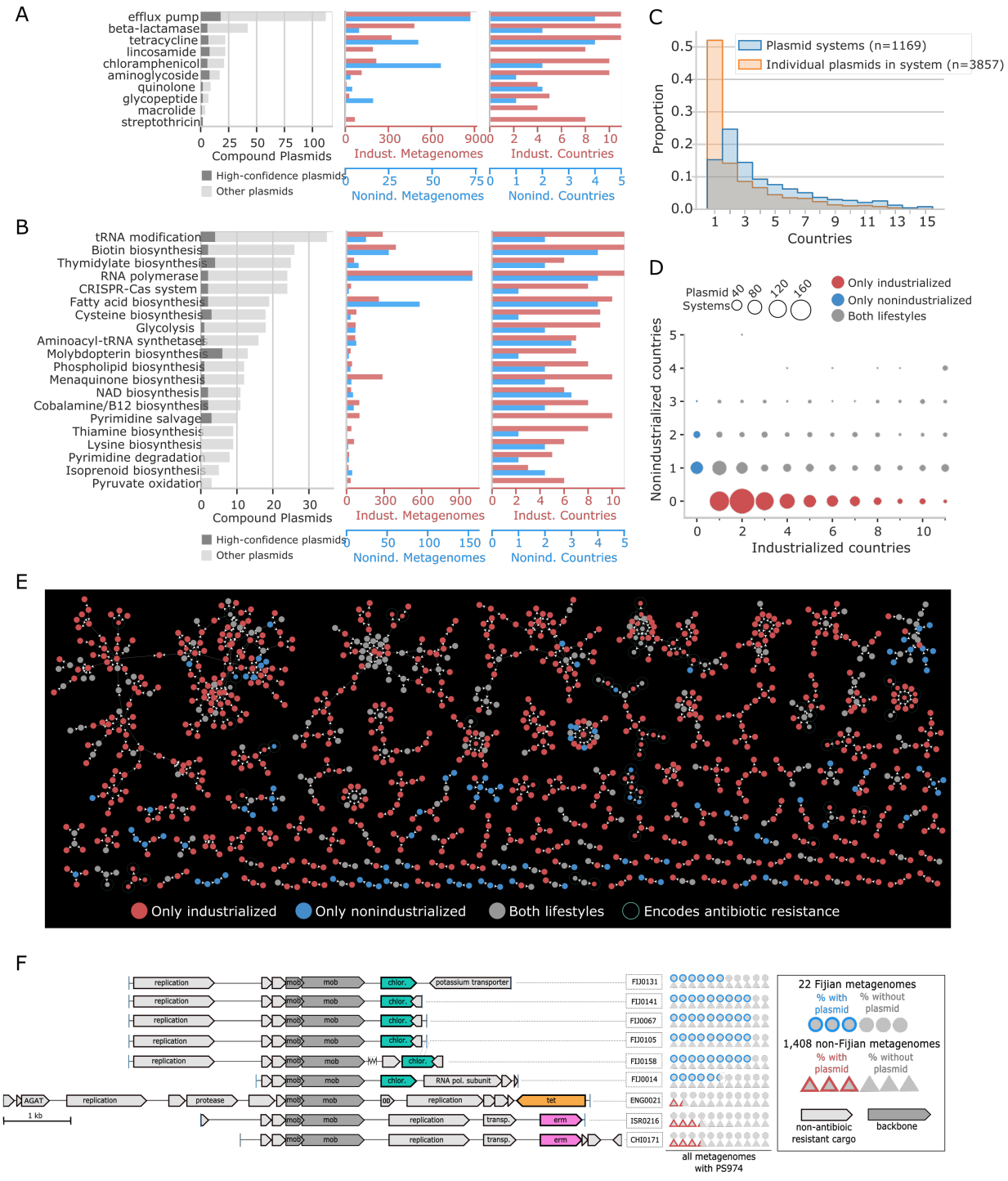

Figure 2.18: Functional and ecological variation of plasmid systems. (A-B) The number of (C) Prevalence of plasmid systems versus the individual plasmids in those systems.

Figure 2.18 continued: (D) Distribution of plasmid systems based on the number of industrialized and non-industrialized countries they are found in. (E) Recoloring of the network of plasmid systems shown in Figure 2.12C. Colors indicate whether a plasmid occurred in only industrialized, only non-industrialized, or both types of countries. A green ring indicates a plasmid encoding antibiotic resistance. (F) Compound plasmids from PS974 that encode for resistance to chloramphenicol (chlor), tetracycline (tet), or erythromycin (erm). 6/9 plasmids are circular. Dark gray genes are the backbone; light gray are cargo not related to antibiotic resistance. AGAT=aminoglycoside adenylyltransferase. OD=Oxaloacetate decarboxylase. PS974 is found in 22 Fijian and 1,408 non-Fijian metagenomes. The pictogram on the right-hand side represents these metagenomes using two shapes: circles (Fijian) and triangles (non-Fijian). For each plasmid, circles are colored blue to represent the proportion of the 22 Fijian-metagenomes that contain the plasmid. Similarly, triangles are colored red to represent the proportion of the 1,408 metagenomes that contain the plasmid.

Other highly prevalent cargo functions included a wide diversity of cellular and metabolic pathways defined in the COG (Figure 2.18B) and KEGG databases (Figure 2.19). The most enriched among these was tRNA modification, encoded in 35 compound plasmids within different systems. For example, the globally prevalent system PS1110 (present in 739 metagenomes) contained 291 compound plasmids (27 non-redundant), three of which encoded an enzyme that performs tRNA Gm18 2'-O-methylation (COG0566) and were collectively present in 498 metagenomes (Figure 2.20, Table 2.5). This enzyme is thought to reduce the immuno-stimulatory nature of bacterial tRNA, which is detected by Toll-like receptors (TLR7) of the mammalian innate immune system [Gehrig et al., 2012, Galvanin et al., 2020]. While plasmids in some pathogens are known to facilitate bacterial evasion of mammalian immune system by regulating surface proteins [Embers et al., 2008], the overwhelming prevalence of tRNA modification enzymes in our data suggests the likely presence of a previously unappreciated role for plasmids to increase the fitness of their bacterial hosts against the surveillance of the human immune system. Distinguishing between the fundamental structure of a plasmid (backbone genes) versus the genetic currency that is exchanged (cargo genes) allowed us to organize the extensive plasmid diversity within plasmid systems and to recognize the recurrent evolution of the same cargo functions across different systems.

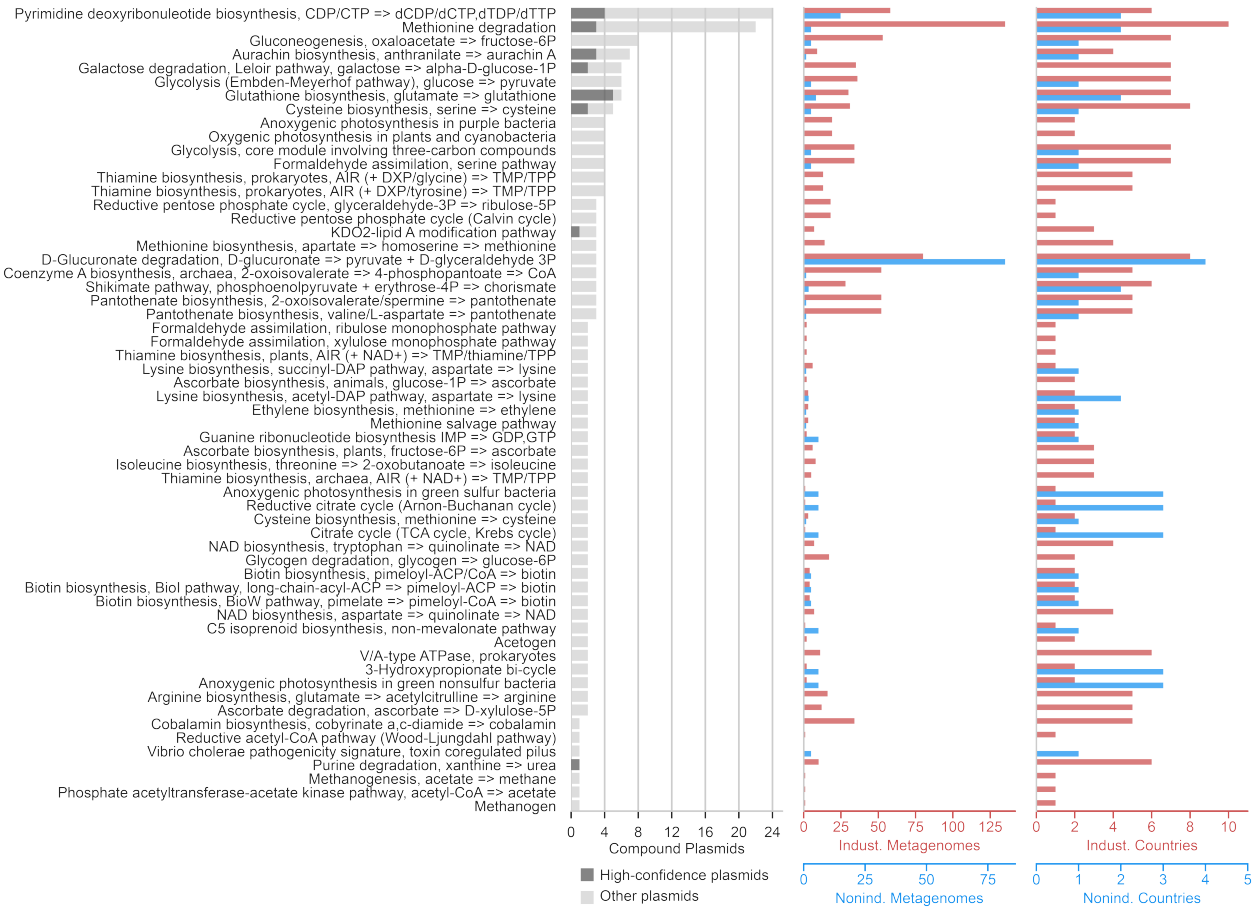

Figure 2.19: Functional annotation of cargo genes to KEGG modules. This plot excludes KEGG modules that occur in only one plasmid system or that occur in cargo genes annotated to antibiotic resistance.

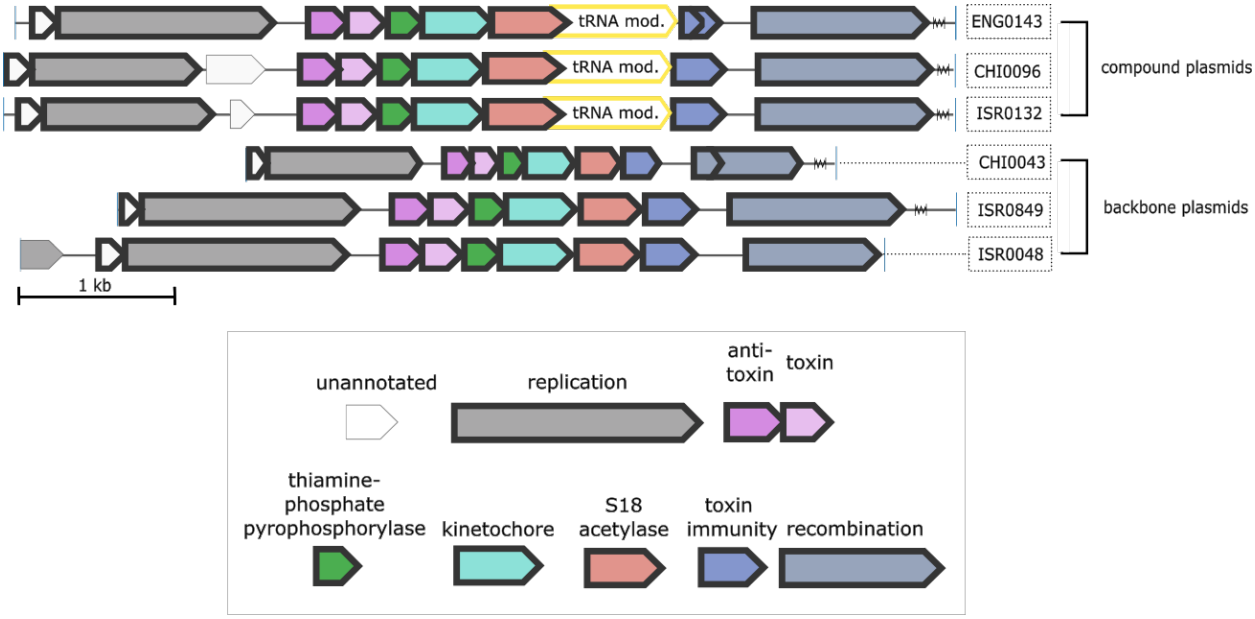

Figure 2.20: Plasmid system PS1110. Compound plasmids in this system contain a gene that encodes two enzymes, a tRNA Gm18 2'-O-methylase (yellow, 'tRNA mod.') and a Ribosomal protein S18 acetylase (red). Backbone plasmids contain a similar gene that encodes the S18 acetylase but lacks the tRNA methylase. Backbone genes have a thick, black outline.

2.4.7 Plasmid systems adapt their cargo genes to specific environments

Thus far, we have observed that our collection of plasmids is highly heterogeneous in their ecological distributions (Figure 2.9B), yet they can also be organized by evolutionary relations into plasmid systems. To understand how ecology and evolution are intertwined phenomena, we asked whether plasmid systems span a single ecological niche or multiple niches. We assumed that every country represents a different niche, as countries are known to differ in microbial composition [Gupta et al., 2017, Yatsunenko et al., 2012, Obregon-Tito et al., 2015, Gomez et al., 2016, Li et al., 2014, Xia et al., 2019, Sonnenburg and Sonnenburg, 2019] and we observed that countries also differ in plasmid composition (Figure 2.10). We found that while individual plasmids are often present in a single country, a plasmid system frequently spans multiple countries (Figure 2.18, Table 2.4). Indeed, 2,005 individual plasmids within a system were unique to a single country, yet 1,794 (89.5%) were part of more geographically diverse systems that were present in at least two countries. In fact, 84.0% (982/1,169) of systems were present in at least two countries, and 9 systems were even in 15 of the 16 countries represented in our data. We also found that plasmid systems are typically mutually exclusive, i.e. plasmids from the same system generally do not occur in the same individual human hosts. Consequently, metagenomes often have at most one plasmid from every system (see Methods, Figure 2.21A).

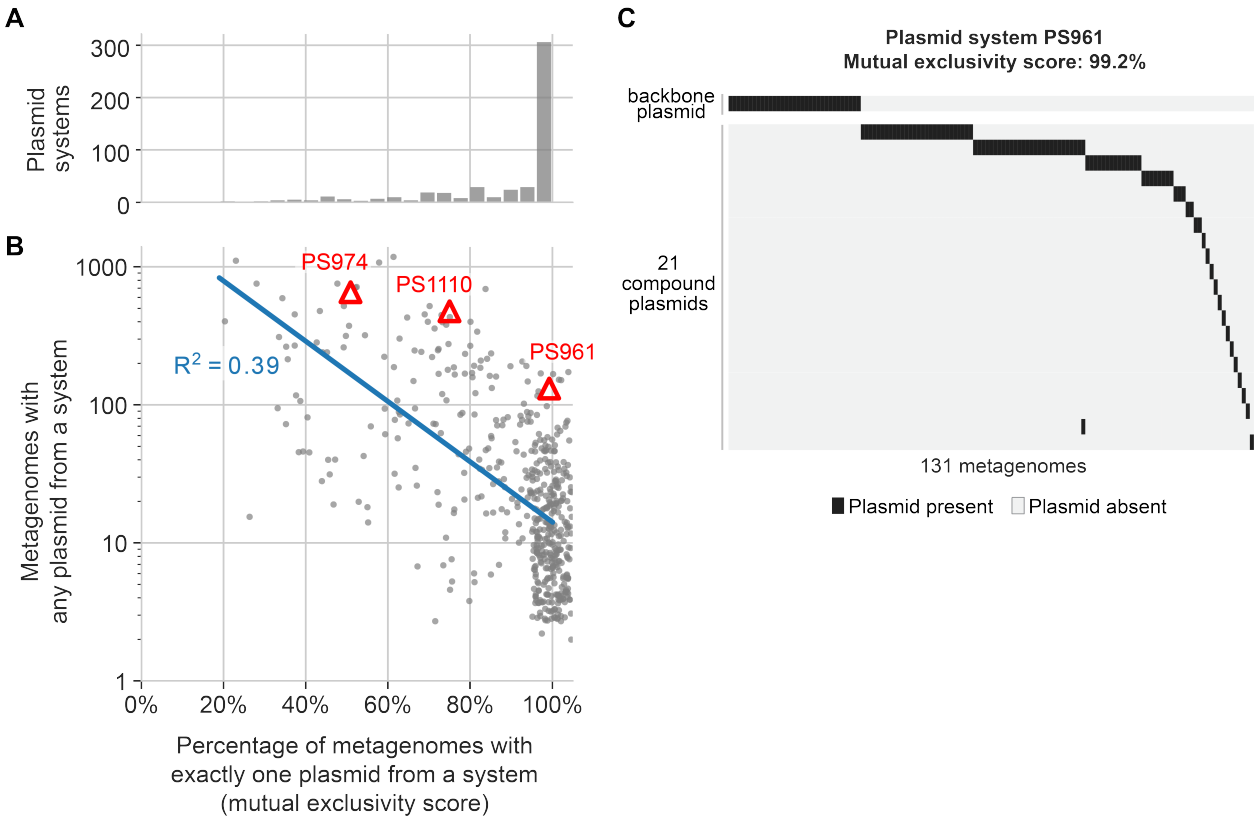

Figure 2.21: Mutual exclusivity of plasmids in the same system. For every plasmid system with two or more compound plasmids, we defined a 'mutual exclusivity score' to quantify how often its plasmids segregated to different metagenomes. We definethis score as the number of metagenomes that have exactly one of the system's plasmids divided by the number of metagenomes that have any of the system's plasmids. (A) Histogram of mutual exclusivity. (B) The inverse relation between the mutual exclusivity and prevalence of a system. Red triangles represent examples of plasmid systems that are highly prevalent (present in >100 metagenomes) but are more mutually exclusive than expected by a linear regression (blue line). For easier visualization, the x- and y-coordinates of systems with >90% mutual exclusivity were randomly jittered within ±5% of the axes lengths. (C) Visualization of mutual exclusivity in PS961, which is one of the examples in B.

One explanation of this mutual exclusivity pattern is that a person may have never been exposed to multiple plasmids from the same system. This scenario is plausible for rare systems (e.g. present in <10 metagenomes). Indeed, we observed a trend where more prevalent systems exhibit less mutual exclusivity (R2=0.39) (Figure 2.21B). However, we identified many prevalent systems (e.g. present in >100 metagenomes) that are more mutually exclusive than expected. For instance, PS961 is almost perfectly mutually exclusive across 131 metagenomes (Figure 2.21C). For such prevalent systems, a more likely explanation lies in either the backbone or cargo content of a system. As 24.1% of backbone genes encoded Pfams and COGs related to replication, transfer, or maintenance of plasmids, the competition for these resources within a cell can lead to incompatibility between plasmids of the same system [Thomas, 2014a]. With a large number of plasmid sequences at hand, our dataset is a new resource that can support experimental investigations of the impact of incompatibility on plasmid replication [Novick, 1987, Velappan et al., 2007], especially its contribution relative to external selective pressure on cargo genes. However, here we focused our attention on investigating ecological associations between environmental selection and cargo gene content.

We propose that our algorithmic definition of plasmid systems can be used to study how ecological pressures on plasmids drive the evolution of cargo genes. Plasmid systems are akin to the concept of a genetic 'delivery van' (backbone) with the flexibility to disseminate variable 'packages' (cargo genes) to microbes. The dynamic pool of cargo genes could serve as a means for a plasmid, or its microbial host, to increase fitness in different environmental conditions. To investigate this hypothesis, we compared the evolution of plasmids in industrialized versus non-industrialized countries, an environmental difference that was reflected in plasmid content (Figure 2.9C). While many plasmids systems were exclusive to one or the other type of country, 396 of them were present in both types (Figure 2.18D). These global systems provide a unique *in silico* framework to explain environmental differences by

variation in plasmid cargo genes.

To demonstrate this framework, we examined antibiotic usage, an extreme environmental difference that is well known to exert selective pressures on microbes, often causing them to maintain plasmids with antibiotic resistance [Svara and Rankin, 2011, Sykes, 2010, Cantón and Morosini, 2011, Baquero, 2001, San Millan et al., 2016, Alonso et al., 2001, Xiong et al., 2015, Ma and Bryers, 2012]. The well-studied nature of antibiotic resistance also provides a testbed to demonstrate that plasmid systems can identify cargo genes under selection. In our data, the evolution of antibiotic resistance in a plasmid system coincided with the ecological variation of compound plasmids in the system. Specifically, we identified 24 high-confidence, compound plasmids that encoded antibiotic resistance in cargo genes and were exclusively present in either non-industrialized or industrialized countries (Figure 2.18E). Among non-industrialized metagenomes, one of the most common types of antibiotic resistance is chloramphenicol (Figure 2.18A). For instance, PS974 is highly diverse with 97 non-redundant plasmids; however, this system possesses chloramphenicol resistance (conferred via an acetyltransferase) only in plasmids assembled from Fiji (Figure 2.18F, Table 2.5). When we searched for these resistance plasmids across the global set of 1,430 metagenomes that contain PS974, we found them in 19/22 Fijian metagenomes but only in 1/1,408 non-Fijian metagenomes ($p=1.1 \times 10^{-13}$, Fisher's exact test) (Figure 2.18F, pictogram). Chloramphenicol is routinely prescribed in Fiji to treat eye infections, central nervous system infections, periodontitis, shigellosis, typhoid and paratyphoid fevers, and diabetic foot infections, but it is rarely used in North America and Europe [Berendsen et al., 2010, Both et al., 2015, Balbi, 2004, Ministry of Health and Medical Services, Government of Fiji, 2019]. Thus, chloramphenicol resistance in this system likely reflects the increased exposure of Fijians to this antibiotic. While we observed that chloramphenicol-resistant plasmids appeared specific to Fiji, more extensive sampling may reveal the presence of these plasmids in other non-industrialized countries that also have high usage of this antibiotic.

Besides chloramphenicol resistance, PS974 also contained non-Fijian compound plasmids that carry tetracycline resistance (171/1,408 metagenomes) or erythromycin resistance (429/1,408 metagenomes) (Figure 2.18F). In an attempt to find an alternative explanation for the distribution of resistance plasmids in this system, we revisited our earlier question "Can plasmid ecology be simply explained by taxonomy?". By searching for these plasmids among known sequences in NCBI, we determined that possible microbial hosts include Firmicutes, such as Blautia hydrogenotrophica. However, none of these hosts nor any other microbial taxon had a similar ecological distribution as any of the resistance plasmids (highest Jaccard index=0.37 across all plasmid-taxon comparisons). These results suggest that compound plasmids in systems have acquired antibiotic resistance to respond to lifestyle-specific usage of antibiotics.

While the connection between antibiotic usage and resistance is expected given previous studies [Svara and Rankin, 2011, Sykes, 2010, Cantón and Morosini, 2011, Baquero, 2001, Alonso et al., 2001, Xiong et al., 2015, Ma and Bryers, 2012], plasmid systems in general can be used to determine if cargo genes are under selection, even when the functions of those genes or the environmental pressures driving the selection are unknown. For example, the cargo gene encoding tRNA modification in system PS1110 may provide some immunoevasive function (Figure 2.20), but there was not a clear geographic or lifestyle-specific association with this function. This is an example where we know the cargo function but not the environmental pressure and motivates collecting additional data about the environment. Overall, our work provides a computational roadmap for generating new hypotheses about plasmid evolution on an omics scale.

2.5 Discussion

Our work greatly expands the number of known plasmids by mining a global collection of metagenomes using machine learning. This expansion provides the community with a new

resource to study fundamental concepts in plasmid biology. While there are many applications of our resource, here we focused on organizing plasmids into cohesive units known as plasmid systems to gain deeper insights into plasmid ecology and evolution. For instance, the diversity captured by our large collection of 1,169 plasmid systems reveals the great extent to which plasmids in complex ecosystems like the human gut are not static entities but actively evolving in response to the environment. By revealing likely determinants of fitness, such as the acquisition of specific antibiotic resistance genes in Fiji as a response to a commonly used antibiotic, plasmid systems serve as a hypothesis generation and testing tool to study forces that drive plasmid evolution and influence the ecology of hosts that carry them. Our study has focused on geographical or lifestyle-based environmental differences, but more generally, our analysis of plasmid systems can be applied to other contexts such as discerning cargo genes that distinguish healthy vs. disease states of the gut microbiome.

During the past few decades, our ability to bypass the limitations of cultivation and study microbial genomes derived from metagenomes has led to key biotechnological insights [Delmont et al., 2018, van Kessel et al., 2015, Edwards et al., 2019, Hug et al., 2016]. The malleability of plasmids is a desirable property in bioengineering and often has motivated the repurposing of naturally occurring plasmids into major tools for genetically modifying organisms. In this vein, we propose computational prediction and analysis of plasmids as an attractive approach to expand the toolkit of available plasmids for genetic engineering, particularly if they can be found in isolates that currently lack tools to make them genetically tractable.

PlasX and other plasmid recognition systems [Zhou and Xu, 2010, Krawczyk et al., 2018, Pellow et al., 2020, Carattoli et al., 2014c, Schwengers et al., 2020, Royer et al., 2018, Hou et al., 2021, Arredondo-Alonso et al., 2018, Gomi et al., 2021], along with MobMess to characterize plasmid systems, present a roadmap for a detailed characterization of naturally occurring plasmids. Historically, plasmids and other genetic elements have been character-

ized on the basis of qualitative properties and descriptions. In contrast, PlasX and machine learning approaches provide an "operational definition" of a plasmid that can be universally and objectively applied. As some of our predicted plasmids contain virus-like or ICE-like signatures, our work can be used to study the spectrum of mobile elements that blur traditional labels and complements recent efforts to characterize viruses [Guo et al., 2021, Al-Shayeb et al., 2020, Shkoporov and Hill, 2019, Antipov et al., 2020] and horizontally transferred elements [Smillie et al., 2011, Groussin et al., 2021, Brito et al., 2016].

To expand the scope of our work, we intentionally designed PlasX using a broad collection of reference sequences, so that it can be applied to study any environment and can include additional training sequences to improve accuracy. These methods provide a complementary approach to frequently used state-of-the-art workflows to study the taxonomic composition or functional potential of environmental or host-associated microbiomes through amplicon sequences or metagenomes. Overall, our findings suggest that high-throughput recognition and characterization of plasmids in microbiome studies are necessary for more complete insights into the ecology of naturally occurring microbial systems.

2.6 Methods

2.6.1 Compiling and annotating a reference set of plasmids and chromosomes

We obtained a list of 16,168 plasmids from the March 5, 2019 version of PLSDB [Galata et al., 2019]. We also downloaded the entire collection of 13,471 complete bacterial genome assemblies from NCBI RefSeq on October 26, 2019, using instructions at https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#allcomplete. The RefSeq assemblies contained 26,376 contigs, of which we discarded 11,350 that are also in PLSDB. The reference set of 16,827 plasmids consisted of 16,168 PLSDB contigs, as well as 659 contigs from the

Refseq assemblies that were labeled as 'Plasmid' in the 'Assigned-Molecule-Location/Type' field of the NCBI assembly report. The reference set of chromosomes was the remaining 14,367 Refeq contigs.

To identify and annotate genes in these sequences, we used the program 'anvi-run-workflow' with '–workflow contigs' implemented [Shaiber et al., 2020] in anvi'o [Eren et al., 2021] v7.1, which uses Snakemake [Köster and Rahmann, 2012] to execute previously defined steps (https://merenlab.org/anvio-workflows/) to generate anvi'o contigs-db files (https://anvio.org/m/contigs-db). These steps include first running Prodigal [Hyatt et al., 2010] to call genes and then running DIAMOND v2.0 [Buchfink et al., 2015] and HMMER v3.3 [Eddy, 2011] on amino acid sequences to determine gene functions against the Cluster of Orthologous Groups of proteins (COGs) [Galperin et al., 2015] and Protein Family Database models (Pfams) v32.0 [El-Gebali et al., 2019], respectively. To minimize noise, we used an e-value cutoff of 10^{-10} for COGs and the default model scores for Pfams.

2.6.2 Modeling de novo gene families

We inferred de novo gene families by running MMseqs2 [Steinegger and Söding, 2018] v10.6d92c on all amino acid sequences in our reference plasmids and chromosomes. First, we ran 'mmseqs clusthash' to collapse identical sequences into a non-redundant set for faster execution of the next step; the collapsing was inverted at the end to annotate all genes. Next, we ran 'mmseqs cluster' to calculate pairwise alignments and then cluster genes that are aligned above a minimum sequence identity threshold (parameter '–min-seq-id'). We ran this program multiple times with different thresholds (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05) to infer a wide range of possible families. Families from different thresholds can be redundant, so we merged nested families, i.e. if family X contains all genes in family Y, then we keep X and discard Y. We also discarded any family that contains only one gene. In theory, families inferred from a higher threshold (e.g. 0.9) should always

nest within a family inferred from a lower threshold (e.g. 0.05), such that we would discard all families from higher thresholds. But in practice, families don't always nest within each other but only overlap partially. After merging, our final model used the following number of families from each threshold.

Identity Threshold	# of De Novo Families
0.05	720,587
0.15	0
0.1	0
0.2	85,042
0.3	71,282
0.25	70,504
0.4	49,106
0.5	23,965
0.6	31,379
0.7	18,837
0.8	10,331
0.9	9,099

In total, our model used 1,090,132 gene families, which annotated 162,783,114 genes. Note that because these gene families can still overlap with each other, a gene may have multiple annotations. This analysis took advantage of MMseqs2's parallelism, taking 6 hours using 256 CPU cores. We refer to a de novo family as a 'subfamily' if 90% or more of its amino acid sequences are also annotated to a specific COG or Pfam. Note that this definition provides a small tolerance such that a subfamily does not need to be a perfect subset of a COG or Pfam. For the example about Pfam PF10609 (Figures 2.1F and 2.1G), we gathered the 253 amino acid sequences annotated to PF10609 and the subfamily mmseqs_5_1535552. We also gathered the 1,391 sequences annotated to PF10609 and the subfamily mmseqs_70_40217271. We collapsed 100% identical sequences to yield a total collection of 142 and 310 sequences from mmseqs_5_1535552 and mmseqs_70_40217271, respectively. We aligned all of these sequences together using muscle v3.8.1551 (default parameters) [Edgar, 2004a] and then constructed a maximum likelihood phylogenetic tree using

IQ-TREEv2.1.2 (parameters -m TEST -bb 1000 -alrt 1000 -T AUTO)(Minh et al. 2020). We then rooted the tree using the midpoint method.

2.6.3 Subtypes and slicing of reference sequences

To group reference sequences into subtypes, we used mash v2.2.2 [Ondov et al., 2016] (command 'mash dist', sketch size 100000, kmer size 21) to calculate a distance score of 0 to 1 between every pair of sequences. Next, we created an undirected graph, where sequences are nodes and sequences are connected if their distance is less than 0.1. We defined a 'subtype' as one of the 7,326 connected components in the graph. 3,935 subtypes contained only plasmids; 3,355 subtypes contained only chromosomes; and 36 subtypes contained both plasmids and chromosomes. We sliced reference sequences into 10kb slices by sliding a 10kb window at 5kb increments. The first window starts at the beginning of the sequence, and the final window stops at the end of the sequence. For instance, a 23kb sequence would be sliced at 0-10kb, 5-15kb, 10-20kb, and 13-23kb. A slice was annotated with any gene that was entirely or partly inside the slice. In total, we generated 10,453,279 slices from the reference chromosomes and 343,246 slices from the reference plasmids.

Assessing model performance in cross-validation. To perform cross-validation, we randomly divided the 10kb slices into four groups. For each fold of cross-validation, three groups formed the training data, and the fourth group formed the testing data. In a naive split, we keep all slices from the same reference sequence together in either training or testing data. In an informed split, we keep all slices from the same subtype together. We assigned weights to the 10kb slices when calculating precision and recall performance (Figure 2.1F and 2.1D). Consider the following notation to represent sequences:

$$S_i = sequence\ i$$

$$P_u = the\ set\ of\ plasmid\ sequences\ in\ subtype\ u$$

$$C_u = the\ set\ of\ chromosome\ sequences\ in\ subtype\ u$$

$$D_i^k = window\ slice\ k\ of\ sequence\ i$$

And consider the following notation to represent weights:

$$w(D_i^k) = weight\ of\ window\ slice\ k\ of\ sequence\ i$$

$$w(S_i) = \sum_k w(D_i^k) = weight\ of\ sequence\ i$$

$$w(P_u) = \sum_{S_i \in P_u} w(S_i) = weight\ of\ plasmid\ sequences\ in\ subtype\ u$$

$$w(C_u) = \sum_{S_i \in C_u} w(S_i) = weight\ of\ chromosome\ sequences\ in\ subtype\ u$$

We defined two different scenarios for assigning weights. Scenario A satisfies the following conditions:

1.) All slices from the same sequence have the same weight

$$w(D_i^s) = w(D_i^t) \quad \forall s, t$$

2.) The weight of every sequence is equal to 1

$$w(S_i) = 1 \quad \forall i$$

Scenario B satisfies the following conditions:

1.) All slices from the same sequence have the same weight

$$w(D_i^s) = w(D_i^t) \quad \forall s, t$$

2.) All plasmid (or chromosome) sequences in the same subtype have equal weight

$$w(S_i) = w(S_j) \quad \forall\ i, j, u\ where\ S_i \in P_u\ and\ S_j \in P_u$$

$$w(S_i) = w(S_j) \quad \forall\ i, j, u\ where\ S_i \in C_u\ and\ S_j \in C_u$$

3.) All subtypes have equal weight

$$w(P_u) = w(P_v) \quad \forall u, v$$

$$w(C_u) = w(C_v) \quad \forall u, v$$

4.) The sum of weights across all slices equals the total number of slices

$$\sum_u w(P_u) = total\ number\ of\ plasmid\ slices\ (i.e.\ 343,246)$$

$$\sum_u w(C_u) = total\ number\ of\ chromosome\ slices\ (i.e.\ 10,453,279)$$

Each scenario implies a unique assignment of weight values. Scenario A requires that every sequence has the same weight. Importantly, this ensures that long sequences, which have disproportionately more slices, have equal weight as shorter sequences. Scenario B further requires that every subtype has the same weight. Importantly, this ensures that subtypes that contain a disproportionately large number of sequences (e.g. subtypes that represent commonly studied bacteria, such as *Escherichia*, *Salmonella*, and *Klebsiella*) have equal weight as subtypes with fewer sequences. We evaluated performance under two different cross-validation and weighting scenarios. Figure 2.1F shows the result of training models using a 'naive' cross-validation split and calculating precision/recall using weights from Scenario A. Figure 2.1D shows the results of training models using an 'informed' cross-validation split and calculating precision/recall using weights from Scenario B. We calculated precision/recall using the function sklearn.metrics.precision_recall_curve from the scikit-learn Python package [Pedregosa et al., 2011], with the parameter sample_weight set to the weights of the slices. We calculated AUCPR with the function sklearn.metrics.average_precision_score.

2.6.4 PlasX implementation

We implemented PlasX as a logistic regression using the SGDClassifier class from scikit-learn [Pedregosa et al., 2011]. Regardless of how we evaluated PlasX, we always trained it with weights defined by Scenario B and based on only slices in the training data. To implement elastic net regularization, we performed a grid search of hyperparameters, with the regularization parameter alpha ranging from 10^{-8} to 10^{-3} in multiplicative increments of 10 and the parameter l1_ratio being 0.0, 0.25, 0.5, 0.75, or 1.0. For each evaluation scenario, we selected the hyperparameters that produced the best performance. We used the best hyperparameters from the 'informed' cross-validation and the weights defined by Scenario B (alpha=3.16×10^{-6}, l1_ratio=0.0) to retrain PlasX on all 10kb slices and create the final model that we used to predict plasmids from metagenomes.

2.6.5 Execution of other plasmid prediction tools

We downloaded PlasClass [Pellow et al., 2020] from https://github.com/Shamir-Lab/PlasClass (v0.1.0-2-gb80a4f4). We downloaded PPR-Meta [Fang et al., 2019] from https://github.com/zhenchengfang/PPR-Meta (v1.0-14-gab99c91). We downloaded Platon [Schwengers et al., 2020] from https://github.com/oschwengers/platon, and then modified the code to more efficiently parallelize across many CPUs (modifications at https://github.com/michaelkyu/platon). We used Platon's RDS score as its final prediction score, ignoring whether it found other features like conjugation and replication genes.

To ensure a fair comparison of models in cross-validation (Figure 2.1F and 2.1D), we retrained PlasClass and Platon using the same training sequences as we used for PlasX in each cross-validation fold. We trained PlasClass on 10kb slices, and we trained Platon on the entire non-sliced sequences. We did not train PlasClass and Platon with sequence weights because they don't take in weights as input, but we did calculate precision and recall with weights. PPR-Meta [Fang et al., 2019] and Deeplasmid [Andreopoulos et al., 2022] do not provide software interfaces for retraining new models, so we ran the pretrained versions of these models that were published in their original studies (and thus were trained on different sequence datasets).

We downloaded the four sequence versions of the Wolbachia plasmid pWCP from https://doi.org/10.6084/m9.figshare.6380015 (Table 2.8). We made predictions of pWCP using the original pretrained and published versions of PlasClass, Platon, PPR-Meta, and Deeplasmid, and we ran the final PlasX model that was trained on all 10kb slices. We downloaded the collection of all ICE sequences (n=552) from ICEberg [Liu et al., 2019] 2.0 at https://db-mml.sjtu.edu.cn/ICEberg/ on September 30, 2022. We also downloaded 455 prophage sequences from the NCBI Virus data portal

(https://www.ncbi.nlm.nih.gov/labs/virus) on September 30, 2022. To download them, we selected the "Bacteriophages" subset from the ">Find Data" menu bar, and then we applied filters of "Only" for the "Provirus" option and "complete" for the "Nucleotide Completeness" option. We made predictions of these ICE's and virus sequences using the original pretrained and published version of Platon, using Platon's default 'accuracy' mode (Tables S9 and S10). We also ran the final PlasX model that was trained on all 10kb slices.

We ran Deeplasmid using the Docker image of the CPU implementation, following instructions at https://github.com/wandreopoulos/deeplasmid (version sha256:10809927e2c8a14cf86231801b804b0bd4bddf600821d17fd8b7e41a15c562c0). While we were able to run Deeplasmid on the Wolbachia plasmid pWCP, it was prohibitively slow to run on the entire set of 10kb slices used for cross-validation evaluation. In particular, we found that Deeplasmid running on a MacOS laptop takes 3 hours for 1,000 slices, so we estimated it would take 3.7 years to run on all slices. While the GPU implementation of Deeplasmid might be able to run faster, we were unable to execute its prebuilt Docker image (version sha256:f3a22993fb765a7f9678b174245b64976e7e52a4dce85570060900b794af5e43). We suspect that this image is incompatible with modern machine setups, like ours, because Deeplasmid depends on software that is several years old. For example, it requires the CNTK library, for which development was abandoned over 3 years ago (https://docs.microsoft.com/en-us/cognitive-toolkit/releasenotes/cntk_2_7_release_notes). We were also unable to build a new Docker image to run the GPU implementation, despite attempts to modify the Docker build file (see the issue we raised at https://github.com/wandreopoulos/deeplasmid/issues/3).

2.6.6 Predicting plasmids from metagenomic assemblies

We downloaded fastq files for 1,782 short-read and paired-end metagenomes from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) using the program 'fastq-dump'. The countries represented are Austria [Feng et al., 2015], Aus-

tralia (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6092), Bangladesh [David et al., 2015], Canada [Raymond et al., 2016], China [Qin et al., 2010, Wen et al., 2017], Denmark [Le Chatelier et al., 2013], England [Xie et al., 2016], Ethiopia [Pasolli et al., 2019], Fiji [Brito et al., 2016], Israel [Zeevi et al., 2015], Italy [Rampelli et al., 2015], Madagascar [Pasolli et al., 2019], Mongolia [Liu et al., 2016], Spain [Li et al., 2014], Tanzania [Rampelli et al., 2015] and USA [Turnbaugh et al., 2007, Obregon-Tito et al., 2015]. Some samples were sequenced multiple times (i.e. multiple records in SRA), in which case we concatenated the fastq files together. We have separated these multiple accessions using the delimiter '|'. We labeled Tanzania, Ethiopia, Bangladesh, Madagascar, and Fiji as non-industrialized and the other countries as industrialized. All steps of quality filtering, metagenomic assembly, read recruitment and profiling were automated using snakemake [Köster and Rahmann, 2012] workflows in anvi'o [Shaiber and Murat Eren, 2018]. The 'illumina-utils' [Murat Eren et al., 2013] commands 'iu-gen-configs' and 'iu-filter-quality-minoche' with the flag '–ignore-deflines' were used to quality filter the raw paired-end reads. Each metagenome was assembled individually using IDBA_UD [Peng et al., 2012] with default settings, except the flag '–min_contig 1000'. We annotated COGs and Pfams in all assembled contigs using the same procedure as the reference plasmids and chromosomes. To annotate de novo families, we first used 'mmseqs result2profile' (default parameters) to represent the sequence conservation in each de novo family as a profile. We then used 'mmseqs search' (default parameters) to search for profiles across all genes. We kept hits where the alignment coverage was greater than 80% of both the gene and the profile and where the alignment identity was at least greater than X-0.05 where X is the minimum identity threshold used to originally construct the family (parameter –min-seq-id). For example, if a family was constructed using an identity threshold of 0.8, then we kept hits with an identity greater than 0.75. Using these gene annotations, we ran PlasX to assign a score to every contig. We kept contigs intact, rather than slicing them into 10kb windows. Contigs with score >0.5 were classified as plasmids.

2.6.7 Detection and circularity of plasmids across metagenomes

We recruited short reads from our collection of metagenomes using Bowtie2 v.2.0.5 [Langmead and Salzberg, 2012]. We used the snakemake workflows in anvi'o to automate execution of bowtie and post-processing to calculate 'detection', i.e. the proportion of a sequence that is covered by at least one read. We ran bowtie2 using the three following combinations of parameters and input files.

First, to identify circular contigs, we recruited each metagenome's reads to a fasta file that contained only the contigs assembled from that metagenome. For computational efficiency, we ran bowtie2 with its default behavior to align every read at most once. We then analyzed the orientation of paired-end reads (Figure 2.3B). During assembly, circular sequences are broken by an artificial breakpoint to represent them as linear contigs. Consequently, DNA sequencing that occurred across this breakpoint will produce paired-end reads that align in a reverse-forward orientation to the ends of the contig. In contrast, if a sequence is not circular, then all paired-end reads are expected to align in a forward-reverse orientation. To illustrate this intuition, suppose the upstream read of a paired-end maps to positions 200-300 of a contig and the downstream read maps to 500-600. If the upstream read maps with a reverse complement strandedness (i.e. 'reverse') and the downstream read maps with the same strandedness as the way the contig is written (i.e. 'forward'), then the paired-end is in a reverse-forward orientation. In other words, if the contig is written 5'-to-3', then the upstream read maps 3'-to-5' and the downstream read maps 5'-to-3'. Inversely, the paired-end is in a forward-reverse orientation if the upstream read maps 5'-to-3' and the downstream read maps 3'-to-5'. Next, we defined the gap (or insert) size of a paired-end to be the distance between the closest (or farthest) aligned positions between its two reads. In our example, the gap size is 600-200=400 and the insert size is 500-300=200. Let D be the contig's length minus three times the median insert size of all forward-reverse paired-ends that are aligned to the contig. Finally, we label a contig as circular if (1) its detection was

greater than 0.95 and (2) it had at least one reverse-forward paired-end with a gap greater than or equal to D. This approach of examining reverse-forward paired-ends was inspired by [Jørgensen et al., 2014].

Second, to study the ecological distribution of all plasmids and plasmid systems at the same time, we recruited each metagenome's reads to a fasta file that contained either the non-redundant set of 68,350 predicted plasmids or the non-redundant set of 11,121 reference plasmids. For computational efficiency, every read was aligned at most once (i.e. the default behavior of bowtie). We designated a plasmid as being present in a metagenome if its detection was greater than 0.95. To compare metagenomes based on their plasmid content in Figure 2.9 and Figure 2.10, we ran UMAP v0.5.1 [McInnes et al., 2018] with parameters 'n_neighbors=30, n_components=2, min_dist=0.15, metric='jaccard', random_state=1'. The heatmaps in Figure 2.9 were generated using the 'heatmap.2' package in R, with agglomerative clustering using median linkage on Euclidean distances.

Third, to study the specific plasmids from PS974 and PS1110 that were shown in Figure 2.18F and Figure 2.20 (see Table 2.5 for contig names), we ran bowtie2 on each sequence separately. This setup allowed every read to align to potentially multiple sequences, resulting in a more complete estimation of which metagenomes contained a plasmid. For the backbone sequences of these systems, we designated them as present in a metagenome if their detection was greater than 0.95. For compound plasmids in PS1110 that encoded a chloramphenicol resistance gene, we designated them as present in a metagenome if they satisfied an additional criterion that greater than 0.95 of the resistance gene was covered by at least one read.

2.6.8 Estimation of microbial taxonomy in metagenomes

We estimated taxonomic abundances in every metagenome by running kraken2 [Wood et al., 2019] v2.1.2 with its standard database (https://github.com/DerrickWood/kraken2) and then refined the abundances using bracken [Lu et al., 2017] v2.5

(https://github.com/jenniferlu717/Bracken) with database parameters '-k 35 -l 96'. We ran bracken (parameter '-r 96') a separate time for every taxonomic rank: S1 (subspecies/strain), S (species), G (genus), F (family), O (order), C (class), P (phylum), D (domain). The output of this analysis is a count of how many reads originated from each taxon. To compare the metagenomic presence/absence of plasmids versus taxa, we calculated M_P the set of metagenomes where a plasmid P is detected at greater than 95%, and M^r_T the set of metagenomes in which at least r reads originated from taxon T. For each plasmid, we attempted to find the best explanation of its ecological distribution by comparing the plasmid to every taxon using the Jaccard index, and by scanning many possible read thresholds. More exactly, we used the following formula to represent the best possible explanation of a plasmid's ecological distribution:

$$\max_T \max_r Jaccard(P, T; r)$$

where

$$Jaccard(P, T; r) = \frac{|M_P \cap M_T^r|}{|M_P \cup M_T^r|}$$

We evaluated 29 values for the threshold r, ranging from 1 read to 10 million reads in multiplicative increments of 101/4. We ignored plasmids that were present in less than 5 metagenomes, i.e. $|MP|<5$, because it was likely that these plasmids would have a high Jaccard similarity to some taxon by chance. For instance, we observed that many pairs of plasmids and taxa occur in exactly one and the same metagenome, and thus they have a Jaccard index of 1. To compare continuous-valued abundances, we defined a plasmid's abundance in a metagenome as the sum of coverage values across all sequence positions divided by sequence length, and we defined a taxon's abundance as the number of reads originating from the taxon. If a plasmid had detection of $<95\%$, then we set its abundance to 0. If a taxon had less than 1000 reads, then we set its abundance to 0. We ignored plasmids and taxa that had non-zero abundances in less than 5 metagenomes. For every pair of plasmid and taxon, we estimated the Pearson correlation between their abundance levels across metagenomes

using FastSpar [Watts et al., 2018] v1.0.0 (https://github.com/scwatts/fastspar), which is an improved implementation of the SparCC [Friedman and Alm, 2012]. This method accounts for the compositional nature of the data—in which abundances reflect relative instead of absolute quantities—by assuming that the amount of correlations in a data is sparse. We ran FastSpar on the non-redundant set of predicted plasmids, and ran it separately on the non-redundant set of reference plasmids.

2.6.9 Additional validation and annotations of plasmids

To determine if a predicted plasmid has canonical plasmid features, we ran MOB-suite [Robertson and Nash, 2018]. This tool searches a sequence for known examples of four types of features: plasmid replicon (e.g. replication genes), relaxase, mating pair formation, and origin of transfer. We installed MOB-suite v3.0.1 using pip, in an Anaconda Python environment that has mash v2.2. We ran the MOB_typer subroutine (command 'mob_typer') using default parameters and followed the execution instructions at https://github.com/phac-nml/mob-suite. To determine if a predictive plasmid is a novel sequence, plasmids were blasted against NCBI using the blast package (v2.9.0, installed from bioconda Anaconda repository). On October 13, 2021, we downloaded version 5 of the NCBI databases non-redundant nucleotide (nt), ref_prok_rep_genomes, ref_viroids_rep_genomes, and ref_viruses_rep_genomes, and then integrated them into a single database using the 'blastdb_aliastool' command. We then searched every predicted plasmids against this combined database, using the 'blastn' tool with the '-task megablast' parameter for efficient searching. For each plasmid, we examined all matching NCBI sequences (called 'subjects') and chose the one with the highest 'qcovs' (query coverage per subject), which represents the fraction of the plasmid sequence that is covered by all high-scoring segment pairs (HSP). Tiebreaking was done by sorting subjects by the maximum bitscore of the HSPs. If the qcovs of the best matching sequence was greater than 90%, then we considered the predicted

plasmid as found in NCBI and further categorized the matching sequence by searching for the keywords 'plasmid', 'virus', 'chromosome' (in that order, disregarding capitalization) in its NCBI description. For example, if the description of the matching sequence contained the word 'plasmid', then we said the predicted plasmid matched a known plasmid on NCBI. Similarly, if the description contained 'chromosome' but not 'plasmid' nor 'virus', then we said that the predicted plasmid matched a known chromosome on NCBI. If the qcovs of the best sequence was <90%, then we labeled the predicted plasmid as not found in NCBI. To visualize pFIJ1037_1 (Figure 2.7A) and pENG0187_1 (Figure 2.8A), we manually imported COG functions into the plasmid maps produced by snapgene (Insightful Science; snapgene.com). We manually curated functions in genes without COGs using NCBI BLASTx.

2.6.10 Keyword analysis of COGs and Pfams for plasmid functions

We labeled COGs and Pfams as being a plasmid-associated function (Figure 2.1E) if its database description contains any of following keywords as a substring: 'plasmid', 'toxin', 'replicat', 'integrase', 'transpos', 'recombinase', 'resolvase', 'relaxase', 'recombination', 'partitioning', 'mobilis', 'mobiliz', 'type iv', 'conjugal', 'conjugat', 'segregat', 'MobA', 'ParA', 'ParB', 'BcsQ'. We labeled backbone and cargo genes as being related to plasmid replication, transfer, or maintenance if they were annotated to any plasmid-associated COG or Pfam (see section "Classification of cargo and backbone genes").

To determine if a predicted plasmid is 'keyword-recognizable' (Figure 2.3C), we searched the plasmids for COGs and Pfams using a more restricted set of keywords (just "plasmid" and "conjugation") instead of the keywords above.

2.6.11 MobMess algorithm to dereplicate plasmids, remove assembly fragments, and discover plasmid systems (see Figure 2.13)

The MobMess algorithm performs three tasks. It de-replicates plasmids that are nearly redundant to each other; it removes plasmids that appear to be assembly fragments; and finally it organizes plasmids together into evolutionary groups called plasmid systems. MobMess consists of several steps described below.

MobMess first performs an all-vs-all pairwise alignment of sequences using the MUMmer alignment package (v4.0.0rc1) [Jain et al., 2018]. All sequences are placed into a single fasta file and then aligned with 'nucmer' (parameters '–maxmatch –minmatch=16') to calculate local alignment blocks. Alignments are specified asymmetrically such that one sequence is designated as the query q and the other is the reference r. For every q and r, the alignment blocks calculated by 'nucmer' are written to a separate file, and then a subset of blocks is identified using 'delta-filter' (parameters '-q -r') to create a one-to-one alignment.

Next, MobMess constructs a directed graph G where vertices are sequences and edges represent the containment of one sequence within another (Figure 2.13A). Formally, consider a query q and reference r. Let |q| be the length of q. For the ith alignment block between q and r, let si, ei, and i be the start position in q, end position in q, and number of alignment mismatches and indels, respectively. The following values summarize the information across all alignment blocks between q and r.

Sum of block lengths $L = \sum_i e^i - s^i$

Number of mismatches and indels $E = \sum_i \delta^i$

Proportion of query positions covered $C = \sum_{j=1}^{|q|} \begin{cases} 1/|q| & \text{if } \exists i \text{ such that } s^i \leq j \leq e^i \\ 0 & \text{otherwise} \end{cases}$

Local sequence identity $I_{local} = (L - E)/L$

Global sequence identity $I_{global} = I_{local} * C$

MobMess creates a directed edge (q,r) in G if I_{local} and C are above user-specified thresholds. In this study, we applied thresholds of I_{local} greater than 0.9 and C greater than 0.9. In Figure 2.14, we re-ran MobMess using various thresholds on I_{local} and C (the same threshold was applied to I_{local} and C at the same time). MobMess clusters sequences according to strongly connected components in G, calculated with igraph v0.8.2 [Csardi et al., 2006] in Python. That is, two sequences x and y are placed in the same cluster if there exists a directed path from x to y and another from y to x in G. Intuitively, a cluster represents a set of sequences that are nearly identical to each other across nearly their entire lengths. MobMess then reduces G to another graph H, called the condensation graph, by contracting every cluster of sequences into a single vertex. A directed edge (u,v) exists in H if and only if there are sequences x $\in$ u and y $\in$ v where edge (x,y) exists in G. Note that H does not have any cycles. As proof by contradiction, if there were a cycle of clusters, then those clusters would have been in the same strongly connected component in G and hence would have been merged into a single, larger cluster. MobMess labels every cluster in H as one of the three following types: (1) a 'backbone cluster' if it has an outgoing edge and at least one of its member sequences is circular, (2) a 'fragment cluster' if it has an outgoing edge but none of its member sequences are circular, or (3) a 'maximal cluster' if it does not have any outgoing edges. Intuitively, a maximal cluster represents the longest version of a plasmid observed in the data. In contrast, a backbone or fragment cluster represents a set of plasmids that

are subsequences of other plasmids in a maximal cluster. The only difference between backbone and fragment clusters is that backbone clusters contain at least one circular plasmid (implying complete assembly), while fragment clusters do not contain any circular plasmids (suggesting they are assembly fragments of the maximal cluster). To dereplicate sequences, MobMess discards all fragment clusters and then chooses a representative sequence from every maximal and backbone cluster. A cluster's representative is the sequence with the highest global sequence identity (I_{global}), averaged across the set of alignments where that sequence is the reference and other sequences in the same cluster are the queries.

MobMess defines a plasmid system as a specific backbone cluster together with its 'compound' clusters, which are the set of non-fragment clusters connected to the backbone in H. Thus, there is a one-to-one correspondence between backbone clusters and plasmid systems. Note that systems can be nested within each other, because backbone clusters can be connected to each other in H. Thus, a backbone cluster can be the backbone that forms a given plasmid system, and at the same time, it can also be a compound cluster with respect to an even smaller backbone that forms a different system. As another note, a maximal cluster can be a 'compound' cluster of a system, but it is also possible that some maximal clusters are not found in any system because they are not connected to any backbone clusters in H.

We ran MobMess to analyze the 226,194 predicted plasmid contigs. MobMess grouped the contigs into a total of 132,616 clusters. 64,266 clusters were 'fragment clusters' that contained 125,475 contigs, which we interpreted as assembly fragments of other predicted plasmids. We discarded these fragments from further analysis. The other 68,350 clusters were non-fragment clusters (i.e. 1,169 backbone and 67,181 maximal clusters) and contained 100,719 contigs, which we further analyzed for the existence of orthogonal support for being plasmids (Figure 2.3C). Finally, MobMess identified 1,169 plasmid systems, which together represent 1,169 backbone and 63,926 maximal clusters (3,255 maximal clusters were excluded). See Figure 2.5 for a diagram of these numbers.

We ran MobMess separately on the 16,827 reference plasmid sequences, yielding 11,121 clusters. We assumed that all reference plasmids were circular, and thus there were no fragment clusters. We visualized networks with Cytoscape [Shannon et al., 2003] v3.8 and laid nodes out using the prefuse directed force layout [Heer et al., 2005]. While we have focused on plasmids, MobMess could be applied to dereplicate and organize other mobile genetic elements into systems.

2.6.12 Classification of cargo and backbone genes

We classified all genes on the backbone plasmids of a plasmid system as backbone genes. For genes on compound plasmids, we tested whether the genes shared any de novo family annotations with the genes on the backbone plasmids. If so, we classified those genes as backbone genes, otherwise as cargo genes. For this analysis, we used the 1,090,132 de novo families that we constructed from reference plasmids and chromosomes in order to train PlasX, and we also used an additional set of 439,584 de novo families that we constructed by running the command MMseqs2 (–min-seq-id 0.05) on the genes from all plasmid sequences in this study (16,827 reference and 226,194 predicted plasmids). These additional families allowed us to capture gene families that might be absent in reference sequences but are conserved in predicted plasmids. Note that the classification of genes as backbone or cargo depends on which plasmid system is being considered. It is possible for a gene to be classified as a backbone gene with respect to one plasmid system and, at the same time, as a cargo gene with respect to another system. This is because a plasmid can be a backbone plasmid of a system and also be a compound plasmid of a different system (see Methods subsection on MobMess algorithm).

For every non-redundant compound plasmid in the system, we calculated the fraction of genes in the plasmid that were cargo genes. We then averaged this fraction across all non-redundant compound plasmids in the system to define the "cargo gene percentage" of

the system (Figure 2.17). Because every gene is either backbone or cargo, the percentage of backbone genes is 100% minus the cargo gene percentage.

For Figure 2.12B and to analyze the content of backbone/cargo genes, we used a non-redundant and unambiguous set of 8,995 backbone and 24,168 cargo genes. To derive these sets of genes, we first considered the 47,172 genes encoded on the 4,424 non-redundant plasmids that were part of at least one plasmid system. Of these 47,172 genes, we used the 8,995 genes that were classified as backbone genes because they were encoded on a backbone plasmid and that were never classified as cargo genes in any plasmid system. 24.1% (2,169/8,995) of these genes had a plasmid-associated keyword in their COG/Pfam annotations (see Methods section "Keyword analysis of COGs and Pfams for plasmid functions"). We also used the 24,168 genes that were always classified as cargo genes and never backbone genes in any plasmid system. 13.4% (3,229/24,168) of these genes had a plasmid-associated keyword. We excluded from analysis the 1,917 genes that were sometimes classified as backbone genes and other times cargo genes, depending on the system. We also excluded 12,092 genes that were on compound plasmids but were classified as backbone genes, as these genes are redundant with the backbone genes that were encoded on the backbone plasmid.

2.6.13 Mutual exclusivity of plasmid systems

To measure the extent of mutual exclusivity in a plasmid system S, we defined two sets of metagenomes. M_S^{any} is the set of metagenomes that contains S, i.e. where one or more plasmids in S has detection at greater than 0.95. M_S^{solo} is the set of metagenomes where one and only one plasmid in S has detection at greater than 0.95. Then, we calculated the mutual exclusivity score E_S $|M_S^{solo}|$ / $|M_S^{any}|$. This score is similar to a test statistic used by methods, such as CoMEt [Leiserson et al., 2015], for studying mutual exclusivity of gene alterations in cancer.

We observed that if a compound plasmid was present in a metagenome (detection greater

than 0.95), then its backbone plasmid was often also present. This could be due to (1) both plasmids being present as separate entities in the same microbiome, or (2) a read recruitment artifact arising from the sequence similarity between the plasmids. To minimize artifacts, we assumed the second scenario is always happening, and we corrected for it by assuming that a backbone plasmid is absent from a metagenome (regardless of its detection) whenever any compound plasmids are present in the metagenome. To formalize this correction procedure, let P_m be the set of plasmids in system S with detection greater than 0.95 in metagenome m. We created the induced subgraph H_m that is formed by subsetting the vertices of P_m in the plasmid similarity graph H (see MobMess section of the Methods). Then, we defined P_m $\subseteq$ P_m as the subset of plasmids that are maximal with respect to H_m (i.e. they don't have outgoing edges in H_m). We used $P_{m'}$, rather than P_m, in order to calculate M_s^{any}, M_s^{solo}, and E_s.

2.6.14 Identification of antibiotic resistance genes

We annotated antibiotic resistance genes using two databases. First, we searched against a database of resistance protein family HMMs from Resfams [Gibson et al., 2015] (v1.2, dated 2015-01-27, 'Core' database at http://www.dantaslab.org/resfams). We used 'anvi-run-hmms' from anvi'o [Eren et al., 2021] to automate running 'hmmsearch' from HMMER(Eddy 2011) 3.3.2 and apply an e-value cutoff of 10-10. Second, we ran rgi (v5.2.0, https://github.com/arpcard/rgi) to search for similarity in the CARD database of resistance genes [Alcock et al., 2020]. We removed CARD hits that were labeled as 'Loose' and kept those labeled as 'Perfect' or 'Strict'. We removed any Resfams or CARD hits that contained the keywords 'transcription', 'regulat', 'modulat' in their database description, to avoid cases (e.g. TetR protein) where the hit is a gene that regulates the expression of another resistance gene but doesn't itself perform the molecular process that confers resistance. We categorized hits into major antibiotic resistance classes by searching for the following keywords in

their functional descriptions: 'lincosamide', 'macrolide', 'erythromycin', 'chloramphenicol', 'aminoglycoside', 'streptothricin', 'glycopeptide', 'efflux pump', 'beta-lactamase', 'nitroimidazole', 'tetraycyline', 'quinolone', 'sulfonamide'. Additionally, we searched the extra keywords 'Van' and 'VanZ' to identify glycopeptide resistance; 'efflux', 'permease', and 'pump' to identify efflux pumps; and 'TetX' to identify tetracycline resistance.

2.6.15 High molecular weight (HMW) DNA extraction, long-read sequencing, and determination of circularity through long-reads

We employed a long-read sequencing strategy on two *Bacteroides fragilis* cultivars from two patients (p-214 and n-216 previously described in Vineis et al(Vineis et al. 2016)). We extracted total genomic HMW DNA by one of two methods. For *B. fragilis* p-214, we used the Qiagen Genomic Tip 20/G procedure (also known as Method #4/GT) as previously described(Trigodet, Lolans, and Fogarty 2022) on a 10 mL overnight BHIS broth culture. For *B. fragilis* n-216, we used a Phenol Chloroform protocol on 25 mL overnight BHIS broth cultures. Libraries were prepared with the Rapid Barcoding Kit (SQK-RBK004) and the standard protocols from Oxford Nanopore Technologies (ONT) with a few modifications. For *B. fragilis* p-214, DNA fragmentation was performed on 6 ug DNA using 5 passes through a 22G needle in a 30 µL volume. The gDNA input was 1.5 µg (Table 2.6), based on sample availability in a 7.5 µL volume, with 2.5 µL Fragmentation mix added. We sequenced for 72 hours using a single R9.4/FLO-MIN106 flow cell (ONT). For *B. fragilis* n-216, DNA fragmentation was performed on 10 ug DNA using 10 passes through a 22G needle in a 250 µL volume. The gDNA input was 0.32 to 0.44 µg, based on sample availability in an 8.5 µL volume, with 1.5 µL Fragmentation mix was added per sample. We sequenced for 72 hours using a single R9.4/FLO-MIN106 flow cell. We used Guppy (v4.0.15) for all post-run base calling, sample de-multiplexing and the conversion of raw FAST5 to FASTQ files.

To determine circularity, we used BLAST to align the long reads with a minimum quality

score of 7 to our predicted plasmid sequences. During assembly, all DNA short reads are assembled as linear sequences even if they are circular elements. Circular elements have an artificial breakpoint to represent them as linear sequences, and this breakpoint can happen anywhere on the sequence depending on the assembly method. We tested for the presence of an artificially introduced breakpoint by aligning 500 long reads and then visualizing these alignments on the sequence as if it were assumed to be a circular element (Figure 2.8). If indeed the sequence is circular, the long reads would overlap each other and "wrap around" the entire circumference of the sequence. In other words, all nucleotide positions of the sequence would be covered by at least one read and there would also exist a read that spans the breakpoint by aligning to both sides of the breakpoint. This property ensures the breakpoint is artificial, and hence the sequence is a circular element. Inversely, this property does not hold when the breakpoint is not artificial (i.e. the sequence is actually an assembly fragment or linear element).

2.6.16 Transfer of predicted plasmid between microbial populations

In duplicate, we streaked *B. fragilis* 214 (donor, erythromycin resistant due to pFIJ0137_1) and *B. fragilis* 638R (recipient, rifampicin resistant) onto plates with brain-heart infusion agar supplemented with hemin and vitamin K (BHIS) and incubated them in 5 mL BHIS media anaerobically at 37℃ for 20 hours. To mate the donor to the recipient, 250 μL of donor cells were pelleted in a centrifuge at 5,000x gravity. We discarded the supernatant and resuspended the donor in 1 mL of the recipient culture. Again, cells were pelleted at 5,000x gravity, then resuspended in 25 μL of BHIS media. Cells were spotted onto BHIS agar plates and incubated anaerobically for 24 hours. After incubation, cells were resuspended in 1 mL BHIS. 250 μL of this suspension was plated onto BHIS plates containing 8 μg/mL rifampicin and 25 μg/mL erythromycin to select for *B. fragilis* 638R recipients of pFIJ0137_1. Duplicate plates each had approximately 300 colonies. Plating the donor or

recipient alone resulted in zero colonies, confirming the transformants were not spontaneous mutants to either antibiotic. Two transformant colonies were restreaked onto fresh BHIS plates containing 8 μg/mL rifampicin and 25 μg/mL erythromycin.

2.6.17 Short-read sequencing of isolate genomes and confirmation of plasmid transfer

We grew 20-hour cultures of *B. fragilis* 214 donor, naive *B. fragilis* 638R, and *B. fragilis* 638R transconjugants containing pFIJ0137_1. Using the QIAseq FX DNA library kit (Qiagen), libraries of these strains were prepared with 100 ng of genomic DNA. DNA was fragmented enzymatically into smaller fragments and desired insert size was achieved by adjusting fragmentation conditions. Fragmented DNA was end repaired and 'A's were added to the 3' ends to stage inserts for ligation. During the ligation step, Illumina compatible Unique Dual Index (UDI) adapters were added to the inserts and the prepared library was PCR amplified. Amplified libraries were cleaned up, and QC was performed using a tapestation. Libraries were sequenced on Illumina MiSeq platform using v2 cassette to generate 2x250bp reads. To confirm the transfer of pFIJ0137_1, we individually recruited reads from the *B. fragilis* 214 donor, naive *B. fragilis* 638R, and *B. fragilis* 638R transconjugants to the pFIJ0137_1 reference sequence. We used anvi'o to create contigs and profile databases (as described above) and visualized these results with the command 'anvi-interactive'. We independently confirmed the presence of pFIJ0137_1 by assembling genomes using SPAdes [Bankevich et al., 2012] with default parameters.

2.7 Supplementary Tables

All supplementary tables were submitted as a separate supplemental file in ProQuest.

Table 2.1: Summary of reference plasmids and chromosomes

Table 2.2: Names, accession numbers, and metadata of metagenomes

Table 2.3: Summary of predicted plasmids (model scores, orthogonal support, circularity, and NCBI blast results)

Table 2.4: Summary of plasmid systems

Table 2.5: Gene sequences for plasmid systems shown in Figure 2.12D, Figure 2.20 and Figure 2.18F

Table 2.6: DNA extraction and sequencing parameters for long read sequencing of isolate genomes

Table 2.7: COGs and Pfams ranked by their PlasX coefficients

Table 2.8: Prediction of a Wolbachia plasmid

Table 2.9: Prediction of ICEs as plasmids by PlasX and Platon

Table 2.10: Prediction of prophages as plasmids by PlasX and Platon

Table 2.11: Prediction of plasmids in the latest version of PLSDB (2020_06_23_v2) by PlasX

Table 2.12: The length and percent circularity of plasmids that are part of a system versus plasmids that are not part of any system.

2.7.1 Data availability

Reproducible Analyses of reference plasmids and chromosomes are available at doi:10.5281/zenodo.5732024. The PlasX model as well as our analyses of known and predicted plasmids are available at doi:10.5281/zenodo.5843600. For all metagenomes, we have compiled the contigs, taxonomic abundances, and PlasX scores at doi:10.5281/zenodo.5730607, gene calls at doi:10.5281/zenodo.5730987, and gene annotations at doi:10.5281/zenodo.5731658. We have deposited long and short sequencing reads from B. fragilis isolates into the NCBI Sequence Read Archive (PRJNA782184).

All supplementary tables were submitted as a separate supplemental file in ProQuest.

2.7.2 Code availability

We have released two open-source packages, PlasX (https://github.com/michaelkyu/plasx) and MobMess (https://github.com/michaelkyu/mobmess), along with detailed installation and usage instructions.

2.7.3 Acknowledgments

We thank Karen Lolans (ORCiD:0000-0003-1903-756X) for performing the long-read sequencing and for providing feedback on the manuscript. We thank Samuel Miller (0000-0002-2836-1401) and Marcus Foo (0000-0003-3436-1632) for their insights into tRNA modification genes. We also thank other members of the Meren Lab at the University of Chicago for their feedback. MKY acknowledges support from Toyota Technological Institute at Chicago.

CHAPTER 3

A HIGHLY CONSERVED AND GLOBALLY PREVALENT CRYPTIC PLASMID IS AMONG THE MOST NUMEROUS MOBILE GENETIC ELEMENTS IN THE HUMAN GUT

This chapter is derived from the following publication:

Emily C. Fogarty, Matthew S Schechter, Karen Lolans, Madeline L. Sheahan, Iva Veseli, Ryan M Moore, Evan Kiefl, Thomas Moody, Phoebe A Rice1,, Michael K Yu, Mark Mimee, Eugene B Chang, Sandra L Mclellan, Amy D Willis, Laurie E Comstock and A. Murat Eren. 2023. "A highly conserved and globally prevalent cryptic plasmid is among the most numerous mobile genetic elements in the human gut" bioRxiv. https://doi.org/10.1101/2023.03.25.534219.

3.1 Author contributions

ECF and AME conceived the study. KL developed methodology. RM, EK, and AME developed computational analysis tools. ECF, MSS, PAR, ADW and AME performed formal analyses. ECF, KL, MS and LEC conducted investigations. TM, MKY, MM and SLM provided resources. ECF, MSS, ADW and AME curated data. ECF and AME prepared the figures. ECF and AME wrote the paper with critical input from all authors. LEC and AME supervised the project. EBC and AME acquired funding.

3.2 Abstract

Plasmids are extrachromosomal genetic elements that often encode fitness enhancing features. However, many bacteria carry 'cryptic' plasmids that do not confer clear beneficial

functions. We identified one such cryptic plasmid, pBI143, which is ubiquitous across industrialized gut microbiomes, and is 14 times as numerous as crAssphage, currently established as the most abundant genetic element in the human gut. The majority of mutations in pBI143 accumulate in specific positions across thousands of metagenomes, indicating strong purifying selection. pBI143 is monoclonal in most individuals, likely due to the priority effect of the version first acquired, often from one's mother. pBI143 can transfer between Bacteroidales and although it does not appear to impact bacterial host fitness *in vivo*, can transiently acquire additional genetic content. We identified important practical applications of pBI143, including its use in identifying human fecal contamination and its potential as an inexpensive alternative for detecting human colonic inflammatory states.

3.3 Introduction

The tremendous density of microbes in the human gut provides a playground for the contact-dependent transfer of mobile genetic elements [Frost et al., 2005] including plasmids. Plasmids are typically defined as extrachromosomal elements that replicate autonomously from the host chromosome [Frost et al., 2005, Kazlauskas et al., 2019, Solar et al., 1998]. In addition to being a workhorse for molecular biology, plasmids have been extensively studied for their ability to expedite microbial evolution [Garoña and Dagan, 2021] and enhance host fitness by providing properties such as antibiotic resistance, heavy metal resistance, virulence factors, or metabolic functions [Jacob and Hobbs, 1974, Moo-Young et al., 2013, Endo et al., 1995, Thouand and Marks, 2016, Al-Shayeb et al., 2022].

Plasmids have been a major focus of microbiology not only for their biotechnological applications to molecular biology [Leonard et al., 2018, Slattery et al., 2018, Rihn et al., 2021, Salvay et al., 2010] but also for their role in the evolution and dissemination of genes for antibiotic resistance [Mutuku et al., 2022, Dimitriu, 2022], which is a growing global public health concern [Prestinaci et al., 2015]. However, outside the spotlight lie a group

of plasmids that appear to lack genetic functions of interest and that do not contain genes encoding obvious beneficial functions to their hosts [Kang et al., 2020, Oliveira et al., 2021]. Such 'cryptic plasmids' are typically small and multi-copy [Shareck et al., 2004], and are often difficult to study as they lack any measurable phenotypes or selectable markers [Attéré et al., 2017, Challacombe et al., 2017], despite their presence in a broad range of microbial taxa [Roberts, 1989, Zillig et al., 1996, Heuer and Smalla, 2012, Vincent et al., 2021]. In the absence of a clear advantage to their hosts, and the presumably non-zero cost of their maintenance, these plasmids are often described as selfish elements [Thomas, 2014b] or genetic parasites [Iranzo et al., 2016]. While they may provide unknown benefits to their hosts, a high transfer rate could also be a factor that enables cryptic plasmids to counteract the negative selection pressure of their maintenance [Levin and Stewart, 1980, Iranzo et al., 2016, Simonsen, 1991].

Analyses of cryptic plasmids are often performed on monocultured bacteria, limiting insights into the ecology of cryptic plasmids in their host's natural environment. However, recent advances in shotgun metagenomics [Quince et al., 2017] and de novo plasmid prediction algorithms [Andreopoulos et al., 2022, Krawczyk et al., 2018, Zhou and Xu, 2010, Pellow et al., 2020, Carattoli et al., 2014c, Robertson and Nash, 2018, Garcillán-Barcia et al., 2009, Pellow et al., 2021, Yu et al., 2020] offer a powerful means to bridge this gap. For instance, in a recent study we characterized over 68,000 plasmids from the human gut [Yu et al., 2020] and observed that the most prevalent known plasmid across geographically diverse human populations was a cryptic plasmid, called pBI143. Here we conduct an in-depth characterization of this cryptic plasmid through 'omics and experimental approaches to study its genetic diversity, host range, transmission routes, impact on the bacterial host, and associations with health and disease states. Our findings reveal the astonishing success of pBI143 in the human gut, where it occurs in up to 92% of individuals in industrialized countries with copy numbers 14 times higher on average than crAssphage, the most abundant phage

in the human gut. We also demonstrate the potential of pBI143 as a cost-effective biomarker to assess the extent of stress that microbes experience in the human gut, and as a sensitive means to quantify the level of human fecal contamination in environmental samples.

3.4 Results

3.4.1 pBI143 is extremely prevalent across industrialized human gut microbiomes

pBI143 (accession ID U30316.1) is a 2,747 bp circular plasmid first identified in 1985 [Smith et al., 1995] in *Bacteroides fragilis* [Smith, 1985], an important member of the human gut microbiome that is frequently implicated in states of health [Tan et al., 2019, Lee et al., 2018, Ochoa-Repáraz et al., 2010] and disease [Purcell et al., 2017, Haghi et al., 2019]. pBI143 encodes only two annotated genes: a mobilization protein (*mobA*) and a replication protein (*repA*) (Figure 3.1A). Due to the desirable features for cloning such as a high copy number and genetic stability, pBI143 has been primarily used as a component of E. coli-*Bacteroides* shuttle vectors (Smith 1985). The absence of any ecological studies of pBI143 prompted us to characterize it further beginning with a characterization of its genetic diversity.

To comprehensively sample the diversity of pBI143, we screened 2,137 individually assembled human gut metagenomes (Supplementary Table 1) for pBI143-like sequences. By surveying all contigs using the known pBI143 sequence as reference, we found three distinct versions of pBI143 (Figure 3.1A), all of which had over 95% nucleotide sequence identity to one another throughout their entire length except at the *repA* gene, where the sequence identity was as low as 75% with a maximum of 81% between Version 1 and Version 2 (Supplementary Table 2).

We then sought to quantify the prevalence of pBI143 across global human populations using a metagenomic read recruitment survey with an expanded set of 4,516 publicly available

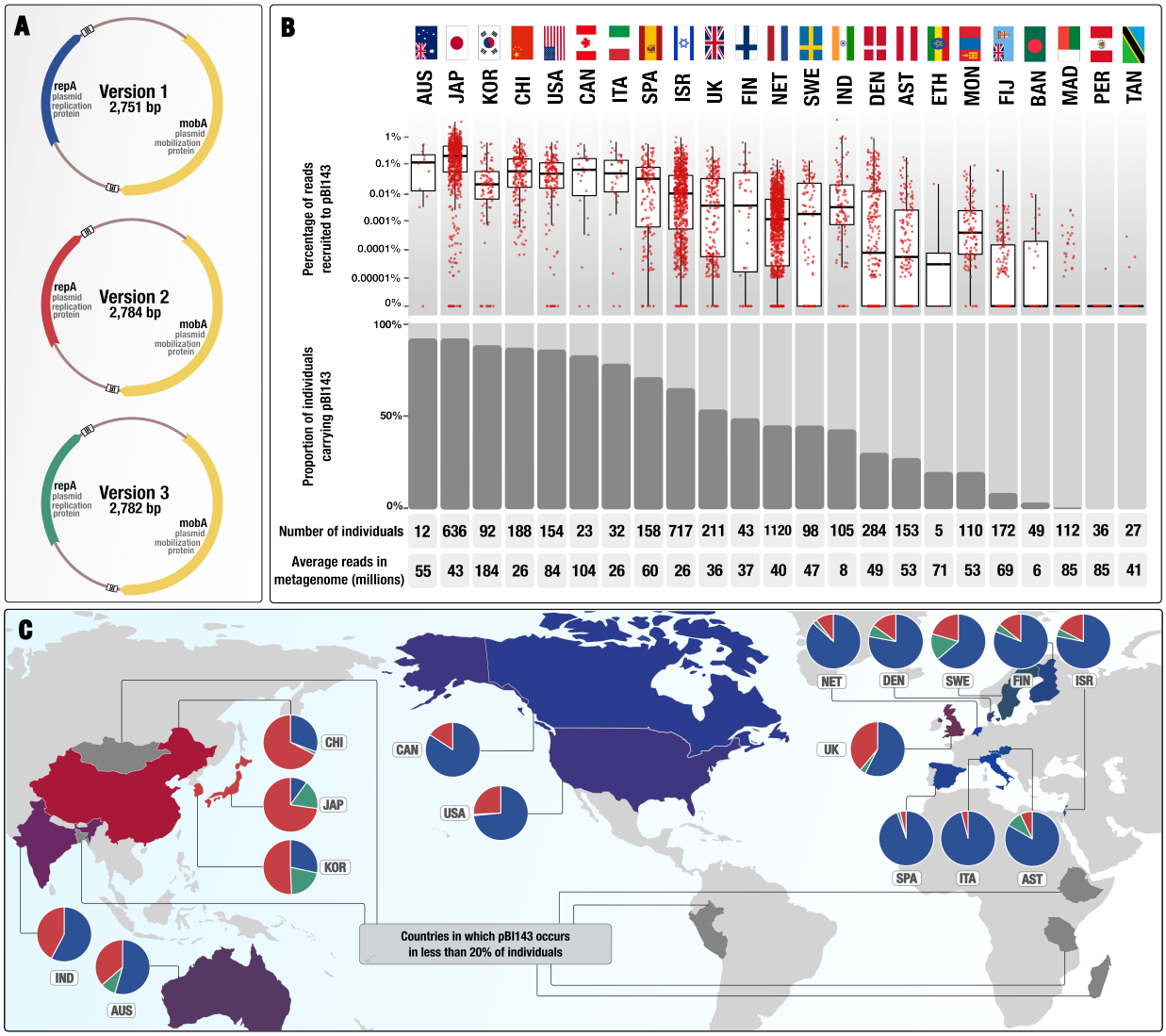

Figure 3.1: pBI143 prevalence and abundance in globally distributed human populations. (A) Plasmid maps of the three distinct versions of pBI143, which differ primarily in the *repA* gene. IR = inverted repeat. The *repA* genes are colored according to Version 1 (blue), Version 2 (red) and Version 3 (green). (B) Read recruitment results from 4,516 metagenomes originating from 23 globally representative countries and mapped to pBI143. Top: The percentage of reads in each metagenome that mapped to pBI143 normalized

Figure 3.1 continued: by number of reads in the metagenome. Bottom: The proportion of individuals in a country that have pBI143 in their gut. Each red dot represents an individual metagenome. (C) Countries that are represented in our collection of 4,516 global adult gut metagenomes. Each country's pie chart is colored based on the version(s) of pBI143 that is most prevalent in that country (Version 1 = blue, Version 2 = red, Version 3 = green). Each country is colored based on the proportion of Version 1, 2 or 3 present in the population, or gray if fewer than 20% of individuals carry pBI143. Pie charts show the proportions of pBI143 versions in all individuals that carry it within a country.

gut metagenomes from 23 countries [Feng et al., 2015, The Human Microbiome Project Consortium, 2012, Obregon-Tito et al., 2015, Li et al., 2014, Bäckhed et al., 2015, Lou et al., 2021, David et al., 2015, Raymond et al., 2016, Qin et al., 2012, Wen et al., 2017, Le Chatelier et al., 2013, Xie et al., 2016, Pasolli et al., 2019, Yassour et al., 2018, Dhakan et al., 2019, Zeevi et al., 2015, Ferretti et al., 2018, Rampelli et al., 2015, Yachida et al., 2019, Kim et al., 2021, Liu et al., 2016, Zhernakova et al., 2016] (Supplementary Table 1). Recruiting metagenomic short reads from each gut metagenome using each pBI143 version independently (Figure 3.2, Supplementary Table 3), we found that pBI143 was present in 3,295 metagenomes, or 73% of all samples (Figure 3.1B, see Methods for the 'detection' criteria). However, the prevalence of pBI143 was not uniform across the globe (Figure 3.1B): pBI143 occurred predominantly in metagenomes of individuals who lived in relatively industrialized countries, such as Japan (92% of 636 individuals) and the United States (86% of 154 individuals). We rarely detected pBI143 in individuals who lived in relatively non-industrialized countries such as Madagascar (0.8% of 112 individuals) or Fiji (8.7% of 172 individuals). This differential coverage is likely due to the non-uniform distribution of *Bacteroides* populations, which tend to dominate individuals who live in relatively more industrialized countries [Gupta et al., 2017]. Within each individual, pBI143 was often highly abundant (Figure 3.1B), and despite its small size, it often recruited 0.1% to 3.5% of all metagenomic reads with a median coverage of over 7,000X (Figure 3.2, Supplementary Table 3). In one extreme example, pBI143 comprised an astonishing 7.5% of all reads in an infant gut metagenome from Italy, with a metagenomic read coverage exceeding 54,000X (Supplementary Table 3).

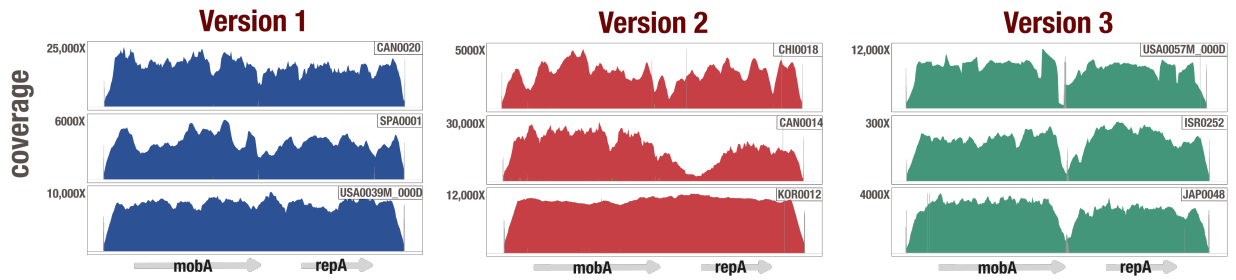

Figure 3.2: Representative coverage plots of global metagenomes mapped to pBI143. Each coverage plot shows the read recruitment results for an individual metagenome to a pBI143 Version 1 (blue), Version 2 (red) and Version 3 (green). Vertical bars show single nucleotide variants (red bar = variant in first or second codon position, green bar = variant in third codon position, gray bar = intergenic variant). The x-axis is the pBI143 reference sequence. 3 coverage plots for each reference version of pBI143 are shown, the remaining 13,539 can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.

The distribution of pBI143 versions across human populations was also not uniform as different versions of pBI143 tended to be dominant in different geographic regions. pBI143 Version 1 (98% identical to the original reference sequence for pBI143 [Smith et al., 1995]) dominated individuals in North America and Europe, and occurred on average in 82.5% of all samples that carry pBI143 from Austria, Canada, Denmark, England, Finland, Italy, Netherlands, Spain, Sweden and the USA (Figure 3.1C, Supplementary Table 3). In contrast, pBI143 Version 2 dominated countries in Asia and occurred in 63.6% of all samples that carry pBI143 in China, Japan, and Korea (Figure 3.1C, Supplementary Table 3). pBI143 Version 3 was relatively rare, comprising only 7.4% of pBI143-positive samples, and mostly occurred in individuals from Japan, Korea, Australia, Sweden, and Israel (Figure 3.1C, Supplementary Table 3).

The extremely high prevalence and coverage of pBI143 suggests that it is likely one of the most numerous genetic elements in the gut microbiota of individuals from industrialized countries. We compared the prevalence and relative abundance of pBI143 to crAssphage, a 97 kbp bacterial virus that is widely recognized as the most abundant family of viruses in the human gut [Yutin et al., 2018]. pBI143 was more prevalent (73% vs 27%) in our

metagenomes than crAssphage, although individual samples differed widely with respect to the abundance of these two elements in a given individual (Supplementary Table 3). The average percentage of metagenomic reads recruited by pBI143 and crAssphage were 0.05% and 0.13%, respectively. However, taking into consideration that crAssphage is approximately 36 times larger than pBI143, and assuming that average coverage is an acceptable proxy to the abundance of genetic entities, these data suggest that on average pBI143 is 14 times more numerous than crAssphage in the human gut. Overall, these data demonstrate that pBI143 is one of the most widely distributed and numerous genetic elements in the gut microbiomes of industrialized human populations world-wide.

3.4.2 pBI143 is specific to the human gut and hosted by a wide range of Bacteroidales species

Interestingly, the detection patterns of pBI143 in metagenomes differed from the detection patterns we observed for its de facto host *Bacteroides fragilis* in the same samples; *B. fragilis* and pBI143 co-occurred in only 41% of the metagenomes. Sequencing depth did not explain this observation, as pBI143 was highly covered (i.e., >50X) in 25% of metagenomes where *B. fragilis* appeared to be absent (Supplementary Table 11), suggesting that the host range of pBI143 extends beyond *B. fragilis*.

To investigate the host range of pBI143, we employed a collection of bacterial isolates from the human gut, which contained 717 genomes that represented 104 species in 54 genera (Supplementary Table 4). We found pBI143 in a total of 82 isolates that resolved to 11 species across 3 genera: *Bacteroides*, *Phocaeicola*, and *Parabacteroides*. Many of the pBI143-carrying isolates of distinct species were from the same individuals, suggesting that pBI143 can be mobilized between species. To confirm this, we inserted a tetracycline resistance gene, *tetQ*, into pBI143 in the *Phocaeicola vulgatus* isolate MSK 17.67 (Figure 3.3, Supplementary Table 4) and tested the ability of this engineered pBI143 to transfer to two strains of two different

families of Bacteroidales, *Bacteroides ovatus* D2 and *Paracteroides johnsonii* CL02T12C29. In these assays, we found that pBI143 was indeed transferred from the donor to the recipient strains at a frequency of 5 x 10^{-7} and 3 x 10^{-6} transconjugants per recipient, respectively (Figure 3.3).

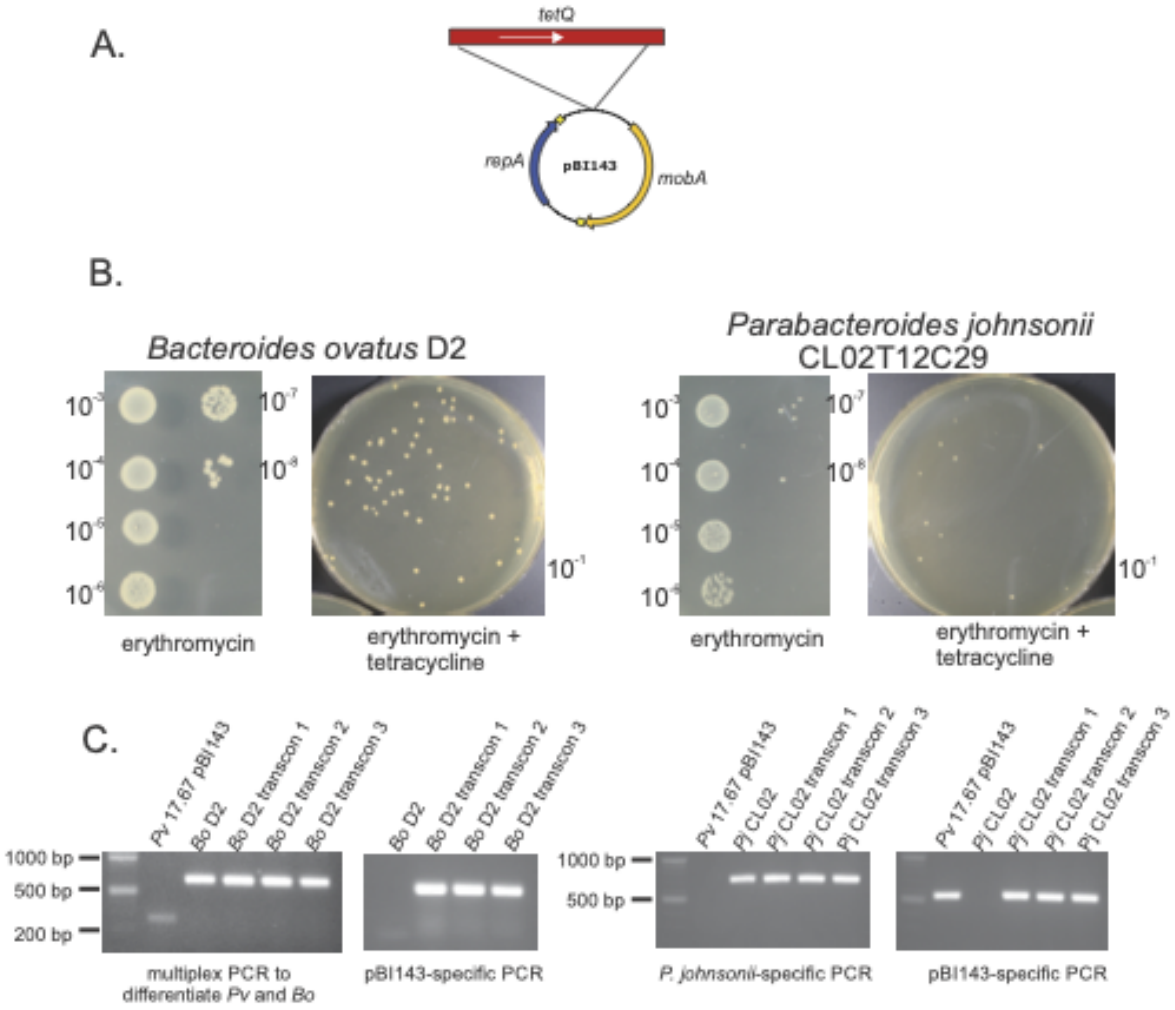

Figure 3.3: pBI143 transfer to other Bacteroidales species. (A) Construct made to select for plasmid transfer. (B) Number of recipients (erythromycin) and number of transconjugants (erythromycin and tetracycline) for transfer of pBI143-*tetQ* to *Bacteroides ovatus* D2 and *Parabacteroides johnsonii* CL02T12C129. (C) PCR to confirm presence of pBI143-*tetQ* in recipient strain.

Given the broad host range of pBI143, one interesting question is whether the ecological niche boundaries of pBI143 hosts exceed a single biome, since the members of the order

Bacteroidales are not specific to the human gut and do occur in a wide range of other habitats from non-human primate guts [Amato et al., 2013] to marine systems [Iino et al., 2014]. To investigate whether pBI143 might exist in non-human environments, we searched for pBI143 in metagenomes from coastal and open ocean samples [Sunagawa et al., 2015, Kopf et al., 2015], captive macaques [Amato et al., 2013], human-associated pets [Coelho et al., 2018], and sewage samples from across the globe [Hendriksen et al., 2019]. The plasmid was absent from all non-human associated samples, but as expected, was present in sewage (Figure 3.4, Supplementary Table 3, Supplementary Text). Given the absence of pBI143 in non-human associated habitats, we also screened metagenomes from human skin and oral cavity [The Human Microbiome Project Consortium, 2012]. Unlike the extremely high presence of pBI143 in the human gut, pBI143 was poorly detected both in samples from skin and the oral cavity (Supplementary Text). Finally, we designed and tested a highly specific qPCR assay for pBI143 (Supplementary Table 5) to confirm its specificity to the human gut. While there was a robust amplification of pBI143 from sewage samples confirming our insights from metagenomic coverages (Figure 3.5), pBI143 was virtually absent in dog, alligator, raccoon, horse, pig, deer, cow, chicken, goose, cat, rabbit, deer, or gull fecal samples (Supplementary Table 6). The only exception was the relatively low copy number (i.e., 73-fold less than human fecal content of sewage) in three of the four cats tested.

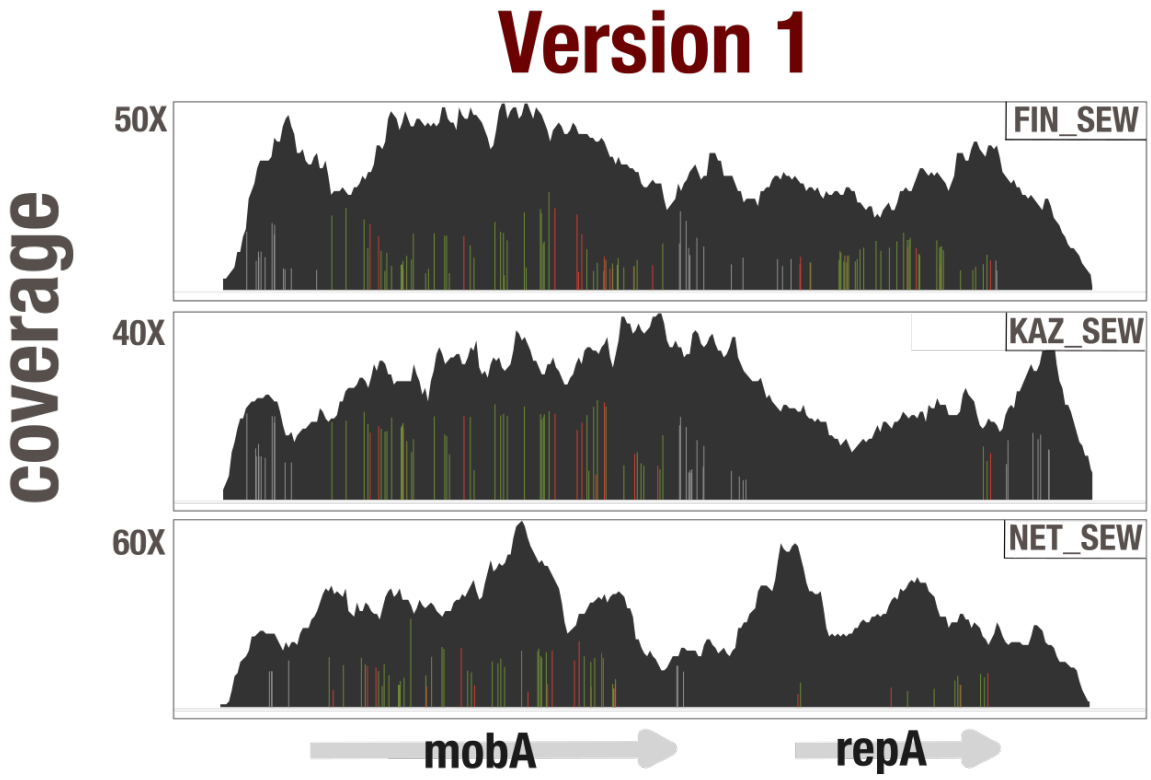

Figure 3.4: Representative coverage plots of sewage metagenomes mapped to pBI143. Each coverage plot shows the read recruitment results for a sewage metagenome to the Version 1 pBI143 reference sequence. Vertical bars show single nucleotide variants (red bar = variant in first or second codon position, green bar = variant in third codon position, gray bar = intergenic variant). The x-axis is the pBI143 reference sequence. 3 sewage coverage plots are shown, the other 435 coverage plots from all non-human environments can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.

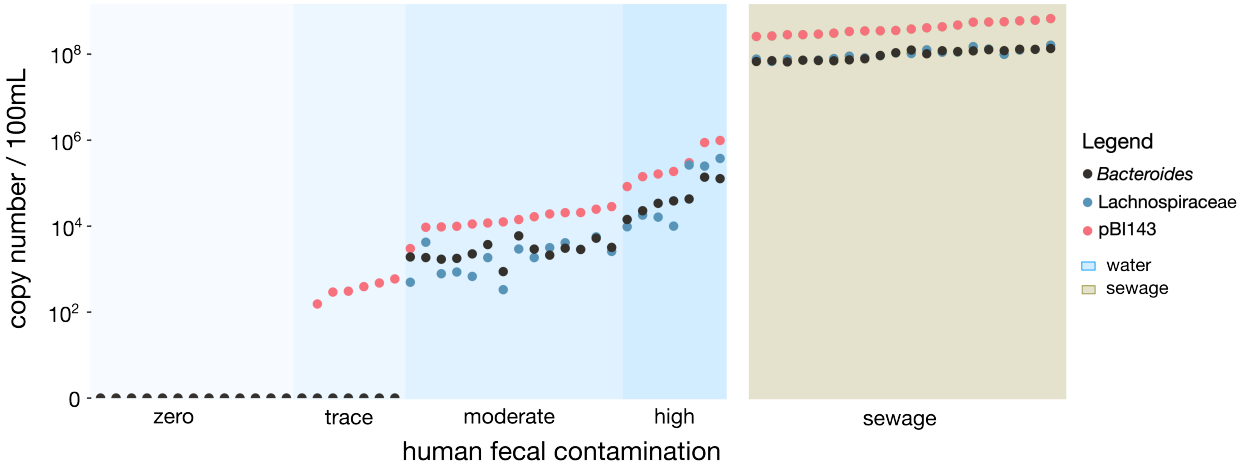

Figure 3.5: Detection of pBI143 and two established human fecal markers in water and sewage samples. Copy number of pBI143, human *Bacteroides* or Lachnospiraceae as measured by qPCR. Zero, trace, moderate, high and sewage categories and sample order designations are determined based on pBI143 copy number. Trace indicates one of the established markers was detected but was below the level of quantification. The blue background indicates water samples and the beige background indicates samples from sewage.

The near-absolute exclusivity of pBI143 to the human gut presents practical opportunities, such as the accurate detection of human fecal contamination outside the human gut. Using the same PCR primers, we also amplified pBI143 from water and sewage samples and compared its sensitivity to the gold standard markers currently used for detecting human fecal contamination in the environment (16S rRNA gene amplification of human *Bacteroides* and Lachnospiraceae) [Feng et al., 2018, Sauer et al., 2011]. pBI143 had higher amplification in all 41 samples where *Bacteroides* and Lachnospiraceae were also detected (Figure 3.5). pBI143 was also amplified in 6 samples with no *Bacteroides* or Lachnospiraceae amplification, suggesting it is a highly sensitive marker for detecting the presence of human-specific fecal material.

Overall, these data show that pBI143 has a broad range of Bacteroidales species, is highly specific to the human gut environment, and can serve as a sensitive biomarker to detect human fecal contamination.

3.4.3 pBI143 is monoclonal within individuals, and its variants across individuals are maintained by strong purifying selection

So far our investigation of pBI143 has focused on its ecology. Next, we sought to understand the evolutionary forces that have conserved the pBI143 sequence by quantifying the sequence variation among the three distinct versions and examining the distribution of single nucleotide variants (SNVs) within and across globally distributed individuals. Across the three versions, both pBI143 genes had low dN/dS values (mobA = 0.11, *repA* = 0.04), suggesting the presence of strong forces of purifying selection acting on *mobA* and *repA* resulting in primarily synonymous substitutions. While the comparison of the three representative sequences provide some insights into the conserved nature of pBI143, it is unlikely they capture its entire genetic diversity across gut metagenomes.

To explore the pBI143 variation landscape, we analyzed metagenomic reads that matched

the Version 1 of *mobA* to gain insights into the population genetics of pBI143 in naturally occurring habitats through single-nucleotide variants (SNVs). Since the *mobA* gene was more conserved across distinct versions of the plasmid compared to the *repA* gene, focusing on *mobA* enabled characterization of variation from all plasmid versions using a single read recruitment analysis. Surprisingly, the vast majority (83.2%) of the nucleotide positions that varied in any metagenome matched a nucleotide position that was variable between at least one pair of the three plasmid versions (Figure 3.6A, Supplementary Table 7). In other words, pBI143 variation across metagenomes was predominantly localized to certain nucleotide positions that differed between the representative sequences of pBI143 for Version 1, 2 and 3, indicating that the three representative versions capture the majority of permissible pBI143 variation within our collection of gut metagenomes. Indeed, only 24.5% of metagenomes had more than three novel SNVs that were not present in at least one plasmid version, and 84.8% of metagenomes had pBI143 sequences that were within 2-nucleotide distance of one of the three versions. In addition to the primarily localized variation of pBI143, we also observed that the vast majority of SNVs were fixed within a metagenome (i.e., a 'departure from consensus' value near 0, see Methods), suggesting that most humans carry a monoclonal population of pBI143 with little to no within-individual variation (Figure 3.6C, Supplementary Table 7).

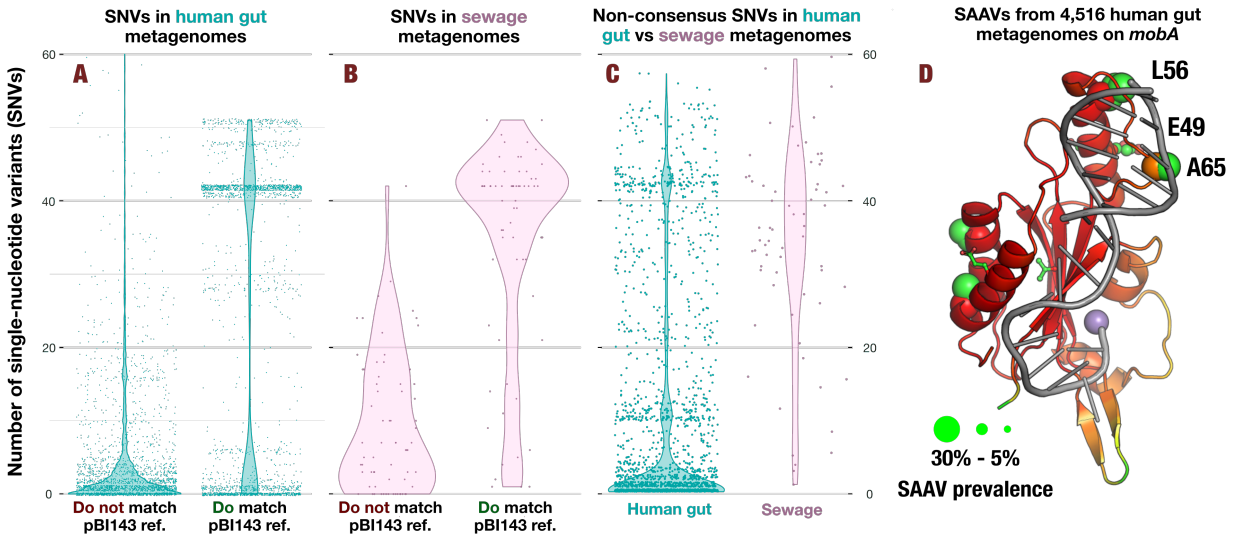

Figure 3.6: The mutational landscape of pBI143 in sewage and the human gut. (A) The proportion of SNVs across 4,516 human gut metagenomes that are present in the same location (match) or different locations (do not match) as variation in one of the versions of pBI143 (turquoise). Each point is a single metagenome. (B) The proportion of SNVs across 68 sewage gut metagenomes that are present in the same location (match) or different locations (do not match) as variation in one of the versions of pBI143 (pink). (C) Non-consensus SNVs present in 4,516 human gut metagenomes and 68 sewage metagenomes. (D) AlphaFold 2 predicted structure of the catalytic domain of MobA with single amino acid variants from all 4,516 human gut metagenomes superimposed as ball-and-stick residues. oriT DNA (gray) and a Mn2+ ion marking the active site (purple) were modeled based on 4lvi.pdb (Radoslaw Pluta et al. 2017). The size of the ball-and-stick spheres indicate the proportion of samples carrying variation in that position (the larger the sphere, the more prevalent the variation at the residue) and the color is in CPK format. The color of the ribbon diagram indicates the pLDDT from AlphaFold 2 with red = very high ($>$ 90 pLDDT) and orange = confident (80 pLDDT).

Next, we sought to investigate the functional context of non-synonymous environmental variants of MobA given its structure. For this, we employed single-amino acid variants [Delmont et al., 2019] (SAAVs) we recovered from gut metagenomes and superimposed them on the AlphaFold 2 [Mirdita et al., 2022, Delmont et al., 2019] predicted structure of MobA using anvi'o structure [Kiefl et al., 2023, Mirdita et al., 2022, Delmont et al., 2019]. The predicted catalytic domain of pBI143 MobA was structurally similar to MobM of the MobV-family (Protein Data Bank accession: 4LVI) encoded by plasmid pMV158 [Pluta et al.,

2017]. We used the structurally similar catalytic domain in MobA to model the binding of the oriT of pBI143 to MobA. We found that there were only 21 SAAVs throughout MobA that were present in greater than 5% of the gut metagenomes (Figure 3.6D). Interestingly, highly prevalent SAAVs occurred exclusively near the DNA binding site (L56, E49, and A64), suggesting that that the non-synonymous variants we observe in the context of MobA may be involved in altering the DNA binding specificity for the oriT sequence [Pluta et al., 2017] demonstrating the coevolution of the oriT with the MobA protein between distinct pBI143 versions. Additionally, we find it likely that the cluster of high prevalence variation at residues V251, A246, V239, T238, I235, and L234 (Figure 3.7B) may be driven by interactions with different host conjugation machinery for plasmid transfer. The functional implications of prevalent SAAVs given the structural context of the MobA gene highlight the role of adaptive processes on the evolution of pBI143 versions.

Unlike the individual gut metagenomes the pBI143 sequences did not occur in a monoclonal fashion in sewage metagenomes as expected (Supplementary Table 7). Sewage metagenomes had, on average, 35 SNVs with a departure from consensus value of lower than 0.9, revealing the polyclonal nature of pBI143 in sewage (Figure 3.6C, Supplementary Table 7). Similar to the individual gut metagenomes, most SNVs in sewage metagenomes (78.8%) occurred at a nucleotide position that was variable across at least one pair of the three pBI143 versions (Figure 3.6B, Supplementary Table 7), suggesting that the majority of the variability in sewage is from the mixing of different versions of pBI143. However, the number of novel SNVs was much higher in sewage: 61.8% of sewage samples had greater than three SNVs that did not match a variable position in one of the three reference plasmids (Figure 3.6B). Given the marked increase in the number of novel SNVs in sewage, it is likely there are additional but relatively rare versions of pBI143 in the human gut.

Overall, these results indicate that pBI143 has a highly restricted mutational landscape in natural habitats, frequently occurs as a monoclonal element in individual gut metagenomes,

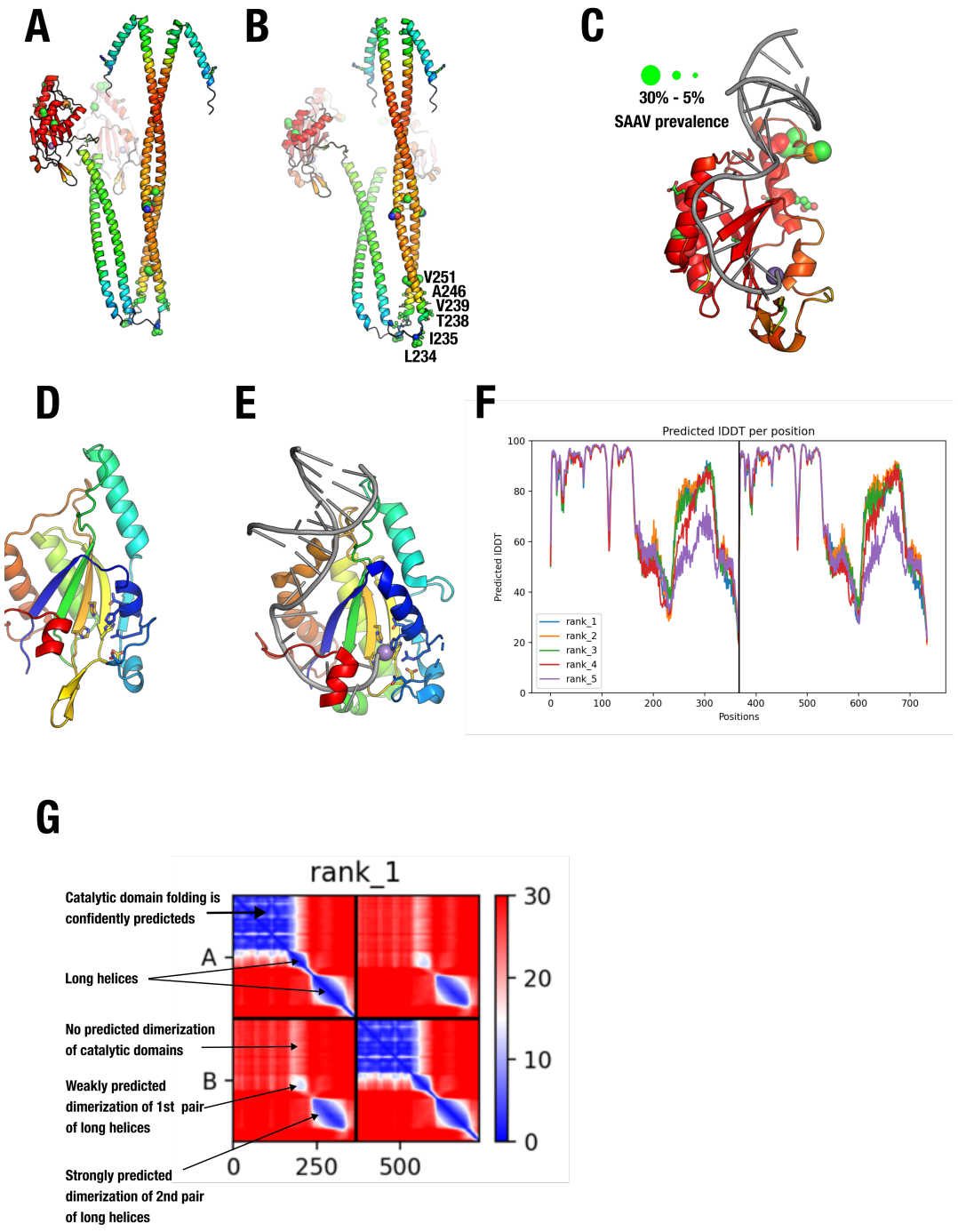

Figure 3.7: The mutational landscape of pBI143 in sewage and the human gut. (A) The proportion of SNVs across 4,516 human gut metagenomes that are present in the same

Figure 3.7 continued: location (match) or different locations (do not match) as variation in one of the versions of pBI143 (turquoise). Each point is a single metagenome. (B) The proportion of SNVs across 68 sewage gut metagenomes that are present in the same location (match) or different locations (do not match) as variation in one of the versions of pBI143 (pink). (C) Non-consensus SNVs present in 4,516 human gut metagenomes and 68 sewage metagenomes. (D) AlphaFold 2 predicted structure of the catalytic domain of MobA with single amino acid variants from all 4,516 human gut metagenomes superimposed as ball-and-stick residues. oriT DNA (gray) and a Mn2+ ion marking the active site (purple) were modeled based on 4lvi.pdb (Radoslaw Pluta et al. 2017). The size of the ball-and-stick spheres indicate the proportion of samples carrying variation in that position (the larger the sphere, the more prevalent the variation at the residue) and the color is in CPK format. The color of the ribbon diagram indicates the pLDDT from AlphaFold 2 with red = very high (>90 pLDDT) and orange = confident (80 pLDDT).

and the non-synonymous variants of MobA in the environment may be responsible for altering its DNA binding.

3.4.4 pBI143 is vertically transmitted, its variants are more specific to individuals than their host bacteria, and priority effects best explain its monoclonality in most individuals

The largely monoclonal nature of pBI143 presents an interesting ecological question: how do individuals acquire it, and what maintains its monoclonality? Multiple phenomena could explain the monoclonality of pBI143 in individual gut metagenomes, including (1) low frequency of exposure (i.e., most individuals are only ever exposed to one version), (2) bacterial host specificity (i.e., some plasmid versions replicate more effectively in certain bacterial hosts), or (3) priority effects (i.e., the first version of pBI143 establishes itself in the ecosystem and excludes others). The sheer prevalence and abundance of pBI143 across industrialized populations renders the 'low frequency of exposure' hypothesis an unlikely explanation. Yet the remaining two hypotheses warrant further investigation.

Bacterial host specificity is a plausible driver for the presence of a singular pBI143 version within an individual, given the interactions between plasmid replication genes and host

replication machinery [Thomas, 2014b, Lu et al., 1998]. However our analysis of 82 bacterial cultures isolated from 10 donors shows that the plasmid is more specific to individuals than it is to certain bacterial hosts (Figure 3.8, Supplementary Table 9). Indeed, identical pBI143 sequences often occurred in multiple distinct taxa isolated from the same individual, in agreement with the monoclonality of pBI143 in gut metagenomes and its ability to transfer within Bacteroidales. If pBI143 monoclonality is not driven by rare exposure or host specificity, it could be driven by priority effects [Debray et al., 2021], where the initial pBI143 version somehow prevents other pBI143 versions from establishing in the same gut community.

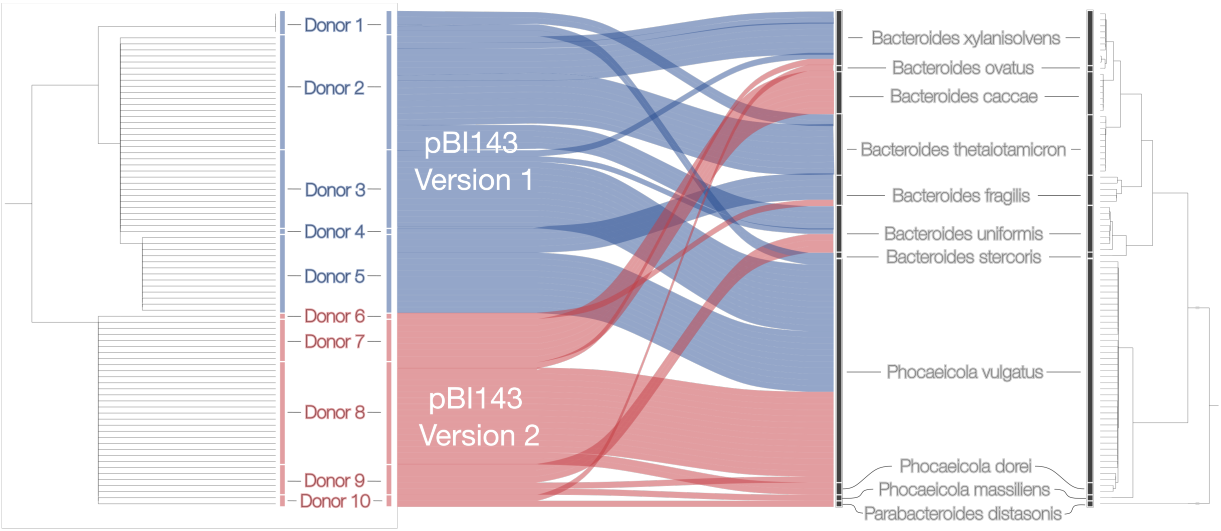

Figure 3.8: Phylogeny of pBI143 in human donors versus the phylogeny of bacterial isolates recovered from the same individuals. pBI143 (left) and bacterial host (right) genome phylogenies. The pBI143 phylogeny was constructed using the MobA and RepA genes; the bacterial phylogeny was constructed using 38 ribosomal proteins (see Methods). Blue alluvial plots are isolates with Version 1 pBI143 and red alluvial plots are isolates with Version 2 pBI143. No isolates had the rarer Version 3.

To examine if priority effects play a role in pBI143 monoclonality, we aimed to determine how pBI143 is acquired. Given that one established route of microbial acquisition is the vertical transmission of microbes from mother to infant [Vatanen et al., 2022], we used our ability to track pBI143 SNVs between environments to investigate if there is evidence for

vertical transmission. We followed the inheritance of identical SNV patterns in pBI143 using 154 mother and infant gut metagenomes from four countries, Finland [Yassour et al., 2018], Italy [Ferretti et al., 2018], Sweden [Bäckhed et al., 2015], and the USA [Lou et al., 2021], where each study followed participants from birth to 3 to 12 months of age. We recruited reads from each metagenome to Version 1 pBI143 (Supplementary Table 1 and 3, Figure 3.9) and identified the location of each SNV in *mobA* (Supplementary Table 10). These data revealed a large number of cases where pBI143 had identical SNV patterns in mother-infant pairs (Figure 3.10A, Supplementary Table 10). A network analysis of shared SNV positions across metagenomes appeared to cluster family members more closely, indicating mother-infant pairs had more SNVs in common than they had with unrelated individuals, which we could further confirm by quantifying the relative distance between each sample to others (Figure 3.11, Supplementary Table 10, Methods).

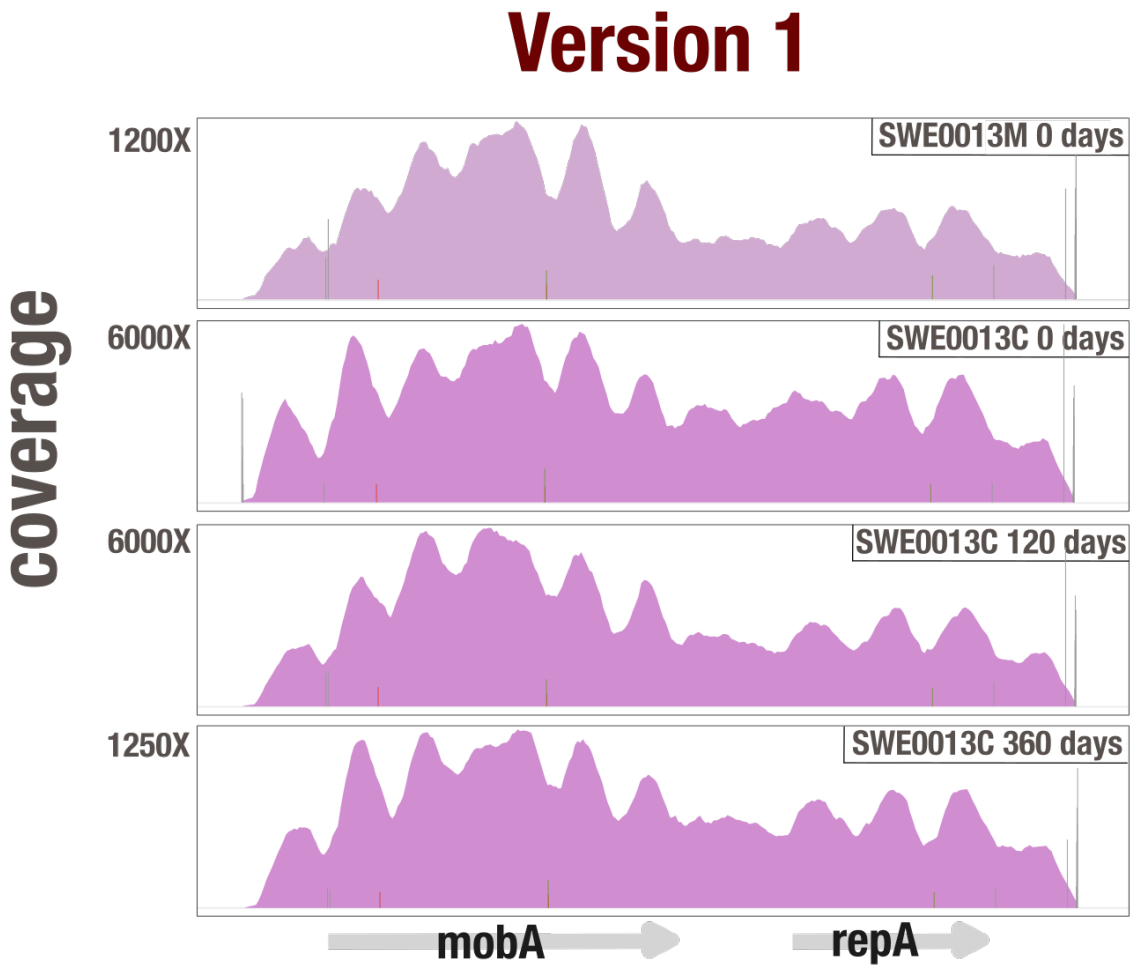

Figure 3.9: Representative mother-infant coverage plots. Each coverage plot shows the read recruitment results for an individual metagenome to the Version 1 pBI143 reference sequence. Vertical bars show single nucleotide variants (red bar = variant in first or second codon position, green bar = variant in third codon position, gray bar = intergenic variant). The x-axis is the pBI143 reference sequence. 4 coverage plots are shown, the other 1,020 can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.

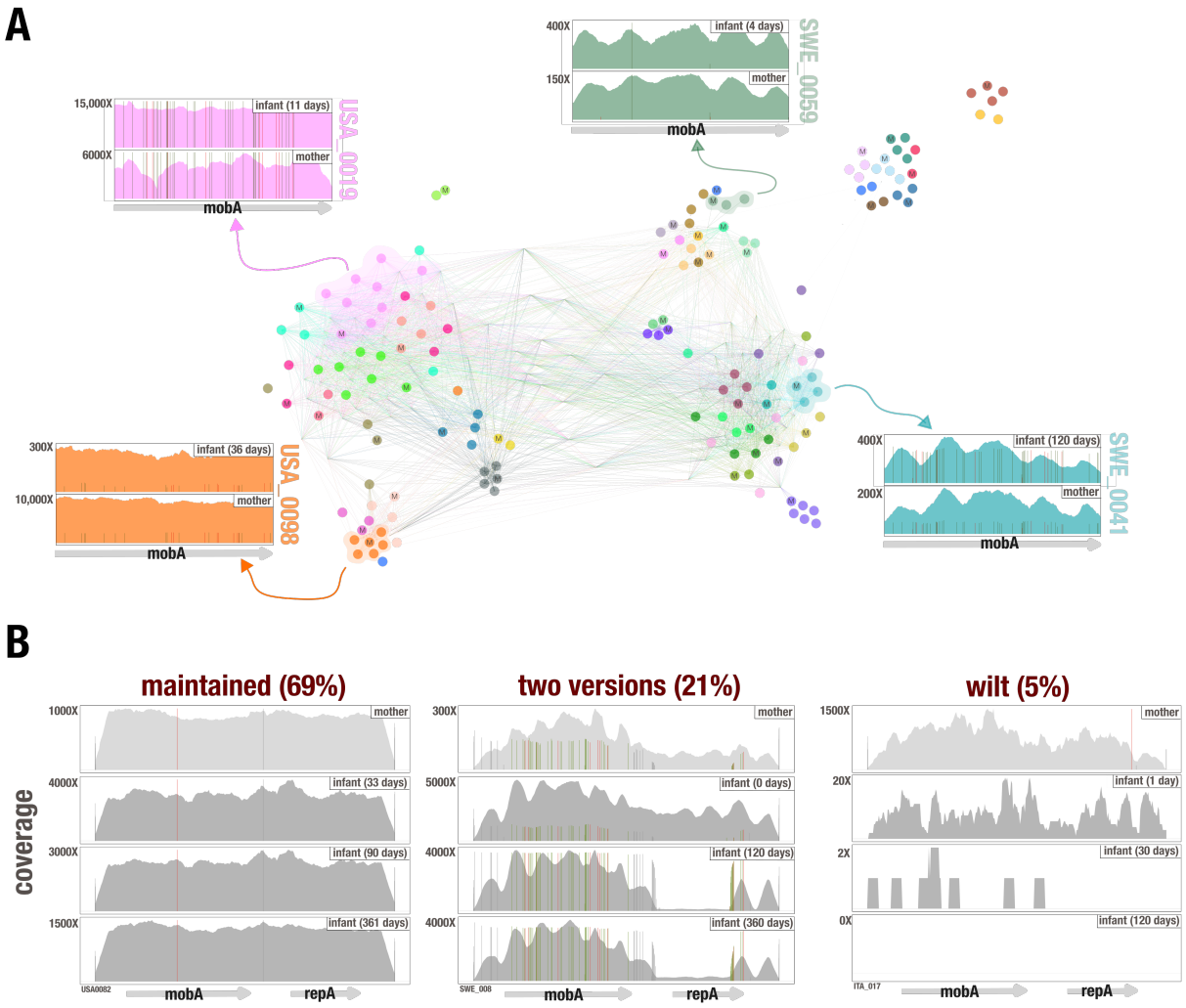

Figure 3.10: Transfer and maintenance of pBI143. (A) The network shows the degree of similarity between pBI143 SNVs across 154 mother and infant metagenomes from Finland, Italy, Sweden and the USA. Each node is an individual metagenome and nodes are colored based on family grouping. The surrounding coverage plots (colored) are visual representations of SNV patterns present in the indicated metagenomes. Nodes labeled with an "M" are mothers; nodes with no labels are infants. (B) Representative coverage plots showing different coverage patterns (maintained, two versions or wilt) observed in plasmids transferred from mothers to infants.

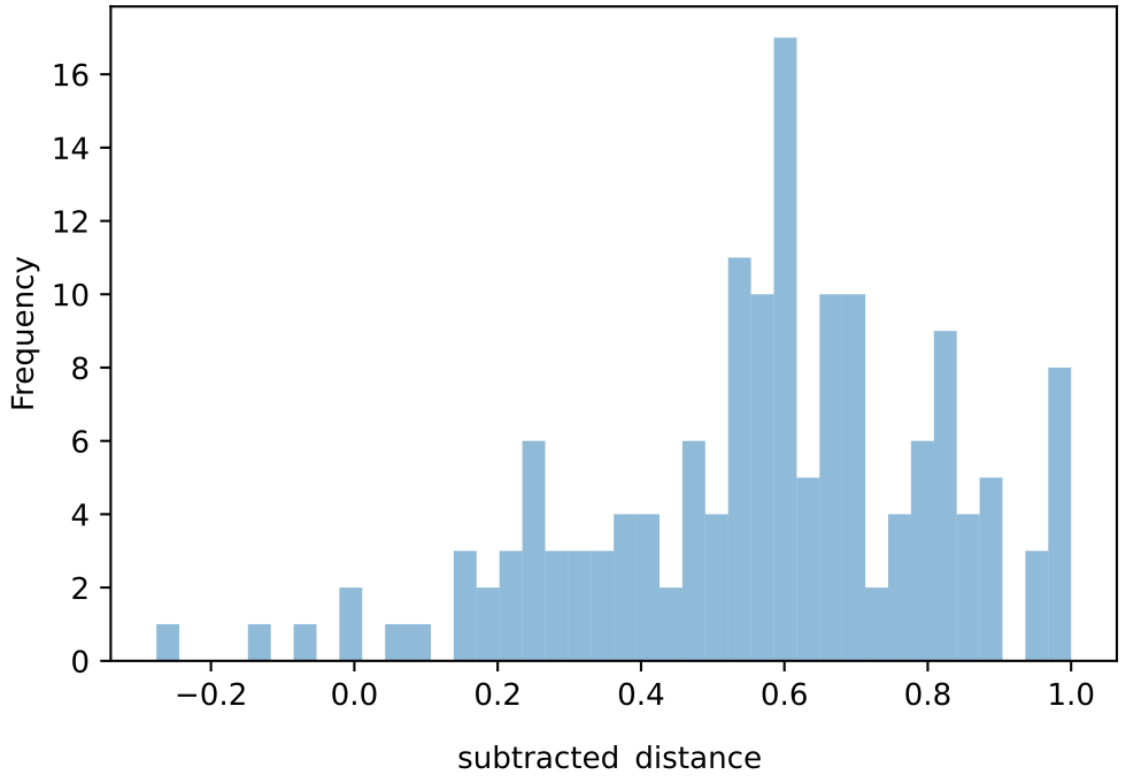

Figure 3.11: Mother-infant network quantification. Quantification of distances between samples in the network, where distance is calculated by converting the network file to a distance matrix using the python 'pdist' function with cosine distances. The "subtracted difference" shows the mean within-family distances subtracted from mean between-family distances for each sample in the mother infant pair network. See methods for more details.

Establishing that pBI143 is often vertically transferred, we next examined the impact of priority effects on pBI143 maintenance over time. We assumed that if priority effects are driving persistence of a single version of pBI143, the first version that enters the infant gut environment should be maintained over time. Indeed, many phage populations are influenced by priority effects where the presence of one phage provides a competitive advantage to the host [Joo et al., 2006] or host immunity to infection with similar phages [Bondy-Denomy et al., 2016, Mavrich and Hatfull, 2019, Chen et al., 2020b]. In our data, we found no instances where pBI143 acquired from the mother was fully replaced in the infant during and up to the first year of life (Figure 3.10, Supplementary Table 10). While 69% of infants

maintained the version received from the mother (Figure 3.10B), we also observed other, less common genotypes. These less common cases included a 'two versions' scenario where the mother possessed two versions of pBI143, both of which were passed to the infant (21%), and a 'wilt' case, where the transferred pBI143 was neither replaced nor persisted until the end of sampling (7%) (Figure 3.10B). Although these less prevalent phenotypes are not necessarily explained by priority effects, 69% maintenance of the initial version of pBI143 suggests that priority effects have an important role in the maintenance of pBI143 in the gut, despite many incoming populations colonizing the infant and likely carrying other pBI143 versions.

Overall, by tracking SNV patterns between environments we established that pBI143 is vertically transferred from mothers to infants and that priority effects likely play a role in maintaining the predominantly monoclonal populations of pBI143.

3.4.5 pBI143 is a highly efficient parasitic plasmid

An intuitive interpretation of the surprising levels of prevalence and abundance of pBI143 across the human population, in addition to its limited variation maintained by strong evolutionary forces, is that it provides some benefit to the bacterial host. However, the two annotated genes in pBI143 appear to serve only the purpose of ensuring its own replication and transfer, contradicting this premise. The coverage of pBI143 and its *Bacteroides*, *Phocaeicola* and *Parabacteroides* hosts in gut metagenomes indeed show a significant positive correlation (R2: 0.5, p-value < 0.001) (Figure 3.12A, Supplementary Table 11), however, these data are not suitable to distinguish whether pBI143 provides a benefit to the bacterial host fitness, or acts as a genetic hitchhiker.

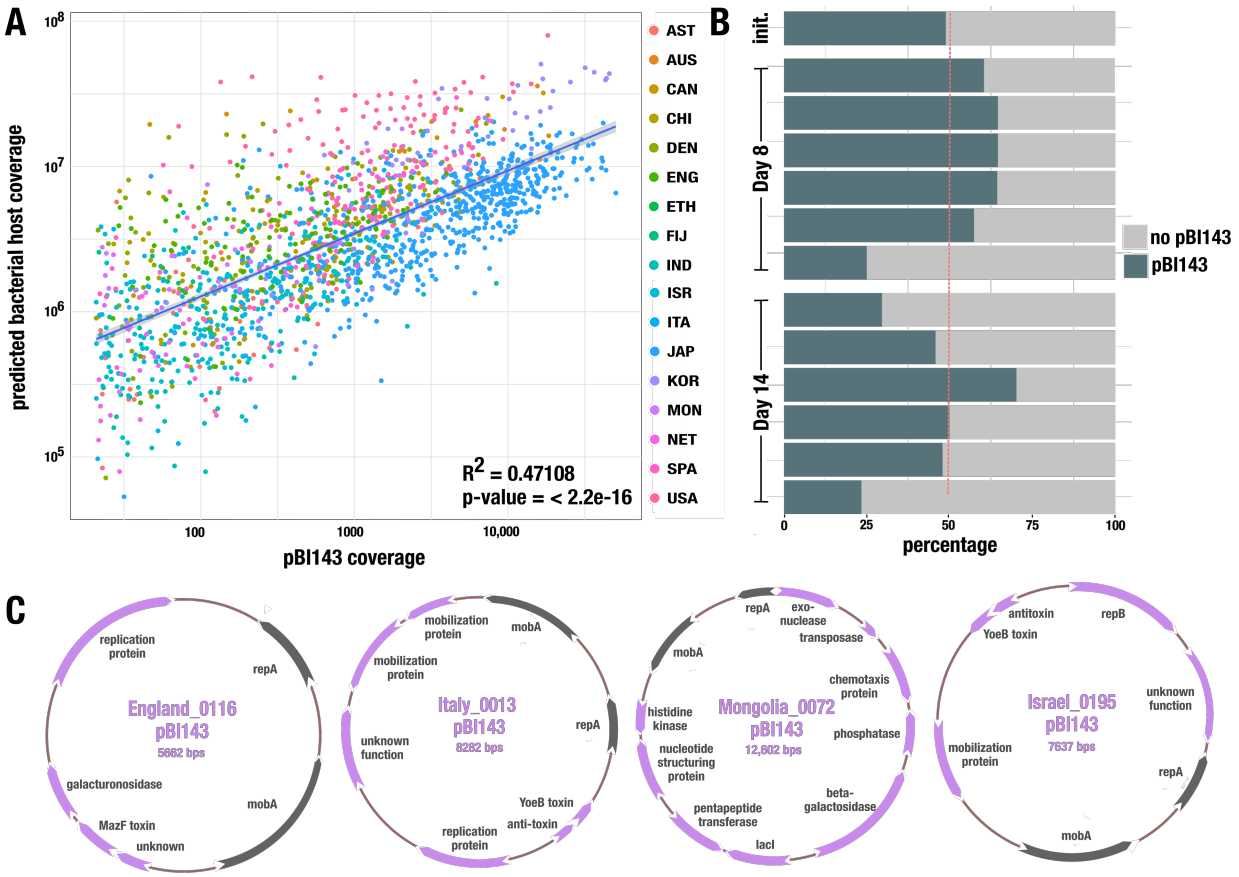

Figure 3.12: The relationship between pBI143 and its bacterial hosts. (A) The average coverage of pBI143 and the corresponding coverage of predicted host genomes (*Bacteroides*, *Parabacteroides* and *Phocaeicola*) in 4,516 metagenomes. (B) Competition experiments in gnotobiotic mice between *B. fragilis* with and without pBI143. The proportion of pBI143-carrying cells in 6 mice in the initial inoculum, at Day 8 and at Day 14 are shown. (C) Four examples of pBI143 assembled from metagenomes that carry additional cargo genes. Gray genes are the canonical *repA* and *mobA* genes of naive pBI143; lilac genes are additional cargo.

To experimentally investigate if pBI143 is advantageous or parasitic, we constructed isogenic pairs of *B. fragilis* 638R and *B. fragilis* 9343 with and without the native Version 1 sequence of pBI143 (Supplementary Methods). To determine if pBI143 is well-adapted to replication in a new *Bacteroides* host, we tested its maintenance in culture. After 7 days of passaging, pBI143 was still present in all colonies of *B. fragilis* 638R and *B. fragilis* 9343 (Supplementary Table 12). Next, we competed the *B. fragilis* 638R (with and without pBI143) of *B. fragilis* 638R in gnotobiotic mice for 2 weeks. At Day 8, 5/6 mice had more *B. fragilis* 638R with pBI143 than without; however this trend did not continue into Day 14, where 4/6 mice had fewer cells with pBI143 (Figure 3.12B, Supplementary Table 12). While we can speculate that these populations may continue to fluctuate, the results at least suggest a negligible negative fitness impact of pBI143 on its bacterial host.

One potential benefit that pBI143 could provide to its host is to act as a natural shuttle vector by transiently acquiring additional genetic material and transferring it between cells in a community. In fact, in our survey of assembled gut metagenomes we observed a few cases that may support such a role for pBI143. In most individuals, we assembled pBI143 in its native form with 2 genes. However, there were 10 instances where the assembled pBI143 sequence from a given metagenome contained additional genes (Figure 3.12C, Supplementary Table 2). Many of the additional genes have no predicted function, but other cargo include toxin-antitoxin genes conferring plasmid stability, as well as those that may confer beneficial functions to the bacterial host, such as galacturonosidase, pentapeptide transferase, phosphatase, and histidine kinase genes. These occasional larger versions of pBI143 share a common backbone of *repA* and *mobA* and thus form a "plasmid system" [Yu et al., 2020], a common plasmid evolutionary pattern suggesting the possibility that pBI143 may dynamically acquire different genes in different environments.

Overall, it does not appear that the native sequence of pBI143 provides a clear benefit to its host cells, however it does appear to positively correlate with these hosts in metagenomic

data, and is maintained in the absence of selection in new hosts *in vitro*.

pBI143 responds to oxidative stress *in vitro*, and its copy number is significantly higher in metagenomes from individuals who are diagnosed with IBD

Mobile genetic elements rely on their hosts for replication machinery, but many have developed mechanisms to increase their rates of replication and transfer during stressful conditions to increase the likelihood of their survival if the host cell dies [Beaber et al., 2004, Comeau et al., 2007, Ubeda et al., 2005, Schumann et al., 1984]. To investigate whether the copy number of pBI143 changes as a function of stress, we first conducted an experiment with *B. fragilis* isolates that naturally carry pBI143.

Given that oxygen exposure upregulates oxidative stress response pathways in the anaerobic *B. fragilis* [Sund et al., 2008], we exposed two different *B. fragilis* cultures, *B. fragilis* R16 (which was isolated from a healthy individual) and *B. fragilis* 214 (which was isolated from a pouchitis patient [Vineis et al., 2016]) to 21% oxygen for increasing periods of time (Figure 3.13A, Figure 3.14, Supplementary Table 13). To calculate the copy number of pBI143 in culture, we quantified the ratio between the total number of plasmids and the total number of cells in culture using a qPCR with primers targeting pBI143 and a *B. fragilis*-specific gene we identified through pangenomics (Methods). As the length of oxygen exposure increased, the copy number of pBI143 per cell also increased. Notably, the copy number was quickly reduced to control levels once the cultures were returned to anaerobic conditions, indicating that copy number fluctuation is a rapid and transient process that is dependent on host stress.

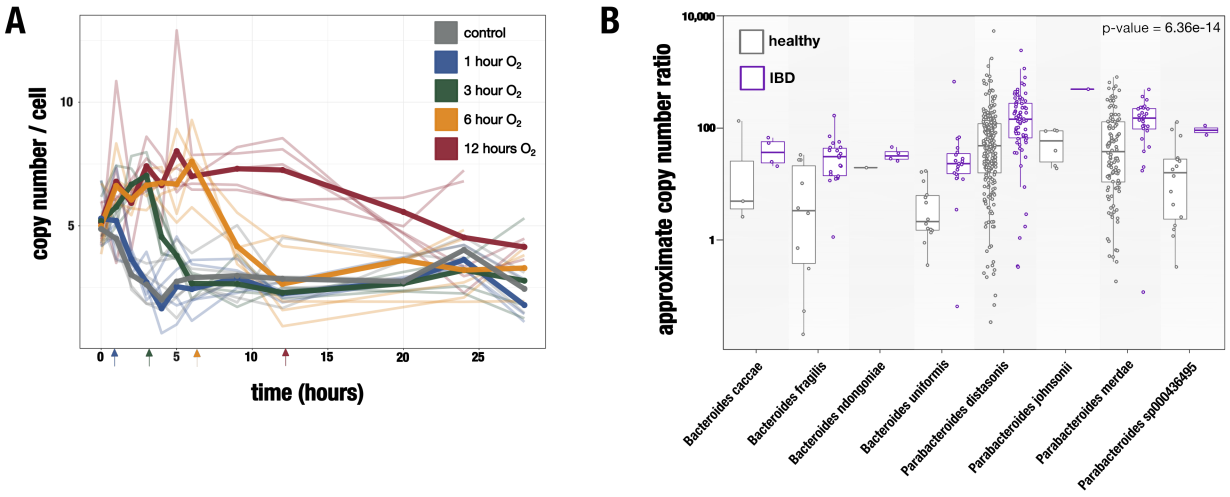

Figure 3.13: pBI143 copy number increases in stressful environments. (A) Copy number of pBI143 in *B. fragilis* 214 cultures with increasing exposure to oxygen. Arrows indicate the time point at which the culture was returned to the anaerobic chamber. The control cultures (gray) were never exposed to oxygen. Opaque lines are the mean of 5 replicates (translucent lines). (B) Host-specific approximate copy number ratio (ACNR) of pBI143 in healthy individuals (gray) versus those with IBD (purple).

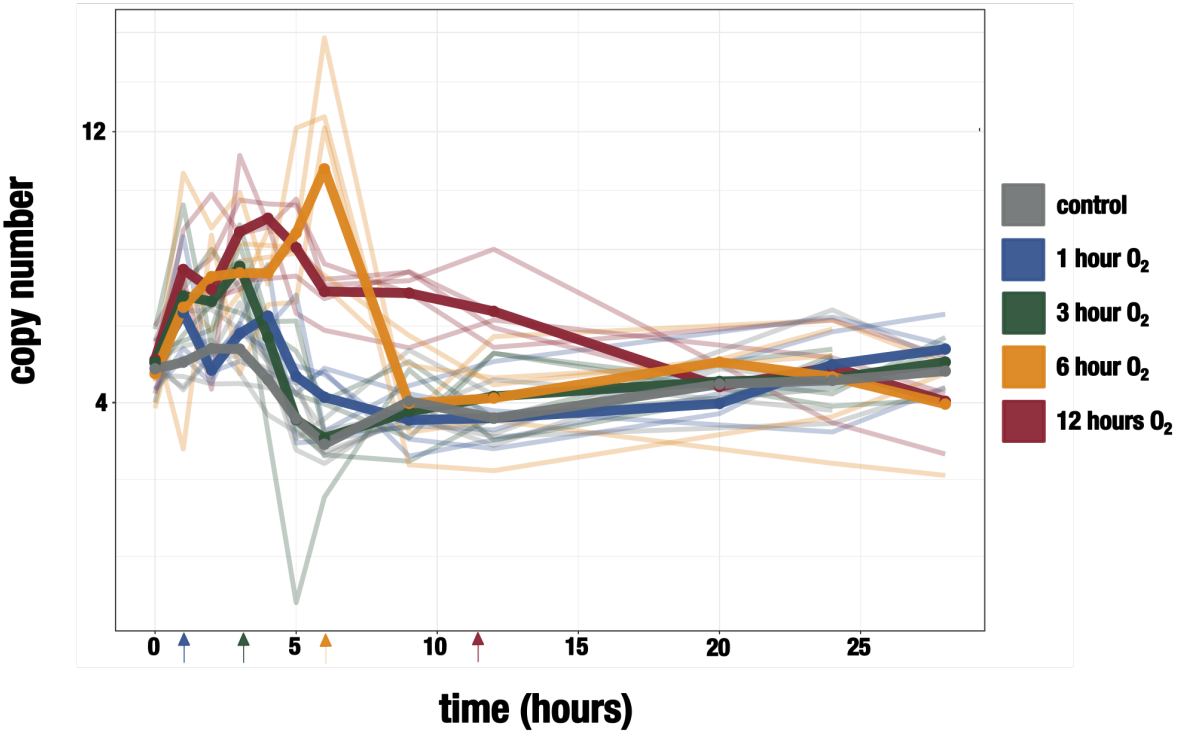

Figure 3.14: R16 oxidative stress experiment. Copy number of pBI143 in *B. fragilis* R16 cultures with increasing exposure to oxygen. Arrows indicate the time point at which the culture was returned to the anaerobic chamber. The control cultures (gray) were never exposed to oxygen. Opaque lines are the mean of 5 replicates (translucent lines).

Oxidative stress is also a signature characteristic of inflammatory bowel disease (IBD), a group of intestinal disorders that cause inflammation of the gastrointestinal tract [Baumgart and Carding, 2007]. The dysregulation of the immune system during IBD typically leads to high levels of oxidative stress in the gut environment [Graham and Xavier, 2020]. We thus hypothesized that, if oxidative stress is among the factors that drive the increased copy number of pBI143 in culture, one should expect a higher copy number of pBI143 in metagenomes from IBD patients compared to healthy controls.

To analyze the copy number of pBI143 in a given metagenome, we calculated the ratio of metagenomic read coverage between pBI143 and its bacterial host in metagenomes where pBI143 could confidently be assigned to a single host. With these considerations, we developed an approach to calculate an 'approximate copy number ratio' (ACNR) for pBI143 and its unambiguous bacterial host in a given metagenome using bacterial single-copy core genes (see Methods). We calculated the ACNR of pBI143 in 3,070 healthy and 1,350 IBD gut metagenomes (Supplementary Table 1, Figure 3.2 and 3.15). Our analyses showed that the geometric mean of the ACNR for pBI143 and its host was 3.72 times larger (robust-Wald 95% CI: 2.66x - 5.20x, p-value $<$ 10-13) in IBD compared to healthy metagenomes, indicating that the pBI143 ACNR was significantly higher in individuals with IBD compared to those who were healthy (Figure 3.13B, Supplementary Table 14).

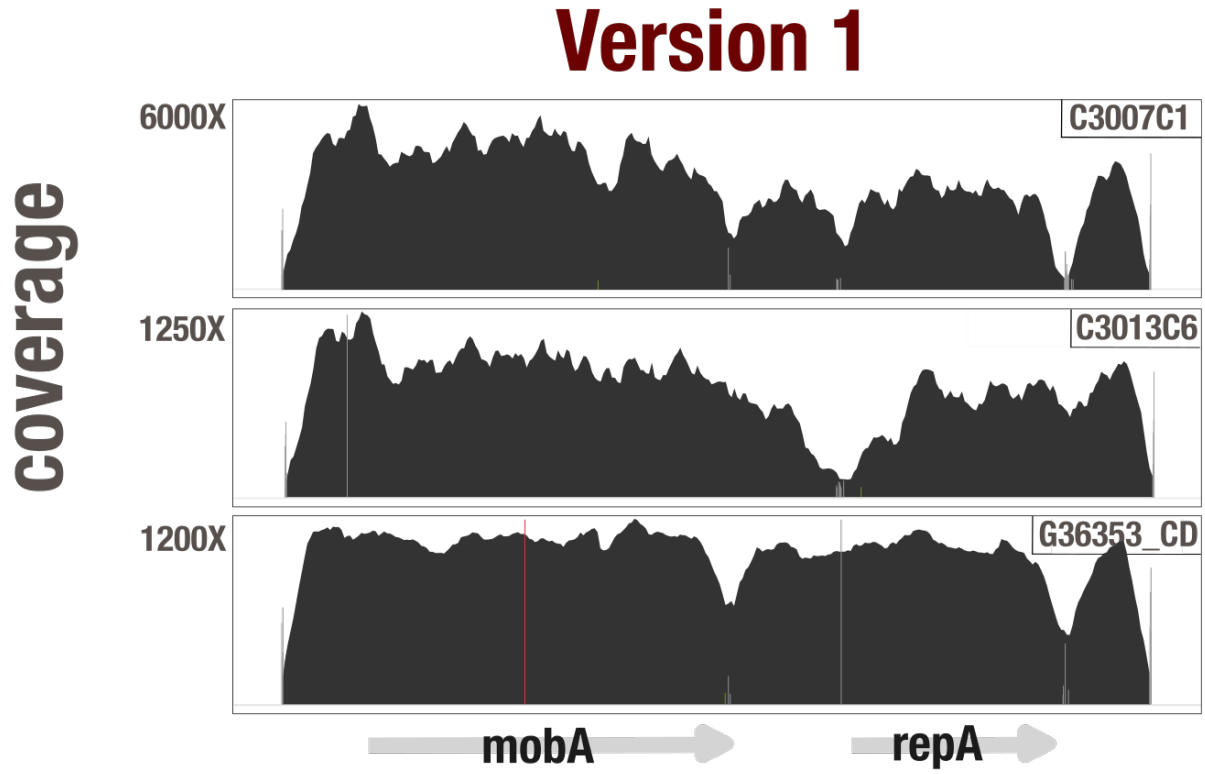

Figure 3.15: Representative IBD gut metagenome coverage plots. Each coverage plot shows the read recruitment results for an individual metagenome to a pBI143 Version 1. Vertical bars show single nucleotide variants (red bar = variant in first or second codon position, green bar = variant in third codon position, gray bar = intergenic variant). The x-axis is the pBI143 reference sequence. 3 coverage plots are shown, the other 3,087 can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.

The copy number ratio of pBI143 to its *B. fragilis* host in culture calculated with qPCR primers was much lower (approximately 5X on average) compared to its approximate copy number ratio in healthy metagenomes (approximately 120X on average). Multiple factors can explain this difference, including biases associated with sequencing steps or the calculation of the coverage, or that the conditions naturally occuring communities experience vastly differ than those conditions encountered in culture media, even in the presence of oxygen. Nevertheless the marked increase of the relative coverages of pBI143 and its host in IBD metagenomes suggest the potential utility of this cryptic plasmid for unbiased measurements

of stress. Overall, these results show that both in metagenomes and experimental conditions, an increased copy number of pBI143 is a consistent phenotype in the presence of host stress.

3.5 Discussion

Our work shed lights on a mysterious corner of life in the human gut. Even though pBI143 is found in greater than 90% of all individuals in some countries, the prevalence of this cryptic plasmid has gone unnoticed for almost four decades since its discovery by Smith, Rollins, and Parker [Smith et al., 1995]. The remarkable ecology, evolution, and potential practical applications of pBI143 that we characterized here through 'omics analyses as well as *in vitro* and *in vivo* laboratory experiments offer a glimpse of the world of understudied cryptic plasmids in the human gut, and elsewhere.

The application of population genetics principles to pBI143 through the recovery of single-nucleotide variants (SNVs) and single-amino acid variants (SAAVs) from gut metagenomes reveals not only the strong forces of purifying selection on the evolution of its sequence, but also hints the presence of adaptive processes at localized amino acid positions that are variable in the critical parts of the DNA-interacting residues of the catalytic domain of its mobilization protein. The presence of pBI143 does not appear to systematically impact bacterial host fitness *in vivo*, which makes this cryptic plasmid seem a mundane parasite, somewhat contradicting the strict evolutionary pressures that maintain its environmental sequence variants.

That said, our observations from naturally occurring gut environments include cases where pBI143 carries additional genes, likely acting as a natural shuttle vector. Although traditionally mobile genetic elements are classified as mutualistic or parasitic with respect to the bacterial host, the fluidity of pBI143 to fluctuate between the cryptic 2-gene state and the larger 3 or more gene state with potentially beneficial functions, suggests that the boundaries between parasitism and mutualism for pBI143 are not clear cut. Instead, pBI143

may act as a 'discretionary parasite', where it has a cryptic form for the majority of its existence in which it could be best described as a parasite, while occasionally being found with additional functions that may be beneficial to its host as a function of environmental pressures. Testing this hypothesis with future experimentation, and if true, investigating to what extent discretionary parasitism applies to cryptic plasmids, may lead to a deeper understanding of the role cryptic plasmids play in microbial fitness to changing environmental conditions.

Our findings show that it has important potential practical applications beyond molecular biology. The first and most straightforward of these applications relies on the prevalence and human specificity of pBI143 to more sensitively detect human fecal contamination in water samples. Human fecal pollution is a global public health problem, and accurate and sensitive indicators of human fecal pollution are essential to identify and remediate contamination sources and to protect public health [McLellan and Eren, 2014]. While culture assays for *E. coli* or *Enterococci* have historically been used to detect human fecal contamination in environmental samples, the common occurrence of these organisms in many different mammalian guts and the poor sensitivity of such assays motivated researchers in the past two decades to utilize PCR amplification of 16S rRNA genes, specifically those from human-specific *Bacteroides* and Lachnospiraceae populations, to detect human-specific fecal contamination with minimal cross-reactivity with animal feces [Feng et al., 2018, Sauer et al., 2011]. Our benchmarking of pBI143 with qPCR revealed that pBI143 is an extremely sensitive and specific marker of human fecal contamination that typically occurs in human fecal samples and sewage in numbers that are several-fold higher than the state-of-the-art markers, which enabled the quantification of fecal contamination in samples where it had previously gone undetected. Another practical application of pBI143 takes advantage of its natural shuttle vector capabilities to incorporate additional genetic material into its backbone. Our demonstration that pBI143 (1) replicates in many abundant gut microbes, (2)

can be stably introduced to new hosts, and (3) naturally acquires genetic material makes this cryptic plasmid an ideal natural payload delivery system for future therapeutics targeting the human gut microbiome. Indeed, our observations of pBI143 with cargo genes in metagenomes indicates that this likely happens in nature. Yet another practical implication of pBI143 is its utility to measure the level of stress in the human gut. Surveying thousands of samples from individuals who are healthy or diagnosed with IBD, our results show that across all bacterial hosts, the approximate copy number of pBI143 increases in individuals with IBD.

From a more philosophical point of view, the prevalence and high conservancy of pBI143 across globally distributed human populations questions the traditional definition of the 'core' microbiome [Neu et al., 2021]. In its aim to define a core microbiome, the field of microbial ecology has primarily focused on bacteria, although sometimes including prevalent archaea or fungi [Aguirre de Cárcer, 2018, Mancabelli et al., 2017, Shetty et al., 2022, Nash et al., 2017]. However, our results indicate that there are mobile genetic elements that fit the standard criteria of prevalence to be defined as core. Broadening the definition of a core microbiome beyond microbial taxa may enable the recognition of other mobile genetic elements (eg. plasmids, phages, transposons) that are prevalent across human populations and fill critical gaps in our understanding of gut microbial ecology.

3.6 Materials and Methods

3.6.1 Genomes and metagenomes

We acquired the original pBI143 genome from the National Center for Biotechnological Information (GenBank: U30316.1). We manually assembled the three reference versions of pBI143 (Version 1, 2 and 3) from metagenomes samples USA0006, CHI0054 and ISR0084. We acquired 717 human gut isolate genomes from the Duchossois Family Institute collection

(Supplementary Table 4). We downloaded 4,516 healthy human adult gut metagenomes from the National Center for Biotechnology Information (NCBI) from (Australia (Accession ID: PRJEB6092), Austria [Feng et al., 2015], Bangladesh [David et al., 2015], Canada [Raymond et al., 2016], China [Qin et al., 2010, Wen et al., 2017], Denmark [Le Chatelier et al., 2013], England [Xie et al., 2016], Ethiopia [Pasolli et al., 2019], Fiji [Brito et al., 2016], Finland [Yassour et al., 2018], India [Dhakan et al., 2019], Israel [Zeevi et al., 2015], Italy [Ferretti et al., 2018, Rampelli et al., 2015], Japan [Yachida et al., 2019], Korea [Kim et al., 2021], Madagascar [Pasolli et al., 2019], Mongolia [Pasolli et al., 2019, Liu et al., 2016], Netherlands [Zhernakova et al., 2016], Peru [Obregon-Tito et al., 2015], Spain [Li et al., 2014], Sweden [Bäckhed et al., 2015], Tanzania [Rampelli et al., 2015], and the USA [Obregon-Tito et al., 2015, Lou et al., 2021, The Human Microbiome Project Consortium, 2012]) (Supplementary Table 1). We acquired 1,096 gut metagenomes from infant-mother pairs from Italy, Finland, Sweden and the USA from NCBI (Supplementary Table 1). We downloaded 935 metagenomes from non-human gut environments (marine ecosystems, pet dog guts, monkey guts, sewage, human oral cavity, and human skin) (Supplementary Table 1).

3.6.2 Metagenomic assembly, read recruitment, and the recovery of coverage and detection statistics

Unless otherwise specified, we performed all metagenomic analyses throughout the manuscript within the open-source anvi'o v7 software ecosystem (https://anvio.org) [Eren et al., 2021]. We automated assembly and read recruitment steps using the anvi'o metagenomics workflow [Shaiber et al., 2020] which used snakemake v5.10 [Köster and Rahmann, 2012]. To quality-filter genomic and metagenomic raw paired-end reads we used illumina-utils v1.4.4 [Murat Eren et al., 2013] program 'iu-filter-quality-minoche' with default parameters, and IDBA_UD v1.1.2 with the flag '–min_contig 1000' to assemble the metagenomes [Peng et al., 2012]. We used Bowtie2 v2.4 [Langmead and Salzberg, 2012] to recruit reads from

the metagenomes to reference sequences and samtools v1.9 [Li et al., 2009] to convert resulting SAM files into sorted and indexed BAM files. We generated anvi'o contigs databases (https://anvio.org/m/contigs-db) using the command 'anvi-gen-contigs-database', during which Prodigal v2.6.3 [Hyatt et al., 2010] identifies open reading frames. We created anvi'o profile databases of the mapping results for each metagenome using 'anvi-profile', which stores coverage and detection statistics, and 'anvi-merge' to combine all profiles together. To recover coverage and detection statistics for a given merged profile database, we used the program 'anvi-summarize' with '–init-gene-coverages' flag.

3.6.3 Criteria for detection of pBI143 and crAssphage in metagenome

Using mean coverage to assess the occurrence of a given sequence in a given sample based on metagenomic read recruitment can yield misleading insights due to non-specific read recruitment (i.e., recruitment of reads from metagenomes to a reference sequence from non-target populations). Thus, we relied upon the detection statistic reported by anvi'o, which is a measure of the proportion of the nucleotides in a given sequence that are covered by at least one short read. We considered pBI143 was present in a metagenome only if its detection value was 0.5 or above. Values of detection in metagenomic read recruitment results often follow a bimodal distribution for populations that are present and absent (see Supplementary Figure 2 in ref. [Utter et al., 2020]). Thus, 0.5 is a conservative cutoff to minimize a false-positive signal to assume presence.

3.6.4 Distinguishing the presence of distinct pBI143 versions in a genome or metagenome

We used the results of individual read recruitments to each known version of pBI143 to measure the coverage of each gene in pBI143 in samples that had a detection of greater than 0.9 and compared the ratio of the coverage of each gene. The pBI143 version where the genes

have the most even coverage ratio was considered the predominant version in that genome or metagenome.

3.6.5 Addition of tetQ *to pIB143*

To study transfer of pBI143 from *Phocaeicola vulgatus* MSK 17.67 to other Bacteroidales species, we added *tetQ* to pBI143. We PCR amplified *tetQ* from *Bacteroides caccae* CL03T12C61 and inserted it at the site shown in Figure 3.3 (all primers are listed in Supplementary Table 15). We PCR amplified the DNA regions flanking each side of this insertion site and the three PCR products were cloned into BamHI-digested pLGB13 [García-Bayona and Comstock, 2019]. We conjugally transferred this plasmid into *Phocaeicola vulgatus* MSK 17.67 and selected cointegrates on gentamycin 200 µg/ml and erythromycin 10 µg/ml. We passaged the cointegrate in non-selective medium and selected the resolvents by plating on anhydrotetracycline (75 ng/ml). We confirmed pIB143 contained *tetQ* by WGS the strain at the DFI Microbiome Metagenomics Facility.

3.6.6 Transfer assays

The recipient strains that received pBI143-*tetQ* were *Parabacteroides johnsonii* CL02T12C29 and *Bacteroides ovatus* D2, both erythromycin resistant and tetracycline sensitive. We grew the donor strain *Phocaeicola vulgatus* MSK 17.67 pBI143-*tetQ* and recipient strains to an OD_{600} of approximately 0.7 and mixed them at a 10:1 ratio (v:v) donor to recipient, and spotted 10 µl onto BHIS plates and grew them anaerobically for 20 h. We resuspended the co-culture spot in 1 mL basal media and cultured 10-fold serial dilutions on plates with erythromycin (to calculate number of recipients) or erythromycin and tetracycline (4.5 µg/ml) (to select for transconjugants). We performed multiplex PCR as described [Zitomersky et al., 2011, Evans et al., 2022] to confirm that TetR ErmR colonies were the recipient strain containing pBI143-*tetQ* (Figure 3.3).

3.6.7 Calculations of purifying selection and characterization of single nucleotide variants across metagenomes

We calculated dN/dS ratios as described previously [Kiefl et al., 2023]; details of which can also be found at https://merenlab.org/data/anvio-structure/chapter-IV/#calculating-dndstextgene-for-1-gene. To determine the mutational landscape of pBI143 across metagenomes, we first identified all variable positions present in the reference pBI143 sequences. We used the program 'anvi-script-gen-short-reads' to generate artificial short reads from the version 2 and version 3 pBI143 sequences and recruited these reads to the version 1 pBI143 sequence to generate data similar to the read recruitment from metagenomes. Then, we took read recruitment data from the global human gut metagenomes and sewage metagenomes mapped to version 1 pBI143. We ran 'anvi-gen-variability-profile' on the artificial read recruitment profile databases as well as on all profile databases from metagenomes with greater than 10X Q2Q3 coverage to identify all SNV positions. We compared the SNV positions in each gut or sewage metagenome to those present in our reference sequences and calculated the number of SNVs in each metagenome that did and did not match SNVs in the references. To calculate the number of non-consensus SNVs in a metagenome, we again ran the command 'anvi-gen-profile-database' on the same metagenomes, this time with the flags '–gene-caller-ids 0', '–min-departure-from-consensus 0.1', '–include-contig-names' and '–quince-mode', which produces a file that describes the variation in every single position across the reference and calculates the departure from consensus for each SNV with a departure from consensus greater than 0.1.

3.6.8 pBI143 structural and polymorphism analysis

To explore the impact of SAAVs on the protein structure of pBI143 MobA, we de novo predicted the monomer and dimer structures using AlphaFold 2 (AF) in ColabFold with default settings [Mirdita et al., 2022]. AlphaFold 2 confidently predicted the structure of the

catalytic domain but had low pLDDT scores for the coil domains and the dimer interactions. However, we explored variants across the whole dimer complex. Next, we integrated the pBI143 MobA AF structure into anvi'o structure by running 'anvi-gen-structure-database' [Delmont et al., 2019]. After that, we summarized SNV data as SAAVs from the metagenomic read recruitment data using 'anvi-gen-variability-profile –engine AA' to create a variability profile (https://anvio.org/m/variability-profile). Subsequently, we superimposed the SAAV data variability profile on the structure with 'anvi-display-structure' which filtered for variants that had at least 0.05 departure from consensus (reducing our metagenomic samples size from 2221 to 1706). Finally, we analyzed SAAVs that were prevalent in at least 5% of remaining samples. This left us with 21 SAAVs to analyze on the monomer. Next, we explored the relationship between SAAVs, relative solvent accessibility (RSA), and ligand binding residues in pBI143 MobA. To do this, we identified the homologous structure PDB 4LVI (MobM) by searching the high pLDDT pBI143 AF domain against the structure database PDB100 2201222 using Foldseek (https://search.foldseek.com/search). We next structurally aligned the pBI143 MobA AF structure to PDB 4LVI (MobM) [Pluta et al., 2014] using PyMol [Delano, 2002]. We chose the MobM structure 4LVI rather than a MobA because it had more structural and sequence homology to the pBI143 MobA catalytic domain AF structure than any PDB MobA structures. Additionally, we leveraged residue conservation values from the pre-calculated 4LVI ConSurf analysis to further explore ligand binding residues [Ben Chorin et al., 2020, Goldenberg et al., 2009].

3.6.9 Phylogenetic tree construction

To construct the pBI143 phylogeny, we identified pBI143 contigs from the isolate genome assemblies (Supplementary Table 4) using BLAST [Altschul et al., 1990]. We ran 'anvi-gen-contigs-database' on each pBI143 contig followed by 'anvi-export-gene-calls' with the flag '–gene-caller prodigal' and concatenated the resulting amino acid sequences. For the

bacterial host phylogeny, we ran 'anvi-gen-contigs-database' on each assembled genome. Then, we extracted ribosomal genes (see Supplementary Methods for details), aligned them with MUSCLE v3.8.1551 [Edgar, 2004a], trimmed the alignments with trimAl [Capella-Gutiérrez et al., 2009] using the flag '-gt 0.5', and computed the phylogeny with IQ-TREE 2.2.0-beta using the flags '-m MFP' and '-bb 1000' [Nguyen et al., 2015]. We visualized the trees with 'anvi-interactive' in '–manual-mode', and used the metadata provided by the Duchossois Family Institute to label the isolates to their corresponding donors. We used the 'geom_alluvium' function in ggplot2 to make the alluvial plots.

3.6.10 Construction and analysis of the network that describes shared single-nucleotide variants across mothers and infants

To investigate whether single-nucleotide variants suggest a vertical transmission of pBI143, we used metagenomic read recruitment results from four independent study that generated metagenomic sequencing of fecal samples collected from mothers and their infants in Finland [Yassour et al., 2018], Italy [Ferretti et al., 2018], Sweden [**?**], and the USA [Lou et al., 2021], against the pBI143 Version 1 reference sequence. The URL https://merenlab.org/data/pBI143 serves a fully reproducible workflow of this analysis. The primary input for this investigation was the anvi'o variability data, which is calculated by the anvi'o program 'anvi-profile', and reported by the anvi'o program 'anvi-gen-variability-profile' (with the flag '–engine NT'). The program 'anvi-gen-variability-profile' (https://anvio.org/m/anvi-gen-variability-profile) offers a comprehensive description of the single-nucleotide variants in metagenomes for downstream analyses. Since the *mobA* gene was conserved enough to represent all three versions of pBI143, for downstream analyses we limited the context to study variants to the *mobA* gene. The total number of samples in the entire dataset with at least one variable nucleotide position was 309, which represented a total of 102 families (Sweden: 52, USA: 24, Finland: 14, Italy: 11). We removed any sample that did not belong to a minimal complete family

(i.e., at least one sample for the mother, and at least one sample of her infant), which reduced the number of families in which both members are represented to 57 families (Sweden: 36, USA: 16, Finland: 3, Italy: 2). We further removed families if the coverage of the *mobA* gene was not 50X or more in at least one mother and one infant sample in the family, which reduced the number of families with both members represented and with a reliable coverage of *mobA* to 49 families (Sweden: 33, USA: 13, Finland: 2, Italy: 1), and from a given family, we only used the samples that had at least 50X for downstream analyses. We subsampled the variability data in R to only include the variable nucleotide position data for the final list of samples. We then used the list of single-nucleotide variants reported in this file to generate a network description of these data using the program 'anvi-gen-variability-network', which reports an 'edge' between any sample pairs that share a SNV with the same competing nucleotides. We then used Gephi [Bastian et al., 2009], an open-source network visualization program, with the ForceAtlas2 algorithm [Jacomy et al., 2014] to visualize the network. To quantify the extent of similarity between family members based on single-nucleotide patterns in the data, we generated a distance matrix from the same dataset using the 'pdist' function in Python's standard library with 'cosine' distances. We calculated the average distance of each sample to all other samples in its familial group ('within distance'), as well as the average distance from each sample to all other samples not present in their familial group ('between distance'). We subtracted the within distance from the between distance to get the 'subtracted distance'.

3.6.11 Metagenomic taxonomy estimation

We used Kraken 2.0.8-beta with the flags '–output', '–report', '–use-mpa-style', '–quick', '–use-names', '–paired' and '–classified-out' to estimate taxonomic composition of each metagenome [Wood et al., 2019]. For the genus-level taxonomic data, we filtered for metagenomes where the total number of reads recruited to a *Bacteroides*, *Parabacteroides* or

Phocaeicola genome was >1000 and the mean coverage of pBI143 was >20X. For the species-level taxonomic data, we used a cutoff of >0.1% percent of reads recruited to designate presence or absence of *B. fragilis* and >0.0001% for pBI143 based on the sizes of the genomes respectively (the *B. fragilis* genome is 3 orders of magnitude larger than pBI143).

3.6.12 Isogenic strain construction

We constructed the plasmid vector pEF108 (as shown in Figure 3.16) by PCR amplifying the desired sections with primers vec_108F, vec_108R, frag1_108F, frag1_108R, frag2_108R and frag2_108R (Supplementary Table 15) from existing plasmids. We assembled the three fragments via Gibson assembly using standard conditions described for NEB Gibson assembly mastermix. We selected for transconjugants on LB-carbenicillin (100ug/mL), then conjugated pEF108 into *B. fragilis* 638R and selected on BHIS + erythromycin 25ug/mL. Then, we counter-selected for recombination events in pEF108 to remove the markers and leave naive pBI143 by growing cells on *Bacteroides* minimal media plates (BMM) with 10mM p-chlorophenylalanine. We screened pBI143 positive, pheS-negative colonies via PCR and confirmed them by WGS. See Supplementary Methods for details.

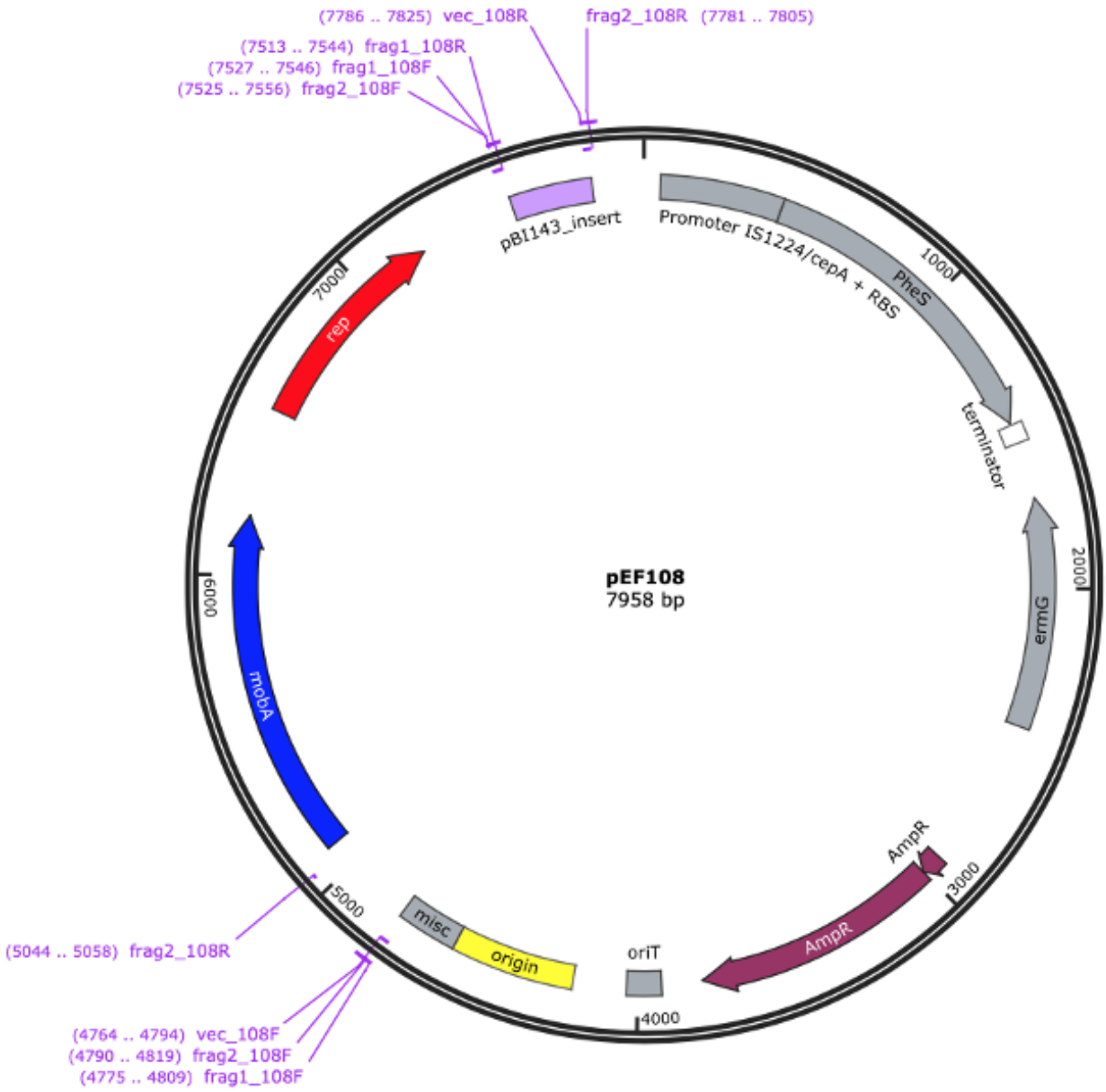

Figure 3.16: pEF108 construct assembled via Gibson Assembly as described above

3.6.13 Mouse competitive colonization assays

All animal experimentation was approved by the Institutional Animal Care and Use Committee at the University of Chicago. We gavaged three male and three female 10-15 week old germ-free C57BL/6J mice with a 1:1 inoculum of *B. fragilis* 638R:B. fragilis 638R pBI143. Males and females were housed separately in isocages and remained gnotobiotic for the duration of the experiment. We collected fecal pellets after eight and 14 days, diluted and plated on BHIS plates. We performed PCR on 48 colonies per mouse using a mixture of four primers (Supplementary Table 15), one set that amplifies a 1248-bp region of the 638R chromosome and a second set that amplifies a 662-bp segment of pBI143. PCR amplicons from all colonies included the 1248-bp region of the 638R chromosome and a subset also contained the amplicon for pBI143, allowing calculation of the ratio over time. The exact starting ratio for gavage was also calculated using this same PCR.

3.6.14 Approximate copy number ratio calculation in metagenomes

The first challenge to use metagenomic coverage values to study pBI143 copy number trends in human gut metagenomes is the unambiguous identification of gut metagenomes that appear to have a single possible pBI143 bacterial host beyond reasonable doubt. To establish insights into the taxonomic make up of the gut metagenomes we previously assembled, we first ran the program 'anvi-estimate-scg-taxonomy' (https://anvio.org/m/anvi-estimate-scg-taxonomy) with the flags '–metagenome-mode' (to profile every single single-copy core gene (SCG) independently) and '–compute-scg-coverages' (to compute coverages of each SCG from the read recruitment results). We also used the flag '–scg-name-for-metagenome-mode' to limit the search space for a single ribosomal protein. We used the following list of ribosomal proteins for this step as they are included among the SCGs anvi'o assigns taxonomy using GTDB, and we merged resulting output files: Ribosomal_S2, Ribosomal_S3_C, Ribosomal_S6, Ribosomal_S7, Ribosomal_S8, Ribosomal_S9, Ribosomal_S11, Riboso-

mal_S20p, Ribosomal_L1, Ribosomal_L2, Ribosomal_L3, Ribosomal_L4, Ribosomal_L6, Ribosomal_L9_C, Ribosomal_L13, Ribosomal_L16, Ribosomal_L17, Ribosomal_L20, Ribosomal_L21p, Ribosomal_L22,ribosomal_L24, and Ribosomal_L27A. For our downstream analyses that relied upon the merged SCG taxonomy and coverage output reported by anvi'o, we considered *Bacteroides*, *Parabacteroides* and *Phocaeicola* as the genera for candidate pBI143 host 'species', and only considered metagenomes in which a single species from these genera was present. Our determination of whether or not a single species of these genera was present in a given metagenome relied on the coverage of species-specific single-copy core genes (SCGs), where the taxonomic assignment to a given SCG resolved all the way down to the level of species unambiguously. We excluded any metagenome from further consideration if three or more candidate host species had positive coverage in any SCG in a metagenome. Due to highly conserved nature of ribosomal proteins and bioinformatics artifacts, it is possible that even when a single species is present in a metagenome, one of its ribosomal proteins may match to a different species in the same genus given the limited representation of genomes in public databases compared to the diversity of environmental populations. So, to minimize the removal of metagenomes from our analysis, we took extra caution with metagenomes before discarding them if only two candidate host species had positive coverage in any SCG. We kept such a metagenome in our downstream analyses only if one species was detected with only a single SCG, and the other one was detected by at least 8. In this case we assumed the large representation of one species (with 8 or more ribosomal genes) suggests the presence of this organism in this habitat confidently, and assumed the single hit to another species within the same genus was likely due to bioinformatics artifacts. It is the most unambiguous case if only a single candidate host species was detected in a given metagenome, but we still removed a given metagenome from further consideration if that single species had 3 or fewer SCGs in the metagenome. These criteria deemed 584 of 2580 metagenomes to have an unambiguous pBI143 host that resolved to 21 distinct species

names. We further removed from our modeling the metagenomes where the candidate host species did not occur in any other metagenome, which removed 5 of these candidate host species from further consideration. Finally, we further removed any metagenome in which the pBI143 coverage was less than 5X. Our final dataset to calculate the "approximate copy number ratio" (ACNR) of pBI143 in metagenomes through coverage ratios contained 579 metagenomes with one of 16 unambiguous pBI143 hosts. We calculated the ACNR by dividing the observed coverage of pBI143 by the empirical mean coverage of the host by averaging the coverage of all host SCGs found in the metagenome. To estimate the multiplicative difference in the geometric mean ACNR, we fit a linear model for the expected value of the logarithm of the ACRN, with disease status and bacterial host as predictors using rigr to construct the interval and estimate [Chen et al., 2022].

3.6.15 Oxidative stress experiments

We grew *B. fragilis* 214 in 5 mL BHIS for 15 hours in an anaerobic chamber. We inoculated 750 µL of this culture into 30mL BHIS in quintuplicate, and grew them for 3 hours. We divided the 30 mL into a further 5 culture flasks of 5 mL BHIS, and exposed each to oxygen with constant shaking for the appropriate time before returning the flask to the anaerobic chamber. At each time point, we took an aliquot of culture to determine the copy number of pBI143 in that sample. We extracted DNA from the cultures using a Thermal NaOH preparation [Conrad et al., 1996] to prepare them for qPCR. Copy number calculated can be found in Supplementary Table 13.

3.6.16 Estimating the pBI143 Plasmid Copy Number by Real-time qPCR

To evaluate plasmid copy number (CN), we developed a real-time TaqMan probe multiplex PCR assay to amplify both pBI143 and a single-copy *B. fragilis*-specific genomic reference gene (referred to as hsp [heat shock protein]) in the same reaction (see Supple-

mentary Information for details). We confirmed the primer and probe specificity to *B. fragilis* with BLAST searches against the NCBI and Ensembl databases, and experimental validation on 45 common gut isolates. For absolute quantification, we constructed a standard curve for each gene of interest by plotting the mean quantification cycle (Cq) values against log[quantity] of a dilution series of known gene of interest amount (range: 3 to 3×10^6 copies/reaction). We calculated the CN of pBI143 per genome equivalent (hsp), by dividing the absolute quantity of plasmid target by the absolute quantity of chromosomal target in the sample using the standard-curve (SC) method of absolute quantification [Lee et al., 2006]. Standard curves were generated with every qPCR run for analysis and to confirm PCR efficiency. Additional details for qPCR, including standard curves and controls, can be found in Supplemental Information. Supplementary Table 5 and Supplementary Table 15 report the relevant data and all primers, respectively.

3.6.17 qPCR analysis of animal, untreated sewage and water samples

Samples were tested with the pBI143 assay and two established assays for human fecal markers that included HF183 and Lachno3 [Olds et al., 2018]. For screening of animal samples to assess the presence of this plasmid in non-human gut microbiomes, archived DNA from a previous study [Feng et al., 2020] was analyzed and included 14 different animals encompassing 81 individual fecal samples. For assessment of fecal contamination of surface waters, archived DNA from 40 samples of river water [Lenaker et al., 2018, Corsi et al., 2021, US-gov] and freshwater beaches [Dila et al., 2022] were analyzed. These water samples were chosen from these previous studies that represented a range of contamination based on HF183 and Lachno3 levels. A total of 20 archived untreated sewage samples as reported in Olds et al. [Olds et al., 2018] were also analyzed for comparison. Since we were using archived samples from previous studies, we retested all the samples for the two human markers to account for any degradation. Additional details for qPCR, including standard curves and

controls, can be found in the Supplemental Information.

3.6.18 Visualizations

We used ggplot2 [Wickham, 2016] to generate all box and scatter plots. We generated coverage plots using anvi'o, with the program 'anvi-script-visualize-split-coverages'. We finalized the figures for publication using Inkscape, an open-source vector graphics editor (available from http://inkscape.org/).

3.7 Supplementary Tables

All supplementary tables are available at `https://doi.org/10.6084/m9.figshare.22336666`.

Table 3.1: The accompanying metadata for all publicly available metagenomes used in this study. This table contains 3 tabs. (1) healthy_gut: all healthy gut metagenomes. (2) IBD: all IBD gut metagenomes. (3) alternative_environment: all non-gut metagenomes.

Table 3.2: The nucleotide sequence and average nucleotide identity (ANI) calculations for all pBI143 contigs. This table has 3 tabs. (1) pBI143_sequences: the nucleotide sequence for the 3 reference versions of pBI143 assembled from metagenomes. (2) ANI information for the 3 reference sequences of pBI143. (3) additional_genes: the nucleotide sequences for pBI143 assembled from metagenomes with additional genetic material.

Table 3.3: Read recruitment data from metagenomes used in this study. This table has 4 tabs. (1) global adult gut metagenomes: coverage and detection data for the reference versions of pBI143 in global adult gut metagenomes (2) mother-infant metagenomes: coverage and detection data for the reference versions of pBI143 in mother and infant metagenomes (3) crassphage_comparison: coverage and detection data for crassphage in global adult gut metagenomes. (4) alternative_environments: coverage and detection data for the reference versions of pBI143 in non-human gut environments.

Table 3.4: The metadata for the Duchossois Family Institute bacterial isolate genomes used in this study.

Table 3.5: pBI143 copy number determination via qPCR. This table includes 2 tabs. (1) Seq DataSource: contains the Gen-Bank accession numbers and other data sources used in primer and probe development. (2) Hsp BLAST result: contains BLASTN results of the hsp nucleotide sequence against the 15 *Bacteroides fragilis* RefSeq complete genomes.

Table 3.6: The data for pBI143 copy number for all animal, environmental and sewage samples as measured via qPCR. This table has 3 tabs. (1) animal_copy_number: contains the data showing sample and copy number of pBI143 in animal fecal samples. (2) environmental_copy_number: contains the data showing sample and copy number of pBI143 in water samples. (3) sewage_copy_number: contains the data showing sample and copy number of pBI143 in sewage samples.

Table 3.7: All the data necessary for quantifying number and type of SNV in gut and sewage metagenomes. Variability profiles are generated by anvi'o to describe the variation found across all contigs of interest; for more information see https://merenlab.org/2015/07/20/analyzing-variability. This table has 9 tabs. (1) artificial_reads_var_profile: The variability profile generated following artificial short read generation and read recruitment of pBI143 Version 2 and 3 to Version 1 (see Methods). (2) global_mg_var_profiles The variability profile generated following read recruitment of all global gut metagenomes to pBI143 Version 1. (3) sewage_mg_var_profiles: The variability profile generated following read recruitment of all global sewage metagenomes to pBI143. (4) matching_SNVs_gut: The number of SNVs that do or do not match one of the reference versions of pBI143 in global gut metagenomes. (5) matching_SNVs_sewage: The number of SNVs that do or do not match one of the reference versions of pBI143 in global sewage metagenomes. (6) gut_var_profile_quince_mode: This file does not fit in excel. Link to online data to regenerate single nucleotide variant data at every position of pBI143 across all global gut metagenomes (for more details on 'quince-mode' see https://merenlab.org/2015/07/20/analyzing-variability/#parameters-to-refine-the-output). (7) sewage_var_profile_quince_mode: single nucleotide variant data at every position of pBI143 across all global sewage metagenomes. (8) gut_non-consensus_SNVs: Data about the plasmid version and number of non-consensus SNVs in each global gut metagenome. (9) sewage_non-consensus_SNVs: Data about the plasmid version and number of non-consensus SNVs in each global sewage metagenome.

Table 3.8: This table contains SNV variability profiles for visualizing SAAVs on the pBI143 AF structure. This table has 3 tabs: (1) merged_variability: contains all SNV variability data calculated with 'anvi-gen-variability-profile –engine AA' which summarized metagenomic read recruitment results to pBI143; (2) merged_variability_filtered: filtered version of merged_variability that reflects the SAAV data visualized on the pBI143 structure in Figure 3.6D; (3) most_prevelent_SAAVs: this tab contains a list of all SAAVs and their residue positions that are prevalent in at least 5% of samples.

Table 3.9: The necessary data to generate pBI143 and isolate genome phylogenies. This table has 5 tabs. (1) amino_acid_repA_mobA_concat: the concatenated MobA and RepA sequences from all 82 isolate genomes. Concatenated genes are separated by 'XXX'. (2) repA_mobA_treefile: the treefile generated from the concatenated *mobA* and *repA* sequences. (3) amino_acid_SCG_concat: the concatenated ribosomal protein sequences from all 82 isolate genomes. Concatenated genes are separated by 'XXX'. (4) SCG_treefile: the treefile generated from the concatenated ribosomal protein sequences. (5) species_donor_information: the associated isolate data that matches the donor, species and pBI143 version.

Table 3.10: The data necessary for generating and quantifying the mother-infant network based on single nucleotide variants. This table has 8 tabs. (1) Finalnd_variability_profile: data for all single nucleotide variants (SNVs) present in pBI143 in Finnish mother and infant metagenomes. (2) Sweden_variability_profile: data for all single nucleotide variants present in pBI143 in Swedish mother and infant metagenomes. (3) Italy_variability_profile: data for all single nucleotide variants present in pBI143 in Italian mother and infant metagenomes. (4) USA_variability_profile: data for all single nucleotide variants present in pBI143 in American mother and infant metagenomes. (5) network_data: data used to generate the network. (6) distance_matrix_cosine: distance matrix calculated from network data used for quantification of distances between samples. (7) subtracted_distance_df: quantified distance between mother and infant samples based on cosine distance matrix. (8) summary of pBI143 version maintenance in infants over the sampling period.

Table 3.11: Kraken data. This table contains 2 tabs. (1) Kraken data for all *Bacteroides* (this includes *Phocaeicola* with old *Bacteroides* genus names) and *Parabacteroides* taxa in global gut metagenomes and the corresponding pBI143 coverage in each of these metagenomes. (2) Kraken data for the number of reads recruited to a *B. fragilis* compared to the coverage of pBI143 in the same metagenomes.

Table 3.12: pBI143 competition experiment additional data. This table contains 2 tabs. (1) The data for the mouse competition experiments. (2) The pBI143 maintenance in culture data.

Table 3.13: The data for pBI143 copy number for each timepoint and condition of the *Bacteroides fragilis* stress experiments in culture as measured via qPCR. This table has 2 tabs. (1) 214_oxidative_stress_qPCR_data: contains data on the copy number for each test condition for the *Bacteroides fragilis* 214 strain. (2) R16_oxidative_stress_qPCR_data: contains data on the copy number for each test condition for the *Bacteroides fragilis* R16 strain.

Table 3.14: The calculated ACNR and necessary data for these calculations. This table has 4 tabs. (1) Coverage_ratio_data: the final ACNR for all predicted single hosts of pBI143 in metagenomes. (2) pBI143_healthy: the coverage of pBI143 in healthy gut metagenomes. (3) pBI143_IBD: the coverage of pBI143 in IBD gut metagenomes. (4) SCG_taxonomy: Link to files containing SCG coverage data.

Table 3.15: The names and sequences of all primers and probes used in this study.

3.8 Supplemental Methods

3.8.1 Phylogenetic tree construction

To construct the pBI143 phylogeny, we identified pBI143 contigs from the isolate genome assemblies (Supp. Table 3.4) using BLAST (cite). We ran 'anvi-gen-contigs-database' on each pBI143 contig followed by 'anvi-export-gene-calls' with the flag '–gene-caller prodigal' and concatenated the resulting amino acid sequences. For the bacterial host phylogeny, we ran 'anvi-gen-contigs-database' on each assembled genome, then extracted ribosomal genes (Ribosomal_L1, Ribosomal_L13, Ribosomal_L14, Ribosomal_L16, Ribosomal_L17, Ribosomal_L18p, Ribosomal_L19, Ribosomal_L2, Ribosomal_L20, Ribosomal_L21p, Ribosomal_L22, Ribosomal_L23, Ribosomal_L27, Ribosomal_L27A, Ribosomal_L28, Ribosomal_L29, Ribosomal_L3, Ribosomal_L32p, Ribosomal_L35p, Ribosomal_L4, Ribosomal_L5, Ribosomal_L6, Ribosomal_L9_C, Ribosomal_S10, Ribosomal_S11, Ribosomal_S13, Ribosomal_S15, Ribosomal_S16, Ribosomal_S17,
Ribosomal_S19, Ribosomal_S2, Ribosomal_S20p, Ribosomal_S3_C, Ribosomal_S6, Ribosomal_S7, Ribosomal_S8, Ribosomal_S9, ribosomal_L24) using the command 'anvi-get-sequences-for-hmm-hits' with the flags '–return-best-hit', '–get-aa-sequences', '–concatenate' and '–min-num-bins-gene-occurs 82' and '–hmm-source Bacteria_71' [Lee et al., 2019]. For both phylogenies, we aligned the genes with MUSCLE v3.8.1551 (Edgar 2004), trimmed the alignments with trimAl [Capella-Gutiérrez et al., 2009] using the flag '-gt 0.5', and computed the phylogeny with IQ-TREE 2.2.0-beta using the flags '-m MFP' and '-bb 1000'. We visu-

alized the trees with 'anvi-interactive' in '–manual-mode', and used the metadata provided by the Duchossois Family Institute to label the isolates to their corresponding donors. All data for the phylogenies can be found in Supp Table 3.9. We used the 'geom_alluvium' function in ggplot2 to make the alluvial plots.

3.8.2 Isogenic B. fragilis *strain construction +/- pBI143*

We constructed the plasmid vector pEF108 (as shown in Supp Figure pEF108_plasmid_map) by PCR amplifying the desired sections with primers vec_108F, vec_108R, frag1_108F, frag1_108R, frag2_108R and frag2_108R (Supplemental Table 15) from existing plasmids. We assembled the three fragments via Gibson assembly using standard conditions described for NEB Gibson assembly mastermix (https://www.neb.com/protocols/2012/12/11/gibson-assembly-protocol-e5510). See pEF108 plasmid map below (Figure 3.16). We transformed the construct into E. coli S17 λpir via electroporation with a BioRad micropulser using 0.1cm cuvettes and selected on LB-carbenicillin 100ug/mL agar plates. We conjugally transferred pEF108 from E. coli S17 λpir into *B. fragilis* 638R. Briefly, we grew the donor and recipient strains in LB-carbenicillin 100ug/mL broth and vitamin K supplemented brain-heart infusion media (BHIS) broth respectively for 12-15 hours. We spun down the cultures and resuspended in BHIS and combined at a 1:5 ratio of donor to recipient. We spotted the donor and recipient mixture onto BHIS plates and incubated for 12 hours aerobically. We scraped the cells off the plate, resuspended in BHIS, then plated on BHIS + erythromycin 25ug/mL. We restreaked the colonies and validated the presence of the construct via PCR and sanger sequencing. Next, we wanted to select for cells where a recombination event had removed the vector containing pheS, ampicillin and erythromycin resistance and left pBI143 in its native form. Colonies with the full sized pEF108 construct were grown in *Bacteroides* minimal media (BMM) with 10mM p-chlorophenylalanine (PCPA) broth for 24 hours, and plated onto BMM + 10mM PCPA.

PCPA prevents the growth of cells that are expressing the pheS gene. Colonies that grew on BMM + 10mM PCPA were screened for presence of pBI143 and absence of the vector containing the pheS negative selection marker.

3.8.3 Primer and probe design for B. fragilis *hsp/pBI143 copy number qPCR*

We aligned the canonical pBI143 plasmid DNA sequence from GenBank, whole genome assemblies and metagenome-assembled genomes (MAGs) as outlined in Supplemental Table 15. The two known pBI143 genes, rep and mob, are common plasmid features across the bacterial kingdom (DelSolar et al., 1998; Wawrzyniak et al, 2017) and use of either gene alone had high potential for cross-amplification from other mobile genetic elements. To ensure pBI143 specificity, we designed our primer set so that the forward primer was located within the 3' region of the rep gene (Table primers pBI143_F) while the reverse primer was located in the intergenic region (Table primers pBI143_R). This required that two conditions would have to be met for amplification to occur: (1) presence of the gene of interest and (2) homology to the pBI143 plasmid backbone. Despite the existence of plasmid variants differing across the rep gene, the 3' region used in the forward primer design is conserved across the source sequences. The 38-yr old canonical pBI143 sequence (U30316.1) demonstrated greater sequence variation in intergenic regions than more contemporary sequences as determined by existing publicly-available metagenomic data. In designing the reverse primer, we strategically excluded U30316.1 in favor of using the more recent pBI143 sequences. The FAM-labeled hydrolysis probe was designed within a conserved plasmid feature, the 56-bp inverted repeat (IR) region and in concert with the designed primers, amplified/detected a 145-bp product (Supplemental Table 15). The choice to use rep, over mob, as our target was based on (1) its conservancy in the 3' gene region and (2) technical difficulties in optimizing a mob-based assay.

To perform relative quantification experiments and to normalize bacterial cell numbers

between samples for the purposes of determining the copy number of pBI143 per genome equivalent required identifying a suitable genomic reference gene. A key prerequisite was identifying a single copy gene present in the genus *Bacteroides*, but absent in other common gastrointestinal (GI) tract organisms. We employed a pangenomic analysis of 12 *Bacteroides* and 15 other human commensal gut microbe genomes to determine potential candidates and used the program 'anvi-run-workflow' with '–workflow pangenomics'. Anvi'o pangenomics workflow is detailed elsewhere [Delmont et al., 2018]. Briefly, the pangenomic analysis used the NCBI's BLAST [Altschul et al., 1990] to quantify similarity between each pair of genes, and the Markov Cluster algorithm (MCL) [Enright et al., 2002] (with inflation parameter of 2) to resolve clusters of homologous genes. The program 'anvi-summarize' created summary tables for pangenomes and 'anvi-display-pan' provided interactive visualizations of pangenomes. Using the criteria of (1) maximum functional homogeneity of 0.99 and (2) maximum geometric homogeneity of 0.99, we identified 35 gene clusters for further interrogation. The corresponding DNA sequences were gathered using the program 'anvi-get-sequences-for-gene-cluster' and aligned using Kalign (https://www.ebi.ac.uk/Tools/msa/kalign) for multiple sequences.

Despite our initial desire to identify a target that could serve as a reference gene across all *Bacteroides* spp., we found that within these 35 gene clusters, the percent sequence identity dropped from >98% in *B. fragilis* to 50-85% in non-B. fragilis sequences. Therefore, we focused on finding a *B. fragilis*-specific target by further requiring 100% coverage and 99.8% - 100% percent identity across all *B. fragilis* genomes. Five gene clusters qualified; a single gene cluster demonstrated 100% identity across all seven *B. fragilis* genomes used in the pangenome. Using this gene clusters' 177-bp nucleotide sequence, we performed BLASTN (Zhang et al., 2000) on the NCBI Reference Sequence (RefSeq) Database (release 99, 3/2/2020), using Megablast (optimize for highly similar sequences) to conduct a systematic and thorough in-silico assessment of *B. fragilis* specificity. A list of the 17,785 complete

genomes was downloaded from RefSeq (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/, accessed 3/31/2020) and found to contain 39 *Bacteroides* genomes; of which, 15 were catalogued as *B. fragilis*. 6 organisms labeled as *B. fragilis* in this collection appeared to be a different species, given their overall ANI to *B. fragilis* genomes was only 84-85% and we disregarded these organisms. The only significant BLAST alignments of the hsp gene were to the nine true *B. fragilis* genomes. These genomes annotated the gene cluster protein product as hypothetical or conserved hypothetical protein (n=2); DUF4250 domain-containing protein (n=5); or heat shock protein (n=2) (Supplemental Table 15).

Based on its specificity to *B. fragilis* only, the previous use of heat shock proteins to discriminate amongst anaerobes [Sakamoto and Ohkuma, 2010], and the documented conservancy of these molecules [Gallie et al., 2002], this gene cluster was chosen as our candidate reference gene and is hereafter referred to as hsp. The primers and Cy5-labeled hydrolysis probe were designed to amplify/detect a 101-bp product (Supplemental Table 15).

3.8.4 qPCR analytical specificity

We assessed the *in vitro* analytical specificity of the hsp qPCR assay using DNA templates extracted from a collection of 41 bacterial isolates (13 aerobes, 28 anaerobes; representing 16 commonly encountered commensal gastrointestinal tract genera). hsp was not detected in any aerobic or anaerobic microorganisms, except for the collections' four *B. fragilis* isolates. The lack of amplification in other *Bacteroides* spp., including B. ovatus (n=3), B. thetaiotamicron (n=2), B. uniformis (n=1) and B. vulgatus (n=3) corroborated the previous *in silico* results.

3.8.5 qPCR experimental conditions

We performed real-time PCR amplification on a LightCycler 480 II system (Roche Diagnostics), using 10 microliter reactions consisting of 2X PrimeTime Gene Expression master mix (Integrated DNA Technologies, Coralville, IA), 0.8 μM pBI143_R, and 0.4 μM of pBI143_F, *B.fragilis*_hsp_F, and *B.fragilis*_hsp_R primers. We used optimized probe concentrations of 0.2 μM HSP and 0.4 μM pBI143 probe. Probe and primer sequences are outlined in Supplemental Table 15. We assessed triplicate PCR reactions using genomic DNA templates (2-ul volume per reaction) and the optimal cycling conditions of an initial denaturation step of 95°C for 3-min, followed by 40 cycles of 95°C for 15-s (denaturation) and 60°C for 60-s (annealing and extension).

3.8.6 qPCR assay performance characteristics

We constructed a single plasmid, by standard recombinant DNA methods, containing both the entire pBI143 plasmid and the reference gene (hsp) DNA and then transformed the plasmid into E. coli EC100D. The DNA concentration of the recombinant plasmid was converted to the number of template copies using the mass of the plasmid molecule [Whelan et al., 2003]. Using a 10-fold serial dilution series of the plasmid DNA standard (ranging from 3 to 3×10^6 copies/reaction), we constructed standard curves for both chromosomal reference gene and the target plasmid.

Each targets' lower limit of detection (LOD) was determined to be 30 copies per reaction, as defined by the first dilution that detects 95% of positive samples [Bustin et al., 2009]. We validated a linear dynamic range of six orders of magnitude for each target, and this range was then used in further assay performance metric calculations.

The primer amplification efficiencies were determined by standard procedure [Bustin et al., 2009] that includes (1) making a log10 dilution series of target DNA, (2) calculating a linear regression based on the targets' mean Cq data points and (3) inferring the efficiency

from the slope of the line. Over 11 experiments, mean Cq values were derived and PCR efficiencies were calculated as 97.8% and 98.98% for pBI143 and hsp, respectively. We demonstrate less than a 5.2% difference when comparing same run target and reference gene efficiency, demonstrating the two genes amplify similarly.

3.8.7 qPCR for animal, water and sewage samples

Quantification of two established human specific markers, HF183 and Lachno3 following methods as published previously (cite olds). The HF183 marker is specific for human *Bacteroides* and is targeted by several assays (green:, seurnick,); including the HB assay (olds) used here. Standard curves were generated based on a minimum of 16 runs (in triplicate) and consisted of linearized plasmids containing the HF183, Lachno3, and pBI143 target sequences. The plasmids used for the standard curves were purified using a Qiagen mini plasmid prep kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. Standard curves were run with DNA serially diluted from 1.5 x 106 to 1.5 x 101 copies/reaction with resulting linear equations and efficiencies as follows:

HF183: Slope: -3.37, Y-intercept 39.363, R2 0.998, Eff% 98.18

Lachno3: Slope: -3.42, Y-intercept 38.13, R2 0.999, Eff% 95.92

pBI143: Slope: -3.43, Y-intercept 39.363, R2 0.999, Eff% 95.90

For each run, two of the standard concentrations (as quality assurance for the standard curve, sterile water (as negative control) and each sample was run in duplicate in a final volume of 25 μL with a final concentration of 1 μM for each primer, 80 nM for the probe, 5 μL of sample DNA, and 12.5 μL of 2X Taqman® Gene Expression Master Mix Kit (Applied Biosystems; Foster City, CA). DNA template was added as undiluted sample for surface water and animal samples, and 1:100 dilutions of sewage samples. Amplification conditions consisted of the following cycles: 1 cycle at 50°C for 2 minutes to activate the uracil-N-glycosylase (UNG); 1 cycle at 95°C for 10 minutes to inactivate the UNG and activate the

Taq polymerase; 40 cycles of 95° C for 15 seconds; and 1 minute at 60°C for HF183 or 1 minute at 64°C for Lachno3 using a StepOne Plus™ instrument (Applied Biosystems; Foster City, CA).

Water samples that amplify after 35 cycles were considered below the standard curve limit of 15 CN/reaction and were therefore considered below limit of quantification (BLQ). For water samples where 400 ml was filter for extraction, this value is 113 CN/100 ml. All no template controls (water) showed no amplification.

3.8.8 pBI143 is specific to the human gut and hosted by a wide range of Bacteroidales populations

pBI143 was absent from all marine samples (n=241, 72 of 63 billion reads match, avg. detection: <0.01, avg. coverage: <0.01X), macaques (n=19, 1885 of 2.6 billion reads match, avg. detection: <0.1, avg. coverage: <0.2X) and pets (n=125, 0 of 1.2 billion reads match, avg. detection: 0, avg. coverage: 0X) (Supplementary Table 3). As expected, pBI143 was present in sewage (n=77; 98,914 of 18 billion reads match, avg. detection: 0.9, avg. coverage: >70X) (Figure 3.4, Supplementary Table 3). Given the near-absolute absence of pBI143 in non-human associated habitats, we also screened metagenomes from human skin and oral cavity 70. Unlike the extremely high presence of pBI143 in the human gut (n=100, 5.1 million of 9 billion reads match, avg. detection: 0.8, avg. coverage: >2,000X), pBI143 was poorly detected both in samples from skin (n=54, 9 thousand of 2.9 billion reads match, avg. detection: 0.4, avg. coverage: 6.4X) and the oral cavity (n=418, 8 thousand of 37 billion reads match, avg. detection: 0.2, avg. coverage: 0.9X) (Supplementary Table 3).

3.8.9 Data Availibility

All genomes and metagenomes are available via the NCBI Sequence Read Archive, and the accession numbers for metagenomes and genomes are reported in Supplementary Table

1 and Supplementary Table 4, respectively. The data object identifier (DOI) 10.6084/m9.figshare.22336666 gives access to Supplementary Table and Supplementary Information files. Additional DOIs for anvi'o data products that describe metagenomic read recruitment results as well as sequences for pBI143 versions and bioinformatics workflows are accessible at the URL https://merenlab.org/data/pBI143 to reproduce our findings. Bacterial cultures for host range investigations, which are listed in Supplementary Table 4, are courtesy of The Duchossois Family Institute (https://dfi.uchicago.edu/). *B. fragilis* strains with pBI143 are available upon request from the Comstock Lab collection (https://comstocklab.uchicago.edu/).

All supplementary tables are available at `https://doi.org/10.6084/m9.figshare.22336666`.

3.8.10 Acknowledgements

We thank the members of the Meren Lab (https://merenlab.org) and Comstock Lab (https://comstocklab.uchicago.edu/) for helpful discussions, Jason Koval for help procuring bacterial cultures, and the Duchossois Family Institute WGS facility for sequencing constructs. We thank Melinda Bootsma for help with the qPCRs on water and sewage samples. ECF acknowledges support from the University of Chicago International Student Fellowship, and ADW acknowledges support from NIGMS R35 GM133420. Additional support for ECF came from an NIH NIDDK grant (RC2 DK122394) to EBC. Authors thank The University of Chicago Center for Data and Computing for their support. This project was funded by University of Chicago start-up funds to AME.

CHAPTER 4

CONCLUSION

4.1 Summary of contributions

Cryptic plasmids have been vastly understudied due to the difficulty of identifying them from naturally occurring environments and the inconveniences they present for designing experiments. The first half of this work demonstrated how new plasmids can be robustly identified from complex samples using a machine learning model trained on gene families from reference plasmids. Through applying this model to predict plasmids from globally distributed human gut metagenomes, I showed that our current datasets have systematically missed many plasmids, and that this new approach increases the number of known human gut plasmids by a factor of ten. With this large dataset, it became possible to identify evolutionary relationships between plasmids based on a shared circular backbone, even when they occurred in individuals living in geographically distant locations. As a result of defining plasmid systems, I was able to identify genes that were incorporated into plasmid backbones and that corresponded to distinct evolutionary pressures present in different geographic locations.

The second half of this work, I dive deep into the analysis of the most prevalent, experimentally verified plasmid across all those we identified in human gut metagenomes. I show that this plasmid, pBI143, is present primarily in industrialized countries and has unprecedented levels of abundance for such a small mobile genetic element. Although globally distributed across individuals from industrialized countries, this plasmid is specific to the human gut, and therefore can be used to identify human fecal contamination in water. I examine the population structure of pBI143 across individuals, and find that while it is primarily under purifying selection, the few prevalent non-synonymous variants likely contribute to the MobA protein binding different origin of transfer sequences. Surprisingly, pBI143 is

also monoclonal within an individual gut, which I determined was most likely due to priority effects of the version usually acquired from the mother. Given the highly conserved nature of pBI143 and the positive correlation with its bacterial hosts in metagenomes, I hypothesized that it played a beneficial role in the lifestyle of its Bacteroidales hosts. However, my competition experiments demonstrated that instead pBI143 most likely acts as a well-adapted parasite. Further exploration of my metagenomic data showed that pBI143 transiently acquires other genes which it likely transfers to new hosts before reverting back to its parasitic state. Genetic parasites, indeed most mobile genetic elements, respond to microbial host stress. In the final part of Chapter 3, I experimentally demonstrate that pBI143 increases its copy number when the host undergoes oxidative stress. The increase in copy number was also apparent in inflammatory bowel disease patients, a naturally occuring form of oxidative stress in the human gut.

4.2 Future directions

My work focuses on identifying human gut plasmids, identifying their evolutionary relationships, and deeply characterizing one plasmid, chosen from 68,350. I chose this plasmid based on its cryptic nature, presence in plasmid systems, and widespread distribution, but as this dissertation shows, there are thousands of plasmids whose functions remain complete mysteries, and plasmid-specific phenomena, such as plasmid systems, that have not been fully investigated. If I were to do another PhD in microbial ecology, one of my primary goals would be to experimentally test the concept of plasmid systems to show active acquisitions of new genes by backbone plasmids to form cargo plasmids. Given the large dataset of plasmid systems available as I describe in Chapter 2, I would choose a system that is present in a tractable organism for ease of genetic manipulations, and encodes for conjugation machinery to transfer between cells. These experiments could utilize a clear selective pressure like sub-lethal concentrations of antibiotics, and determine if the backbone plasmid from

a plasmid system is capable of uptaking antibiotic resistance present on the bacterial host chromosome or on a non-transmissible MGE, then transferring it to neighboring cells that previously lacked resistance.

Another goal for my hypothetical second PhD would be to examine the functions carried on larger versions of pBI143 assembled from gut metagenomes. I would synthesize these larger plasmids, create isogenic strain sets similar to those described in Chapter 3, then use biolog assays (plates with large arrays of environmental conditions and nutrients) to test the conditions in which these additional functions could benefit host cells.

The final goal for this second PhD would be to validate the practical applications of pBI143. To determine its potential to be used as a drug delivery system for complex microbial communities, I would insert a common resistance marker into the pBI143 backbone, introduce this into Bacteroides, and colonize an SPF mouse with this strain. To determine if the antibiotic-carrying pBI143 transferred and replicated in new hosts, I would use a combination of Hi-C sequencing and cultivation on antibiotic-containing media. Another application of pBI143 is to measure gut inflammatory stress, which is promising according to metagenomic and in vitro data, but lacks validation in a mouse model. I would use a murine model of inflammatory bowel disease to validate that oxidative stress also increases the copy number of pBI143 in vivo. In parallel, I would also analyze metagenomic data from patients with other gut stressors like antibiotic treatment, colorectal cancer or Clostridium difficile infections to determine if pBI143 copy number is elevated in all stressful conditions or if it is limited to oxidative stress caused by inflammatory bowel disease.

Collectively this dissertation demonstrates the power of combining data-driven insights with tailored experiments to bring a simple observation, initially made possible by state-of-the-art plasmid classification, to a highly characterized component of the human gut microbiome, and finally, provides a foundation for future exploration of cryptic human gut plasmids.

REFERENCES

A framework for human microbiome research. *Nature*, 486(7402):215–221, June 2012.

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

Daniel Aguirre de Cárcer. The human gut pan-microbiome presents a compositional core formed by discrete phylogenetic units. *Sci. Rep.*, 8(1):14069, September 2018.

Basem Al-Shayeb, Rohan Sachdeva, Lin-Xing Chen, Fred Ward, Patrick Munk, Audra Devoto, Cindy J Castelle, Matthew R Olm, Keith Bouma-Gregson, Yuki Amano, Christine He, Raphaël Méheust, Brandon Brooks, Alex Thomas, Adi Lavy, Paula Matheus-Carnevali, Christine Sun, Daniela S A Goltsman, Mikayla A Borton, Allison Sharrar, Alexander L Jaffe, Tara C Nelson, Rose Kantor, Ray Keren, Katherine R Lane, Ibrahim F Farag, Shufei Lei, Kari Finstad, Ronald Amundson, Karthik Anantharaman, Jinglie Zhou, Alexander J Probst, Mary E Power, Susannah G Tringe, Wen-Jun Li, Kelly Wrighton, Sue Harrison, Michael Morowitz, David A Relman, Jennifer A Doudna, Anne-Catherine Lehours, Lesley Warren, Jamie H D Cate, Joanne M Santini, and Jillian F Banfield. Clades of huge phages from across earth's ecosystems. *Nature*, 578(7795):425–431, February 2020.

Basem Al-Shayeb, Marie C Schoelmerich, Jacob West-Roberts, Luis E Valentin-Alvarado, Rohan Sachdeva, Susan Mullen, Alexander Crits-Christoph, Michael J Wilkins, Kenneth H Williams, Jennifer A Doudna, and Jillian F Banfield. Borgs are giant genetic elements with potential to expand metabolic capacity. *Nature*, 610(7933):731–736, October 2022.

Brian P Alcock, Amogelang R Raphenya, Tammy T Y Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, Sally Y Min, Anatoly Miroshnichenko, Hiu-Ki Tran, Rafik E Werfalli, Jalees A Nasir, Martins Oloni, David J Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N Sharma, Emily Bordeleau, Andrew C Pawlowski, Haley L Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L Winsor, Robert G Beiko, Fiona S L Brinkman, William W L Hsiao, Gary V Domselaar, and Andrew G McArthur. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, 48(D1):D517–D525, January 2020.

Ana Alonso, Patricia Sanchez, and Jose L Martinez. Environmental selection of antibiotic resistance genes. minireview, 2001.

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3), October 1990.

Katherine R Amato, Carl J Yeoman, Angela Kent, Nicoletta Righini, Franck Carbonero, Alejandro Estrada, H Rex Gaskins, Rebecca M Stumpf, Suleyman Yildirim, Manolito

Torralba, Marcus Gillis, Brenda A Wilson, Karen E Nelson, Bryan A White, and Steven R Leigh. Habitat degradation impacts black howler monkey (alouatta pigra) gastrointestinal microbiomes. *ISME J.*, 7(7):1344–1353, March 2013.

William B Andreopoulos, Alexander M Geller, Miriam Lucke, Jan Balewski, Alicia Clum, Natalia N Ivanova, and Asaf Levy. Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. *Nucleic Acids Res.*, 50(3):e17, February 2022.

Dmitry Antipov, Nolan Hartwick, Max Shen, Mikhail Raiko, Alla Lapidus, and Pavel A Pevzner. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*, 32(22):3380–3387, November 2016.

Dmitry Antipov, Mikhail Raiko, Alla Lapidus, and Pavel A Pevzner. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.*, 29(6):961–968, June 2019.

Dmitry Antipov, Mikhail Raiko, Alla Lapidus, and Pavel A Pevzner. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics*, 36(14):4126–4129, August 2020.

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. *Microb Genom*, 3(10):e000128, October 2017.

Sergio Arredondo-Alonso, Malbert R C Rogers, Johanna C Braat, Tess D Verschuuren, Janetta Top, Jukka Corander, Rob J L Willems, and Anita C Schürch. Mlplasmids: A user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb. Genom.*, 4(11), November 2018.

Sabrina A Attéré, Antony T Vincent, Mégane Paccaud, Michel Frenette, and Steve J Charette. The role for the small cryptic plasmids as moldable vectors for genetic innovation in aeromonas salmonicida subsp. salmonicida, 2017.

Fredrik Bäckhed, Josefine Roswall, Yangqing Peng, Qiang Feng, Huijue Jia, Petia Kovatcheva-Datchary, Yin Li, Yan Xia, Hailiang Xie, Huanzi Zhong, Muhammad Tanweer Khan, Jianfeng Zhang, Junhua Li, Liang Xiao, Jumana Al-Aama, Dongya Zhang, Ying Shiuan Lee, Dorota Kotowska, Camilla Colding, Valentina Tremaroli, Ye Yin, Stefan Bergman, Xun Xu, Lise Madsen, Karsten Kristiansen, Jovanna Dahlgren, and Jun Wang. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe*, 17(6):852, June 2015.

H J Balbi. Chloramphenicol: A review, 2004.

Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A Alekseyev, and Pavel A Pevzner. SPAdes: A new genome assembly algorithm and its applications to Single-Cell sequencing. *J. Comput. Biol.*, 19(5):455, May 2012.

F Baquero. Low-level antibacterial resistance: a gateway to clinical resistance. *Drug Resist. Updat.*, 4(2):93–105, April 2001.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. *ICWSM*, 3(1):361–362, March 2009.

Daniel C Baumgart and Simon R Carding. Inflammatory bowel disease: cause and immunobiology. *Lancet*, 369(9573):1627–1640, May 2007.

John W Beaber, Bianca Hochhut, and Matthew K Waldor. SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature*, 427(6969):72–74, January 2004.

Adi Ben Chorin, Gal Masrati, Amit Kessel, Aya Narunsky, Josef Sprinzak, Shlomtzion Lahav, Haim Ashkenazy, and Nir Ben-Tal. ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. *Protein Sci.*, 29(1): 258–267, January 2020.

Bjorn Berendsen, Linda Stolker, Jacob de Jong, Michel Nielen, Enkhtuya Tserendorj, Ruuragchas Sodnomdarjaa, Andrew Cannavan, and Christopher Elliott. Evidence of natural occurrence of the banned antibiotic chloramphenicol in herbs and grass. *Anal. Bioanal. Chem.*, 397(5):1955, 2010.

C T Bergstrom, M Lipsitch, and B R Levin. Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics*, 155(4):1505–1519, August 2000.

Bryan Bishé, Arnaud Taton, and James W Golden. Modification of RSF1010-Based Broad-Host-Range plasmids for improved conjugation and cyanobacterial bioprospecting. *iScience*, 20:216–228, October 2019.

Ben E Black. *Centromeres and Kinetochores: Discovering the Molecular Mechanisms Underlying Chromosome Inheritance*. Springer, August 2017.

Louis-Marie Bobay and Howard Ochman. Biological species in the viral world. *Proc. Natl. Acad. Sci. U. S. A.*, 115(23):6040–6045, June 2018.

Joseph Bondy-Denomy, Jason Qian, Edze R Westra, Angus Buckling, David S Guttman, Alan R Davidson, and Karen L Maxwell. Prophages mediate defense against phage infection through diverse mechanisms. *ISME J.*, 10(12):2854–2866, December 2016.

Leonard Both, Radu Botgros, and Marco Cavaleri. Analysis of licensed over-the-counter (OTC) antibiotics in the european union and norway, 2012. *Euro Surveill.*, 20(34):30002, 2015.

Jean-Yves Bouet and Barbara E Funnell. Plasmid localization and partition in enterobacteriaceae. *EcoSal Plus*, 8(2), June 2019.

I L Brito, S Yilmaz, K Huang, L Xu, S D Jupiter, A P Jenkins, W Naisilisili, M Tamminen, C S Smillie, J R Wortman, B W Birren, R J Xavier, P C Blainey, A K Singh, D Gevers, and E J Alm. Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435, July 2016.

Aya Brown Kav, Goor Sasson, Elie Jami, Adi Doron-Faigenboim, Itai Benhar, and Itzhak Mizrahi. Insights into the bovine rumen plasmidome. *Proc. Natl. Acad. Sci. U. S. A.*, 109 (14):5452–5457, April 2012.

Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12(1):59–60, January 2015.

Antonio Pedro Camargo, Simon Roux, Frederik Schulz, Michal Babinski, Yan Xu, Bin Hu, Patrick S G Chain, Stephen Nayfach, and Nikos C Kyrpides. You can move, but you can't hide: identification of mobile genetic elements with genomad. March 2023.

Rafael Cantón and María-Isabel Morosini. Emergence and spread of antibiotic resistance following exposure to antibiotics. *FEMS Microbiol. Rev.*, 35(5):977–991, September 2011.

Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25 (15):1972–1973, August 2009.

Alessandra Carattoli, Alessia Bertini, Laura Villa, Vincenzo Falbo, Katie L Hopkins, and E John Threlfall. Identification of plasmids by PCR-based replicon typing, 2005.

Alessandra Carattoli, Ea Zankari, Aurora García-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Møller Aarestrup, and Henrik Hasman. In SilicoDetection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing, 2014a.

Alessandra Carattoli, Ea Zankari, Aurora García-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Møller Aarestrup, and Henrik Hasman. *In Silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing, 2014b.

Alessandra Carattoli, Ea Zankari, Aurora García-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Møller Aarestrup, and Henrik Hasman. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, 58(7):3895–3903, July 2014c.

Victoria R Carr, Andrey Shkoporov, Colin Hill, Peter Mullany, and David L Moyes. Probing the mobilome: Discoveries in the dynamic microbiome. *Trends Microbiol.*, 29(2):158–170, February 2021.

Amanda C Carroll and Alex Wong. Plasmid persistence: costs, benefits, and the plasmid paradox. *Can. J. Microbiol.*, 64(5):293–304, May 2018.

Centers for Disease Control and Prevention. Antibiotic use in the united states, 2021 update: Progress and opportunities. Technical report, 2021.

Jean F Challacombe, Segaran Pillai, and Cheryl R Kuske. Shared features of cryptic plasmids from environmental and pathogenic francisella species. *PLoS One*, 12(8):e0183554, August 2017.

Indranil Chattopadhyay, Ruby Dhar, Karthikeyan Pethusamy, Ashikh Seethy, Tryambak Srivastava, Ramkishor Sah, Jyoti Sharma, and Subhradip Karmakar. Exploring the role of gut microbiome in colon cancer. *Appl. Biochem. Biotechnol.*, 193(6):1780–1799, January 2021.

Beibei Chen, Zhao Chen, Yuchen Wang, Han Gong, Linshan Sima, Jiao Wang, Shushan Ouyang, Wenqiang Gan, Mart Krupovic, Xiangdong Chen, and Shishen Du. ORF4 of the temperate archaeal virus SNJ1 governs the Lysis-Lysogeny switch and superinfection immunity. *J. Virol.*, 94(16), July 2020a.

Lin-Xing Chen, Karthik Anantharaman, Alon Shaiber, A Murat Eren, and Jillian F Banfield. Accurate and complete genomes from metagenomes. *Genome Res.*, 30(3):315–333, March 2020b.

Yiqun T Chen, Brian D Williamson, Taylor Okonek, Charles J Wolock, Andrew J Spieker, Travis Y Hee Wai, James P Hughes, Scott S Emerson, and Amy D Willis. rigr: Regression, inference, and general data analysis tools in R, 2022.

Zhenhua Chen, Li Zhong, Meijuan Shen, Ping Fang, and Zhongjun Qin. Characterization of streptomyces plasmid-phage pFP4 and its evolutionary implications. *Plasmid*, 68(3): 170–178, November 2012.

Luis Pedro Coelho, Jens Roat Kultima, Paul Igor Costea, Coralie Fournier, Yuanlong Pan, Gail Czarnecki-Maulden, Matthew Robert Hayward, Sofia K Forslund, Thomas Sebastian Benedikt Schmidt, Patrick Descombes, Janet R Jackson, Qinghong Li, and Peer Bork. Similarity of the dog and human gut microbiomes in gene content and response to diet. *Microbiome*, 6(1):1–11, April 2018.

André M Comeau, Françoise Tétart, Sabrina N Trojet, Marie-Françoise Prère, and H M Krisch. Phage-Antibiotic synergy (PAS): β-Lactam and quinolone antibiotics stimulate virulent phage growth. *PLoS One*, 2(8):e799, August 2007.

S Conrad, M Oethinger, K Kaifel, G Klotz, R Marre, and W V Kern. gyra mutations in high-level fluoroquinolone-resistant clinical isolates of escherichia coli. *J. Antimicrob. Chemother.*, 38(3):443–455, September 1996.

Steven R Corsi, Laura A De Cicco, Angela M Hansen, Peter L Lenaker, Brian A Bergamaschi, Brian A Pellerin, Debra K Dila, Melinda J Bootsma, Susan K Spencer, Mark A Borchardt, and Sandra L McLellan. Optical properties of water for prediction of wastewater contamination, Human-Associated bacteria, and fecal indicator bacteria in surface water at three watershed scales. *Environ. Sci. Technol.*, 55(20):13770–13782, October 2021.

Gabor Csardi, Tamas Nepusz, and Others. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

Alba Cuecas, Wirojne Kanoksilapatham, and Juan M Gonzalez. Evidence of horizontal gene transfer by transposase gene analyses in fervidobacterium species. *PLoS One*, 12(4): e0173961, April 2017.

Lawrence A David, Ana Weil, Edward T Ryan, Stephen B Calderwood, Jason B Harris, Fahima Chowdhury, Yasmin Begum, Firdausi Qadri, Regina C LaRocque, and Peter J Turnbaugh. Gut microbial succession follows acute secretory diarrhea in humans. *MBio*, 6(3), 2015.

Willem M de Vos, Herbert Tilg, Matthias Van Hul, and Patrice D Cani. Gut microbiome and health: mechanistic insights. *Gut*, 71(5):1020–1032, May 2022.

Reena Debray, Robin A Herbert, Alexander L Jaffe, Alexander Crits-Christoph, Mary E Power, and Britt Koskella. Priority effects in microbiome assembly. *Nat. Rev. Microbiol.*, 20(2):109–121, August 2021.

Gloria del Solar, Rafael Giraldo, María Jesús Ruiz-Echevarría, Manuel Espinosa, and Ramón Díaz-Orejas. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.*, 62(2):434, June 1998.

Sarah Delaney, Richard Murphy, and Fiona Walsh. A comparison of methods for the extraction of plasmids capable of conferring antibiotic resistance in a human pathogen from complex broiler cecal samples. *Front. Microbiol.*, 9:1731, August 2018.

W L Delano. The PyMOL molecular graphics system. *http://www.pymol.org/*, 2002.

Tom O Delmont, Christopher Quince, Alon Shaiber, Özcan C Esen, Sonny Tm Lee, Michael S Rappé, Sandra L McLellan, Sebastian Lücker, and A Murat Eren. Nitrogen-fixing populations of planctomycetes and proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol*, 3(7):804–813, July 2018.

Tom O Delmont, Evan Kiefl, Ozsel Kilinc, Ozcan C Esen, Ismail Uysal, Michael S Rappé, Steven Giovannoni, and A Murat Eren. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife*, 8, September 2019.

D B Dhakan, A Maji, A K Sharma, R Saxena, J Pulikkan, T Grace, A Gomez, J Scaria, K R Amato, and V K Sharma. The unique composition of indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience*, 8(3):giz004, January 2019.

Deborah K Dila, Emily R Koster, Jill McClary-Guterriez, Bahram Khazaei, Hector R Bravo, Melinda J Bootsma, and Sandra L McLellan. Assessment of regional and local sources of contamination at urban beaches using hydrodynamic models and Field-Based monitoring. *ACS EST Water*, 2(10):1715–1724, October 2022.

Tatiana Dimitriu. Evolution of horizontal transmission in antimicrobial resistance plasmids. *Microbiology*, 168(7):001214, July 2022.

Terje Dokland. Molecular piracy: Redirection of bacteriophage capsid assembly by mobile genetic elements. *Viruses*, 11(11), November 2019.

Sean R Eddy. Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195, October 2011.

Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, March 2004a.

Robert C Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, August 2004b.

Robert A Edwards, Alejandro A Vega, Holly M Norman, Maria Ohaeri, Kyle Levi, Elizabeth A Dinsdale, Ondrej Cinek, Ramy K Aziz, Katelyn McNair, Jeremy J Barr, Kyle Bibby, Stan J J Brouns, Adrian Cazares, Patrick A de Jonge, Christelle Desnues, Samuel L Díaz Muñoz, Peter C Fineran, Alexander Kurilshikov, Rob Lavigne, Karla Mazankova, David T McCarthy, Franklin L Nobrega, Alejandro Reyes Muñoz, German Tapia, Nicole Trefault, Alexander V Tyakht, Pablo Vinuesa, Jeroen Wagemans, Alexandra Zhernakova, Frank M Aarestrup, Gunduz Ahmadov, Abeer Alassaf, Josefa Anton, Abigail Asangba, Emma K Billings, Vito Adrian Cantu, Jane M Carlton, Daniel Cazares, Gyu-Sung Cho, Tess Condeff, Pilar Cortés, Mike Cranfield, Daniel A Cuevas, Rodrigo De la Iglesia, Przemyslaw Decewicz, Michael P Doane, Nathaniel J Dominy, Lukasz Dziewit, Bashir Mukhtar Elwasila, A Murat Eren, Charles Franz, Jingyuan Fu, Cristina Garcia-Aljaro, Elodie Ghedin, Kristen M Gulino, John M Haggerty, Steven R Head, Rene S Hendriksen, Colin Hill, Heikki Hyöty, Elena N Ilina, Mitchell T Irwin, Thomas C Jeffries, Juan Jofre, Randall E Junge, Scott T Kelley, Mohammadali Khan Mirzaei, Martin Kowalewski, Deepak Kumaresan, Steven R Leigh, David Lipson, Eugenia S Lisitsyna, Montserrat Llagostera, Julia M Maritz, Linsey C Marr, Angela McCann, Shahar Molshanski-Mor, Silvia Monteiro, Benjamin Moreira-Grez, Megan Morris, Lawrence Mugisha, Maite Muniesa, Horst Neve, Nam-Phuong Nguyen, Olivia D Nigro, Anders S Nilsson, Taylor O'Connell, Rasha Odeh, Andrew Oliver, Mariana Piuri, Aaron J Prussin, Ii, Udi Qimron, Zhe-Xue Quan, Petra Rainetova, Adán Ramírez-Rojas, Raul Raya, Kim Reasor, Gillian A O Rice, Alessandro Rossi, Ricardo Santos, John Shimashita, Elyse N Stachler, Lars C Stene, Ronan Strain, Rebecca Stumpf, Pedro J Torres, Alan Twaddle, Maryann Ugochi Ibekwe, Nicolás Villagra, Stephen Wandro, Bryan White, Andy Whiteley, Katrine L Whiteson, Cisca Wijmenga, Maria M Zambrano, Henrike Zschach, and Bas E Dutilh. Global phylogeography and ancient evolution of the widespread human gut virus crassphage. *Nat Microbiol*, 4(10): 1727–1736, October 2019.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto,

and Robert D Finn. The pfam protein families database in 2019. *Nucleic Acids Res.*, 47 (D1):D427–D432, January 2019.

Monica E Embers, Xavier Alvarez, Tara Ooms, and Mario T Philipp. The failure of immune response evasion by linear plasmid 28-1-deficient borrelia burgdorferi is attributable to persistent expression of an outer surface protein. *Infect. Immun.*, 76(9):3984–3991, September 2008.

Ginro Endo, Guangyong Ji, and Simon Silver. Heavy metal resistance plasmids and use in bioremediation, 1995.

A Murat Eren, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E Miller, Matthew S Schechter, Isaac Fink, Jessica N Pan, Mahmoud Yousef, Emily C Fogarty, Florian Trigodet, Andrea R Watson, Özcan C Esen, Ryan M Moore, Quentin Clayssen, Michael D Lee, Veronika Kivenson, Elaina D Graham, Bryan D Merrill, Antti Karkman, Daniel Blankenberg, John M Eppley, Andreas Sjödin, Jarrod J Scott, Xabier Vázquez-Campos, Luke J McKay, Elizabeth A McDaniel, Sarah L R Stevens, Rika E Anderson, Jessika Fuessel, Antonio Fernandez-Guerra, Lois Maignien, Tom O Delmont, and Amy D Willis. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol*, 6(1):3–6, January 2021.

Jordan C Evans, Valentina Laclare McEneany, Michael J Coyne, Elizabeth P Caldwell, Madeline L Sheahan, Salena S Von, Emily M Coyne, Rodney K Tweten, and Laurie E Comstock. A proteolytically activated antimicrobial toxin encoded on a mobile plasmid of bacteroidales induces a protective response. *Nat. Commun.*, 13, 2022.

Yong Fan and Oluf Pedersen. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.*, 19(1):55–71, September 2020.

Zhencheng Fang, Jie Tan, Shufang Wu, Mo Li, Congmin Xu, Zhongjie Xie, and Huaiqiu Zhu. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience*, 8(6), June 2019.

Qiang Feng, Suisha Liang, Huijue Jia, Andreas Stadlmayr, Longqing Tang, Zhou Lan, Dongya Zhang, Huihua Xia, Xiaoying Xu, Zhuye Jie, Lili Su, Xiaoping Li, Xin Li, Junhua Li, Liang Xiao, Ursula Huber-Schönauer, David Niederseer, Xun Xu, Jumana Yousuf Al-Aama, Huanming Yang, Jian Wang, Karsten Kristiansen, Manimozhiyan Arumugam, Herbert Tilg, Christian Datz, and Jun Wang. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.*, 6(1):1–13, March 2015.

Shuchen Feng, Melinda Bootsma, and Sandra L McLellan. Human-Associated lachnospiraceae genetic markers improve detection of fecal pollution sources in urban waters. *Appl. Environ. Microbiol.*, 84(14), July 2018.

Shuchen Feng, Warish Ahmed, and Sandra L McLellan. Ecological and technical mechanisms for Cross-Reaction of human fecal indicators with animal hosts. *Appl. Environ. Microbiol.*, 86(5), February 2020.

Raúl Fernández-López, M Pilar Garcillán-Barcia, Carlos Revilla, Miguel Lázaro, Luis Vielva, and Fernando de la Cruz. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. *FEMS Microbiol. Rev.*, 30(6):942–966, November 2006.

Raul Fernandez-Lopez, Santiago Redondo, M Pilar Garcillan-Barcia, and Fernando de la Cruz. Towards a taxonomy of conjugative plasmids. *Curr. Opin. Microbiol.*, 38:106–113, August 2017.

Pamela Ferretti, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Valentina Gorfer, Sabina Fedi, Federica Armanini, Duy Tin Truong, Serena Manara, Moreno Zolfo, Francesco Beghini, Roberto Bertorelli, Veronica De Sanctis, Ilaria Bariletti, Rosarita Canto, Rosanna Clementi, Marina Cologna, Tiziana Crifò, Giuseppina Cusumano, Stefania Gottardi, Claudia Innamorati, Caterina Masè, Daniela Postai, Daniela Savoi, Sabrina Duranti, Gabriele Andrea Lugli, Leonardo Mancabelli, Francesca Turroni, Chiara Ferrario, Christian Milani, Marta Mangifesta, Rosaria Anzalone, Alice Viappiani, Moran Yassour, Hera Vlamakis, Ramnik Xavier, Carmen Maria Collado, Omry Koren, Saverio Tateo, Massimo Soffiati, Anna Pedrotti, Marco Ventura, Curtis Huttenhower, Peer Bork, and Nicola Segata. Mother-to-Infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe*, 24(1):133–145.e5, July 2018.

Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.*, 8(9):e1002687, September 2012.

Laura S Frost, Raphael Leplae, Anne O Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, 3(9):722–732, September 2005.

Valentina Galata, Tobias Fehlmann, Christina Backes, and Andreas Keller. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.*, 47(D1):D195–D202, January 2019.

Michael Y Galperin, Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, 43(Database issue):D261–9, January 2015.

Adeline Galvanin, Lea-Marie Vogt, Antonia Grober, Isabel Freund, Lilia Ayadi, Valerie Bourguignon-Igel, Larissa Bessler, Dominik Jacob, Tatjana Eigenbrod, Virginie Marchand, Alexander Dalpke, Mark Helm, and Yuri Motorin. Bacterial tRNA 2'-o-methylation is dynamically regulated under stress conditions and modulates innate immune response. *Nucleic Acids Res.*, 48(22):12833–12844, December 2020.

Leonor García-Bayona and Laurie E Comstock. Streamlined genetic manipulation of diverse bacteroides and parabacteroides isolates from the human gut microbiota. *MBio*, 10(4), 2019.

M Pilar Garcillán-Barcia, Belén Ruiz del Castillo, Andrés Alvarado, Fernando de la Cruz, and Luis Martínez-Martínez. Degenerate primer MOB typing of multiresistant clinical isolates of e. coli uncovers new plasmid backbones. *Plasmid*, 77:17–27, January 2015.

María Pilar Garcillán-Barcia, María Victoria Francia, and Fernando de la Cruz. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.*, 33(3):657–687, May 2009.

Maria Pilar Garcillán-Barcia, Andrés Alvarado, and Fernando de la Cruz. Identification of bacterial plasmids based on mobility and plasmid population biology. *FEMS Microbiol. Rev.*, 35(5):936–956, September 2011.

Ana Garoña and Tal Dagan. Darwinian individuality of extrachromosomal genetic elements calls for population genetics tinkering. *Environ. Microbiol. Rep.*, 13(1):22–26, February 2021.

Stefanie Gehrig, Mariel-Esther Eberle, Flavia Botschen, Katharina Rimbach, Florian Eberle, Tatjana Eigenbrod, Steffen Kaiser, Walter M Holmes, Volker A Erdmann, Mathias Sprinzl, Guillaume Bec, Gérard Keith, Alexander H Dalpke, and Mark Helm. Identification of modifications in microbial, native tRNA that suppress immunostimulatory activity. *J. Exp. Med.*, 209(2):225–233, February 2012.

Molly K Gibson, Kevin J Forsberg, and Gautam Dantas. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, 9(1): 207–216, January 2015.

Ofir Goldenberg, Elana Erez, Guy Nimrod, and Nir Ben-Tal. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.*, 37(Database issue):D323–7, January 2009.

Andres Gomez, Klara J Petrzelkova, Michael B Burns, Carl J Yeoman, Katherine R Amato, Klara Vlckova, David Modry, Angelique Todd, Carolyn A Jost Robinson, Melissa J Remis, Manolito G Torralba, Elise Morton, Juan D Umaña, Franck Carbonero, H Rex Gaskins, Karen E Nelson, Brenda A Wilson, Rebecca M Stumpf, Bryan A White, Steven R Leigh, and Ran Blekhman. Gut microbiome of coexisting BaAka pygmies and bantu reflects gradients of traditional subsistence patterns. *Cell Rep.*, 14(9):2142–2153, March 2016.

Ryota Gomi, Kelly L Wyres, and Kathryn E Holt. Detection of plasmid contigs in draft genome assemblies using customized kraken databases. *Microb Genom*, 7(4), April 2021.

D M Gordon. Rate of plasmid transfer among escherichia coli strains isolated from natural populations. *J. Gen. Microbiol.*, 138(1):17–21, January 1992.

Daniel B Graham and Ramnik J Xavier. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature*, 578(7796):527–539, February 2020.

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, April 2021.

Jiarong Guo, Ben Bolduc, Ahmed A Zayed, Arvind Varsani, Guillermo Dominguez-Huerta, Tom O Delmont, Akbar Adjie Pratama, M Consuelo Gazitúa, Dean Vik, Matthew B Sullivan, and Simon Roux. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, 9(1):37, February 2021.

Vinod K Gupta, Sandip Paul, and Chitra Dutta. Geography, ethnicity or Subsistence-Specific variations in human microbiome composition and diversity, 2017.

Fakhri Haghi, Elshan Goli, Bahman Mirzaei, and Habib Zeighami. The association between fecal enterotoxigenic b. fragilis with colorectal cancer. *BMC Cancer*, 19(1):879, September 2019.

Jo Handelsman. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, 68(4):669–685, December 2004.

Jeffrey Heer, Stuart K Card, and James A Landay. prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 421–430, New York, NY, USA, April 2005. Association for Computing Machinery.

Donald R Helinski. A brief history of plasmids. *EcoSal Plus*, 10(1):eESP00282021, December 2022.

Rene S Hendriksen, Patrick Munk, Patrick Njage, Bram van Bunnik, Luke McNally, Oksana Lukjancenko, Timo Röder, David Nieuwenhuijse, Susanne Karlsmose Pedersen, Jette Kjeldgaard, Rolf S Kaas, Philip Thomas Lanken Conradsen Clausen, Josef Korbinian Vogt, Pimlapas Leekitcharoenphon, Milou G M van de Schans, Tina Zuidema, Ana Maria de Roda Husman, Simon Rasmussen, Bent Petersen, Clara Amid, Guy Cochrane, Thomas Sicheritz-Ponten, Heike Schmitt, Jorge Raul Matheu Alvarez, Awa Aidara-Kane, Sünje J Pamp, Ole Lund, Tine Hald, Mark Woolhouse, Marion P Koopmans, Håkan Vigre, Thomas Nordahl Petersen, and Frank M Aarestrup. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.*, 10(1): 1–12, March 2019.

Matthew T Henke, Douglas J Kenny, Chelsi D Cassilly, Hera Vlamakis, Ramnik J Xavier, and Jon Clardy. , a member of the human gut microbiome associated with crohn's disease, produces an inflammatory polysaccharide. *Proc. Natl. Acad. Sci. U. S. A.*, 116(26):12672–12677, June 2019.

Holger Heuer and Kornelia Smalla. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol. Rev.*, 36(6):1083–1104, November 2012.

J Hinnebusch and K Tilly. Linear plasmids and chromosomes in bacteria. *Mol. Microbiol.*, 10(5):917–922, December 1993.

Kathryn E Holt, Nicholas R Thomson, John Wain, Minh Duy Phan, Satheesh Nair, Rumina Hasan, Zulfiqar A Bhutta, Michael A Quail, Halina Norbertczak, Danielle Walker, Gordon Dougan, and Julian Parkhill. Multidrug-resistant salmonella enterica serovar paratyphi a harbors IncHI1 plasmids similar to those found in serovar typhi. *J. Bacteriol.*, 189(11): 4257–4264, June 2007.

Shengwei Hou, Siliangyu Cheng, Ting Chen, Jed A Fuhrman, and Fengzhu Sun. DeepMicrobeFinder sorts metagenomes into prokaryotes, eukaryotes and viruses, with marine applications. October 2021.

Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Banfield. A new view of the tree of life. *Nat Microbiol*, 1: 16048, April 2016.

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*, 11:119, March 2010.

Takao Iino, Koji Mori, Takashi Itoh, Takuji Kudo, Ken-Ichiro Suzuki, and Moriya Ohkuma. Description of mariniphaga anaerophila gen. nov., sp. nov., a facultatively aerobic marine bacterium isolated from tidal flat sediment, reclassification of the draconibacteriaceae as a later heterotypic synonym of the prolixibacteraceae and description of the family marinifilaceae fam. nov. *Int. J. Syst. Evol. Microbiol.*, 64(Pt_11):3660–3667, November 2014.

Jaime Iranzo, Pere Puigbò, Alexander E Lobkovsky, Yuri I Wolf, and Eugene V Koonin. Inevitability of genetic parasites, 2016.

A E Jacob and S J Hobbs. Conjugal transfer of plasmid-borne multiple antibiotic resistance in streptococcus faecalis var. zymogenes. *J. Bacteriol.*, 117(2):360–372, February 1974.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS One*, 9(6):e98679, June 2014.

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

Adam S B Jalal and Tung B K Le. Bacterial chromosome segregation by the ParABS system. *Open Biol.*, 10(6):200097, June 2020.

Timothy J Johnson and Lisa K Nolan. Pathogenomics of the virulence plasmids of escherichia coli. *Microbiol. Mol. Biol. Rev.*, 73(4):750–774, December 2009.

Brian V Jones and Julian R Marchesi. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods*, 4(1):55–61, January 2007.

Jaewook Joo, Michelle Gunny, Marisa Cases, Peter Hudson, Réka Albert, and Eric Harvill. Bacteriophage-mediated competition in bordetella bacteria. *Proc. Biol. Sci.*, 273(1595): 1843–1848, July 2006.

Tue Sparholt Jørgensen, Zhuofei Xu, Martin Asser Hansen, Søren Johannes Sørensen, and Lars Hestbjerg Hansen. Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS One*, 9(2):e87924, February 2014.

Xingyu Kang, Chunqiu Li, and Yi Luo. Cloning of pAhX22, a small cryptic plasmid from aeromonas hydrophila, and construction of a pAhX22-derived shuttle vector. *Plasmid*, 108:102490, March 2020.

Darius Kazlauskas, Arvind Varsani, Eugene V Koonin, and Mart Krupovic. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat. Commun.*, 10(1):1–12, July 2019.

S A Khan. Rolling-circle replication of bacterial plasmids, 1997.

Evan Kiefl, Ozcan C Esen, Samuel E Miller, Kourtney L Kroll, Amy D Willis, Michael S Rappé, Tao Pan, and A Murat Eren. Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution. *Sci Adv*, 9(8):eabq4632, February 2023.

Chan Yeong Kim, Muyoung Lee, Sunmo Yang, Kyungnam Kim, Dongeun Yong, Hye Ryun Kim, and Insuk Lee. Human reference gut microbiome catalog including newly assembled genomes from under-represented asian metagenomes. *Genome Med.*, 13(1):134, August 2021.

U Klümper, L Riber, A Dechesne, A Sannazzarro, L H Hansen, S J Sørensen, and B F Smets. Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.*, 9(4), March 2015.

Verena Kohler, Ankita Vaishampayan, and Elisabeth Grohmann. Broad-host-range inc18 plasmids: Occurrence, spread and transfer mechanisms. *Plasmid*, 99:11–21, September 2018.

Anna Kopf, Mesude Bicak, Renzo Kottmann, Julia Schnetzer, Ivaylo Kostadinov, Katja Lehmann, Antonio Fernandez-Guerra, Christian Jeanthon, Eyal Rahav, Matthias Ullrich, Antje Wichels, Gunnar Gerdts, Paraskevi Polymenakou, Giorgos Kotoulas, Rania Siam, Rehab Z Abdallah, Eva C Sonnenschein, Thierry Cariou, Fergal O'Gara, Stephen Jackson, Sandi Orlic, Michael Steinke, Julia Busch, Bernardo Duarte, Isabel Caçador, João Canning-Clode, Oleksandra Bobrova, Viggo Marteinsson, Eyjolfur Reynisson, Clara Magalhães Loureiro, Gian Marco Luna, Grazia Marina Quero, Carolin R Löscher, Anke Kremp, Marie E DeLorenzo, Lise Øvreås, Jennifer Tolman, Julie LaRoche, Antonella Penna, Marc Frischer, Timothy Davis, Barker Katherine, Christopher P Meyer, Sandra Ramos, Catarina Magalhães, Florence Jude-Lemeilleur, Ma Leopoldina Aguirre-Macedo, Shiao Wang, Nicole Poulton, Scott Jones, Rachel Collin, Jed A Fuhrman, Pascal Conan, Cecilia Alonso, Noga Stambler, Kelly Goodwin, Michael M Yakimov, Federico Baltar, Levente Bodrossy, Jodie Van De Kamp, Dion Mf Frampton, Martin Ostrowski, Paul Van Ruth, Paul Malthouse, Simon Claus, Klaas Deneudt, Jonas Mortelmans, Sophie Pitois, David Wallom, Ian Salter, Rodrigo Costa, Declan C Schroeder, Mahrous M Kandil, Valentina Amaral, Florencia Biancalana, Rafael Santana, Maria Luiza Pedrotti, Takashi Yoshida, Hiroyuki Ogata, Tim Ingleton, Kate Munnik, Naiara Rodriguez-Ezpeleta, Veronique Berteaux-Lecellier, Patricia Wecker, Ibon Cancio, Daniel Vaulot, Christina Bienhold, Hassan Ghazal, Bouchra Chaouni, Soumya Essayeh, Sara Ettamimi, El Houcine Zaid, Noureddine Boukhatem, Abderrahim Bouali, Rajaa Chahboune, Said Barrijal, Mohammed Timinouni, Fatima El Otmani, Mohamed Bennani, Marianna Mea, Nadezhda Todorova, Ventzislav Karamfilov, Petra Ten Hoopen, Guy Cochrane, Stephane L'Haridon, Kemal Can Bizsel, Alessandro Vezzi, Federico M Lauro, Patrick Martin, Rachelle M Jensen, Jamie Hinks, Susan Gebbels, Riccardo Rosselli, Fabio De Pascale, Riccardo Schiavon, Antonina Dos Santos, Emilie Villar, Stéphane Pesant, Bruno Cataletto, Francesca Malfatti, Ranjith Edirisinghe, Jorge A Herrera Silveira, Michele Barbier, Valentina Turk, Tinkara Tinta, Wayne J Fuller, Ilkay Salihoglu, Nedime Serakinci, Mahmut Cerkez Ergoren, Eileen Bresnan, Juan Iriberri, Paul Anders Fronth Nyhus, Edvardsen Bente, Hans Erik Karlsen, Peter N Golyshin, Josep M Gasol, Snejana Moncheva, Nina Dzhembekova, Zackary Johnson, Christopher David Sinigalliano, Maribeth Louise Gidley, Adriana Zingone, Roberto Danovaro, George Tsiamis, Melody S Clark, Ana Cristina Costa, Monia El Bour, Ana M Martins, R Eric Collins, Anne-Lise Ducluzeau, Jonathan Martinez, Mark J Costello, Linda A Amaral-Zettler, Jack A Gilbert, Neil Davies, Dawn Field, and Frank Oliver Glöckner. The ocean sampling day consortium. *Gigascience*, 4:27, June 2015.

Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, August 2012.

Pawel S Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.*, 46(6):e35, April 2018.

M S Kris A. Wetterstrand. DNA sequencing costs: Data. `https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data`, March 2019. Accessed: 2023-4-1.

Deepanshu Kumar, Hemant Kumar Prajapati, Anjali Mahilkar, Chien-Hui Ma, Priyanka Mittal, Makkuni Jayaram, and Santanu K Ghosh. The selfish yeast plasmid utilizes the condensin complex and condensed chromatin for faithful partitioning. *PLoS Genet.*, 17 (7):e1009660, July 2021.

Ruiting Lan, Gordon Stevenson, and Peter R Reeves. Comparison of two major forms of the shigella virulence plasmid pINV: Positive selection is a major force driving the divergence. *Infect. Immun.*, 71(11):6298, November 2003.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357, 2012.

E Le Chatelier, T Nielsen, J Qin, E Prifti, F Hildebrand, G Falony, M Almeida, M Arumugam, J M Batto, S Kennedy, P Leonard, J Li, K Burgdorf, N Grarup, T Jørgensen, I Brandslund, H B Nielsen, A S Juncker, M Bertalan, F Levenez, N Pons, S Rasmussen, S Sunagawa, J Tap, S Tims, E G Zoetendal, S Brunak, K Clément, J Doré, M Kleerebezem, K Kristiansen, P Renault, T Sicheritz-Ponten, W M de Vos, J D Zucker, J Raes, T Hansen, P Bork, J Wang, S D Ehrlich, and O Pedersen. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464), August 2013.

Changsoo Lee, Jaai Kim, Seung Gu Shin, and Seokhwan Hwang. Absolute and relative QPCR quantification of plasmid copy number in escherichia coli, 2006.

Yun Kyung Lee, Parpi Mehrabian, Silva Boyajian, Wei-Li Wu, Jane Selicha, Steven Vonderfecht, and Sarkis K Mazmanian. The protective role of *Bacteroides fragilis* in a murine model of Colitis-Associated colorectal cancer, 2018.

Mark D M Leiserson, Hsin-Ta Wu, Fabio Vandin, and Benjamin J Raphael. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.*, 16:160, August 2015.

Peter L Lenaker, Steven R Corsi, Sandra L McLellan, Mark A Borchardt, Hayley T Olds, Deborah K Dila, Susan K Spencer, and Austin K Baldwin. Human-Associated indicator bacteria and Human-Specific viruses in surface water: A spatial assessment with implications on fate and transport. *Environ. Sci. Technol.*, 52(21):12162–12171, November 2018.

Sean P Leonard, Jiri Perutka, J Elijah Powell, Peng Geng, Darby D Richhart, Michelle Byrom, Shaunak Kar, Bryan W Davies, Andrew D Ellington, Nancy A Moran, and Jeffrey E Barrick. Genetic engineering of bee gut microbiome bacteria with a toolkit for modular assembly of Broad-Host-Range plasmids. *ACS Synth. Biol.*, 7(5):1279–1290, May 2018.

B R Levin and F M Stewart. The population biology of bacterial plasmids: a priori conditions for the existence of mobilizable nonconjugative factors. *Genetics*, 94(2):425–443, February 1980.

Bing Li, Yong Qiu, Jing Zhang, Xia Huang, Hanchang Shi, and Huabing Yin. Real-Time study of rapid spread of antibiotic resistance plasmid in biofilm using microfluidics. *Environ. Sci. Technol.*, 52(19):11132–11141, October 2018.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

Junhua Li, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, Jens Roat Kultima, Edi Prifti, Trine Nielsen, Agnieszka Sierakowska Juncker, Chaysavanh Manichanh, Bing Chen, Wenwei Zhang, Florence Levenez, Juan Wang, Xun Xu, Liang Xiao, Suisha Liang, Dongya Zhang, Zhaoxi Zhang, Weineng Chen, Hailong Zhao, Jumana Yousuf Al-Aama, Sherif Edris, Huanming Yang, Jian Wang, Torben Hansen, Henrik Bjørn Nielsen, Søren Brunak, Karsten Kristiansen, Francisco Guarner, Oluf Pedersen, Joel Doré, S Dusko Ehrlich, Peer Bork, and Jun Wang. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, 32(8):834–841, July 2014.

Joshua Lilly and Manel Camps. Mechanisms of theta plasmid replication. *Microbiol Spectr*, 3(1):PLAS–0029–2014, February 2015.

Meng Liu, Xiaobin Li, Yingzhou Xie, Dexi Bi, Jingyong Sun, Jun Li, Cui Tai, Zixin Deng, and Hong-Yu Ou. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.*, 47(D1):D660–D665, January 2019.

Wenjun Liu, Jiachao Zhang, Chunyan Wu, Shunfeng Cai, Weiqiang Huang, Jing Chen, X I Xiaoxia, Zebin Liang, Qiangchuan Hou, Bing Zhou, Nan Qin, and Heping Zhang. Unique features of ethnic mongolian gut microbiome revealed by metagenomic analysis. *Sci. Rep.*, 6, 2016.

Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, David Casero, Holly Courtney, Antonio Gonzalez, Thomas G Graeber, A Brantley Hall, Kathleen Lake, Carol J Landers, Himel Mallick, Damian R Plichta, Mahadev Prasad, Gholamali Rahnavard, Jenny Sauk, Dmitry Shungin, Yoshiki Vázquez-Baeza, Richard A White, 3rd, IBDMDB Investigators, Jonathan Braun, Lee A Denson, Janet K Jansson, Rob Knight, Subra Kugathasan, Dermot P B McGovern, Joseph F Petrosino, Thaddeus S Stappenbeck, Harland S Winter, Clary B Clish, Eric A Franzosa, Hera Vlamakis, Ramnik J Xavier, and Curtis Huttenhower. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, May 2019.

Jon Lorsch. *Laboratory Methods in Enzymology: DNA*. Elsevier, September 2013.

Yue Clare Lou, Matthew R Olm, Spencer Diamond, Alexander Crits-Christoph, Brian A Firek, Robyn Baker, Michael J Morowitz, and Jillian F Banfield. Infant gut strain persis-

tence is associated with maternal origin, phylogeny, and traits including surface adhesion and iron acquisition. *Cell Reports Medicine*, 2(9), September 2021.

Jennifer Lu, Florian P Breitwieser, Peter Thielen, and Steven L Salzberg. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.*, 3(e104):e104, January 2017.

Y B Lu, H J Datta, and D Bastia. Mechanistic studies of initiator-initiator interaction and replication initiation. *EMBO J.*, 17(17):5192–5200, September 1998.

Hongyan Ma and James D Bryers. Non-invasive determination of conjugative transfer of plasmids bearing antibiotic-resistance genes in biofilm-bound bacteria: effects of substrate loading and antibiotic selection. *Appl. Microbiol. Biotechnol.*, 97(1):317–328, June 2012.

R Craig MacLean and Alvaro San Millan. Microbial evolution: Towards resolving the plasmid paradox. *Curr. Biol.*, 25(17):R764–R767, August 2015.

R Craig MacLean and Alvaro San Millan. The evolution of antibiotic resistance. *Science*, 365(6458):1082–1083, September 2019.

Leonardo Mancabelli, Christian Milani, Gabriele Andrea Lugli, Francesca Turroni, Chiara Ferrario, Douwe van Sinderen, and Marco Ventura. Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. *Environ. Microbiol.*, 19(4): 1379–1390, April 2017.

Ohad Manor, Chengzhen L Dai, Sergey A Kornilov, Brett Smith, Nathan D Price, Jennifer C Lovejoy, Sean M Gibbons, and Andrew T Magis. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.*, 11(1):1–12, October 2020.

Travis N Mavrich and Graham F Hatfull. Evolution of superinfection immunity in cluster a mycobacteriophages, 2019.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. February 2018.

Sandra L McLellan and A Murat Eren. Discovering new indicators of fecal pollution. *Trends Microbiol.*, 22(12):697–706, December 2014.

F Meinhardt, R Schaffrath, and M Larsen. Microbial linear plasmids. *Appl. Microbiol. Biotechnol.*, 47(4):329–336, April 1997.

D Meletzus, A Bermphol, J Dreier, and R Eichenlaub. Evidence for plasmid-encoded virulence factors in the phytopathogenic bacterium clavibacter michiganensis subsp. michiganensis NCPPB382. *J. Bacteriol.*, 175(7):2131–2136, April 1993.

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, 37(5):1530–1534, May 2020.

Ministry of Health and Medical Services, Government of Fiji. Fiji antibiotic guidelines. Technical report, Government of Fiji, November 2019.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nat. Methods*, 19 (6):679–682, June 2022.

Tanya M Monaghan, Tim J Sloan, Stephen R Stockdale, Adam M Blanchard, Richard D Emes, Mark Wilcox, Rima Biswas, Rupam Nashine, Sonali Manke, Jinal Gandhi, Pratishtha Jain, Shrejal Bhotmange, Shrikant Ambalkar, Ashish Satav, Lorraine A Draper, Colin Hill, and Rajpal Singh Kashyap. Metagenomics reveals impact of geography and acute diarrheal disease on the central indian human gut microbiome. *Gut Microbes*, 12(1): 1752605, November 2020.

Murray Moo-Young, W A Anderson, and A M Chakrabarty. *Environmental Biotechnology: Principles and Applications*. Springer Science & Business Media, June 2013.

A Murat Eren, Joseph H Vineis, Hilary G Morrison, and Mitchell L Sogin. A filtering method to generate high quality short reads using illumina Paired-End technology. *PLoS One*, 8 (6):e66643, June 2013.

Christopher Mutuku, Zoltan Gazdag, and Szilvia Melegh. Occurrence of antibiotics and bacterial resistance genes in wastewater: resistance mechanisms and antimicrobial resistance control approaches. *World J. Microbiol. Biotechnol.*, 38(9):1–27, July 2022.

Andrea K Nash, Thomas A Auchtung, Matthew C Wong, Daniel P Smith, Jonathan R Gesell, Matthew C Ross, Christopher J Stewart, Ginger A Metcalf, Donna M Muzny, Richard A Gibbs, Nadim J Ajami, and Joseph F Petrosino. The gut mycobiome of the human microbiome project healthy cohort. *Microbiome*, 5(1):153, November 2017.

Alexander T Neu, Eric E Allen, and Kaustuv Roy. Defining and quantifying the core microbiome: Challenges and prospects. *Proc. Natl. Acad. Sci. U. S. A.*, 118(51), December 2021.

Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32(1):268–274, January 2015.

Peter Norberg, Maria Bergström, Vinay Jethava, Devdatt Dubhashi, and Malte Hermansson. The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination. *Nat. Commun.*, 2:268, April 2011.

S J Norris, C J Carter, J K Howell, and A G Barbour. Low-passage-associated proteins of borrelia burgdorferi b31: characterization and molecular cloning of OspD, a surface-exposed, plasmid-encoded lipoprotein. *Infect. Immun.*, 60(11):4662–4672, November 1992.

R P Novick. Plasmid incompatibility, 1987.

Artem S Novozhilov, Georgy P Karev, and Eugene V Koonin. Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol.*, 22(8):1721–1732, May 2005.

Alexandra J Obregon-Tito, Raul Y Tito, Jessica Metcalf, Krithivasan Sankaranarayanan, Jose C Clemente, Luke K Ursell, Zhenjiang Zech Xu, Will Van Treuren, Rob Knight, Patrick M Gaffney, Paul Spicer, Paul Lawson, Luis Marin-Reyes, Omar Trujillo-Villarroel, Morris Foster, Emilio Guija-Poma, Luzmila Troncoso-Corzo, Christina Warinner, Andrew T Ozga, and Cecil M Lewis. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.*, 6(1):1–9, March 2015.

J Ochoa-Repáraz, D W Mielcarz, Y Wang, S Begum-Haque, S Dasgupta, D L Kasper, and L H Kasper. A polysaccharide from the human commensal bacteroides fragilis protects against CNS demyelinating disease. *Mucosal Immunol.*, 3(5):487–495, September 2010.

Hayley T Olds, Steven R Corsi, Deborah K Dila, Katherine M Halmo, Melinda J Bootsma, and Sandra L McLellan. High levels of sewage contamination released from urban areas after storm events: A quantitative survey with sewage specific bacterial indicators. *PLoS Med.*, 15(7):e1002614, July 2018.

M Oliva, C Calia, M Ferrara, P D'Addabbo, M Scrascia, G Mulè, R Monno, and C Pazzani. Antimicrobial resistance gene shuffling and a three-element mobilisation system in the monophasic salmonella typhimurium strain ST1030, 2020.

M A Oliva, A J Martin-Galiano, Y Sakaguchi, and J M Andreu. Tubulin homolog TubZ in a phage-encoded partition system. *Proc. Natl. Acad. Sci. U. S. A.*, 109(20), May 2012.

Vanessa Oliveira, Ana R M Polónia, Daniel F R Cleary, Yusheng M Huang, Nicole J de Voogd, Ulisses N da Rocha, and Newton C M Gomes. Characterization of putative circular plasmids in sponge-associated bacterial communities using a selective multiply-primed rolling circle amplification. *Mol. Ecol. Resour.*, 21(1):110–121, January 2021.

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 Biol.*, 17(1):132, June 2016.

Alex Orlek, Nicole Stoesser, Muna F Anjum, Michel Doumith, Matthew J Ellington, Tim Peto, Derrick Crook, Neil Woodford, A Sarah Walker, Hang Phan, and Anna E Sheppard. Plasmid classification in an era of Whole-Genome sequencing: Application in studies of antibiotic resistance epidemiology. *Front. Microbiol.*, 0, 2017.

Alana Palomino, Danya Gewurz, Lela DeVine, Ujana Zajmi, Jenifer Moralez, Fatima Abu-Rumman, Robert P Smith, and Allison J Lopatkin. Metabolic genes on conjugative plasmids are highly prevalent in escherichia coli and can protect against antibiotic treatment. *ISME J.*, October 2022.

Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662.e20, January 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and Others. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

David Pellow, Itzik Mizrahi, and Ron Shamir. PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.*, 16(4):e1007781, April 2020.

David Pellow, Alvah Zorea, Maraike Probst, Ori Furman, Arik Segal, Itzhak Mizrahi, and Ron Shamir. SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome*, 9(1):144, June 2021.

Yu Peng, Henry C M Leung, S M Yiu, and Francis Y L Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, April 2012.

Eugen Pfeifer, Jorge A Moura de Sousa, Marie Touchon, and Eduardo P C Rocha. Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.*, 49(5):2655–2673, February 2021.

R Pluta, D R Boer, and M Coll. MobM relaxase domain (MOBV; Mob_Pre) bound to plasmid pMV158 orit DNA (22nt). mn-bound crystal structure at ph 4.6, September 2014.

Radoslaw Pluta, D Roeland Boer, Fabián Lorenzo-Díaz, Silvia Russi, Hansel Gómez, Cris Fernández-López, Rosa Pérez-Luque, Modesto Orozco, Manuel Espinosa, and Miquel Coll. Structural basis of a histidine-DNA nicking/joining mechanism for gene transfer and promiscuous spread of antibiotic resistance. *Proc. Natl. Acad. Sci. U. S. A.*, 114(32): E6526–E6535, August 2017.

C Poyart-Salmeron, C Carlier, P Trieu-Cuot, A L Courtieu, and P Courvalin. Transferable plasmid-mediated antibiotic resistance in listeria monocytogenes. *Lancet*, 335(8703):1422–1426, June 1990.

Francesca Prestinaci, Patrizio Pezzotti, and Annalisa Pantosti. Antimicrobial resistance: a global multifaceted phenomenon. *Pathog. Glob. Health*, 109(7):309–318, September 2015.

Rachel V Purcell, John Pearson, Alan Aitchison, Liane Dixon, Frank A Frizelle, and Jacqueline I Keenan. Colonization with enterotoxigenic bacteroides fragilis is associated with early-stage colorectal neoplasia. *PLoS One*, 12(2):e0171602, February 2017.

Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, Peer Bork, S Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, March 2010.

Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu, Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle LeChatelier, Pierre Renault, Nicolas Pons, Jean-Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, Jian Wang, S Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jun Wang. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, September 2012.

Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, 35(9):833–844, September 2017.

Simone Rampelli, Stephanie L Schnorr, Clarissa Consolandi, Silvia Turroni, Marco Severgnini, Clelia Peano, Patrizia Brigidi, Alyssa N Crittenden, Amanda G Henry, and Marco Candela. Metagenome sequencing of the hadza Hunter-Gatherer gut microbiota. *Curr. Biol.*, 25(13):1682–1693, June 2015.

Frédéric Raymond, Amin A Ouameur, Maxime Déraspe, Naeem Iqbal, Hélène Gingras, Bédis Dridi, Philippe Leprohon, Pier-Luc Plante, Richard Giroux, Ève Bérubé, Johanne Frenette, Dominique K Boudreau, Jean-Luc Simard, Isabelle Chabot, Marc-Christian Domingo, Sylvie Trottier, Maurice Boissinot, Ann Huletsky, Paul H Roy, Marc Ouellette, Michel G Bergeron, and Jacques Corbeil. The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.*, 10(3):707, March 2016.

Santiago Redondo-Salvo, Raúl Fernández-López, Raúl Ruiz, Luis Vielva, María de Toro, Eduardo P C Rocha, M Pilar Garcillán-Barcia, and Fernando de la Cruz. Pathways for

horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.*, 11(1):3602, July 2020.

Julie Reveillaud, Sarah R Bordenstein, Corinne Cruaud, Alon Shaiber, Özcan C Esen, Mylène Weill, Patrick Makoundou, Karen Lolans, Andrea R Watson, Ignace Rakotoarivony, Seth R Bordenstein, and A Murat Eren. The wolbachia mobilome in culex pipiens includes a putative plasmid. *Nat. Commun.*, 10(1):1–11, March 2019.

Suzannah J Rihn, Andres Merits, Siddharth Bakshi, Matthew L Turnbull, Arthur Wickenhagen, Akira J T Alexander, Carla Baillie, Benjamin Brennan, Fiona Brown, Kirstyn Brunker, Steven R Bryden, Kerry A Burness, Stephen Carmichael, Sarah J Cole, Vanessa M Cowton, Paul Davies, Chris Davis, Giuditta De Lorenzo, Claire L Donald, Mark Dorward, James I Dunlop, Matthew Elliott, Mazigh Fares, Ana da Silva Filipe, Joseph R Freitas, Wilhelm Furnon, Rommel J Gestuveo, Anna Geyer, Daniel Giesel, Daniel M Goldfarb, Nicola Goodman, Rory Gunson, C James Hastie, Vanessa Herder, Joseph Hughes, Clare Johnson, Natasha Johnson, Alain Kohl, Karen Kerr, Hannah Leech, Laura Sandra Lello, Kathy Li, Gauthier Lieber, Xiang Liu, Rajendra Lingala, Colin Loney, Daniel Mair, Marion J McElwee, Steven McFarlane, Jenna Nichols, Kyriaki Nomikou, Anne Orr, Richard J Orton, Massimo Palmarini, Yasmin A Parr, Rute Maria Pinto, Samantha Raggett, Elaine Reid, David L Robertson, Jamie Royle, Natalia Cameron-Ruiz, James G Shepherd, Katherine Smollett, Douglas G Stewart, Meredith Stewart, Elena Sugrue, Agnieszka M Szemiel, Aislynn Taggart, Emma C Thomson, Lily Tong, Leah S Torrie, Rachel Toth, Margus Varjak, Sainan Wang, Stuart G Wilkinson, Paul G Wyatt, Eva Zusinaite, Dario R Alessi, Arvind H Patel, Ali Zaid, Sam J Wilson, and Suresh Mahalingam. A plasmid DNA-launched SARS-CoV-2 reverse genetics system and coronavirus toolkit for COVID-19 research. *PLoS Biol.*, 19(2):e3001091, February 2021.

Alice Risely. Applying the core microbiome to understand host–microbe systems, 2020.

M C Roberts. Plasmids of neisseria gonorrhoeae and other neisseria species, 1989.

James Robertson and John H E Nash. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*, 4(8), August 2018.

G Royer, J W Decousser, C Branger, M Dubois, C Médigue, E Denamur, and D Vallenet. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom*, 4(9), September 2018.

Roye Rozov, Aya Brown Kav, David Bogumil, Naama Shterzer, Eran Halperin, Itzhak Mizrahi, and Ron Shamir. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics*, 33(4):475–482, February 2017.

D M Salvay, M Zelivyanskaya, and L D Shea. Gene delivery by surface immobilization of plasmid to tissue-engineering scaffolds. *Gene Ther.*, 17(9):1134–1141, September 2010.

Alvaro San Millan, Jose Antonio Escudero, Danna R Gifford, Didier Mazel, and R Craig MacLean. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol*, 1(1):10, November 2016.

Elizabeth P Sauer, Jessica L Vandewalle, Melinda J Bootsma, and Sandra L McLellan. Detection of the human specific bacteroides genetic marker provides evidence of widespread sewage contamination of stormwater in the urban environment. *Water Res.*, 45(14):4081–4091, August 2011.

Melanie Schirmer, Ashley Garner, Hera Vlamakis, and Ramnik J Xavier. Microbial genes and pathways in inflammatory bowel disease, 2019.

Patrick D Schloss. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio*, 9(3), June 2018.

J Petra Schumann, David T Jones, and David R Woods. Induction of proteins during phage reactivation induced by UV irradiation, oxygen and peroxide in bacteroides fragilis. *FEMS Microbiol. Lett.*, 23(2-3):131–135, July 1984.

Oliver Schwengers, Patrick Barth, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty, and Alexander Goesmann. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom*, 6(10), October 2020.

Diya Sen, Geraldine A Van der Auwera, Linda M Rogers, Christopher M Thomas, Celeste J Brown, and Eva M Top. Broad-host-range plasmids from agricultural soils have IncP-1 backbones with diverse accessory genes. *Appl. Environ. Microbiol.*, 77(22):7975–7983, November 2011.

Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.*, 14(8):e1002533, August 2016.

Alon Shaiber and A Murat Eren. Anvi'o snakemake workflows. `http://merenlab.org/2018/07/09/anvio-snakemake-workflows/`, July 2018. Accessed: 2020-10-30.

Alon Shaiber, Amy D Willis, Tom O Delmont, Simon Roux, Lin-Xing Chen, Abigail C Schmid, Mahmoud Yousef, Andrea R Watson, Karen Lolans, Özcan C Esen, Sonny T M Lee, Nora Downey, Hilary G Morrison, Floyd E Dewhirst, Jessica L Mark Welch, and A Murat Eren. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.*, 21(1):292, December 2020.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, November 2003.

Julie Shareck, Young Choi, Byong Lee, and Carlos B Miguez. Cloning vectors based on cryptic plasmids isolated from lactic acid bacteria: their characteristics and potential applications in biotechnology. *Crit. Rev. Biotechnol.*, 24(4):155–208, 2004.

Manvi Sharma, Yuanyuan Li, Matthew L Stoll, and Trygve O Tollefsbol. The epigenetic connection between the gut microbiome in obesity and diabetes. *Front. Genet.*, 10, January 2020.

Sudarshan A Shetty, Ben Kuipers, Siavash Atashgahi, Steven Aalvink, Hauke Smidt, and Willem M de Vos. Inter-species metabolic interactions in an in-vitro minimal human gut microbiome of core bacteria, 2022.

Andrey N Shkoporov and Colin Hill. Bacteriophages of the human gut: The "known unknown" of the microbiome. *Cell Host Microbe*, 25(2):195–209, February 2019.

L Simonsen. The existence conditions for bacterial plasmids: Theory and reality. *Microb. Ecol.*, 22(1):187–205, December 1991.

Samuel S Slattery, Andrew Diamond, Helen Wang, Jasmine A Therrien, Jeremy T Lant, Teah Jazey, Kyle Lee, Zachary Klassen, Isabel Desgagné-Penix, Bogumil J Karas, and David R Edgell. An expanded Plasmid-Based genetic toolbox enables cas9 genome editing and stable maintenance of synthetic pathways in phaeodactylum tricornutum. *ACS Synth. Biol.*, 7(2):328–338, February 2018.

Kornelia Smalla, Sven Jechalke, and Eva M Top. Plasmid detection, characterization and ecology. *Microbiology spectrum*, 3(1), February 2015.

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, October 2011.

C J Smith. Development and use of cloning systems for bacteroides fragilis: cloning of a plasmid-encoded clindamycin resistance determinant. *J. Bacteriol.*, 164(1):294–301, October 1985.

C J Smith, L A Rollins, and A C Parker. Nucleotide sequence determination and genetic analysis of the bacteroides plasmid, pBI143. *Plasmid*, 34(3):211–222, November 1995.

Gloria del Solar, Gloria del Solar, Rafael Giraldo, Mariia Jesuus Ruiz-Echevarrria, Manuel Espinosa, and Ramon Diiaz-Orejas. Replication and control of circular bacterial plasmids, 1998.

Erica D Sonnenburg and Justin L Sonnenburg. The ancestral and industrialized gut microbiota and implications for human health. *Nat. Rev. Microbiol.*, 17(6):383–390, June 2019.

Masahiro Sota, Hirokazu Yano, Julie M Hughes, Gary W Daughdrill, Zaid Abdo, Larry J Forney, and Eva M Top. Shifts in the host range of a promiscuous plasmid through parallel evolution of its replication initiation protein. *ISME J.*, 4(12):1568–1580, December 2010.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, 2017.

Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nat. Commun.*, 9(1):1–8, June 2018.

S Sukupolvi and C D O'Connor. TraT lipoprotein, a plasmid-specified mediator of interactions between gram-negative bacteria and their environment. *Microbiol. Rev.*, 54(4): 331–341, December 1990.

David Summers. *The Biology of Plasmids*. 1993.

Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, Francisco M Cornejo-Castillo, Paul I Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G Acinas, and Peer Bork. Ocean plankton. structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, May 2015.

Christian J Sund, Edson R Rocha, Arthur O Tzianabos, W Greg Wells, Jason M Gee, Michael A Reott, Dorcas P O'Rourke, and C Jeffrey Smith. The bacteroides fragilis transcriptome response to oxygen and H2O2: the role of OxyR and its effect on survival and virulence. *Mol. Microbiol.*, 67(1):129–142, January 2008.

Fabian Svara and Daniel J Rankin. The evolution of plasmid-carried antibiotic resistance. *BMC Evol. Biol.*, 11:130, May 2011.

Richard Sykes. The 2009 garrod lecture: the evolution of antimicrobial resistance: a darwinian perspective. *J. Antimicrob. Chemother.*, 65(9):1842–1852, September 2010.

Huizi Tan, Jianxin Zhao, Hao Zhang, Qixiao Zhai, and Wei Chen. Novel strains of bacteroides fragilis and bacteroides ovatus alleviate the LPS-induced inflammation in mice. *Appl. Microbiol. Biotechnol.*, 103(5):2353–2365, March 2019.

The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

Christopher M Thomas. *Horizontal Gene Pool: Bacterial Plasmids and Gene Spread*. CRC Press, September 2003.

Christopher M Thomas. Plasmid incompatibility, 2014a.

Christopher M Thomas. Evolution and population genetics of bacterial plasmids, 2014b.

Gérald Thouand and Robert Marks. *Bioluminescence: Fundamentals and Applications in Biotechnology - Volume 3*. Springer, January 2016.

F Trigodet, K Lolans, E Fogarty, and others. High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Mol. Ecol.*, 2022.

Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, October 2007.

Carles Ubeda, Elisa Maiques, Erwin Knecht, Iñigo Lasa, Richard P Novick, and José R Penadés. Antibiotic-induced SOS response promotes horizontal dissemination of pathogenicity island-encoded virulence factors in staphylococci. *Mol. Microbiol.*, 56(3):836–844, May 2005.

US-gov. Usgs water data for the nation. `https://waterdata.usgs.gov/nwis`. Accessed: 2023-3-15.

Daniel R Utter, Gary G Borisy, A Murat Eren, Colleen M Cavanaugh, and Jessica L Mark Welch. Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol.*, 21(1):293, December 2020.

Maartje A H J van Kessel, Daan R Speth, Mads Albertsen, Per H Nielsen, Huub J M Op den Camp, Boran Kartal, Mike S M Jetten, and Sebastian Lücker. Complete nitrification by a single microorganism. *Nature*, 528(7583):555–559, December 2015.

Tommi Vatanen, Karolina S Jabbar, Terhi Ruohtula, Jarno Honkanen, Julian Avila-Pacheco, Heli Siljander, Martin Stražar, Sami Oikarinen, Heikki Hyöty, Jorma Ilonen, Caroline M Mitchell, Moran Yassour, Suvi M Virtanen, Clary B Clish, Damian R Plichta, Hera Vlamakis, Mikael Knip, and Ramnik J Xavier. Mobile genetic elements from the maternal microbiome shape infant gut microbial assembly and metabolism. *Cell*, 185(26):4921–4936.e15, December 2022.

Nileena Velappan, Daniele Sblattero, Leslie Chasteen, Peter Pavlik, and Andrew R M Bradbury. Plasmid incompatibility: more compatible than previously thought? *Protein Eng. Des. Sel.*, 20(7):309–313, March 2007.

Antony T Vincent, Nava Hosseini, and Steve J Charette. The *Aeromonas salmonicida* plasmidome: a model of modular evolution and genetic diversity, 2021.

J H Vineis, D L Ringus, H G Morrison, T O Delmont, S Dalal, L H Raffals, D A Antonopoulos, D T Rubin, A M Eren, E B Chang, and M L Sogin. Patient-Specific bacteroides genome variants in pouchitis. *MBio*, 7(6), November 2016.

Corneliu Ovidiu Vrancianu, Laura Ioana Popa, Coralia Bleotu, and Mariana Carmen Chifiriuc. Targeting plasmids to limit acquisition and transmission of antimicrobial resistance. *Front. Microbiol.*, 11:761, May 2020.

Zhenmao Wan, Joseph Varshavsky, Sushma Teegala, Jamille McLawrence, and Noel L Goddard. Measuring the rate of conjugal plasmid transfer in a bacterial population using quantitative PCR, 2011.

Guan H Wang, Bao F Sun, Tuan L Xiong, Yan K Wang, Kristen E Murfin, Jin H Xiao, and Da W Huang. Bacteriophage WO can mediate horizontal gene transfer in endosymbiotic wolbachia genomes. *Front. Microbiol.*, 0, 2016.

Stephen C Watts, Scott C Ritchie, Michael Inouye, and Kathryn E Holt. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics*, 35(6):1064–1066, August 2018.

Chengping Wen, Zhijun Zheng, Tiejuan Shao, Lin Liu, Zhijun Xie, Emmanuelle Le Chatelier, Zhixing He, Wendi Zhong, Yongsheng Fan, Linshuang Zhang, Haichang Li, Chunyan Wu, Changfeng Hu, Qian Xu, Jia Zhou, Shunfeng Cai, Dawei Wang, Yun Huang, Maxime Breban, Nan Qin, and Stanislav Dusko Ehrlich. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.*, 18, 2017.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, June 2016.

Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome Biol.*, 20(1):1–13, November 2019.

World Health Organization. WHO report on surveillance of antibiotic consumption: 2016-2018 early implementation. Technical report, 2018.

Rachel A F Wozniak and Matthew K Waldor. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.*, 8(8): 552–563, August 2010.

Yao Xia, Yanshan Zhu, Qier Li, and Jiahai Lu. Human gut resistome can be country-specific. *PeerJ*, 7:e6389, March 2019.

Hailiang Xie, Ruijin Guo, Huanzi Zhong, Qiang Feng, Zhou Lan, Bingcai Qin, Kirsten J Ward, Matthew A Jackson, Yan Xia, Xu Chen, Bing Chen, Huihua Xia, Changlu Xu, Fei Li, Xun Xu, Jumana Yousuf Al-Aama, Huanming Yang, Jian Wang, Karsten Kristiansen, Jun Wang, Claire J Steves, Jordana T Bell, Junhua Li, Timothy D Spector, and Huijue Jia. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell systems*, 3(6):572, December 2016.

Wenguang Xiong, Yongxue Sun, Xueyao Ding, Mianzhi Wang, and Zhenling Zeng. Selective pressure of antibiotics on ARGs and bacterial communities in manure-polluted freshwater-sediment microcosms. *Front. Microbiol.*, 0, 2015.

Shinichi Yachida, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Takeshi Nakajima, Taku Sakamoto, Hikaru Watanabe, Keigo Masuda, Yuichiro Nishimoto, Masaru Kubo, Fumie Hosoda, Hirofumi Rokutan, Minori Matsumoto, Hiroyuki Takamaru, Masayoshi Yamada, Takahisa Matsuda, Motoki Iwasaki, Taiki Yamaji, Tatsuo Yachida, Tomoyoshi Soga, Ken Kurokawa, Atsushi Toyoda, Yoshitoshi Ogura, Tetsuya Hayashi, Masanori Hatakeyama, Hitoshi Nakagama, Yutaka Saito, Shinji Fukuda, Tatsuhiro Shibata, and Takuji Yamada. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.*, 25(6):968–976, June 2019.

Moran Yassour, Eeva Jason, Larson J Hogstrom, Timothy D Arthur, Heli Siljander Surya Tripathi, Jenni Selvenius, Sami Oikarinen, Heikki Hyöty, Jorma Ilonen Suvi M Virtanen, Pamela Ferretti, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Nicola Segata, Hera Vlamakis, Eric S Lander, Curtis Huttenhower, Mikael Knip, and Ramnik J Xavier. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host Microbe*, 24(1):146–154.e4, July 2018.

Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, Andrew C Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J Gregory Caporaso, Catherine A Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, May 2012.

Michael K Yu, Emily C Fogarty, and A Murat Eren. The genetic and ecological landscape of plasmids in the human gut. November 2020.

Natalya Yutin, Kira S Makarova, Ayal B Gussow, Mart Krupovic, Anca Segall, Robert A Edwards, and Eugene V Koonin. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol*, 3(1):38–46, January 2018.

David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, Jotham Suez, Jemal Ali Mahdi, Elad Matot, Gal Malka, Noa Kosower, Michal Rein, Gili Zilberman-Schapira, Lenka Dohnalová, Meirav Pevsner-Fischer, Rony Bikovsky, Zamir Halpern, Eran Elinav, and Eran Segal. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094, November 2015.

Alexandra Zhernakova, Alexander Kurilshikov, Marc Jan Bonder, Ettje F Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila, Gwen Falony,

Sara Vieira-Silva, Jun Wang, Floris Imhann, Eelke Brandsma, Soesma A Jankipersadsing, Marie Joossens, Maria Carmen Cenit, Patrick Deelen, Morris A Swertz, LifeLines cohort study, Rinse K Weersma, Edith J M Feskens, Mihai G Netea, Dirk Gevers, Daisy Jonkers, Lude Franke, Yurii S Aulchenko, Curtis Huttenhower, Jeroen Raes, Marten H Hofker, Ramnik J Xavier, Cisca Wijmenga, and Jingyuan Fu. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–569, April 2016.

Fengfeng Zhou and Ying Xu. cbar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, 26(16): 2051–2052, August 2010.

W Zillig, D Prangishvilli, C Schleper, M Elferink, I Holz, S Albers, D Janekovic, and D Götz. Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic archaea. *FEMS Microbiol. Rev.*, 18(2-3):225–236, May 1996.

Naamah Levy Zitomersky, Michael J Coyne, and Laurie E Comstock. Longitudinal analysis of the prevalence, maintenance, and IgA response to species of the order bacteroidales in the human gut. *Infect. Immun.*, 79(5):2012, May 2011.