Winter 2020

Bacterial Genome and Population Dynamics with Implications for Public Health

Cooper John Park
University of New Hampshire, Durham

Bacterial Genome and Population Dynamics with Implications for Public Health

BY

Cooper J. Park

B.S. Microbiology, Weber State University, 2017

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

In

Molecular and Evolutionary Systems Biology

December 2020

This dissertation was examined and approved in partial fulfillment of the requirements for the degree of Ph.D. in Molecular and Evolutionary Systems Biology by:

Dissertation Director, Cheryl P. Andam, Ph.D.,
Primary advisor
Affiliate Faculty, Molecular, Cellular, and Biomedical Sciences
University of New Hampshire
(Assistant Professor, Biological Sciences
University at Albany, State University of New York)

W. Kelley Thomas, Ph.D.
Secondary advisor
Professor, Molecular, Cellular, and Biomedical Sciences
University of New Hampshire

Louis Tisa, Ph.D.,
Professor, Molecular, Cellular, and Biomedical Sciences
University of New Hampshire

Matthew MacManes, Ph.D.,
Associate Professor, Molecular, Cellular, and Biomedical Sciences
University of New Hampshire

Anissa Poleatewich, Ph.D.,
Assistant Professor, Agriculture, Nutrition, and Food Systems
University of New Hampshire

On November 20, 2020

Approval signatures are on file with the University of New Hampshire Graduate School.

ACKNOWLEDGEMENTS

I first want to thank my advisor, Dr. Cheryl Andam, for her incredible support and mentorship during my time at UNH. Nearly every professional opportunity I have had over the last few years is due to your constant encouragement and unrelenting commitment to my success. I will forever be grateful for everything you taught me. I would also like to thank the rest of my committee; Dr. Anissa Poleatewich, Dr. Lou Tisa, Dr. Matt MacManes, and Dr. W. Kelley Thomas. Each of you were generous with your time and unwavering in your support.

I owe more gratitude than I could ever express to my partner, Abby Owen. Through every moment of this experience you've been by my side to help me endure the failures and, more importantly, celebrate the successes. I'm so excited to continue our adventurous life together. I love you.

To my family, I literally don't have enough space in this page to thank each of you properly. Mom and Tony, your weekly phone calls and undying support are the foundation of my accomplishments. Nothing I've achieved would be possible without you. Dad, you have been my escape from some of the toughest moments in this experience with our nightly gaming. Through all the ups and downs (of our K/D and my PhD) you have been there with your motivation and advice to keep going. Thank you so much. To all of my siblings; Sydney, Braxton, A.J., Kyler, Jaydn, Chris, and Alyssa, you have been a constant source of humor and support through the years. Seeing each of you develop your unique personality and achievements has been my own motivation to constantly improve. I'm proud of each of you, and thankful for what you've taught me. To Abby's parents, Jim and Valerie Owen, thank you for being my unexpected New Hampshire family. You have been crucial to making New Hampshire feel like home, and I appreciate everything you've done for me and Abby.

Lastly, I want to thank everyone I met through UNH for their support. Most importantly, Toni Westbrook for being the reason I'm a bioinformatician, Kayti Belknap for being my best friend and constant source of support through the graduate student process, and Josh Smith and Isaiah Lee for being phenomenal friends and lab mates that have always been supportive and constructive in improving my research.

I would like to thank the resources that funded my research at UNH, including the New Hampshire Agricultural Experiment Station, the National Science Foundation, the New Hampshire Space Grant Consortium, the Edward F. Landry Scholarship, the Dickie Family Scholarship, the UNH graduate school travel grants, and the Robert & Ruth Zsigray Academic Enrichment Fund.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

ABSTRACT

BACTERIAL GENOME AND POPULATION DYNAMICS WITH IMPLICATIONS FOR PUBLIC HEALTH

by

Cooper J. Park

University of New Hampshire

Bacterial populations are extraordinarily heterogeneous. Despite growing clonally, these populations are often composed of multiple lineages distinguished by both phenotypic and genetic differences that are caused by both allelic and whole gene variation. Such genomic mosaicism and within-species variation can significantly impact a species' response to selective pressures from antibiotic use, vaccination, immune responses and host environment. One important process that contributes to this phenomenon is recombination, the exchange of very similar DNA sequences between strains which can result to either the addition or replacement of homologous genes. Current models of microbial recombination incorporate the null expectation that recombination is a homogeneous process across a species, whereby different lineages of the same species and different genes within a genome exhibit the same rates of DNA donation and receipt. However, recent work has demonstrated that intra-species recombination rates can differ even between strains. This dissertation attempts to elucidate the extent of, and the processes underlying, heterogeneity in genomic content in microbial species and populations relevant to human health. The first chapter addresses the best-known producer of the tetracycline class of antibiotics, *Streptomyces rimosus*. Results suggest that even strains appearing nearly identical in

a core-genome phylogeny have divergent biosynthetic gene cluster content, emphasizing the importance of analyzing entire populations in drug discovery protocols. The second chapter explores the population dynamics of one of the most common causes of foodborne illness in the world, *Salmonella enterica* with results that indicate the evolution of ecologically unique subspecies of *S. enterica* are intricately linked by heterogeneous recombination. The third and fourth chapters demonstrate similar patterns of genomic diversity and recombination of clinically relevant genes in populations of *Campylobacter jejuni* and *S. enterica* collected from hospitals in New Hampshire in 2017. Finally, the fifth chapter describes a novel bioinformatic program called HERO which rapidly identifies and visualizes donor-recipient recombination pairs from a bacterial population. It also reports measures of heterogeneity in the population's total recombination including events per donor-recipient pair, recombined DNA fragment length and the number of events per gene. Collectively, these results contribute to the growing evidence that intra-species heterogeneity plays a role in the evolution and management of bacterial species associated with public health.

INTRODUCTION

Microbial species are a critical component of human health and disease. A taxonomically and functionally diverse community of microbes are implicated in the emergence and management of infectious diseases. Dramatic improvements in technological and computational techniques for genome sequencing have shown that this functional and genetic diversity can exist even within individual bacterial species. Such variation is often described by the contents of a species' pangenome, defined as the collection of all unique gene families present within it (McInerney, McNally, and O'Connell 2017; Tettelin et al. 2005). Many bacterial species are characterized by large pangenomes that are made up of relatively small core genomes (i.e., genes found ubiquitously among representative individuals) and larger accessory genomes (i.e., rarer genes found in one or few individuals) (McInerney, McNally, and O'Connell 2017). For example, a study of 2,085 *Escherichia coli* genomes, a species with roughly 5,000 genes per strain, revealed a pangenome of ~90,000 genes, dominated by the accessory genome (Land et al. 2015). Such genomic mosaicism and within-species variation can significantly impact a species' response to selective pressures from antibiotic use, vaccination, immune responses and host environment (Sela et al. 2018; Leventhal et al. 2018; Brüggemann et al. 2018).

There is a critical need to understand the dynamics of the genomic variation that exists within bacterial species and the mechanisms which regulate it. One important process that contributes to this phenomenon is recombination, the exchange of very similar DNA sequences between strains which can result to either the addition or replacement of homologous genes

(Didelot and Maiden 2010; Didelot et al. 2012). Current models of microbial recombination incorporate the null expectation that recombination is a homogeneous process across a species, whereby different lineages of the same species and different genes within a genome exhibit the same rates of DNA donation and receipt (Vos et al. 2015). However, recent work has demonstrated that intra-species recombination rates can differ between strains (Beiko, Harlow, and Ragan 2005; Chewapreecha et al. 2014). These highways of recombination are likely to represent specific lineages that function as hubs of gene flow, facilitating the rapid spread of genes necessary for rapid adaptation to ecological changes (Chewapreecha et al. 2014). Therefore, elucidating the influence of intra-species recombination on clinically relevant bacterial populations should provide new insights on the prevention and treatment of bacterial pathogens.

In this dissertation, I describe six studies which aimed to evaluate the overall genomic variability and characterize the rates and patterns of intra-species recombination within several bacterial populations of clinical importance. In the first two chapters I describe two studies that explore the evolutionary impact of recombination on the antibiotic producing species *Streptomyces rimosus* (Chapter 1) and a major worldwide foodborne pathogenic species *Salmonella enterica* (Chapter 2). In the next two chapters I describe two additional studies which apply the concepts of intra-species variation and recombination to analyze the genomic epidemiology of two bacterial populations sampled by the New Hampshire Department of Health and Human Services during 2017, *Campylobacter jejuni* (Chapter 3) and *Salmonella enterica* (Chapter 4). Finally, Chapter 5 describes the invention of an open source bioinformatic program called HERO (Highways Enumerated by Recombination Observations). HERO takes a collection

of predicted recombination events within a bacterial population to identify donor-recipient relationships and related metrics of recombination. The results from these studies demonstrate that within-species variation genome structure and recombination dynamics in both clinically beneficial and detrimental bacterial species is an important driver of bacterial evolution and adaptation. In conclusion, this dissertation provides new evidence that within-species genomic diversity plays a significant role in our management of future public health crises, including disease outbreaks and multidrug resistance.

CHAPTER 1

Within-species genomic variation and variable patterns of recombination in the tetracycline producer *Streptomyces rimosus*

Cooper J. Park, Cheryl P. Andam

Article published in *Frontiers in Microbiology*

Presented here with permission from publisher (see Apendix 1)

ABSTRACT

Streptomyces rimosus is best known as the primary source of the tetracycline class of antibiotics, most notably oxytetracycline, which have been widely used against many gram-positive and gram-negative pathogens and protozoan parasites. However, despite the medical and agricultural importance of *S. rimosus*, little is known of its evolutionary history and genome dynamics. In this study, we aim to elucidate the pan-genome characteristics and phylogenetic relationships of 32 *S. rimosus* genomes. The *S. rimosus* pan-genome contains more than 22,000 orthologous gene clusters, and approximately 8.8% of these genes constitutes the core genome. A large part of the accessory genome is composed of 9,646 strain-specific genes. *S. rimosus* exhibits an open pan-genome (decay parameter $\alpha = 0.83$) and high gene diversity between strains (genomic fluidity $\varphi = 0.12$). We also observed strain-level variation in the distribution and abundance of biosynthetic gene clusters (BGCs) and that each individual *S. rimosus* genome has a unique repertoire of BGCs. Lastly, we observed variation in recombination, with some strains donating or receiving DNA more often than others, strains that tend to frequently recombine with specific partners, genes that often experience recombination more than others, and variable sizes of recombined

DNA sequences. We conclude that the high levels of inter-strain genomic variation in *S. rimosus* is partly explained by differences in recombination among strains. These results have important implications on current efforts for natural drug discovery, the ecological role of strain-level variation in microbial populations, and addressing the fundamental question of why microbes have pan-genomes.

INTRODUCTION

The gram-positive genus *Streptomyces* (phylum Actinobacteria) constitutes a highly diverse group that is widely distributed in nature. Streptomyces are prolific producers of bioactive specialized metabolites that have adaptive functions in nature and have found extensive utility in human medicine(Xu et al. 2016; Kinashi 2011; Cruz-Morales et al. 2016). They are known as the major source of naturally derived antibiotics and many pharmaceutically relevant compounds (e.g., antifungals, antitumor, antihelminths, antiprotozoans, immunosuppressants) (Kinashi 2011). Many invertebrates such as wasps and ants also use the antibiotics produced by their *Streptomyces* symbionts to protect themselves against infection (Seipke, Kaltenpoth, and Hutchings 2012; Kaltenpoth et al. 2005). In contrast to most bacteria, *Streptomyces* species are characterized by complex secondary metabolism and a fungal-like morphological differentiation that involves the formation of branching, filamentous vegetative growth and aerial hyphae bearing long chains of reproductive spores (Flärdh and Buttner 2009); hence they were originally misclassified as fungi. The formation of aerial mycelium corresponds to the production of secondary metabolites such as antibiotics (Barka et al. 2016). Current estimate of the number of known *Streptomyces* species is approximately 650 (Labeda et al. 2012), making it one of the largest genera in the bacterial domain.

Whole genome sequencing of closely related, locally co-occurring microbial strains has revealed the existence of tremendous diversity within a species, arising from both allelic and gene content differences (Chang et al. 2018; Croucher et al. 2014; A. Zhu et al. 2015; Levade et al. 2017). Hence, using traditional taxonomic methods, it is difficult to delineate two lineages

that are considered the same species yet vary substantially in gene content (Jaspers and Overmann 2004; Land et al. 2015; Segerman 2012). For example, fuzzy species i.e., those that that do not form clear, distinct species boundaries due to frequent gene exchange through recombination, have been reported in *Neisseria meningitidis* (William P. Hanage, Fraser, and Spratt 2005). Hybrid lineages as in the case of *Klebsiella pneumoniae* sequence type [ST] 258 have been formed via a large chromosomal replacement event (Chen et al. 2014). Such genomic mosaicism and within-species variation can significantly impact a species' response to selective pressures from antibiotic use, vaccination, immune responses and host environment (Brüggemann et al. 2018; Leventhal et al. 2018; Sela et al. 2018). Within-species genomic variation has also been reported to impact species divergence (Papke et al. 2007; Youngblut et al. 2015), metabolic diversity and versatility (Silby et al. 2011), and symbiotic relationships (De Maayer et al. 2014) in microbes, with medically relevant implications. For example, hyper-recombinant strains of *Streptococcus pneumoniae* are associated with the highest levels of drug resistance (William Paul Hanage et al. 2009). One important process that generates genomic variation in microbial species is recombination, the exchange of very similar DNA sequences between strains, and which can result to either the addition or replacement of homologous genes (Didelot and Maiden 2010; Didelot et al. 2012). Most studies dealing with within-species genomic variation has been focused on antibiotic resistant pathogens [for example, (Andam et al. 2017; Grad et al. 2014; Grinberg et al. 2017; Lam et al. 2018)], yet rarely do we find investigations on antibiotic producers. In *Streptomyces*, genomic diversity between species has been widely investigated (Andam et al. 2016; Doroghazi and Metcalf 2013; Huguet-Tapia et al. 2016; J. N. Kim et al. 2015), but the extent, origins and functional role of genomic variation among closely related strains of the same species remains poorly understood.

In this study, we focus on *Streptomyces rimosus*, which is best known as the primary source of the tetracycline class of antibiotics, most notably oxytetracycline (Petković et al. 2006). Tetracyclines are noted for their broad spectrum antibacterial activity and since the 1940s, have been used against a wide range of both gram-positive and gram-negative pathogens, mycoplasmas, chlamydiae, rickettsiae and protozoan parasites (Chopra and Roberts 2001). Oxytetracycline, a well-studied polyketide natural product, is a bacteriostatic antibiotic that inhibits bacterial growth by reversibly binding to the 30S ribosomal subunit, thus inhibiting protein synthesis (Petković et al. 2006; Schnappinger and Hillen 1996). *S. rimosus* is also known to produce the polyene antifungal rimocidin (Davisson et al., 1951). Although the precise mechanism of action of rimocidins is still not well understood, antifungal activity seems to be due to polyene molecules causing the sterol-containing cell membrane to become permeable (Seco et al. 2005). Despite the medical and agricultural importance of *S. rimosus* and the variety of antibiotics it produces, little is known of its evolutionary history and genome characteristics. Here, we explore the pan-genome characteristics and phylogenetic relationships of 32 *S. rimosus* genomes. We report high levels of inter-strain genomic variation, including the differential distribution and abundance of biosynthetic gene clusters (BGCs) among strains. BGCs represent a collection of genes that, together are responsible for the production of a specific secondary metabolite, such as antibiotics. We also observed high frequency of recombination which may partly explain the large genomic variation among strains; however, recombination is biased, with some strains exhibiting more frequent donation or receipt of DNA than other strains. These results have important implications on current efforts for natural drug discovery, the ecological

role of strain-level genomic variation in microbial populations, and addressing the fundamental question of why microbes have pan-genomes.

METHODS

Dataset

A total of 32 genomes of *S. rimosus* available in November 2018 were downloaded from the RefSeq database of the National Center for Biotechnology Information (NCBI). Accession numbers and genomic information (genome size, % GC content, number of genes, number of protein-coding genes) are shown in Supplementary Table S1. To maintain consistency in gene annotations, the genomes were re-annotated using Prokka with default parameters (Torsten Seemann 2014).

Pan-genome and phylogenetic analysis

Core and accessory genes were identified using Roary with default settings (Page et al. 2015). Roary iteratively pre-clusters protein sequences using CD-HIT (Li and Godzik 2006), a fast program for clustering and comparing, which results to a substantially reduced set of data. Sequences in this reduced dataset were compared using all-against-all BLASTP (Altschul et al. 1990) and were then clustered the second time using Markov clustering (Enright, Van Dongen, and Ouzounis 2002). Each orthologous gene family from the merged CD-HIT and MCL were aligned using MAFFT (Kazutaka Katoh et al. 2002). We used Phandango (Hadfield et al. 2018) to visualize the presence-absence of genes per strain. The gene sequence alignments of each

identified core gene family were concatenated to give a single core alignment, and a maximum-likelihood phylogeny was then generated using the program RAxML v.8.2.11 (Stamatakis 2006) with a general time reversible (GTR) nucleotide substitution model (Tavaré, 1986), four gamma categories for rate heterogeneity and 100 bootstrap replicates. The phylogenetic tree was visualized using the Interactive Tree of Life [iToL] (Letunic and Bork 2016).

We used the program micropan (Snipen and Liland 2015) implemented in R (R Core Team 2019) to calculate the pan-genome's decay parameter (α) (Tettelin et al. 2008) and genomic fluidity (φ) (Kislyuk et al. 2011). The decay parameter measures the number of new gene clusters observed when genomes are ordered in a random way, which provides an indication of the openness or closeness of a pan-genome (Tettelin et al. 2008). An open pan-genome indicates that the number of new genes to be observed in future genomes is large, while a closed pan-genome indicates that after a certain number of sequenced genomes are added, the number of new genes discovered reaches a plateau (Tettelin et al. 2008). The genomic fluidity is a measure of the dissimilarity of genomes based on the degree of overlap in gene content and is defined as the number of unique gene families divided by the total number of gene families (Kislyuk et al. 2011). Both metrics are used to evaluate within-species genomic variation. Genome-wide average nucleotide identity (ANI) of all orthologous genes shared between any two genomes was calculated for all possible pairs of genomes (Jain et al. 2018). ANI is a robust similarity metric that has been widely used to resolve inter- and intra-strain relatedness. The threshold value of 95% has been widely used as a cutoff for comparisons belonging to the same or different species (Jain et al. 2018).

BGCs encoding secondary metabolites were predicted and annotated using the standalone version of antiSMASH 4.1 (Weber et al. 2015). antiSMASH predicts BGCs using signature profile Hidden Markov Models (pHMMs) derived from multiple sequence alignments of experimentally characterized signature proteins or protein domains of known BGCs (Blin et al., 2017). It then aligns the identified regions at the gene cluster level to their nearest relatives from a database containing all other known gene clusters (Weber et al. 2015). BGCs that encode for oxytetracycline and rimocidin were identified by searching all the genomes for homologs of each of the genes comprising the two BGCs using BLASTP (Altschul et al. 1990) with a minimum e-value of 10^{-10}. Individual genes in a BGC obtained from previous studies (Seco et al. 2005; W. Zhang et al. 2006) were used as query sequences. Presence of the BGC was ascertained if there were BLASTP hits for at least 90% of the genes within the BGC. Sequences for the individual genes of the two BGCs were obtained from the Database of BIoSynthesis cluster CUrated and InTegrated (DoBISCUIT) (Ichikawa et al. 2013) based on previous studies of the oxytetracycline and rimocidin BGCs (Seco et al. 2005; W. Zhang et al. 2006).

Recombination detection

We used three approaches to detect recombination in the population. First, the pairwise homoplasy index or PHI (Φw) test was used to determine the statistical likelihood of recombination being present in our dataset (Bruen, Philippe, and Bryant 2006). This statistic measures the genealogical correlation or similarity of adjacent sites. Under the null hypothesis of no recombination, the genealogical correlation of adjacent sites is invariant to permutations of the sites as all sites have the same history (Bruen, Philippe, and Bryant 2006). Significance of the observed Φw was obtained using a permutation test. We then visualized potential recombination

events using Splitstree v.4.14.4, which integrates reticulations due to recombinations in phylogenetic relationships rather than forcing the data to be represented in a bifurcating tree (Huson 1998). Next, we ran fastGEAR (Mostowy et al. 2017) with default parameters to detect genome-wide mosaicism. Using the individual sequence alignments of all core and shared accessory genes, we first identified sequence clusters were first identified using BAPS (Cheng, Rong, and Huang 2016) implemented in fastGEAR. fastGEAR infers the population structure of individual alignments using a Hidden Markov Model to identify lineages in an alignment. Lineages are defined as groups which are genetically distinct in at least 50% of the alignment. Within each lineage, recombinations are identified by comparing every nucleotide site in the target sequence to all remaining lineages and asks whether it is more similar to something else compared to other strains in the same lineage. In other words, fastGEAR infers recombination by searching for similar nucleotide segments between diverse sequence clusters. To test the significance of the inferred recombinations and identify false-positive recombinations, fastGEAR uses a diversity test, wherein the diversity of the fragment in question is different compared to its background. To predict the origin of the recently recombined regions, the sequences on which the recombination event was predicted to have occurred were first extracted from the genome data. The recombined regions were then used as query sequences in BLASTN (Altschul et al. 1990) searches against all possible genomes from the identified donor lineage as well as from the non-redundant (nr) nucleotide database in NCBI. The top BLAST hit with the highest bit score was considered as the potential donor, provided that the hit covered at least 50% of the recombination fragment length and had a minimum of 99% nucleotide identity.

RESULTS

Pan-genome characteristics of S. rimosus

We used a total of 32 *S. rimosus* genomes downloaded from the RefSeq database of NCBI (Supplementary Table S1). Genome sizes range from 8.14-10.02 Mb (mean = 9.20 Mb), while the number of predicted genes per genome ranges from 7,071 – 8,666 (mean = 8,020). The % G+C content also varies among genomes, ranging from 71.7 – 72.1%. We used Roary (Page et al. 2015) to calculate the *S. rimosus* pan-genome, defined as the totality of genes present in a group of genomes (Page et al. 2015). Roary classifies orthologous gene families into core genes and accessory genes. Core genes are present in 99% ≤ strains ≤ 100% (Supplementary Tables S2 and S3). To take sequencing and assembly errors into account, Roary also calculates the number of soft core genes which are present in 95% ≤ strains < 99%. Accessory genes comprise the shell genes which are present in 15% ≤ strains < 95% and cloud genes which are present in < 15% of strains (Figure 1a). We found a considerably small core genome (1,945 genes) comprising 8.8% of the pan-genome (22,114 genes). Broadening our definition of the core genome to incorporate the soft core genes still only represented approximately 17% of the total pan-genome. The core genome comprises 22.44 - 27.51% of each individual genome. It is also notable that the vast majority of accessory genes (9,646, representing 44% of the pan-genome) are unique to a single strain. In microbes, large accessory genomes and high number of strain-specific genes are often associated with horizontal gene transfer [HGT] (Pohl et al. 2014; Vos and Eyre-Walker 2017; B. Zhu et al. 2016).

The size of the pan-genome and its increase/decrease in size upon addition of new strains can be used to predict the future rate of discovery of novel genes in a species (Medini et al. 2005; Tettelin et al. 2008). We used the program micropan to estimate the openness of the *S. rimosus* pan-genome by using the Heap's power law function (Tettelin et al. 2008) for all possible permutations of all *S. rimosus* genomes. We calculated the decay parameter α, wherein an $\alpha > 1.0$ indicates that the size of the pan-genome approaches a constant as more genomes are sampled (i.e., the pan-genome is closed), while $\alpha < 1.0$ indicates that the size of the pan-genome is increasing and unbounded by the number of genomes considered (i.e., the size of the pan-genome follows Heaps' law and the pan-genome is open) (Medini et al. 2005; Tettelin et al. 2008). We obtained an $\alpha = 0.83$ using 100 permutations in *S. rimosus* and suggests an open pan-genome; hence, we are likely to find new genes as more genomes are sequenced in the future. The openness of pan-genome reflects the diversity of the gene pool within bacterial species, and is often associated with bacterial species that inhabit multiple environments or have different mechanisms and opportunities for gene exchange (Rouli et al. 2015; Brito et al. 2018). We find that the pan-genome of *S. rimosus* increases with the addition of new genomes, while the core genome decreases and begins to plateau at approximately 20 genomes (Figure 1b). The number of new, previously unseen, genes found as each genome is added to the plot averages 450 (Figure 1c). Finally, we also show the number of unique genes overall that have been observed exactly once continues to increase as each genome is added (Figure 1c).

To estimate the degree of overlap with respect to gene cluster content between any two genomes, we also calculated the genomic fluidity (φ), which provides an overview of gene-level similarity between genomes and is defined as the number of unique gene families divided by the

total number of gene families (Kislyuk et al. 2011). Fluidity values range from 0-1, with 0.0 to indicate that the two genomes contain identical gene clusters, while 1.0 if the two genomes are non-overlapping (Kislyuk et al. 2011). Hence, a fluidity value of 0.2 for example implies that 20% of the genes are unique to their host genome and the remaining 80% are shared between genomes (Halachev, Loman, and Pallen 2011). We obtained a genomic fluidity value of 0.12, which suggests that *S. rimosus* has a high degree of genomic diversity and is within the range found in other bacterial species (Halachev, Loman, and Pallen 2011; Kislyuk et al. 2011).

To determine the degree of genomic relatedness and hence clarify whether these 32 genomes belong to the same species, we calculated the pairwise ANI for all possible pairs of genomes. ANI calculates the average nucleotide identity of all orthologous genes shared between any two genomes and organisms belonging to the same species typically exhibit ≥95% ANI (Jain et al. 2018). The distribution of pairwise ANI values reveal that the *S. rimosus* genomes are within the 95% cutoff and should therefore considered the same species (Figure 1d, e and Supplementary Table S4). Strain NRRL WC-3904 exhibits a slightly lower ANI value of 94% when compared to the rest of the genomes in the dataset. To further visualize the distribution of genes among the strains, we generated a pan-genome matrix using Roary and Phandango (Figure 1f). We find that NRRL WC-3904 exhibits a highly divergent accessory genome profile compared to the remaining 31 genomes, which may explain its slightly lower ANI values.

Strain-level variation in the distribution and abundance of BGCs

Streptomyces are renowned for their ability to produce structurally diverse natural products (called secondary metabolites), many of which are widely used in medicine, agriculture and bioenergy processes. Secondary metabolites differ from primary metabolites in that they are not involved in essential metabolic activities required for normal growth and reproduction of the organism, but may contribute significantly to an individual's fitness and ecological adaptation (Zotchev 2014). Mining bacterial genomes has shown that their potential for producing secondary metabolites and other bioactive compounds is much higher than what is observed in the laboratory (Doroghazi and Metcalf 2013), and hence has important implications in discovering novel bioactive compounds.

Biosynthesis of secondary metabolites is typically governed by 10–30 genes organized as clusters in the genome, allowing the coordinated expression of the genes involved in their biosynthesis, resistance and efflux (Zotchev 2014). We used antiSMASH 4.1 (Weber et al. 2015) to identify BGCs present in each *S. rimosus* genome. Each genome harbors 35-71 BGCs, with more than half of the BGCs predicted to produce polyketide (PKS) and non-ribosomal peptide synthetase (NRPS), or hybrids of the two (Figure 2a). This range in BGC content in *S. rimosus* is consistent with results from previous BGC surveys in other *Streptomyces* species (Choudoir, Pepe-Ranney, and Buckley 2018; Seipke 2015; Seipke et al. 2011; Vicente et al. 2018) and the widely studied actinobacterium *Salinispora* (Letzel et al. 2017; Udwary et al. 2007), although many BGCs often remain "silent" under standard laboratory conditions (Bentley et al. 2002; Ikeda et al. 2003; Ohnishi et al. 2008). Hybrid BGCs contain genes that code for more than one type of scaffold-synthesizing enzymes (Cimermancic et al. 2014; Zotchev 2014). Many of the

NRPS or PKS hybrids are found in one or few genomes: lantipeptide-t1pks-nrps hybrid in two genomes, melanin-t1pks hybrid in two genomes, phosphonate in one genome, t1pks-lassopeptide-nrps hybrid in two genomes, terpene-t2pks-t1pks-lassopeptide hybrid in two genomes, and terpene-t2pks-t1pks-lassopeptide hybrid in one genome. Aside from NRPS and PKS, other commonly shared BGCs are bacteriocin, butyrolactone, ectoine, lantipeptide, lassopeptide, melanin, nucleoside, siderophore, and terpene. Other BGCs are also differentially distributed among the 32 genomes: indoles in five genomes, ladderane in two genomes, phosphonate in one genome, and thiopeptide in one genome. Interestingly, Type II PKS and its hybrids were detected in 29 strains. Type II PKS synthesize tetracyclines and other aromatic polyketides such as anthracyclines, angucyclines and pentangular polyphenols, which are also widely used as antibiotics or chemotherapeutics (Hertweck et al. 2007; J. Kim and Yi 2012). Overall, we find that each individual *S. rimosus* genome harbor a unique combination of BGCs, further highlighting the extent of inter-strain genomic variation in *S. rimosus*. We note, however, that the reported numbers have likely been overestimated due to the low quality of some of the genome assemblies, which can affect the accurate BGC prediction in antiSMASH.

S. rimosus is particularly well known for its production of the antibiotics oxytetracycline and rimocidin (Petković et al. 2006). To determine the presence of BGCs that encode for these two antibiotics, we used BLASTP to search the 32 genomes for the individual genes of each BGC (Figure 2b, Tables S5 and S6). We found that, except for a single genome (R6-500MV9-R8), all *S. rimosus* genomes carry one or both BGCs. A total of 30 genomes had nearly 100% matches for each of the 21 genes found in the oxytetracycline

BGC, while two showed a match for only a single gene. On the other hand, the rimocidin BGC was detected in 28 genomes.

Frequent but biased recombination between strains

In *Streptomyces*, recombination is known to have greatly contributed to shaping its evolution and diversity, with some taxonomically recognized species exhibiting significant genetic mosaicism (Andam et al. 2016; Cheng, Rong, and Huang 2016; Doroghazi and Buckley 2010). To infer the phylogenetic relationships of the 32 *S. rimosus* genomes, the 1,945 core genes were aligned and concatenated, giving a total length of 2,017,766 bp. The core genome phylogeny reveals four clusters (Figure 3). Under the null hypothesis of no recombination, we calculated the PHI statistic (Bruen, Philippe, and Bryant 2006) and detected evidence for significant recombination in the core genome (p value = 0.0). Recombination in *S. rimosus* core genome can be visualized using Neighbor Net implemented in SplitsTree4, which shows the reticulations in their phylogenetic relationships (Huson 1998) (Figure 3a). To further characterize the extent of genome-wide recombination in *S. rimosus*, we ran fastGEAR (Mostowy et al. 2014) on individual sequence alignments of core and shared accessory genes. Each predicted recombination fragment was then used as a query in BLASTN (Altschul et al. 1990) against the predicted lineage of donors to identify the most likely donor-recipient linkages. We found that recombination is frequent, with a total of 2,148 genes that had experienced recombination. However, when we mapped the donor-recipient recombination partners, we found that although recombination is frequent, it does not impact all genomes similarly (Figure 3b). A total of 12 genomes were not identified to be either a donor or recipient of recombined DNA. Of those genomes wherein recombination was detected, there were genomes that appear to accept more

recombined DNA than others. We calculated the number of recombination events for any genome pair that is at least one standard deviation above the group's average of 36 recombination events. We identified five genomes (NRRL WC-3869, NRRL WC-3927, NRRL WC-3924, NRRL WC-3896, NRRL B-16073) that have received more recombined DNA than others. We observed that although recombination is frequent, it does not impact all 32 genomes similarly. We find that recombination is biased, with some strains receiving more recombined DNA more often (NRRL WC-3896 and B-16073), while others exhibit preferences to specific exchange partners (Figure 3b).

The strength of fastGEAR is its ability to identify both recent (affecting a few strains) and ancestral (affecting entire lineages) recombinations (Mostowy et al. 2017). Of the recent recombination events identified, we observed a total of 91 unique donor-recipient pairs and five of these pairs contributed 49% or more of the total recombination events (Figure 3b). A total of 30 recent recombination events originate from donors outside of the *S. rimosus* dataset. Of these, half came from other *Streptomyces* species and two from other genera in Actinobacteria (*Micromonospora* and *Rhodococcus*). The taxonomic origins of the remaining recombination events could not be precisely determined due to the short length of the recombined sequences. Finally, we found that, of the 22,114 genes that comprise the pan-genome, a total of 2,149 genes have had a history of recombination (Figure 3c). Of these, 1,147 genes were involved in recent recombination and 386 genes in ancestral recombination (Table S2). The most frequently recombined genes include those associated with antibiotic biosynthesis (*lgrB, tycC*), transmembrane transport (*ygbN, efpA*) and transferase (*aftD*) (Figure 3c and Supplementary Table S7).

The lengths of the recombined regions have an approximately exponential distribution, with majority of recombination events being small (<500 bp) and large events occurring relatively infrequently (Figure 3d). The median length of recombined fragments is 230 bp and the largest recombination event is 11,934 bp in strain NRRL B-16073. Our finding of a heterogeneous model of recombination is consistent with those reported in other bacterial species, such as the pathogens *Streptococcus pneumoniae* (Chewapreecha et al. 2014) and *Legionella pneumophila* (David et al. 2017), and our results demonstrate that it also holds true for non-pathogenic species. The observed heterogeneity in recombination sizes has been previously described and classified into micro-recombinations (i.e., short, frequent sequence replacements) and macro-recombinations (i.e., rarer, multi-fragment, saltational sequence replacements) (Mostowy et al. 2014), and our results are consistent with these. Overall, our analysis of recombination in *S. rimosus* reveals inter-strain variation in terms of the frequency of DNA donation or receipt, genes that experience the most frequent recombination and the size of recombination events.

DISCUSSION

The tremendous diversity and ability of *Streptomyces* to inhabit numerous ecological niches and produce diverse clinically useful compounds have been attributed to their large pan-genomes (J. Kim and Yi 2012; Zhan Zhou et al. 2012). In a recent study of 122 *Streptomyces* genomes comprising multiple species, a mere 2.63% (n=1,048 genes present in ≥95% of all genomes) of the 39,893 gene families present constitutes the core genome while the remaining genes are classified as accessory genes (McDonald and Curriea 2017). At the species level, our results on 32 *S. rimosus* genomes reveal similar patterns of having a small fraction of core genes

(n=1,945 genes) which make up 8.8% of a much larger pan-genome (22,114 genes). When we include the soft-core genes (genes present in at least 95% of the strains) numbering 1,874 genes, the core genome still represents only 17% of the pan-genome. While sequencing errors and the draft nature of the genomes used here may partly explain the low number of core genes in *S. rimosus*, the observation of a small core genome in microbial species is not uncommon and has been reported in other species (McInerney, McNally, and O'Connell 2017), including Actinobacteria. For example, the core genome of 28 *Bifidobacterium longum* subsp. *longum* strains consists of 1,160 genes from a pan-genome of 4,169 genes (Chaplin et al. 2015). In 18 strains of *Corynebacterium pseudotuberculosis*, the core genome consists of 1,355 genes and a pan-genome of 3,183 genes (Baraúna et al. 2017). In an analysis of 2,085 *Escherichia coli* genomes, the largest pan-genome analysis to date, a total of 3,188 genes comprises the core genome and is a remarkably small number compared to the stunning 90,000 genes that comprise the *E. coli* pan-genome, with a third of these genes occurring in only one genome (Land et al. 2015). The open pan-genome of *S. rimosus* means that the sequencing of new genomes will possibly add new genes not described in this current pan-genome study. Lastly, while it is difficult to speculate on the causes of why one strain (NRRL WC-3904) has an ANI of 94% compared to the other genomes (slightly below the 95% cutoff for species delineation), previous ANI-based studies have found similar results and may reflect the edge of a genetic discontinuum between species (Caro-Quintero, Rodriguez-Castaño, and Konstantinidis 2009; Jain et al. 2018). However, using the 83% ANI cutoff to delineate different species (Jain et al. 2018), WC-3904 cannot be classified as a separate species.

Compared to other Actinobacteria species and other bacterial phyla, *Streptomyces* also harbors the highest numbers of secondary metabolite BGCs from a large variety of classes and often with little overlap between strains (Doroghazi and Metcalf 2013). Here, each *S. rimosus* genome harbors a unique repertoire of BGCs ranging from 35-71 BGCs per genome, including many NRPS, PKS and hybrid clusters. These results highlight the importance of sampling multiple strains of the same species in improving efforts for natural drug discovery. Antibiotics with new inhibitory mechanisms or cellular targets are urgently needed as resistance to our existing arsenal of drugs is growing and multidrug resistance becomes widespread. While emergence of resistance to and decreased effectiveness of existing tetracyclines as front-line antibiotics have grown over the years (Chopra and Roberts 2001), our genomic analyses suggest that the potential of *S. rimosus* as producers of novel antibiotics has not been fully explored and many natural products are yet to be discovered from this species.

Only recently with whole genome sequencing do we come to recognize the extent in which, within each bacterial species, different strains may vary in the set of genes they encode (Konstantinidis, Ramette, and Tiedje 2006; Leonard et al. 2016; Seipke 2015; Truong et al. 2017). Recently, a polyphasic analyses was conducted on ten strains closely related to *Streptomyces cyaneofuscatus*, with all strains having identical 16S rRNA sequences (Antony-Babu et al. 2017). Authors reported significant differences in morphological, phenotypical and metabolic characteristics, and could in fact be distinguished as five different species (Antony-Babu et al. 2017). Such variation is not uncommon and has been reported to influence functions relevant to the structure and dynamics of the entire microbial community, adaptation to changes in the environment, and interactions with the eukaryotic host (Greenblum, Carr, and

Borenstein 2015). However, the large pan-genome size of a microbial species remains intriguing. Efforts to elucidate the factors that shape and maintain the existence of a multitude of genes in a few strains have recently demonstrated the contributions of selection, drift, recombination, migration and effective population (Andreani, Hesse, and Vos 2017; L.-M. M. Bobay and Ochman 2018; McInerney, McNally, and O'Connell 2017; Vos and Eyre-Walker 2017). While the relative contributions of these processes across multiple microbial species remain unclear, it is likely that one or few of these processes may explain the large pan-genome size of *S. rimosus*.

Equally intriguing is our observation of heterogeneity in the frequency and characteristics of recombination. We observed that some strains donate or receive DNA more often than others, while some strains that tend to frequently recombine with specific partners. Such a pair of strains or lineages exchanging DNA more often between them than with others is said to be linked by a highway of gene sharing (Bansal et al. 2013; Beiko, Harlow, and Ragan 2005). A highway of recombination between a pair of genomes, wherein they exchange DNA more often between them than with others, are likely to represent specific lineages that function as hubs of gene flow, facilitating the rapid spread of genes (for example, those associated with antibiotic resistance, metabolic genes, niche-specific genes) (Chewapreecha et al. 2014). These highways have been previously identified at higher taxonomic groups (domains, phyla, families) (Bansal et al. 2013; Beiko, Harlow, and Ragan 2005; Zhaxybayeva et al. 2009), but have only recently been reported at the sub-species level (Chewapreecha et al. 2014). However, the drivers of heterogeneity in the frequency and characteristics of recombination among members of the same species is poorly understood. Biases in recombination partners and other forms of genetic exchange have been reported to arise from phylogenetic relatedness (including compatible mismatch repair

systems), geographical or physical proximity, shared ecological niches, or common set of mobile elements (Andam and Gogarten 2011; Beiko, Harlow, and Ragan 2005; Skippington and Ragan 2012; Smillie et al. 2011). However, it is unclear whether this variation in recombination is adaptive or not at the population level, to what extent strains that less often recombine benefit from the population, and how the population evolves with a mix of strains that vary in recombination frequencies and partners. In the future, a possible approach to further understand the variation in the recombination process in microbial genomes is to integrate evolutionary game theory with genome sequencing of closely related bacterial strains, composed of recombining ("cooperators") and non-recombining ("cheaters") that can be modeled over hundreds of generations (Rauch, Kondev, and Sanchez 2017; Van Dyken et al. 2013; Zomorrodi and Segrè 2017).

The principal caveat in this analysis is that that the quality of the *S. rimosus* genomes we examined are of varying quality, with some genomes having several hundred contigs. The draft nature of the genomes can have a significant impact on the antiSMASH output, particularly so in the identification of hybrid BGCs. There are two reasons for this. First, antiSMASH is conservative in terms of predicting the borders of BGCs and second, most strains harbor BGC islands on the arms of linear chromosomes (as in *Streptomyces* (Kinashi 2011)), which antiSMASH can misidentify as hybrid BGCs. Another important limitation is that NCBI did not have information about the specific ecological and/or geographical origins of these strains (Supplementary Table S1). Moreover, only 32 genomes were considered. Because the size of the core and accessory genomes is a function of the number and characteristics of the dataset, improved sequencing quality as well as the sequencing of additional genomes is likely to alter

some of our results. In our study, we found evidence of an open *S. rimosus* pan-genome (i.e., the number of new genes discovered increases with the number of additionally sequenced strains) even with the use of draft genomes. Hence, we may expect to find a larger core genome and additional accessory genes if these 32 strains are re-sequenced and complete genomes are generated. We also expect to find additional new and unique *S. rimosus* genes from strains inhabiting diverse environments. While they are most prevalent in soil and decaying vegetation, many *Streptomyces* species have also been identified in extreme environments and the gut of insects (Barka et al. 2016; van der Meij et al. 2017). In these places, we are likely to find niche-specific genes (Croucher et al. 2014; Gupta et al. 2015; B. Zhu et al. 2016), further expanding the size of the accessory genomes of *Streptomyces* species. Much of the work on *Streptomyces* isolation have only concentrated on soil environments, but future work should increase sampling efforts of *S. rimosus* in previously unexplored niches.

CONCLUSION

In this study, we focus on elucidating the pan-genome characteristics and phylogenetic relationships of 32 *S. rimosus* genomes, which is best known as the primary source of the tetracyclines used against many species of pathogens and parasites. There are two major conclusions from this study. First, *S. rimosus* exhibits tremendous inter-strain genomic and biosynthetic variation, which suggests that their potential as an antibiotic producer remains to be fully explored. Second, we observed high levels of recombination between strains; however, recombination is not a homogenous process in this species. Our findings contribute to addressing the puzzle of why microbes have pan-genomes (Andreani, Hesse, and Vos 2017; L.-M. M. Bobay and Ochman 2018; McInerney, McNally, and O'Connell 2017; Vos and Eyre-Walker

2017) and the contributions of biased gene exchange to maintaining gene content variability within a species (Andam and Gogarten 2011; Bansal et al. 2013; Beiko, Harlow, and Ragan 2005; Chewapreecha et al. 2014).

FIGURES

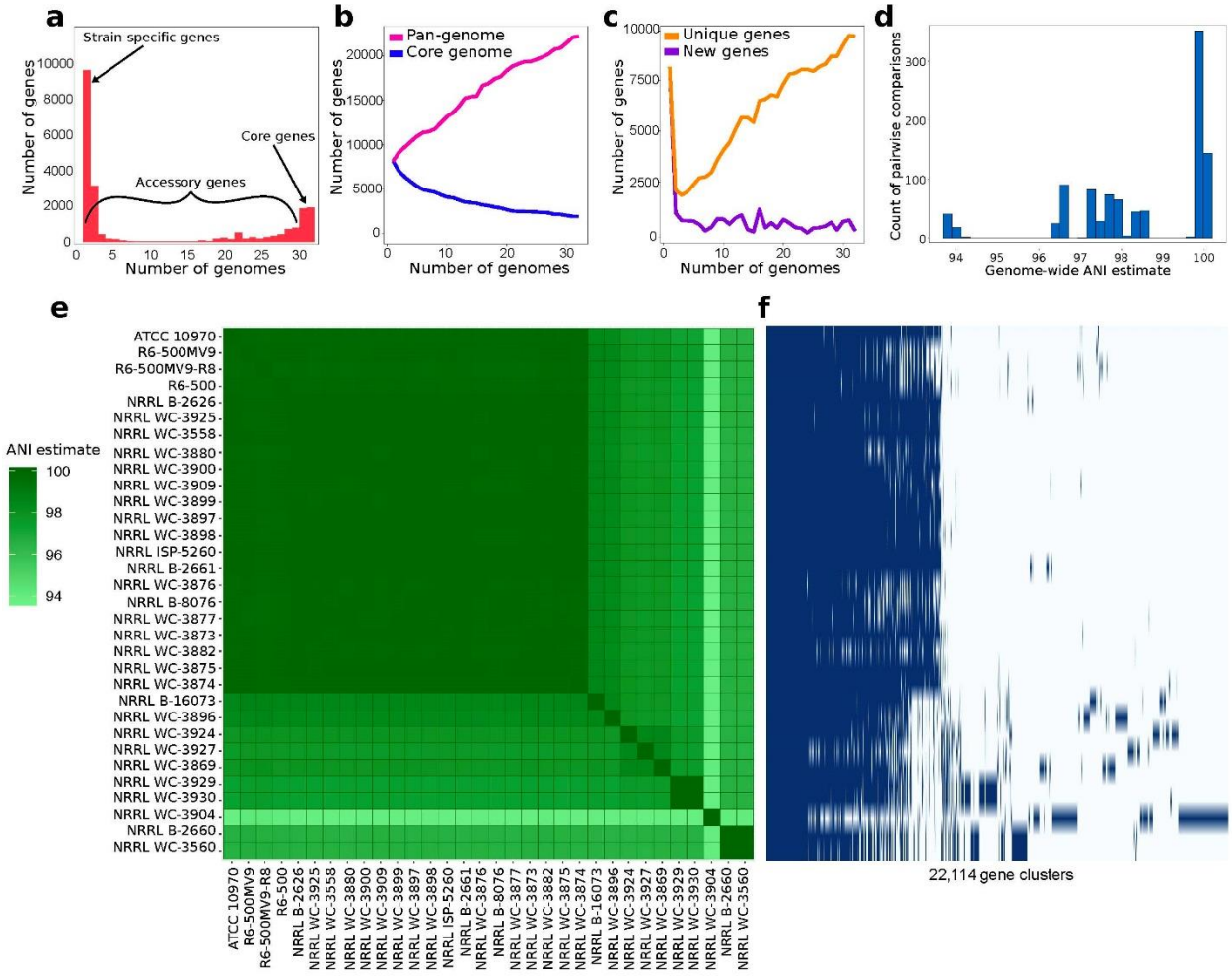

Figure 1. Pan-genome analysis of 32 *S. rimosus* strains. (a) The number of unique genes that are shared by any given number of genomes or unique to a single genome. Numerical values for each gene category are shown in Supplementary Table S2. (b) The size of the core genome, i.e., genes that are present in at least 31 of the 32 strains (blue line) and pan-genome, i.e., the totality of unique genes present in the population (pink line) in relation to numbers of genomes compared. The list of core genes is listed in Supplementary Table S3. (c) The number of unique genes, i.e., genes unique to individual strains (green line) and new genes, i.e., genes not found in the previously compared genomes (purple line) in relation to numbers of genomes compared. (d) Distribution of pairwise average nucleotide identity (ANI) values. ANI calculates the average nucleotide identity of all orthologous genes shared between any two genomes. The 95% ANI cutoff is a frequently used standard for species demarcation. (e) Pairwise whole genome ANI comparison. Percentage values are shown in Supplementary Table S4. (f) Gene presence-absence matrix showing the distribution of genes present in a genome. Each row corresponds to a strain in panel e. Each column represents an orthologous gene family. Dark blue blocks represent the presence of a gene, while light blue blocks represent the absence of a gene

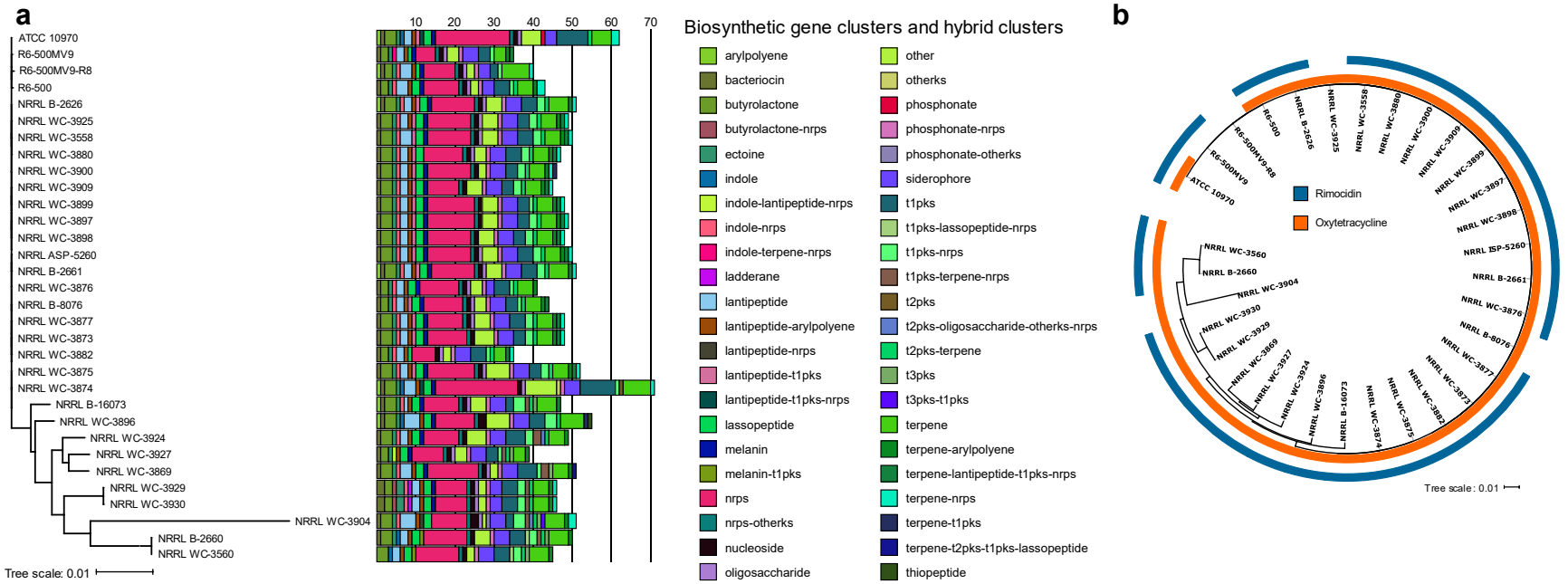

Figure 2. Distribution of BGCs per genome. (a) BGCs and hybrid clusters were identified using antiSMASH. The maximum likelihood phylogenetic tree was reconstructed using concatenated alignments of 1,945 core genes. Scale bar of phylogenetic tree represents nucleotide substitutions per site. Acronyms: nrps – non-ribosomal peptide synthase, t1pks –Type 1 polyketide synthase, t2pks - Type II polyketide synthase, t3pks - Type III polyketide synthase, ks - ketosynthase. (b) Phylogenetic distribution of the oxytetracycline and rimocidin BGCs. Colored rings outside the tree show the presence/absence of BGCs known to encode for oxytetracycline and rimocidin. The two BGCs were identified by searching all the genomes for homologs of each of the genes comprising the BGCs using BLASTP (Altschul et al. 1990) with a minimum e-value of 10^{-10}. Individual genes in a BGC obtained from previous studies (Seco et al. 2004; W. Zhang et al. 2006) were used as query sequences. Presence of the BGC was inferred if there were significant BLASTP hits for at least 90% of the individual genes within the BGC.

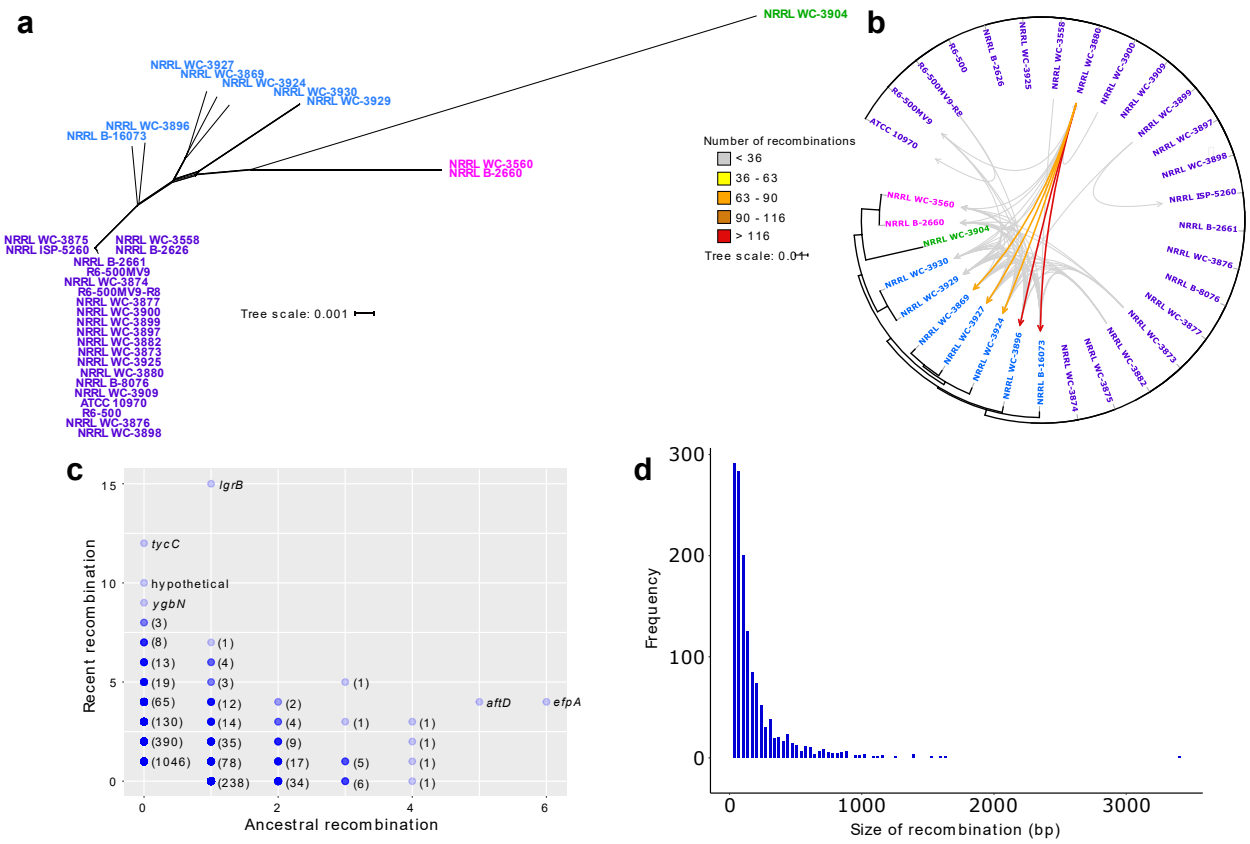

Figure 3. Genetic relationships among *S. rimosus* strains are influenced by homologous recombination. (a) A phylogenetic network of the *S. rimosus* core genome generated using SplitsTree. The strain names were colored according to clustering results using BAPS. (b) Donor-recipient linkages of major recombination events (i.e., highways of recombination) identified using fastGEAR and BLASTN. Scale bar represents nucleotide substitutions per site. Each arrow represents a certain number of recombination events between a pair of genomes, with different colors representing the range of of numbers. (c) Genes that have undergone recent or ancestral recombination. Horizontal axis shows the estimated number of ancestral recombinations and vertical axis shows the estimated number of recent recombinations. Names of some of the genes are shown. Numbers in parenthesis indicate the number of genes represented by overlapping dots found on the same position. (d) Frequency histogram of the size of recombination events of all genes in the pan-genome.

ACKNOWLEDGEMENTS

The authors thank the University of New Hampshire Resource Computing Center and Anthony Westbrook for providing technical and bioinformatics assistance. The study was supported by the New Hampshire Agricultural Experiment Station (NHAES) grant number NH00653 to CA. CP is funded by the New Hampshire NASA Space Grant Graduate Student Fellowship from the New Hampshire Space Grant Consortium.

APPENDIX 1

Supplementary Tables for this chapter can be found at:
https://www.frontiersin.org/articles/10.3389/fmicb.2019.00552/full#h12

CHAPTER 2

Distinct but intertwined evolutionary histories of multiple *Salmonella enterica* subspecies

Cooper J. Park, Cheryl P. Andam

Article published in *mSystems*

Presented here with permission from publisher (see Apendix 2)

ABSTRACT

Salmonella is responsible for many non-typhoidal foodborne infections and enteric (typhoid) fever in humans. Of the two *Salmonella* species, *Salmonella enterica* is highly diverse and includes ten known subspecies and approximately 2,600 serotypes. Understanding the evolutionary processes that generate the tremendous diversity in *Salmonella* is important in reducing and controlling the incidence of disease outbreaks and the emergence of virulent strains. In this study, we aim to elucidate the impact of homologous recombination in the diversification of *S. enterica* subspecies. Using a dataset of previously published 926 *Salmonella* genomes representing the ten *S. enterica* subspecies and *Salmonella bongori*, we calculated a genus-wide pan-genome composed of 84,041 genes and the *S. enterica* pan-genome of 81,371 genes. The size of the accessory genomes varies between 12,429 genes in *S. e. arizonae* (IIIa) to 33,257 genes in *S. e. enterica* (I). A total of 12,136 genes in the *Salmonella* pan-genome have had a history of recombination, representing 14.44% of the pan-genome. We identified genomic hotspots of recombination that include genes associated with flagellin and the synthesis of methionine and thiamine pyrophosphate, which are known to influence host adaptation and virulence. Lastly, we uncovered within-species heterogeneity in rates of recombination and preferential genetic exchange between certain donor and recipient strains. Frequent but biased

recombination within a bacterial species may suggest that lineages vary in their response to environmental selection pressure. Certain lineages, such as the more uncommon non-*enterica* subspecies, may also act as a major reservoir of genetic diversity for the wider population.

INTRODUCTION

Salmonella is widely known for causing non-typhoidal foodborne infections and enteric (typhoid) fever in humans (Gal-Mor, Boyle, and Grassl 2014; Eng et al. 2015; Crump et al. 2015). It is a major public health concern, causing 93.8 million illnesses and 155,000 deaths per year globally (Eng et al. 2015). Salmonellosis in humans manifests itself as diarrhea, fever and abdominal pain within 12-72 hours after infection (Crump et al. 2015). Aside from being able to colonize almost all warm- and cold-blooded animals (Hoelzer, Switt, and Wiedmann 2011; Elmberg et al. 2017; Branchu, Bawn, and Kingsley 2018), *Salmonella* is also prevalent in environmental reservoirs (H. Liu, Whitehouse, and Li 2018; Underthun et al. 2018). In the United States, food products such as vegetables, fruits, and meat have been identified as vehicles of *Salmonella*-associated foodborne outbreaks in the past decade (CDC 2018). The emergence of antimicrobial resistant *Salmonella* lineages further exacerbates the burden caused by this pathogen and compromises our ability to treat clinical infections (Klemm et al. 2018; Hawkey et al. 2019).

Salmonella consists of two species, *Salmonella bongori* and *Salmonella enterica*, with the latter further classified into ten subspecies: *enterica* (I), *salamae* (II), *arizonae* (IIIa), *diarizonae* (IIIb), *houtenae* (IV), *indica* (VI), unnamed subsp. VII, and three novel subspecies A, B, and C (Alikhan et al. 2018). *S. enterica* consists of approximately 2,600 different serotypes (Lan, Reeves, and Octavia 2009; Andino and Hanning 2015), but only a few serotypes cause the majority of gastroenteritis (food poisoning) cases (2). Approximately 99% of salmonellosis is due to *S. e. enterica* (I) serotypes, with 70% caused by only 12 serotypes (Lan, Reeves, and Octavia 2009; Andino and Hanning 2015). In the United States, the two most common serotypes

are *S. e.* ser. Enteritidis and *S. e.* ser. Typhimurium (CDC 2018). *S. e. enterica* (I) represents the vast majority of *Salmonella* strains isolated from humans and warm-blooded animals, while all the other subspecies and *S. bongori* are more typically isolated from cold-blooded animals (Eng et al. 2015; Lamas et al. 2018).

There is a critical need to define the processes that shape how the success of *S. enterica* result from the combination of intrinsic genomic factors, evolutionary processes and the selective environment (ecology), which favors the emergence of new lineages or those with novel characteristics that enhance their resistance, virulence or transmission. One important process that contributes to a pathogen's success is recombination, which can rapidly spread adaptive alleles and novel genes across the population (Didelot and Maiden 2010; William P. Hanage 2016). Hence, recombination can significantly impact the pathogen's response to selective pressures from clinical interventions such as antibiotic use, host immune responses, and extra-host environments (Sela et al. 2018; Brüggemann et al. 2018; Leventhal et al. 2018). Previous studies have shown that frequent recombination and the acquisition of novel genes have contributed to the ecology, evolution and pathogenicity of *S. enterica* (Didelot et al. 2011; Desai et al. 2013), with evidence of recombination affecting the diversity of the lipopolysaccharide antigenic factor (Davies et al. 2013), animal host range (Langridge et al. 2015), and antimicrobial resistance (Klemm et al. 2018; Hawkey et al. 2019). Understanding the role of recombination in *Salmonella* diversity will be particularly crucial in reducing and controlling incidence of disease outbreaks and the emergence of antimicrobial resistance in this pathogen.

In this study, we aim to compare the genomic content and elucidate the impact of homologous recombination on the diversification of the different *S. enterica* subspecies. Using a dataset of 926 previously published *Salmonella* genomes, representing the ten *S. enterica* subspecies and *S. bongori*, we report marked differences in core and accessory genome content between subspecies. We identified genomic hotspots of recombination that include genes associated with flagellin and the synthesis of methionine and thiamine pyrophosphate. Lastly, we uncovered heterogeneity and biases in rates and patterns of recombination. We interpret these findings as indicating the presence of genetic or ecological influences that facilitate the creation of hubs of gene flow between lineages and barriers between other lineages. Our results also highlight the role of the more uncommon non-*enterica* subspecies as a major reservoir of genetic diversity for the wider population. Our study offers important insights into within-species diversification, ecological adaptation and co-circulation of multiple *Salmonella* lineages.

METHODS

Dataset

Our dataset consisted of 926 *Salmonella enterica* genomes downloaded from EnteroBase (Alikhan et al. 2018; Zhemin Zhou et al. 2020). It consists of 297 genomes of *S. enterica* subsp. *enterica* (I), 116 *S. enterica* subsp. *salamae* (II), 116 *S. enterica* subsp. *arizonae* (IIIa), 187 *S. enterica* subsp. *diarizonae* (IIIb), 136 *S. enterica* subsp. *houtenae* (IV), 36 S. *bongori* (V), 16 *S. enterica* subsp. *indica* (VI), six *S. enterica* subsp VII, three *S. enterica* subsp. *A*, six *S. enterica* subsp. *B* and seven *S. enterica* subsp. *C* genomes. Classification of the genomes into subspecies was based on delineation of the core SNPs reported by Alikhan et al. (Alikhan et al. 2018). To

maintain consistency in gene annotations, all genomes were re-annotated using Prokka v1.12 (Torsten Seemann 2014) with default parameters.

Pan-genome analyses

To determine the degree of genomic relatedness and clarify the relationships between the subspecies, we calculated the genome-wide ANI for all possible pairs of genomes using the program FastANI v.1.0 (Jain et al. 2018). ANI is a robust similarity metric that has been widely used to resolve inter- and intra-strain relatedness. The threshold value of 95% has been often used as a cutoff for comparisons belonging to the same or different species (Jain et al. 2018). We used Roary v3.11 with default parameters (95% identity and 99% presence for core genome inclusion) (Page et al. 2015) to characterize the pan-genome at the genus, species and subspecies levels. Roary classifies genes into core, soft core, shell, and cloud genes by iteratively pre-clustering protein sequences using CD-HIT (Fu et al. 2012), all-against-all BLASTP (Altschul et al. 1990) and Markov clustering (Enright, Van Dongen, and Ouzounis 2002). A strength of Roary is that it treats paralogous genes as independent gene families and splits the paralogs into separate clusters by examining the synteny (i.e., the physical co-localization of genes) of flanking genes. We used this clustering output in all downstream analyses, including the pan-genome characterization and recombination detection. Visualization of the pan-genome was done using the post-processing scripts provided by Roary. Gene functions were inferred using the Gene Ontology Consortium's Enrichment Analysis (Ashburner et al. 2000). For the plasmid analysis, we downloaded the *S. e.* subsp. *enterica* serovar Typhimurium st. LT2 genome and its plasmid sequence from the NCBI RefSeq database (Accession ID: GCF_000006945.2) to be used as a reference. Plasmid-associated genes were identified by using BLASTN (Altschul et al.

1990) to compare genes in the reference plasmid against all genes in the *Salmonella* pan-genome with a conservative e-value threshold of 1e-10. Operons were identified by running the *S. enterica* reference genome through the Operon-mapper web-based pipeline (Taboada et al. 2018).

Phylogeny reconstruction

Nucleotide sequences of each single-copy orthologous gene family obtained from Roary was aligned using MAFFT v.7.305b (Kazutaka Katoh et al. 2002). Sequence alignments of core genes were concatenated to give a single core alignment and a maximum-likelihood phylogeny was then generated using the program Randomized Axelerated Maximum Likelihood (RAxML) v.8.2.11 (Stamatakis 2006) with a general time-reversible (GTR) nucleotide substitution model (Tavaré 1986), four gamma categories for rate heterogeneity and 100 bootstrap replicates. All phylogenies were visualized using the Interactive Tree of Life (Letunic and Bork 2016). Pairwise SNP differences in the core genome alignment were identified using the R script available in https://github.com/MDU-PHL/pairwise_snp_differences.

Detection of homologous recombination

Using the core genome alignments, we also calculated the pairwise homoplasy index (PHI) test to determine the statistical likelihood of recombination being present in the entire dataset and within each subspecies (Bruen, Philippe, and Bryant 2006). This statistic measures the genealogical correlation or similarity of adjacent sites. Under the null hypothesis of no recombination, the genealogical correlation of adjacent sites is invariant to permutations of the

sites as all sites have the same history (Bruen, Philippe, and Bryant 2006). Significance of the observed PHI was estimated using a permutation test.

To calculate and compare rates of recombination between subspecies, we ran mcorr, which uses a coalescent-based model of evolution to calculate the probability that a pair of genomes differs at one locus conditional on having differences at another locus (Lin and Kussell 2019). As input to mcorr, we used the core genes identified by Roary (Page et al. 2015) of each subspecies. The recombination parameters estimated by mcorr include: θ - the average number of mutations per locus; ϕ - the average number of recombinations per locus; the ratio of ϕ/θ - the number of recombination events per mutation in a population and is comparable to γ/μ; d - the amount of diversity in a sample brought on by the effects of both recombination and clonal evolution; *c* - the fraction of the sample diversity derived from recombination.

To identify the most frequently recombining genes across the genomes, we used fastGEAR (Mostowy et al. 2017) with default parameters on individual core and shared accessory genes identified by Roary. The program fastGEAR predicts recombination events by first clustering sequences into lineages using a Hidden Markov Model implemented in BAPS (Pritchard, Stephens, and Donnelly 2000). These lineages are defined as groups which are genetically divergent by at least 50% of the sequence alignment. Within each lineage, each genome was then examined using a Hidden Markov Model which iteratively compares polymorphic sites in the strain's sequence (relative to other members of its own lineage) against the same nucleotide site in other lineages. The comparison is made over multiple iterations of the

model, each with updated parameters from the prior run. At the conclusion of the simulation, if a nucleotide site of a strain is found to more similar to the same site in strains of another lineage, it is considered to be a recombination event. To test the significance of these inferred recombinations and identify false-positives, fastGEAR uses a diversity test that compares the diversity of the recombined fragment in question to its background. Recombinations were visualized using R (R Core Team 2019) and the post-processing scripts provided by fastGEAR.

For every recent recombination event identified by fastGEAR, we inferred its donor strain by extracting the nucleotide sequence of the predicted recombined fragment and used it as a query in a BLASTN (Altschul et al. 1990) search against all possible genomes from the identified donor lineage, following the methodology used to identify recombination donors in *S. pneumoniae* (Chewapreecha et al. 2014). The top BLAST hit with the highest bit score was considered the potential donor and given a probability score of 1 for that event, provided that it had an e-value of at least 10^{-10} and at least 95% nucleotide identity. The e-value and nucleotide identity values were chosen to maintain a strict conservative relationship between the donor and recipient. Following a recent recombination event, we expect that the nucleotide similarity between donor and recipient will be remarkably high, and in many cases identical. While our chosen threshold values were arbitrary from a biological perspective, they were chosen to reflect that expectation. In the event of a tie where the e-value and nucleotide identity values were the same across multiple donors, the probability score for that event was divided evenly among each donor (i.e., a probability score of 0.25 was assigned in a four-way tie). This approach involves calculating the sum of a potential donor's probability score across every recombination event in every gene as its likelihood of being a recombination donor. We then assigned the role of most

probable donor in each recombination event to the strain with the highest cumulative donor probability score. Events with potential donors of equal cumulative scores were considered to have originated from the most recent common ancestor of the donors and was discarded from the analysis as an ancestral recombination event.

Data availability

The genomes analyzed in this study were downloaded from and are available in the Enterobase database (https://enterobase.warwick.ac.uk/species/index/senterica) (Zhemin Zhou et al. 2020). Accession numbers are listed in Table S1.

RESULTS

Pan-genome characteristics of Salmonella

To investigate the relative contributions of homologous recombination to the genomic diversity of *S. enterica* subspecies, we compiled a total of 926 representative genomes downloaded from EnteroBase (Table S1) (Zhemin Zhou et al. 2020; Alikhan et al. 2018). We also included *S. bongori* because we hypothesized that recombination also occurs between the two species. Of the ten *S. enterica* subspecies, three were reported to be novel [referred to as subsp. A, B, C (Alikhan et al. 2018)] (Fig. 1a). The core genome-based phylogenetic relationships of these 926 genomes and the discovery of the novel subspecies have been published elsewhere (Alikhan et al. 2018). Subspecies classification in this dataset was based on core single nucleotide polymorphisms (SNPs), which revealed ten distinct *S. enterica* subspecies

(Alikhan et al. 2018). Across the entire dataset, genome size varied between 4.01-5.76 Mb (mean = 4.8 Mb) and the number of predicted genes ranged from 3,745 - 5,593 (mean = 4,564) (Table S1).

We used Roary (Page et al. 2015) to estimate the pan-genome of the entire *Salmonella* dataset and of each subspecies. Roary classifies orthologous gene families into core genes (present in 99% ≤ strains ≤ 100%), soft core genes (present in 95% ≤ strains < 99%), shell genes (present in 15% ≤ strains < 95%) and cloud genes (present in < 15% of strains) (Table S1, Fig. S1). At the genus level, we found a considerably small core genome composed of 1,596 genes, which represents a mere 1.90% of the entire pan-genome (84,041 genes; Table S1). For *S. enterica*, core genes make up 2.28% (1,858 genes) of the species pan-genome (81,371 genes; Table S1). It is also notable that the vast majority of accessory genes of *S. enterica* (75,631 genes, representing 92.95% of the pan-genome) are present in less than 15% of the genomes, with most accessory genes also being unique to a strain (33,474 genes, representing 41.14% of the pan-genome). Comparing the five largest *S. enterica* subspecies (I, II, IIIa, IIIb, IV), we found that the sizes of their core genomes are comparable, ranging from 2,636 genes in *S. e. enterica* (I) to 3,292 genes in *S. e. arizonae* (IIIa). However, we found major differences in the size of their accessory genomes. Combining the shell and cloud genes, the accessory genomes comprise 71.82% [12,429 genes in *S. e. arizonae* (IIIa)] to 90.48% [33,257 genes in *S. e. enterica* (I)] of the pan-genome of each subspecies. (Table S1). A remarkable component of the accessory genome of *S. enterica* [31,809 genes, 40% of the accessory genome] is composed of strain-specific and ORFan genes (i.e., genes with no known homology to genes in other taxonomically or evolutionary lineages (Tautz and Domazet-Lošo 2011)), which have been recently reported to

be significantly associated with pathogenicity in nine bacterial genera (Entwistle, Li, and Yin 2019). Sequencing and annotation errors may also partly explain the large number of accessory genes in *Salmonella*.

To determine the degree of genomic relatedness and hence clarify the distinction among the *S. enterica* subspecies, we calculated the pairwise average nucleotide identity (ANI) for all possible pairs of genomes. ANI estimates the average nucleotide identity of all orthologous genes shared between any two genomes and organisms belonging to the same species typically exhibit ≥ 95% ANI (Jain et al. 2018). The ten *S. enterica* subspecies can be delineated based on their ANI (Fig. 1a) and can be clearly differentiated from *S. bongori* with a mean ANI between the two species of 89.95% (range: 89.20 - 90.53%) (Fig.1b). Mean ANI across all pairs of *S. enterica* genomes is 94.68% (92.62 - 97.26%), while mean ANI within each *S. enterica* subspecies is 98.81% (range: 96.92 - 99.99%).

We also compared the core and accessory genomes within and among *S. enterica* subspecies. We first calculated the number of core SNP differences between any pair of genomes. Within *S. e. salamae* (II), we found the greatest range of pairwise SNPs (between 3 and 15,846), while *S. e. diarizonae* (IIIb) showed significantly less variation (between 1 and 4,386) despite it being one of the largest clusters in the study. As expected, we found considerably fewer SNPs within subspecies than between subspecies, with a maximum pairwise SNP count of 16,624 among genomes in subsp. A (Fig. 1c). Comparing the two *Salmonella* species, we obtained a mean of 66,486 core SNPs that differentiate the them (range: 64,131 -

69,571 SNPs) (Fig. S2). We also compared the number of accessory genes per genome among the different subspecies. *S. e. diarizonae* (IIIb) exhibited the highest mean as well as the greatest variability in the accessory gene content, ranging from 2,509 and 3,678 accessory genes per genome (Fig. 1d). However, pan-genome estimates are greatly influenced by the size of the dataset being examined (Lapierre and Gogarten 2009) and it is thus challenging to compare subspecies of different sizes.

Lineage-specific rates of homologous recombination

Within-species variation in rates of recombination has been previously reported in other bacterial pathogens, such as *Streptococcus pneumoniae* (Chewapreecha et al. 2014; Andam et al. 2017) and *Staphylococcus aureus* (Castillo-Ramírez et al. 2012). We therefore sought to determine whether this is also true for *Salmonella*. We compared rates of recombination among the different *Salmonella* subspecies because variable recombination rates between subspecies may reflect a differential response to environmental selection pressure and different capacities for adaptation (Chewapreecha et al. 2014). Because the number of genomes in each subspecies are greatly dissimilar, ranging from 3 genomes in novel subsp. A to 297 in *S. e. enterica* (I), we restricted our recombination analyses to the five largest subspecies. Under the null hypothesis of no recombination, we calculated the pairwise homoplasy index (PHI) statistic. We found significant evidence for the presence of recombination in *S. e. enterica* (I), *S. e. arizonae* (IIIa), *S. e. diarizonae* (IIIb) and *S. e. houtenae* (IV) (p-value < 0.01 for each subspecies).

Next, using the program mcorr, we calculated the probability that a pair of genomes differs at one locus conditional on having differences in another locus, which defines the correlation profile (Lin and Kussell 2019). In the absence of recombination, the correlation profile will be constant (flat), while recombination will generate monotonically decaying correlations as a function of the distance between loci (Lin and Kussell 2019). This decay is due to each recombination event creating a sequentially identical fragment between the genomes of the donor and recipient; hence, a higher recombination rate results in a faster decay rate (Lin and Kussell 2019). The correlation profiles for each of the five subspecies exhibit a monotonic decay, with recombination rates decreasing as a function of the size of the homologous fragment (Fig. S3). Similar decaying correlation profiles have been calculated in other recombining pathogenic bacteria, such as *Helicobacter pylori* and *Pseudomonas aeruginosa* (Lin and Kussell 2019).

We also used mcorr (Lin and Kussell 2019) to calculate five recombination parameters based on the correlation profiles of synonymous substitutions for pairs of homologous sequences (Fig. 2 and Table S2). As input, we used the core genes of each *S. enterica* subspecies and 100 bootstrap replicates. Sample diversity (d), which is generated from both recombination and accumulation of mutations of the clonal lineage, ranged from 4.3×10^{-3} in *S. e. diarizonae* (IIIb) to 0.016 in *S. e. enterica* (I). For comparison, other pathogenic species of *Gammaproteobacteria* exhibit a sample diversity of 3.3×10^{-4} (*Yersinia pestis*), 0.014 (*P. aeruginosa*) and 0.031 (*Acinetobacter baumanii* and *Klebsiella pneumoniae*) (Lin and Kussell 2019). The mutational divergence (θ), which refers to the mean number of mutations per locus since the divergence of a pair of homologous sites, ranged from 0.012 in *S. e. houtenae* (IV) to 0.023 in *S. e. enterica* (I). For comparison, mutational divergence in global collections of *Y. pestis, P. aeruginosa, A.*

baumanii and *K. pneumoniae* are 0.0091, 0.027, 0.087, and 0.13, respectively (Lin and Kussell 2019). Recombinational divergence (ϕ) ranged from 0.066 in *S. e. diarizonae* (IIIb) to 0.225 in *S. e. enterica* (I). The same parameter was reported to be 0.027, 0.29, 0.11, and 0.56 in *Y. pestis, P. aeruginosa, A. baumanii* and *K. pneumoniae*, respectively (Lin and Kussell 2019). The ratio ϕ/θ (or γ/μ), which gives the relative rate of recombination to mutation, ranged from 3.38 in *S. e. arizonae* (IIIa) to 9.75 in *S. e. enterica* (I). For comparison, γ/μ is estimated to be 3.0, 11, 4.2, and 1.3 in *Y. pestis, P. aeruginosa, A. baumanii* and *K. pneumoniae*, respectively (Lin and Kussell 2019). Lastly, the recombination coverage (c), which indicates the fraction of the genome whose diversity was derived from recombination events since its last common ancestor and ranges from 0 (clonal evolution) to 1 (complete recombination) (Lin and Kussell 2019), ranged from 0.248 in *S. e. arizonae* (IIIa) to 0.714 in *S. e. enterica* (I). This parameter is reported to be 0.033 in *Y. pestis*, 0.52 in *P. aeruginosa*, 0.40 in *A. baumanii* and 0.27 in *K. pneumoniae* (Lin and Kussell 2019). Comparing the five subspecies across each parameter, we found significant differences (p-value < 0.01 for each parameter; Kruskall-Wallis test). Overall, we found that the degree in which the *S. enterica* subspecies differ from each other in terms of the five recombination parameters is comparable to those found when comparing different bacterial species.

Heterogeneity and biases in patterns of homologous recombination

Recent population genomic studies have reported variation not only in rates of recombination among members of a single bacterial species but also in other characteristics of recombination (Chewapreecha et al. 2014; Lin and Kussell 2019; Park and Andam 2019). One

such variation can be found in the length of recombined DNA sequences. In bacterial genomes, two distinct modes of recombination have been proposed to occur: micro-recombination (frequent exchange of short DNA fragments) and macro-recombination (occasional larger replacements, usually associated with major phenotypic changes) (Mostowy et al. 2014). To determine the size distribution of recombined DNA segments, we ran fastGEAR (Mostowy et al. 2017) on individual sequence alignments of core and shared accessory genes. In the entire *Salmonella* dataset, the lengths of the recombination fragments greatly varied, ranging in size from 101 bp to 2,712 bp in the core genome and from 101 bp to 7,606 bp in the accessory genome (Fig. 3a). Among the five largest subspecies, the number of recombination events range from 1,604 in *S. e. houtenae* (IV) to 5,260 in *S. e. enterica* (I). Overall, the sizes of recombination events follow a geometric distribution, with majority of recombination events encompassing short DNA segments of <1000 bp. Large recombination events (>1,000 bp) occurred less frequently, with the longest recombination block detected in a genome from novel subsp. A (7,606 bp). For comparison, macro-recombination in other bacterial species such as the highly recombining *S. pneumoniae* has been reported to reach up to 100,000 bp (Andam et al. 2017).

The strength of fastGEAR is its ability to identify both recent (affecting a few strains) and ancestral (affecting entire lineages) recombinations (Mostowy et al. 2017). We found that, of the 84,041 genes that comprise the *Salmonella* pan-genome, a total of 12,136 genes have had a history of recombination, representing 14.44% of the pan-genome (Fig. 3b and Table S2). Of these, 6,722 genes were involved only in recent recombination, 1,071 genes only in ancestral recombination and 4,343 genes in both recent and ancestral recombination. Of the 12,136

recombining genes, 1,475 are core genes and the remaining 10,661 are accessory genes. Some of the most frequently recombining genes have unknown or hypothetical functions, while those genes with the highest frequencies of recombination and which also have known functions include *fliC*, *thiH*, *metE*, and *metH* and will be highlighted here (Fig. 3b). The flagellin gene *fliC* encodes the *Salmonella* phase 1 antigen and, along with *fliB* (which encodes the phase 2 antigen), is considered as a *Salmonella* serotype determinant gene (Y. Liu et al. 2017). Flagellin genes contribute to ecological adaptation of *Salmonella* by allowing the cell to adjust their expression through phase variation when it encounters a new niche (De Maayer and Cowan 2016) and in the generation of new serotypes (Smith, Beltran, and Selander 1990). Flagellar motility plays a role in host colonization, surface adhesion and biofilm formation; hence they are also important virulence factors in *Salmonella* (Horstmann et al. 2017). The *thiH* gene is involved in the biosynthesis of thiamine pyrophosphate, an essential cofactor for several enzymes in central metabolism and amino acid biosynthesis (Martinez-Gomez, Robers, and Downs 2004). The specific contribution of thiamine pyrophosphate in *Salmonella* pathogenicity is unclear; however, it has been reported that thiamine acquisition is a critical step in the replication and proliferation of *Listeria monocytogenes* within host cells during the infection process (Schauer et al. 2009). The products of *metE* and *metH* are transmethylases that function in cobalamin-independent and cobalamin-dependent reactions, respectively, during the last step of methionine biosynthesis (Weissbach and Brot 1991). While the specific role of MetE and MetH in *S. enterica* infection remains unclear, these genes have been reported to contribute to metabolic adaptation to physiological host conditions and pathogenicity in *Ralstonia solanacearum* during plant infection (Plener et al. 2012). Other recombining genes detected by fastGEAR are listed in Table S2. The phylogenies of genes *metE, metH* and *thiH* show that

strains of the same subspecies often cluster together and rarely do we find strains from one subspecies grouping within another subspecies (Fig. S4). In contrast, the *fliC* gene tree reveals numerous instances of phylogenetic incongruence, with multiple strains from one subspecies grouping with members of other subspecies. We also observed that paralogous gene families exhibit different number of recombination events. For example, fastGEAR identified 173, 3, 53 and 1 recent recombination events in the flagellin genes *fliC*, *fliC*_1, *fliC*_2 and *fliC*_5, respectively and 7, 2, 67, 0 and 2 recent recombination events in the aldehyde-alcohol dehydrogenase genes *adhE*, *adhE*_1, *adhE*_2, *adhE*_3 and *adhE*_5, respectively (Table S2). We also explored evidence for recombination in the 115 plasmid-associated genes in the plasmid sequence of *S. e.* subsp. *enterica* serovar Typhimurium st. LT2 genome that we used as a reference. A total of 112/753 plasmid-associated genes (i.e., 753 genes from the *Salmonella* pan-genome with an e-value of 1e-10 or lower when compared to any of the 115 reference plasmid genes using BLASTN) have experienced recombination (Fig. S5, Fig. S6, Table S3). We also observed that the genes that comprise an operon do not show similar frequencies of recombination (Fig. S7, Table S3).

Highways of recombination, whereby a pair of strains or lineages frequently recombine with each other more often than they do with others, have been previously reported in the Gram-positive *S. pneumoniae* (Chewapreecha et al. 2014). Here, we aim to determine whether such highways of recombination also exist in *Salmonella*. To achieve this, we first identified the recombining pairs of donor and recipient genomes. Using the method developed in the *S. pneumoniae* study (Chewapreecha et al. 2014), we first calculated the sum of a potential donor's probability score across every recombination event in every gene as its probability of being a

recombination donor. We then assigned the role of the most probable donor in each recombination event to the genome with the highest cumulative donor probability score. For each pair, we characterized it as one linked by a highway of recombination when the number of recombination events from donor to recipient was at least one standard deviation above the average number of recombination events per recombining pair across the entire dataset. We also considered the direction of recombination events, which means that any pair of recombining genomes can be linked by a highway in either direction. We identified a total of 38,105 unique recombining pairs of genomes in the entire *Salmonella* dataset, of which 2,190 fit our definition of a highway. Of these, a total of 1,784 are highways that linked genomes from different subspecies (Fig. 3c). Lastly, we also found that 86% of strains in the dataset acted as a DNA donor, while every genome has received recombined DNA at least once.

DISCUSSION

S. enterica continues to threaten animal and human health worldwide. While *S. e. enterica* (I) accounts for majority of clinical infections, little is known of how other subspecies contribute to the entire species' virulence and adaptive potential. To elucidate its success as a pathogen, analyses of the genomic structure and phylogenetic relationships among the different *S. enterica* subspecies is critical. Here, we show that recombination within and between subspecies has played a major role in shaping the evolution and genome structure of *S. enterica*. Widespread recombination within the species means that new adaptations arising in one lineage can be rapidly transferred to another distantly related lineage (Didelot and Maiden 2010; William P. Hanage 2016).

The major finding in this study is that while the different *S. enterica* subspecies can be distinguished from each other based on their core and accessory genomes, variation in recombination frequencies occurs between the different subspecies. Our findings greatly expand on the results of a previous study that reported an uneven role of recombination among *S. e. enterica* (I) lineages based on sequencing approximately 10% of their core genome (Didelot et al. 2011). In that study, the authors report that some lineages displayed evidence of more frequent recombination than others, and that recombination has occurred predominantly between members of the same lineage, thus suggesting barriers to recombination (Didelot et al. 2011). More recently, a recombination analysis of 73 *S. enterica* genomes using co-ancestry and hybridization methods also show variation in recombination across the species, resulting in the formation of hybrid groups within the genus (Criscuolo et al. 2019). Variability in gene content and in patterns of recombination may be considered effective strategies for a species to maintain potentially useful adaptive alleles and novel genes that can rapidly be shared among specific members of the species. This variation also means that a species can prevent the likelihood that a gene is lost from the population by ensuring that some strains, even rare ones, carry them. Within-species differences in recombination also suggests that lineages within a species respond to selective pressures and environmental changes in different ways (Chewapreecha et al. 2014). Our results also imply that recombinations are not random events that impact all members of a species in a uniform manner. Genetic or ecological influences likely exist that facilitate the creation of hubs of gene flow between certain lineages as well as barriers between other lineages. We interpret these findings as indicating the existence of both biases and barriers of recombination between multiple lineages, which can shape the phylogenetic distribution of different genetic elements independent of the organisms that harbor them (Fondi et al. 2016).

Several factors can potentially explain within-species variation in rates of recombination and biases in donor-recipient linkages. First, minimal niche overlap can impact opportunities for recombination between strains and subspecies. Non-*enterica* subspecies are often sampled from cold-blooded animals (e.g., turtles, snakes, lizards, crocodiles), while *S. e. enterica* (I) is frequently found in humans and warm-blooded animals consumed by humans (i.e., poultry, cattle and pigs) (Lamas et al. 2018). Such ecological barriers may explain the fewer highways of recombination observed between *S. e. enterica* (I) and the non-*enterica* subspecies compared to recombination between the different non-*enterica* subspecies. However, *S. e. enterica* (I) and the non-*enterica* subspecies are not exclusively isolated from each other, and both can sometimes be found together in cold- and warm-blooded animals. Hence, another possible explanation for the variation in recombination is that different *Salmonella* subspecies occupy distinct micro-ecological niches (Fung et al. 2019), which may even be separated by a few millimeters, within a human or animal host and therefore reduce the opportunity for genetic exchange. The existence of cryptic niches and their role in structuring bacterial populations has been previously reported. Two generalist *Campylobacter jejuni* lineages inhabiting the same animal host show no evidence of recombination between them even though they freely recombine with other lineages and with each other in laboratory setting (Sheppard et al. 2014).

Certain genomic elements can also influence the success of a recombination event, thus contributing to the biases and barriers to recombination. One example is the functional linkage of multiple genes in operons. Functional similarity, and in some cases dependency, of operon-linked genes may likely limit the potential for recombination to impact individual genes in a region under positive selection and hence promote the horizontal gene transfer (HGT) of entire

operons (Lawrence and Roth 1996; Omelchenko et al. 2003; Kominek et al. 2019). However, it has been reported that a remarkable 35% of operons that show evidence of HGT is made up of genes with different phylogenetic affinities, occurring through *in situ* xenologous displacement through recombination (Omelchenko et al. 2003), and thus may partly explain our result of differential recombination within an operon. Frequent homologous replacement of genes within an operon allows the bacterium to maintain operon integrity (i.e., without causing disruption of operon organization and function) in the face of strong positive selection (Omelchenko et al. 2003). Plasmids and other mobile elements can also facilitate and influence patterns of recombination and virulence in enteric pathogens (Pilla and Tang 2018). In *Salmonella*, only a small number of recombining genes are associated with plasmids; hence other mechanisms of recombination likely play a more substantial role. Future work should therefore explore the contributions of a variety of mechanisms (transduction, transformation, conjugation, other types of mobile genetic elements) in mobilizing different components of the *Salmonella* pan-genome. Additionally, incompatible restriction-modification (R-M) systems act as genetic barriers that can limit extensive recombination and incorporation of longer DNA segments (Brown et al. 2003). A previous study of *S. e. enterica* (I) showed mosaicism in the *mutS* gene, which encodes a key component of the methyl-directed mismatch repair (MMR) system, with mutant alleles in *mutS* able to enhance the recombination between lineages (LeClerc et al. 1996; Zahrt and Maloy 1997). It is possible that minute R-M differences and MMR defects can facilitate frequent recombination between certain subspecies but not with others. Future work focusing on in vitro recombination assays of strains from different *S. enterica* subspecies may provide important insights into whether genetic, mechanistic or ecological barriers can explain biases in recombination partners.

The major limitation in this study is the high variability in the number of genomes in each of the ten subspecies, making it difficult to elucidate and compare the novel but less well-known subspecies with the more prevalent *S. e. enterica* (I). The non-*enterica* subspecies have been less studied, mainly because they are often associated with cold-blooded animals (Guyomard-Rabenirina et al. 2019; Pulford et al. 2019) and cases of human salmonellosis are almost entirely limited to serotypes of *S. e. enterica* (I) (Desai et al. 2013; Eng et al. 2015). There is therefore a stark gap in sampling and genome sequencing work that has been done to date on non-*enterica* subspecies. Previous reports indicate that non-*enterica* subspecies have lower invasive capacity, virulence, and levels of resistance to common antibiotics, and human infections have been mostly those involving weakened immune systems (Lamas et al. 2018; Giner-Lamia et al. 2019). However, as we have shown in this study, there is frequent recombination between subspecies, hence these less well-known subspecies likely act as reservoirs of novel allelic variants or genes that human-associated lineages can sample from when needed (e.g., as a response to environmental change or host immune system). Future genome sequencing endeavors may shed important insights on the genomic diversity on many non-*enterica* subspecies from various hosts and habitats. Lastly, the draft nature of these genomes, potential sequencing errors and mis-annotation may also have influenced our analysis of genome structure, including the characterization of core and accessory genes, detection of recombination events, and identification of donors and recipients.

Recombination, either through homologous or illegitimate means, plays a fundamental role in the evolution and species diversification of bacterial genomes (Didelot and Maiden 2010; Dixit, Pang, and Maslov 2017; Marttinen and Hanage 2017). For many bacterial pathogens,

including *Salmonella*, recombination has been implicated in the emergence of highly virulent lineages (Klemm et al. 2018; William Paul Hanage et al. 2009; Paul et al. 2013). Our results provide crucial insights into the contributions of recombination into the diversification and adaptive capabilities of *S. enterica* as a species. Understanding the extent of genomic variation within a species, and the ecological and evolutionary underpinnings of this variation, will enable successful surveillance of emerging infectious agents. It will also facilitate the development of effective clinical interventions to limit the emergence of new pathogenic clones and of accurate predictions of how specific lineages will respond to environmental changes.

FIGURES

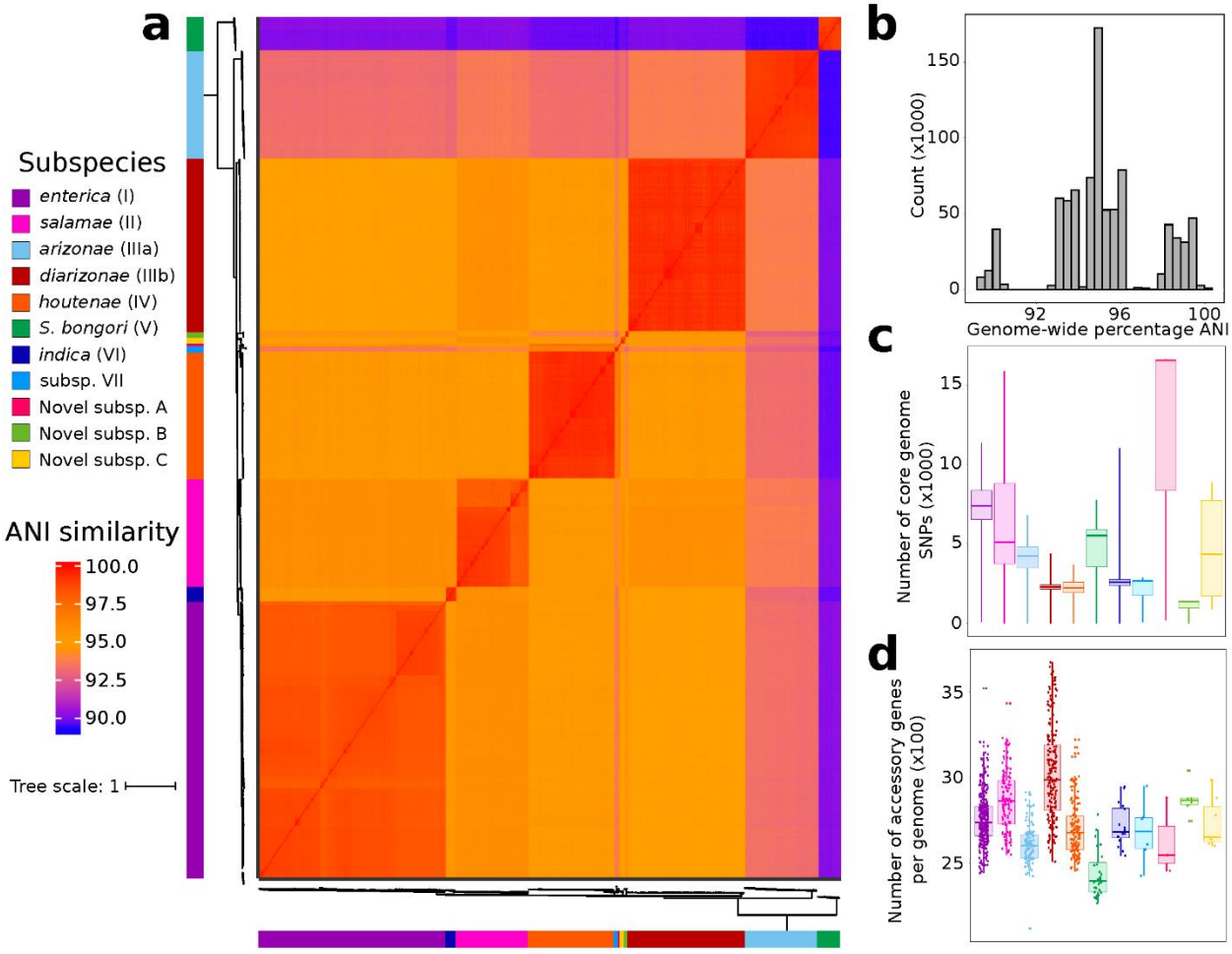

Figure 1. Genomic differences among *Salmonella* genomes. (a) Pairwise genome-wide ANI values. ANI calculates the average nucleotide identity of all orthologous genes shared between any two genomes. The phylogeny was reconstructed using the concatenated alignment of 1,596 genus-wide core genes. Scale bar represents nucleotide substitutions per site. (b) Frequency distribution of all pairwise ANI values. The 95% ANI cutoff is a frequently used standard for species demarcation. (c) Number of SNPs in the core genome alignment per subspecies. The box shows the median SNP count, and lower and upper quartiles. The whiskers represent the minimum and maximum SNP counts. (d) Number of accessory genes per genome for each subspecies. Subspecies classification is based on core genome variation calculated by Alikhan et al. (Alikhan et al. 2018).

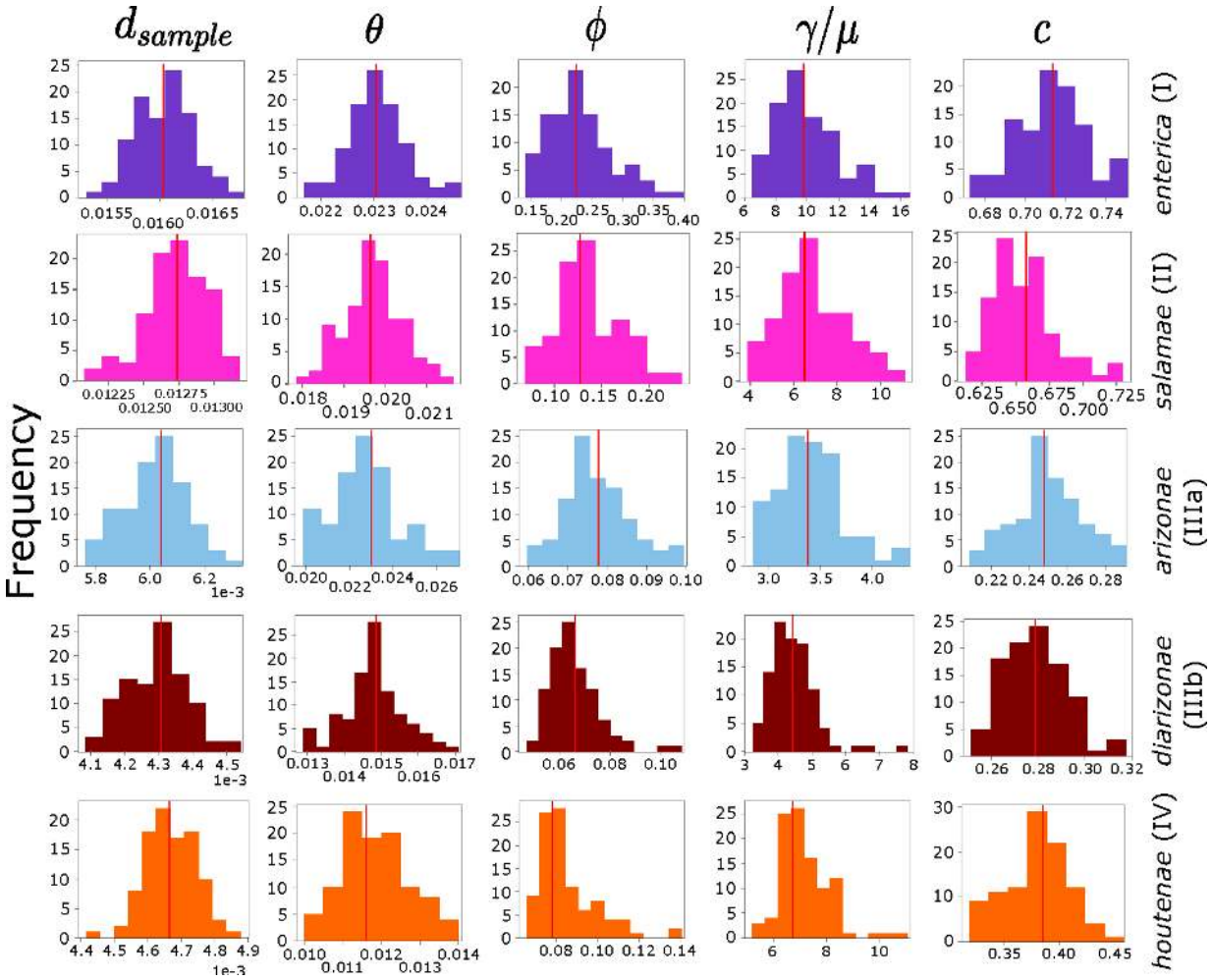

Figure 2. Recombination parameters of the five largest *S. enterica* subspecies calculated using mcorr (Lin and Kussell 2019). Histograms show the frequency distribution of each recombination parameter for all pairs of genomes.

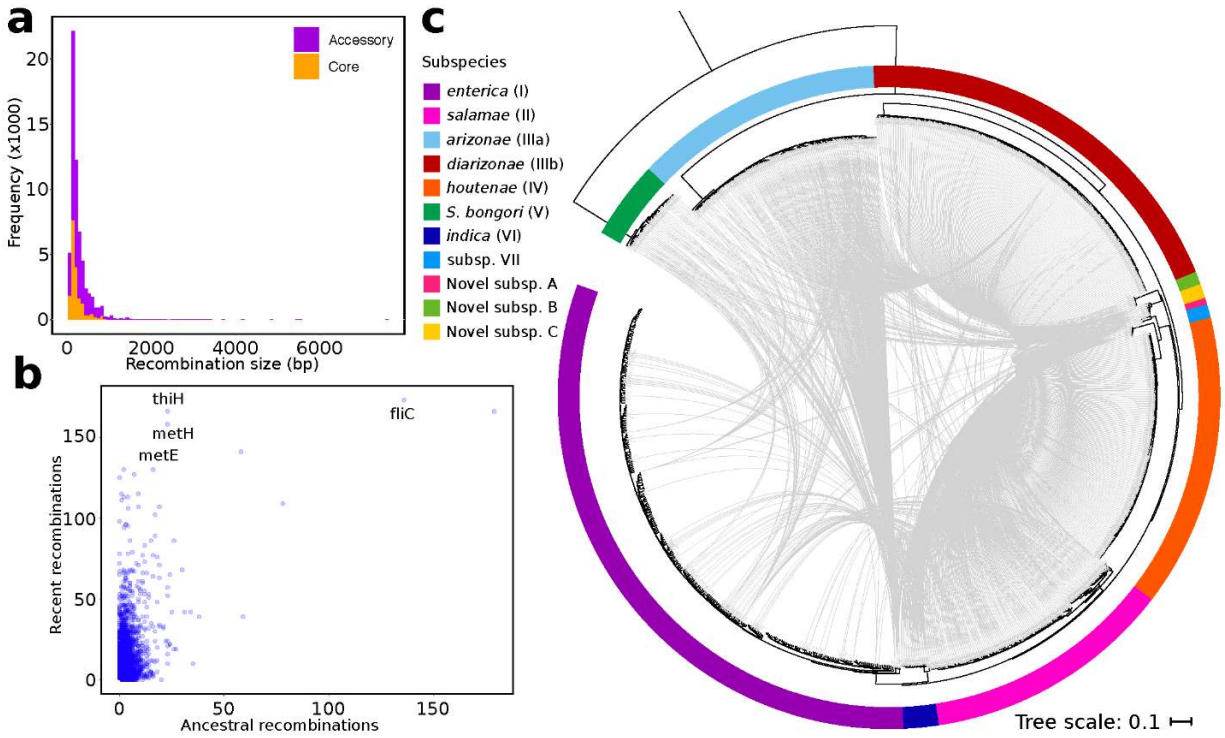

Figure 3. Variable patterns of recombination. (a) Size distribution of lengths of recombined core and accessory DNA fragments. (b) Genes that have undergone recent or ancestral recombination. Horizontal axis shows the estimated number of ancestral recombinations, and vertical axis shows the estimated number of recent recombinations. For clarity, names of some of the most frequently recombined genes with known functions are shown. (c) The maximum likelihood phylogenetic tree was calculated using the concatenation of 1,596 core genes present in all 926 genomes and rooted using *S. bongori*. Scale bar represents nucleotide substitutions per site. The outer ring shows the different subspecies identified in Alikhan et al. (Alikhan et al. 2018). For visual clarity, only inter-subspecies highways of recombination events identified by fastGEAR are shown (as gray arrowlines) and non-highway recombination pairs are not shown for visual clarity. Inferred recipient genomes are indicated by the arrowheads.

ACKNOWLEDGEMENTS

We thank the UNH Resource Computing Center where all bioinformatics analyses were performed. We thank Anthony Westbrook for providing technical and bioinformatics assistance. This research was funded by the USDA/NHAES Hatch Multi-State Award No. NH00687.

APPENDIX 2

Supplementary Tables for this publication can be found at:

https://msystems.asm.org/content/5/1/e00515-19/figures-only#fig-data-supplementary-materials

CHAPTER 3

Genomic epidemiology and evolution of diverse lineages of clinical *Campylobacter jejuni* cocirculating in New Hampshire, USA, 2017

Cooper J. Park[a], Jinfeng Li[b], Xinglue Zhang[b], Fengxiang Gao[b], Christopher S. Benton[b], Cheryl P. Andam[a]

Article is published in *Journal of Clinical Microbiology*

Presented here with permission from publisher (see Appendix 3)

[a] University of New Hampshire, Department of Molecular, Cellular and Biomedical Sciences, Durham, *New Hampshire,* USA

[b] *New Hampshire* Department of Health and Human Services, 29 Hazen Drive, Concord, *New Hampshire,* USA

ABSTRACT

Campylobacter jejuni is one of the leading causes of bacterial gastroenteritis worldwide. In the United States, New Hampshire was one of the 18 states that reported cases in the 2016-2018 multistate outbreak of multidrug resistant *C. jejuni*. Here, we aimed to elucidate the baseline diversity of the wider New Hampshire *C. jejuni* population during the outbreak. We used genome sequences of 52 clinical isolates sampled in New Hampshire in 2017, including one of the two isolates from the outbreak. Results revealed a remarkably diverse population composed

of at least 28 sequence types, which are mostly represented by one or few strains. Comparison with 249 clinical *C. jejuni* from other states showed frequent phylogenetic intermingling, suggesting lack of geographical structure and minimal local diversification within the state. Multiple independent acquisitions of resistance genes from five classes of antibiotics characterize the population, with 47/52 (90.4%) of the genomes carrying at least one horizontally acquired resistance gene. Frequently recombining genes include those associated with heptose biosynthesis, colonization and stress resistance. We conclude that the diversity of clinical *C. jejuni* in New Hampshire in 2017 was driven mainly by the co-existence of phylogenetically diverse antibiotic resistant lineages, widespread geographical mixing, and frequent recombination. This study provides an important baseline census of the standing pan-genomic variation and drug resistance to aid the development of a statewide database for epidemiological studies and clinical decision making. Continued genomic surveillance will be necessary to accurately assess how the population of *C. jejuni* changes over the long term.

INTRODUCTION

Campylobacter jejuni is a major foodborne pathogen and the most commonly reported bacterial cause of gastroenteritis (campylobacteriosis) in the United States and worldwide (Kirk et al. 2015; Tack et al. 2019). Severe cases of *C. jejuni* infections can also lead to invasive infections such as bacteremia (Hussein et al. 2016). Infection with *C. jejuni* is also considered one of the main precedents for the development of the autoimmune condition Guillain-Barré Syndrome (GBS), a serious demyelinating neuropathy (Yu, Usuki, and Ariga 2006). The World Health Organization estimates that *Campylobacter* spp. has resulted to 166 million illnesses and 37,604 deaths in 2010 worldwide (Kirk et al. 2015). In the United States, the Centers for Disease Control and Prevention (CDC) estimates a total of 1.5 million infections and $270 million in direct medical costs every year caused by *Campylobacter* infections, mostly involving *C. jejuni* (CDC 2019). *Because Campylobacter* naturally colonizes the gastrointestinal tract of food-producing, companion and wild animals, disease outbreaks have often been linked to consumption of raw, undercooked, or contaminated water, food and food products as well as through direct contact with animals (Kaakoush et al. 2015).

Due to the self-limiting characteristic of campylobacteriosis, antimicrobial therapy is not routinely recommended; however, in acute or persistent infections, immunocompromised cases or those patients with comorbidities, antibiotics are commonly prescribed (Kaakoush et al. 2015). The emergence and spread of *Campylobacter* isolates exhibiting resistance to antibiotics commonly used to treat severe infections have been alarmingly increasing in the past two decades (CDC 2019; Whitehouse, Zhao, and Tate 2018; Yang et al. 2019). In CDC's 2019 report on Antibiotic Resistance Threats in the United States, antibiotic resistant *Campylobacter* is

listed as one of the 11 serious threats to public health that require prompt and sustained action (CDC 2019). The CDC estimates that 28% of *Campylobacter* isolates from 2015-2017 have decreased susceptibility to ciprofloxacin (fluoroquinolone), 4% with decreased susceptibility to azithromycin (macrolide), and 2% with decreased susceptibility to both ciprofloxacin and azithromycin (CDC 2019). The public health threat of antibiotic resistance (ABR) in this pathogen was recently brought to light when a multistate outbreak of multidrug resistant *C. jejuni* infections occurred in the United States from January 2016 to February 2018 (Montgomery et al. 2018). The source of the outbreak were puppies from breeders, distributors and pet stores (Montgomery et al. 2018). Antibiotic susceptibility testing showed that the outbreak isolates were resistant to all antibiotics commonly used to treat *Campylobacter* infections (Montgomery et al. 2018). The state of New Hampshire was one of the 18 states that reported cases in the 2016-2018 *C. jejuni* outbreak, with two of the 118 cases reported (Montgomery et al. 2018). In our study, we aimed to elucidate the genetic diversity of the wider New Hampshire *C. jejuni* population during the period of the outbreak, how resistance and virulence determinants are distributed among strains, and the evolutionary processes that have shaped the local population. This study provides an important baseline census of the standing *C. jejuni* pan-genomic diversity and drug resistance characteristics in New Hampshire to aid in the development of a statewide database for epidemiological studies and clinical decision making. Continued genomic surveillance of the background diversity will be necessary to accurately assess how the population of *C. jejuni* changes over the long term, in response to changes in the selective landscape, and during disease outbreaks.

METHODS

Bacterial Isolates

Isolates were submitted to the Public Health Laboratories, *New Hampshire* Department of Health and Human Services (NH DHHS) in Concord, New Hampshire, USA in 2017. These isolates were received from New Hampshire health care providers and were recovered primarily from stool specimens collected from individuals with *Campylobacter* infection. The state of New Hampshire considers *Campylobacter* infections as a reportable disease and the NH DHHS strongly encourages isolate submission to the Public Health Laboratories. However, submission of isolates is not mandatory. No identifiable information is associated with the isolates submitted by the health care providers. In total, our dataset comprised 52 isolates.

DNA extraction and genome sequencing

Sequencing of *Campylobacter* isolates is part of the PulseNet surveillance program, a United States national laboratory network that connects foodborne illness cases to detect outbreaks (Tolar et al. 2019). DNA extraction, library preparation and whole genome sequencing were done following the PulseNet USA standard operating procedures (https://www.cdc.gov/pulsenet/pathogens/wgs.html). Briefly, DNA extraction procedures were conducted using the Qiagen DNeasy Blood & Tissue Kit (Qiagen, Valencia CA). DNA quality and concentration were measured using Qubit fluorometer and NanoDrop spectrometer. A total of 1 ng of genomic DNA from each isolate was used to construct sequencing libraries using the Illumina Nextera XT DNA Library Preparation Kit (Illumina, Inc. San Diego, CA) per the manufacturer's instructions. Samples were sequenced as multiplexed libraries on the Illumina

MiSeq platform operated per the manufacturer's instructions for 500 cycles to produce paired-end reads of 250 bp in length. The MiSeq sequencer is housed at the NH DHHS Public Health Laboratories.

De novo genome assembly, annotation, pan-genome and phylogenetic analyses

We used the Nullarbor pipeline v2.0 (https://github.com/tseemann/nullarbor) to perform read trimming, quality assessment, contig assembly, gene annotation, pan-genome, sequence type (ST) identification, sequence alignment and phylogenetic analysis of the entire dataset. The Nullarbor pipeline can be described as follows: Adapters were trimmed using Trimmomatic v0.38 (Tolar et al. 2019). Trimmed reads were assembled into contigs using SKESA v2.3.0 (Souvorov, Agarwala, and Lipman 2018) using a *C. jejuni* subsp. *jejuni* reference genome obtained from the NCBI's RefSeq database (Accession ID: GCF_000009085.1). Quality of genome assemblies was assessed using Quast (Gurevich et al. 2013). Assembled genomes were annotated using Prokka v1.13.3 (Torsten Seemann 2014) with default parameters. Roary v3.12.0 (Page et al. 2015) was used to characterize the pan-genome of the New Hampshire *C. jejuni* dataset and to classify genes into core, soft core, shell, and cloud genes. Each orthologous gene family was aligned using MAFFT v.7.407 (K. Katoh and Standley 2013). The ST of each isolate was determined using the program multilocus sequence typing (MLST) (https://github.com/tseemann/mlst), which extracts the sequences of seven housekeeping genes (*aspA, glnA, gltA, glyA, pgm, tkt, uncA*) from the Illumina raw data and compares them to the *C. jejuni* MLST database (www.mlst.net) (Jolley, Chan, and Maiden 2004). Single nucleotide polymorphisms (SNPs) from the core genes were identified and aligned using Snippy v4.3.6 (https://github.com/tseemann/snippy) and were used to generate a maximum likelihood

phylogeny using the program IQ-TREE v1.6.9 (Nguyen, Lam-Tung, Schmidt, Heiko A., Haeseler, Arndt von, Minh 2015).

To determine the degree of overall genomic relatedness between genomes, we calculated the genome-wide average nucleotide identity (ANI) for all possible pairs of genomes using the program FastANI v.1.0 (Jain et al. 2018). ANI estimates the average nucleotide identity of all orthologous genes shared between any two genomes (Jain et al. 2018). Organisms belonging to the same species typically exhibit ≥95% ANI (Jain et al. 2018). Pairwise ANI values were visualized using an heatmap generated in R (R Core Team 2019) and the ggplot2 package (Wickham 2016).

In order to place the New Hampshire isolates within a country-wide context, we queried the genome sequences of 48,987 clinical *C. jejuni* isolates that were included in the 100K Pathogen Project as of March 2020 (Weimer 2017). Of these, we selected 367 isolates that were derived only from the United States, from human samples, from clinical specimens, as well as those that have information on their state of origin. These were filtered further to only include those genomes that are within the 95% ANI threshold that defines a bacterial species (Jain et al. 2018). A total of 249 genomes representing 13 other states were used for comparison with the New Hampshire genomes (Table S1). After annotating with Prokka (Torsten Seemann 2014) and identifying the pan-genome using Roary (Page et al. 2015), we generated a core genome phylogeny using RAxML v8.2.11 (Stamatakis 2006) with a general time reversible nucleotide substitution model, four gamma categories for rate heterogeneity and 100 bootstrap replicates.

In silico identification of ABR genes, virulence genes and plasmids

We screened all genomes for known resistance and virulence genes using a direct read mapping method called ABRicate v.0.8.10 (https://github.com/tseemann/abricate) implemented in Nullarbor. ABRicate identifies ABR genes using BLASTN comparison search (Altschul et al. 1990) against the Resfinder database (Zankari, Ea, Hasman, Henrik, Cosentino, Salvatore, Vestergaard, Martin, Rasmussen, Simon, Lund, Ole, Aarestrup, Frank M., Larsen 2012). ABRicate only identifies horizontally acquired resistance genes and not resistance due to chromosomal mutations. Virulence genes were identified using BLASTN against the Virulence Factor Database (VFDB) (Liu, Bo, Zheng, Dandan, Jin, Qi, Chen, Lihong, Yang 2019). Some of these predicted genes may be complete, exact matches or incomplete; hence ABRicate classifies the predicted genes based on the proportion of the gene that is covered. These categories are present (≥95% sequence coverage), questionable (<95% sequence coverage) and absent, which provide a level of confidence on ABRicate's predictions. We also used PlasmidFinder with default parameters to perform an *in silico* detection and characterization of plasmid sequences (Carattoli et al. 2014).

Recombination detection

Using the core genome alignment, we calculated the pairwise homoplasy index test implemented in SplitsTree v.4.14.8 (Huson 1998) to determine the statistical likelihood of recombination being present in the entire dataset (Bruen, Philippe, and Bryant 2006). This statistic measures the genealogical correlation or similarity of adjacent nucleotide sites. Under the null hypothesis of no recombination, the genealogical correlation of adjacent sites is invariant

to permutations of the sites because all sites should have the same evolutionary history (Bruen, Philippe, and Bryant 2006). Significance of the observed index was estimated using a permutation test. We then visualized potential recombination events using SplitsTree, which integrates reticulations due to recombination in a phylogeny (Huson 1998). To identify the most frequently recombining genes across the genomes, we used fastGEAR (Mostowy et al. 2017) with default parameters on individual core and shared accessory genes identified by Roary. To test the significance of the inferred recombination events and identify false positives, we used the diversity test implemented in fastGEAR, which compares the diversity of the recombined fragment in question to its sequence background. Recombinations were visualized using R (R Core Team 2019) and the post-processing scripts provided by fastGEAR. We used EggNOG-mapper v2 to perform orthology assignment for functional annotation of the recombined genes (Huerta-Cepas et al. 2017). The reference sequences of recombined genes were used as input to obtain the gene ontology IDs. We restricted our search only within the subphylum Epsilon-proteobacteria to which *Campylobacter* belongs. These IDs were then used as input in the webtool PANTHER (Mi et al. 2019) to perform a statistical overrepresentation test to determine if the recombined genes were biased towards a specific ontological process. PANTHER classifies the ontological function of each recombined gene using different categories: Molecular Function, Cellular Component, Biological Process, Protein Class.

Parameters used for all programs are listed in Table S1.

Data Availability

All *Campylobacter* genomic sequences generated under PulseNet USA surveillance (Tolar et al. 2019) are uploaded in real-time to the sequence read archive (SRA) hosted by the National Center for Biotechnology Information (NCBI). The genomes analyzed in this study are available in BioProject PRJNA239251. The genomes obtained from the 100K Pathogen Project were obtained from BioProject PRJNA186441. Accession numbers and Biosample IDs are listed in Table S1.

RESULTS

Genomic characteristics of C. jejuni in New Hampshire

We sequenced the genomes of 52 clinical *C. jejuni* isolates collected in New Hampshire, USA in 2017 (Table S2). The genome sequences contain between 21-78 contigs and N50 values range between 34,459 - 197,591. *De novo* genome assemblies generated sequences of sizes ranging from 1.57-1.81 Mb (mean = 1.70 Mb). We used PlasmidFinder to determine if the variation in genome size could be attributed to the presence or absence of plasmids. No plasmids were detected in any of the New Hampshire genomes. We next used Roary to estimate the pan-genome of the entire *C. jejuni* dataset (Figure S1 and Table S3). Of the 4,335 gene families identified in the pan-genome, a total of 1,176 genes comprised the core genome (genes present in 99% ≤ strains ≤ 100%), which represents approximately 27% of the pan-genome. The maximum likelihood phylogenetic tree based on the alignment of 83,210 core SNPs revealed lineages that have relatively little structure relative to the location of the healthcare provider (county) or date of collection (Fig. 1A). Genome-wide ANI values for every possible pair of *C. jejuni* genomes

ranged from 96.7 - 99.99% (mean = 98.26%) (Fig. 1B,C and Table S4). Together, the core genes (n = 1,176 genes) and the soft-core genes (n = 111 genes; genes present in 95% ≤ strains < 99%) constituted only 29.69% of the entire population's pan-genome. Accessory genes can be categorized into shell (n = 881; genes present in 15% ≤ strains < 95%) and cloud genes (n = 2,167; genes present in < 15% of strains). Together, both categories of accessory genes constituted 70.31% of the population's pan-genome. There was substantial strain-level variation in the New Hampshire population in terms of gene content. The number of protein-coding genes per genome ranged from 1,575 – 1,918 (mean = 1,743) (Fig. 1D). The number of accessory genes per genome ranged from 385-724 (mean = 539.8) (Fig. 1E). Many accessory genes were unique to individual strains (1,059 genes, representing 24.42% of the pan-genome), with 1-166 singleton genes present per genome (Fig. 1F).

Our results from *in silico* MLST showed that the *C. jejuni* isolates belonged to 28 unique known STs (Fig. 1A, Table S5). Four novel STs found in five strains have MLST profiles that did not match known STs in the MLST database (Jolley, Chan, and Maiden 2004). We did not identify any one genome that dominated the entire population; instead, the population was composed of multiple STs represented only by one or few strains. The most common were STs 48 and 50 which were represented by six and five strains, respectively. In this dataset, we also included a genome (SRR6152533) from one of the two isolates from New Hampshire that was part of the 2016-2018 multistate puppy-associated outbreak of multidrug resistant *C. jejuni* (Montgomery et al. 2018). This isolate has been identified as ST 2109.

Relationship of the New Hampshire C. jejuni isolates to the wider United States population

To place the genetic diversity and population structure of the New Hampshire *C. jejuni* isolates within the broader United States *C. jejuni* population, we used a genome dataset consisting of 249 clinical *C. jejuni* isolates primarily from stool specimens from the 100K Pathogen Project (Table S2) (Weimer 2017). These genomes represented 13 other states in the country. Pairwise genomic comparison in this merged dataset (i.e., 52 from New Hampshire and 249 from the 100K Pathogen Project) revealed ANI values that ranged between 96.06 - 100% (mean = 98.23%) (Fig. 2, Table S4). Pan-genome analysis using Roary showed a total of 10,763 genes in the pan-genome in the merged dataset, which was 2.48x more than the New Hampshire pan-genome alone. We identified only 937 core and 203 soft-core genes, which were 0.2x fewer and 1.8x more than the New Hampshire pan-genome, respectively. We also identified a total of 423 genes (representing 3.93% of the pan-genome of the merged dataset) that were found exclusively in the New Hampshire population compared to the 6,150 (representing 57.1% of the pan-genome) found exclusively outside the state. A maximum likelihood tree generated using the alignment of the core genes showed that the phylogenetic clustering of isolates was independent of the state of origin and that the New Hampshire genomes were intermingled with those from other states (Fig. 2).

Distribution of horizontally acquired ABR genes

Frequent horizontal gene transfer (HGT) characterize the evolutionary history of numerous bacterial species (Soucy, Huang, and Gogarten 2015), including *Campylobacter* (Sheppard and Maiden 2015). In many bacterial pathogens, HGT has greatly contributed to the

emergence and spread of many “superbugs” that have acquired resistance to a broad spectrum of antibiotics (Juhas 2015). We used the program ABRicate to determine the presence of horizontally acquired genes known to encode resistance to a range of different classes of antibiotics. We identified a total of 14 unique genes associated with ABR and which represent five different major classes of antibiotics (aminoglycosides, β-lactams, chloramphenicol-florfenicol, streptothricin and tetracycline) (Fig. 3 and Table S6). Multiple independent acquisitions of resistance genes from the five major classes of antibiotics characterized the New Hampshire *C. jejuni* population, with 47/52 (90.4%) of the genomes carrying at least one horizontally acquired resistance gene. Five genomes (representing 9.6% of the population) carried at least one of the six genes that encode resistance against aminoglycosides. Of the five genes that encode for β-lactam resistance, one gene ($bla_{OXA\text{-}605}$) was found in 38 genomes, representing 73% of the population. Four other genomes harbored three other unique genes that encode β-lactam resistance. Overall, we found that resistance to β-lactams is most common in the population, with a remarkable 80.77% of the population carrying at least one of the five β-lactam resistance genes detected. Two genomes carried the *sat4* gene, which confers streptothricin resistance, while 17 genomes harbor the *tetO* gene which confers tetracycline resistance. One genome (SRR5859317) contained at least one resistance gene for each of the four classes (aminoglycosides, β-lactams, streptothricin and tetracycline), while three genomes carried genes that encode resistance against three major classes of antibiotics. Notably, the isolate from the puppy outbreak shared at least three distinct ABR genes with the rest of the local population (*aph(3')-IIIa,* $bla_{OXA\text{-}605}$*, sat4*) in addition to three other ABR genes that were unique to it (*aad9, aadE, aph(2'')-Ih*). It has been postulated that antibiotic use in puppies may have led to the emergence and transmission of multi-drug resistant *C. jejuni* isolates during the 2016-2018

outbreak (Montgomery et al. 2018). We also identified the likely presence of the multidrug resistance phenotype mediated by the plasmid-borne gene that encodes for Cfr rRNA methyltransferase, which confers resistance to phenicols, lincosamides, oxazolidinones, pleuromutilins, and streptogramin A antibiotics (Long et al. 2006; Tang et al. 2017), in five genomes. Lastly, we did not detect the presence of any one acquired resistance gene in five genomes (9.6% of the population). Overall, we found that many of the clinical *C. jejuni* isolates in the local population were carriers of a diverse suite of resistance genes that can be horizontally exchanged between strains. The outbreak isolate was not the only one that was multidrug resistant; at least six other isolates carry transferrable genes that encode resistance against multiple classes of antibiotics.

Distribution of virulence determinants

We also used ABRicate to determine the presence of virulence genes in *C. jejuni* (Fig. 3 and Table S6). In all, we detected a total of 126 virulence-related genes. A total of 78 virulence genes were most common in the population and were found in at least 50 out of 52 genomes. The most common virulence genes in the New Hampshire *C. jejuni* population were those that encode for traits related to capsule, lipooligosaccharide, flagella-mediated motility, bacterial adherence to intestinal mucosa, invasive capability, toxin production and type four secretion system. Genes associated with adherence included those that function in capsule variation, binding to fibronectin, lipooligosaccharide and major outer membrane protein (porin) (Liu, Bo, Zheng, Dandan, Jin, Qi, Chen, Lihong, Yang 2019). Some virulence genes were particularly noteworthy and will be discussed here.

The cytolethal distending toxin (cdt) is one of the well-characterized virulence factors of *C. jejuni* and is reported to be associated with local acute inflammation in enterocolitis (Hickey et al. 2000), hyper-invasion (Baig et al. 2015) and colorectal tumorigenesis (He et al. 2019). The *C. jejuni cdt* operon, consisting of *cdtA*, *cdtB*, and *cdtC*, encodes a multi-subunit holotoxin that has DNAse activity and induces DNA double-strand breaks (Lara-Tejero and Galán 2001; Bezine, Vignard, and Mirey 2014). While the presence of a single *cdt* gene does not have any effect on the virulence of *C. jejuni*, it has been reported that the presence of all three *cdt* genes results in the release of a functional cytotoxin (Lara-Tejero and Galán 2001). It is therefore not surprising that all three genes were found in at least 90% of the New Hampshire population, which consists solely of human clinical isolates. The *cdt* genes were present at high frequencies: *cdtA* in 51/52 genomes, *cdtB* in 47/52 genomes, and *cdtC* in 52/52 genomes. However, 1/52 and 5/52 genomes also possess *ctdA* and *cdtB*, respectively, but have <95% sequence coverage that may be due to sequencing errors.

C. jejuni is the most frequent pathogen associated with acute immune-mediated neuropathies GBS and Miller-Fisher Syndrome, which can cause acute flaccid paralysis in humans (Taboada et al. 2018; Yu, Usuki, and Ariga 2006). It has been previously reported that ganglioside mimicry by the *C. jejuni* lipooligosaccharide is a critical factor in eliciting the two neuropathies (Yu, Usuki, and Ariga 2006). The gene *wlaN* encodes β-1,3 galactosyltransferase, which is involved in the biosynthesis of ganglioside-mimicking lipooligosaccharide in *C. jejuni* (Linton et al. 2000). We detected *wlaN* in two genomes in the New Hampshire population. Previous studies on the prevalence of *wlaN* in *C. jejuni* from other geographical regions report similar low frequencies (e.g., 13-17% in 624 *C. jejuni* isolates from humans and poultry in

Poland (45); 7.5% in 58 stool isolates in Bangladesh (Talukder et al. 2008); 10% in 111 human, animal and environmental isolates in Brazil (Frazão et al. 2017)). In contrast, another study reports that, of the 40 isolates of *C. jejuni* from human, bovine and turkey sources, *wlaN* was more prevalent and was detected in 46.7% of strains that exhibit no or weak colonization and invasion capacity and in 60% of strains with strong colonization and invasion capacity (Müller et al. 2007). Sialylated lipooligosaccharide has been reported to have the potential to also produce ganglioside mimics and induce GBS (Neal-McKinney et al. 2018). The gene *cstIII*, which encodes a lipooligosaccharide sialyltransferase, is reported to be also associated with neuropathy (Neal-McKinney et al. 2018). In the New Hampshire *C. jejuni* population, a total of nine strains carried the *cstIII* gene. For comparison, previous studies report the presence of *cstIII* in 30.8% of 266 isolates of human, chicken, bovine and turkey origin in Germany (Zautner et al. 2012) and in 18.9% of 827 genomes analyzed by the Food and Drug Administration Pacific Northwest Laboratory (Neal-McKinney et al. 2018).

Glycosylation of *Campylobacter* flagellins with pseudaminic acid and its derivative has been previously shown to be essential for flagellar assembly and motility, which are required for colonization of the mucus lining of the gastrointestinal tract (Guerry 2007; Chidwick and Fascione 2020). The genes *pseA-I* are required for the biosynthesis and/or transfer of pseudaminic acid to the flagellin (Guerry 2007; Chidwick and Fascione 2020). In the New Hampshire population, we found that these genes were differentially distributed among genomes: 52/52 genomes have *pseB, pseC, pseF, pseG* and *pseI*; 51/52 genomes have *pseA*; 7/52 genomes have *pseD*; 42/52 genomes have *pseE*; and 48/52 genomes have *pseH*. Such variation in the distribution of individual genes of an operon among closely related strains is not uncommon and

may be indicative of frequent *in situ* gene displacement through gene gain and loss, which does not often result to losing the integrity and function of the operon (Omelchenko et al. 2003). The differential distribution of these genes may also contribute to the generation of variation in flagellin glycosylation among strains that can influence antigenic diversity in *C. jejuni* (Guerry 2007).

Reticulated evolution due to frequent recombination in New Hampshire genomes

Recombination plays an important role in the evolutionary history of *C. jejuni* (Wilson et al. 2009; Woodcock et al. 2017). Here, we aimed to elucidate to what extent recombination contributes to the genomic structure of *C. jejuni* at the local scale. Using the pairwise homoplasy index statistic, we detected evidence for significant recombination in the core genome (p-value << 0.01). Recombination in *C. jejuni* core genome can be visualized using NeighborNet implemented in SplitsTree4 (Huson 1998), which showed the phylogenetic reticulations due to recombination (Fig. 4A). We then used fastGEAR to estimate recombination in core genes and shared accessory genes (Mostowy et al. 2017) (Table S7). In the New Hampshire *C. jejuni* population, the lengths of the recombination fragments greatly varied. Overall, the sizes of recombination events followed a geometric distribution, with majority of the recombination encompassing short DNA segments and a median size of 116 bp (Fig. 4B). Large recombination events (>2,000 bp) occurred less frequently, with the longest recombination blocks detected in three genomes (SRR5278283 [ST 475], SRR6014507 [ST 48], SRR6014981 [ST 475]). Similar patterns of frequent micro-recombinations and rare macro-recombinations (Mostowy et al. 2014) have been reported in other bacterial pathogens, such as *Streptococcus pneumoniae* and *Salmonella enterica* (Mostowy et al. 2014; Park and Andam 2020). Such patterns have been

reported to greatly contribute to shaping the genomic and phenotypic heterogeneity, including resistance and pathogenicity characteristics, of a pathogen species (Mostowy et al. 2014; Park and Andam 2020; David et al. 2017).

We also used fastGEAR to identify the genes that were frequently recombined. A total of 1,071 genes representing 24.7% of the pan-genome have experienced recombination (Fig. 4C and Table S7). Of these genes, 1,020 were involved in recent recombination (i.e., recombination affecting a few strains) and 224 in ancestral recombination (i.e., recombination affecting entire lineages) (Fig. 4C). Some of the most frequently recombining genes with known function that fastGEAR detected included those that may contribute to virulence and adaptation. The gene product MutS2 has been reported to be associated with the overall function of preserving genomic integrity by inhibiting homologous recombination (Pinto et al. 2005). The gene products of *hddA* (D-glycero-D-*manno*-heptose 7-phosphate kinase) and *gmhA* (phosphoheptose isomerase) are involved in heptose biosynthesis (Liang et al. 2016). Modifications in capsular heptose have been shown to contribute to *C. jejuni* colonization and persistence in the gastrointestinal tract (A. Wong et al. 2015). The gene product of *nspC* (carboxynorspermidine decarboxylase) is involved in the biosynthesis of the polyamine norspermidine, which functions in biofilm formation (Wotanis et al. 2017). The carbamoyltransferase encoded by *hypF* aids in the maturation of [NiFe] hydrogenases in *Escherichia coli* (Paschos et al. 2002). *hypF* mutants have been shown to exhibit loss of resistance against extreme acidic conditions (Hayes et al. 2006) as in the case during passage through the stomach (Reid et al. 2008). Lastly, it is curious that *dltA* was identified as frequently recombining in the gram-negative *C. jejuni*. The *dlt* operon functions in the D-alanylation of teichoic acids in gram-positive bacteria and has been shown to

confer resistance to antimicrobial peptides (Kovács et al. 2006). A previous study reported the presence of the *dlt* operon in three gram-negative genera (*Erwinia, Bordetella* and *Photorhabdus*) and was thought to have been acquired by HGT (Abi Khattar et al. 2009).

To further elucidate the general functions of the recombined genes, we used EggNOG-mapper v2 and PANTHER to perform orthology prediction and functional annotation. Of the 1,071 genes inferred by Roary to have had experienced recombination, EggNOG-mapper v2 did not retrieve gene ontology results for 795 genes. Using PANTHER, we classified the remaining 276 genes based on different functional categories: molecular function, biological process, cellular component and protein class (Table S8 and Figure S2). A total of 149 genes can be classified as having catalytic activity. A total of 131 genes were associated with metabolic processes. A total of 69 genes were associated with a variety of cellular components or the cytoplasm and 19 genes associated with the cell membrane. Lastly, 137 genes were associated with metabolic interconversion enzymes. Overall, our recombination analysis shows that even within a single year of sampling, the standing pan-genomic variation in a local population is amplified through frequent but variable recombination of genes associated with a variety of functions, which can greatly contribute to *C. jejuni*'s potential to evolve rapidly (Sheppard and Maiden 2015).

DISCUSSION

Rapid advances and declining costs in whole genome sequencing are transforming the public health system. Pathogen genomics is expected to become an integral part of a systematic

surveillance required to monitor emerging trends in disease epidemiology, including campylobacteriosis, which will allow for earlier detection and more precise investigations of outbreaks, transmission, virulence and drug resistance (Grad and Lipsitch 2014; Gaiarsa et al. 2015). Pathogen genomic surveillance should include long-term monitoring of the standing pathogen diversity in any local population at a fine-scale resolution to provide a baseline census of antibiotic resistant and other high-risk clones circulating within a region, from local to global scales. Such information is integral in epidemiological studies and clinical decision making in managing *Campylobacter* infections. In this study, we analyzed the genomic diversity of 52 clinical isolates of *C. jejuni* in the state of New Hampshire in 2017. This dataset was selected in order to assess the background genomic variation in *C. jejuni* during the 2016-2018 puppy-associated outbreak of multidrug resistant *C. jejuni* in the United States. Our analysis included one of the two outbreak isolates that were reported in the state. Results revealed a remarkably high phylogenetic and genomic diversity of strains co-circulating in the wider New Hampshire *C. jejuni* population. Our results showed lack of geographical structure and minimal local diversification within the state. We did not detect evidence for clonal expansion shaping the local population structure; the co-circulation of multiple STs suggest multiple introduction and widespread dissemination of divergent *C. jejuni* lineages between multiple counties in New Hampshire as well as between states, which may be facilitated by the constant movement of agricultural products, animals and people.

The rapid evolution and diversification of *C. jejuni* within only a single year has also been facilitated by frequent recombination and HGT, which has been often observed in previous studies of *C. jejuni* (Vegge et al. 2012; Sheppard et al. 2014; Mourkas et al. 2019). We present

different lines of evidence to demonstrate the contribution of these processes in shaping the genomic structure of the New Hampshire population. First, we found that accessory genes are differentially distributed among strains, likely due to rapid gene gain and loss, which contributes to the overall genomic diversity of the local population. The variable distribution of accessory genes between strains is often attributed to adaptation to specific ecological niches (McInerney, McNally, and O'Connell 2017; Chaudhry and Patil 2020), even within the same host (Stoesser et al. 2015; Chung et al. 2017). For example, mobile integrated elements and plasmids were reported to be more common in fecal than blood *C. jejuni* isolates, while a hybrid capsule locus was more common in blood than fecal isolates (Skarp et al. 2017). Here, we show that even among fecal isolates, there is substantial heterogeneity in accessory gene content, which may indicate either neutral evolution due to random processes (Haegeman and Weitz 2012) or the existence of cryptic ecological niches (Sheppard et al. 2014) in the gastrointestinal tract that selects for certain adaptive genes. Second, the population harbors numerous horizontally acquired resistance determinants from five major classes of antibiotics. The origins and direction of transfer of these genes remain uncertain, but it is safe to assume that their acquisition and mobility may have greatly contributed to the overall distribution of ABR genes in the local population. The outbreak isolate has been previously characterized as multidrug resistant (Montgomery et al. 2018). Our analysis shows that it harbors six horizontally acquired resistance genes, three of which were unique to it and another three that were shared with other New Hampshire genomes. Yet it is remarkable that the genome sequences of the rest of the population revealed that many of the isolates were also drug resistant, with resistance to beta-lactams the most common. A few multidrug resistant genomes were also detected. Hence, while the outbreak isolate did not spread through clonal expansion within the state, the risk of widespread

dissemination of resistance genes through HGT among *C. jejuni* lineages is a serious public health threat and must be considered in the implementation of control measures and antibiotic stewardship practices. Lastly, frequently recombining genes include those associated with heptose biosynthesis, colonization and stress resistance, all of which can have a substantial impact on the pathogen's adaptive potential. This includes the rapid emergence of novel phenotypes (Sheppard and Maiden 2015; Golz et al. 2020), such as multidrug resistance (Lopes et al. 2019) and the ability to colonize a specific host (i.e., specialists) or multiple hosts (i.e., generalists) (Woodcock et al. 2017). Because increased genetic variation leads to more rapid adaptation (Arber 2000), populations have a broader reservoir of mobile accessory genomic variants that can be mixed and matched in individual genomes through frequent recombination, which would suggest that individual strains each has a unique suite of capabilities to adapt to their environment.

Defining the baseline genomic diversity of a pathogen in a local population is integral to elucidating the ecological factors that sustain the co-circulation of diverse and drug resistant lineages. It will aid in the development of a statewide database for epidemiological studies and clinical decision in response to changing selective pressures and during disease outbreaks (Grad and Lipsitch 2014; Gaiarsa et al. 2015). This is particularly important to precisely identify and trace high-risk clones in the local population that can disseminate easily or accumulate additional resistance mechanisms. While the 2017 genomes were phylogenetically diverse, represented by 28 unique known STs, it remains unclear whether there are certain lineages that will become more successful over the long term, e.g., hyper-virulent, hyper-recombinant, highly transmissible or multidrug resistant. Only continuous genomic surveillance of the local population over many years will allow us to determine the bacterial population dynamics within the state. Nevertheless,

our study provides the initial genomic surveillance of *C. jejuni* for New Hampshire, which can be built on in future years to track the evolutionary changes that underlie phenotypic and population shifts of high-risk or super-fit clones over time.

A few limitations need to be acknowledged. First, bacterial samples were based on what were received by the NH DHHS from local health providers and may not fully reflect the clinical *C. jejuni* diversity present in the entire state. It is likely that numerous and genetically distinct lineages in the clinical setting circulate in New Hampshire but remain undiscovered or undetected (e.g., if a strain causes less severe symptoms in a patient during infection and thus may not seek medical intervention). The broad phylogenetic and pan-genomic diversity of the New Hampshire population paired with a low sample size in this study suggests that we have merely touched on the existing diversity of this pathogen within the state. It is possible that one or few of the 28 STs are already undergoing clonal expansion and more predominant in certain regions in New Hampshire, yet remain invisible to current surveillance schemes. Hence, future genomic studies should involve a more systematic sampling and active surveillance of patients from healthcare providers across the state in order to target certain counties and localities if needed (e.g., during outbreaks). Such statewide strategy across the country will also allow us to precisely define the phylogenetic relationships of *C. jejuni* co-circulating across the country and map the geographical dispersal of specific clones of interest. Unfortunately, our dataset does not include an extensive amount of clinical, phenotypic or other epidemiological information for each isolate because of how the sampling scheme was set up in the state. This is another important lesson we can learn from this study and apply to future genomic surveillance systems within the state. We strongly advocate for sampling and surveillance schemes of infectious

diseases, including *Campylobacter* infections, in the state of New Hampshire that include such pertinent information. Second, only clinical isolates were included in this study, which certainly posed limitations on elucidating the statewide diversity of the pathogen. Asymptomatic individuals may carry a genetically distinct *C. jejuni* population that remains to be characterized (G. G. Perron et al. 2012; Chisholm et al. 2018). Moreover, because campylobacteriosis is often associated with contaminated food products and exposure to animals (Kaakoush et al. 2015), whole genome sequencing of isolates from various sources (agricultural and food production settings, domestic animals, wild animals, environment) should be a major component of studies of disease ecology and epidemiology. Many reservoirs of *C. jejuni* are yet to be identified and bacterial populations from these sources undoubtedly contain many lineages that are yet to be described. Sampling and sequencing from non-clinical sources will provide valuable insights into the sources of horizontally acquired ABR genes, routes and mechanisms of transmission from agricultural and environmental reservoirs to humans, and genetic bases of bacterial adaptation to specific ecological niches (e.g., host versus non-host). Widespread application of whole genome sequencing of foodborne pathogens and other zoonotic diseases across the entire spectrum of the One Health paradigm (Destoumieux-Garzón et al. 2018) will therefore greatly facilitate public health interventions across multiple sectors. Lastly, next generation sequencing methods remain imperfect. *In silico* identification of any genetic elements, including resistance genes, relies on high-quality sequencing output. Genome sequencing failures are known to occur with any sequencing platform. Possible sources of errors include low number of reads, high incidence of unidentified or unreliable nucleotide calls (represented by "N"), high positional bias within the flowcell, and poor overall sequence qualities. The New Hampshire genomes used in our study all have <100 contigs, which is generally satisfactory in many bacterial genome studies. Application

of whole genome sequencing in public health laboratories is expected to improve the quality of sequences given the ongoing and rapid development in sequencing, DNA library preparation and bioinformatics technologies.

Whole genome sequencing is a powerful tool that provides timely, accurate and granular information about a pathogen that can be translated to public health action. The NH DHHS has only recently started sequencing bacterial genomes of select pathogens. This study presents some of the initial results of the state's initiative to implement whole genome sequencing in public health laboratories. It is expected that our results will reinforce the need to incorporate pathogen genomics as an integral component of New Hampshire's disease surveillance, control, clinical decisions and policy making. Here, we present an analysis of the standing pan-genomic variation of clinical *C. jejuni* within a local region in the United States. We conclude that the diversity of clinical *C. jejuni* in New Hampshire in 2017 was driven mainly by the co-existence of phylogenetically diverse antibiotic resistant lineages, widespread geographical mixing, and frequent recombination. Continued genomic surveillance will be necessary to assess how the local population of *C. jejuni* changes over the long term and in response to changing selective landscapes within the state.

FIGURES

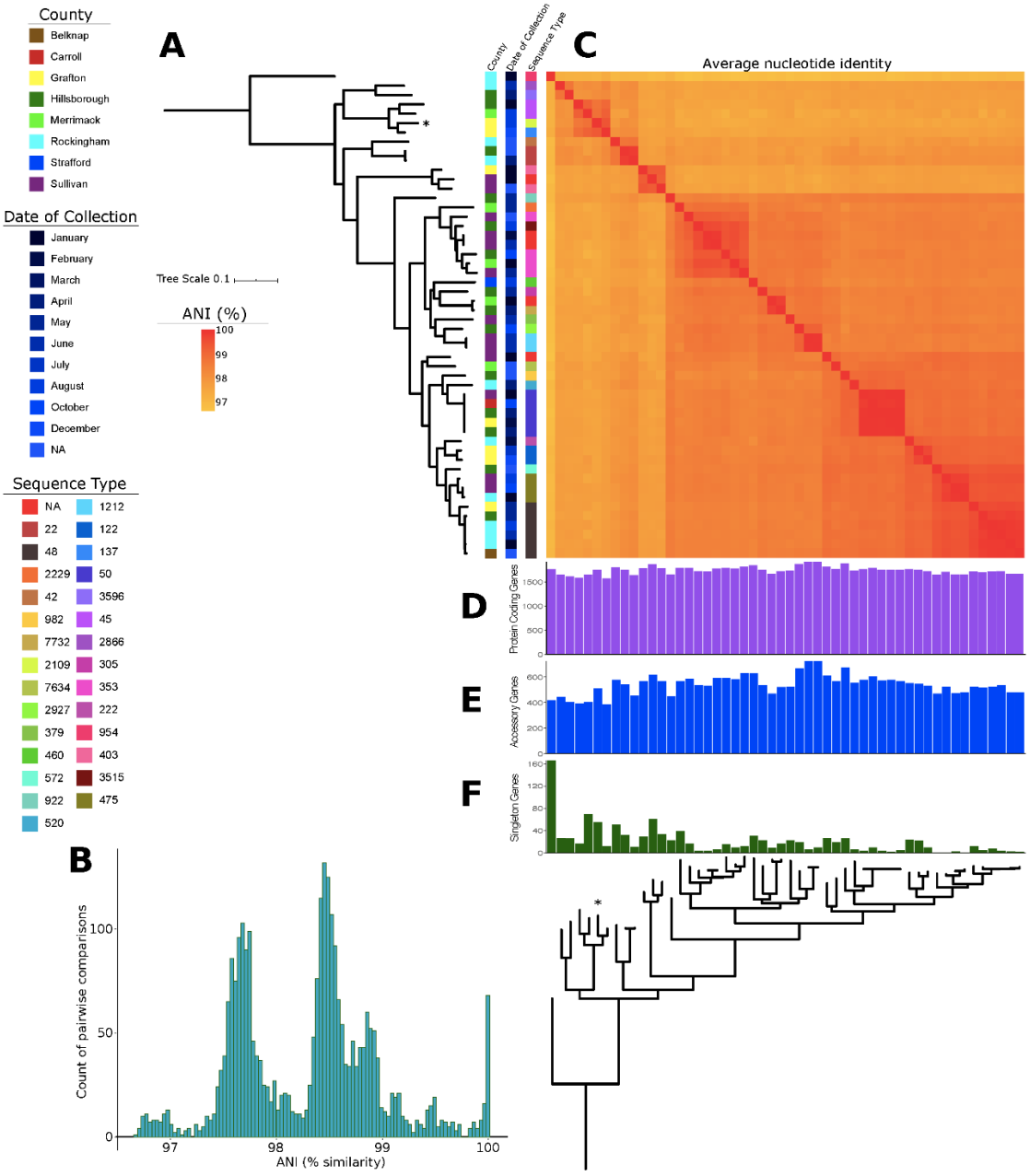

Figure 1. Phylogenetic relationships and pan-genome characteristics of the 52 *C. jejuni* isolates. (A) The phylogeny was reconstructed using 83,210 core SNPs. Scale bar represents the number of nucleotide substitutions per site. Asterisk indicates the genome of the *C. jejuni* from the multi-state puppy outbreak. (B) Frequency distribution of all pairwise ANI values. (C) ANI values were calculated for every pair of genomes in the entire dataset. Bar plots show the number of (D) protein coding genes, (E) accessory genes and (F) singleton genes per genome. Singleton genes those that are unique to an individual genome.

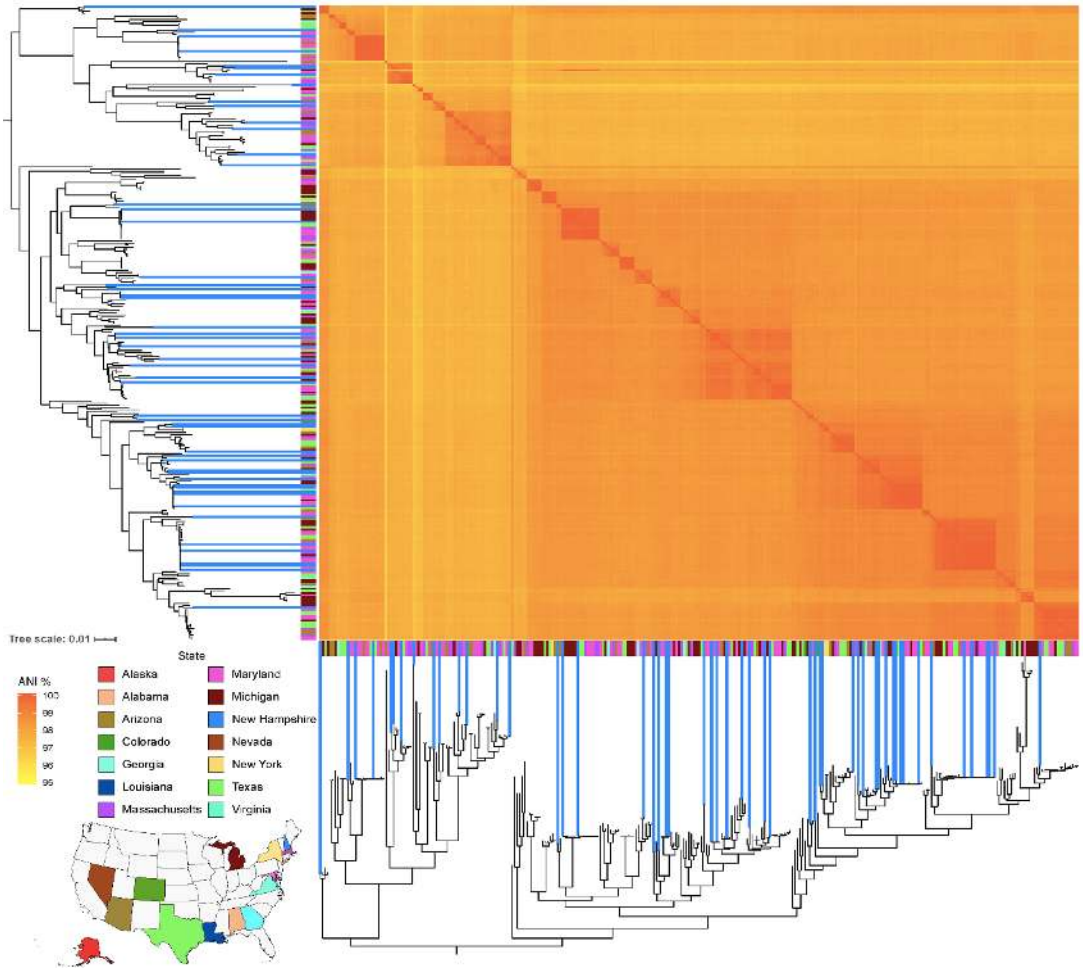

Figure 2. Phylogenetic relationships of 52 *C. jejuni* isolates combined with 249 isolates from 13 other states in the United States. The genome sequences of the latter were obtained from the 100K Pathogen Project. The phylogeny was constructed from the alignment of 937 core genes. Scale bar represents the number of nucleotide substitutions per site. ANI values were calculated for every pair of genomes in the entire dataset. Colored strip represents the state of origin for each isolate. Colored strips representing New Hampshire are elongated to distinguish them from the rest of the United States population.

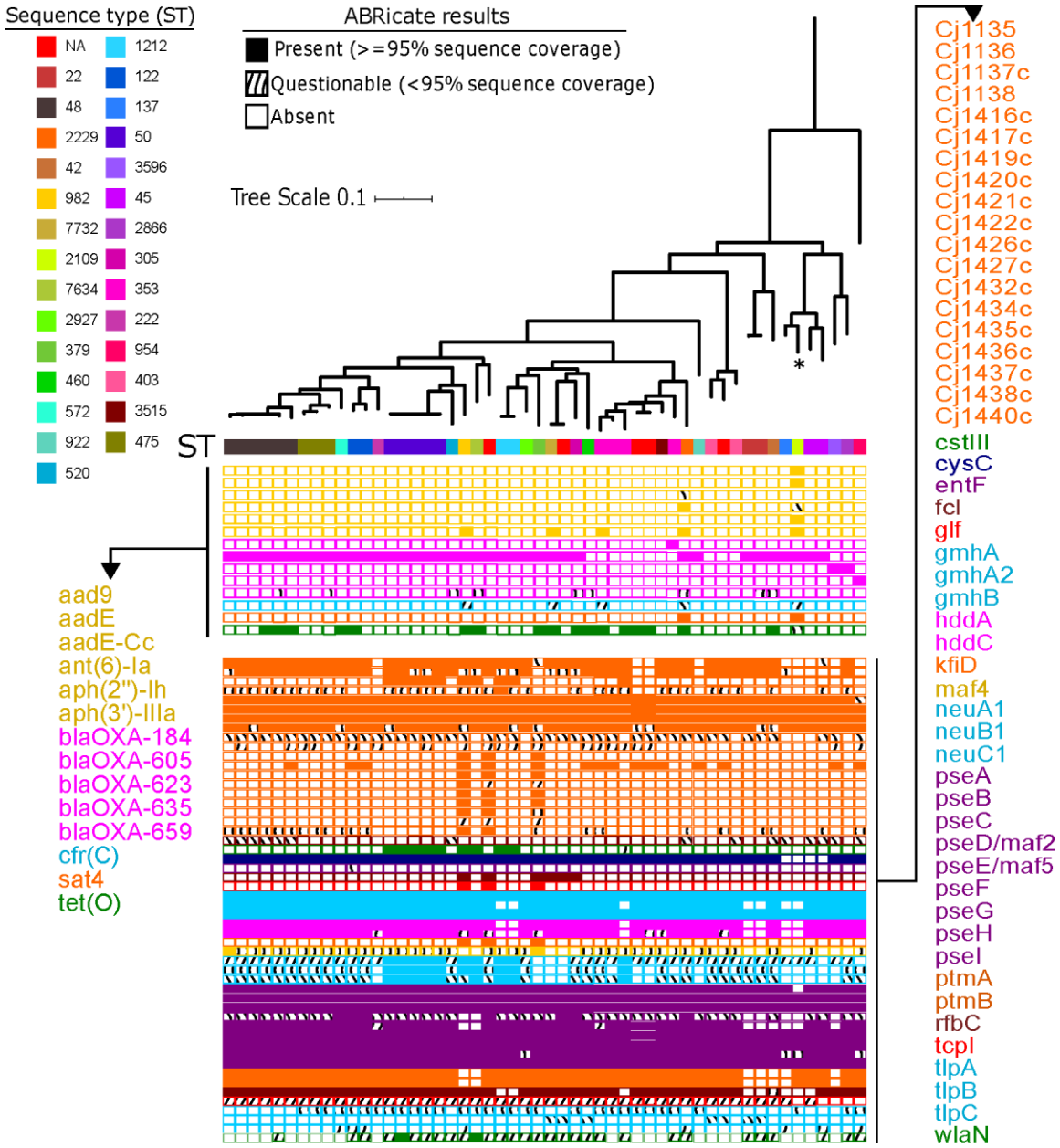

Figure 3. Summary of ABR and virulence profiles of individual C. jejuni genomes. Names of horizontally acquired resistance genes are on the left and colored by antibiotic class. Names of virulence genes are listed on the right. Solid blocks indicate presence of gene (≥95% sequence coverage), wavy blocks indicate questionable presence (<95% sequence coverage), and empty boxes indicate the absence of the gene. The tree is identical to that in Fig. 1. Only those virulence genes that are differentially distributed among strains are shown here. A comprehensive list of all virulence genes identified in each strain is shown in Table S6.

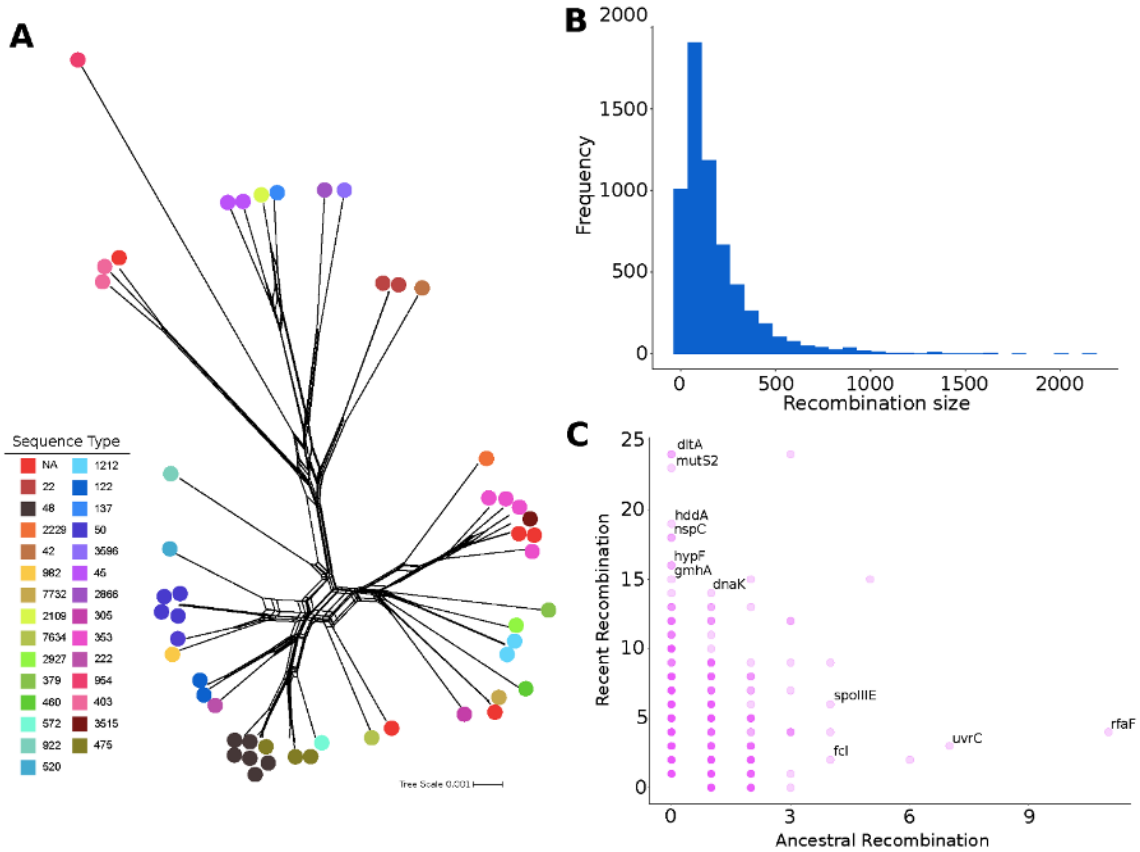

Figure 4. Recombination characteristics of the New Hampshire C. jejuni. (A) Phylogenetic SplitsTree network generated from the core genome alignment. Scale bar represents nucleotide substitutions per site. (B) Frequency distribution of the size of recombined DNA segments. (C) Genes that have undergone recent and/or ancestral recombination. For clarity, names of some of the most frequently recombined genes with known functions are shown. A list of all recombination events is presented in Table S7.

ACKNOWLEDGMENTS

We thank the UNH Resource Computing Center where all bioinformatics analyses were performed. We thank Anthony Westbrook for providing technical and bioinformatics assistance. We thank all the healthcare providers in New Hampshire who provided the bacterial samples to NH DHHS. The authors declare no conflict of interest relevant to the study.

APPENDIX 3

Supplementary Tables for this publication can be found at:

https://jcm.asm.org/content/58/6/e02070-19/figures-only#fig-data-additional-files

CHAPTER 4

Diverse lineages of multidrug resistant clinical *Salmonella enterica* in New Hampshire, USA revealed from a year-long genomic surveillance

Cooper J. Park[a], Jinfeng Li[b], Xinglu Zhang[b], Fengxiang Gao[b], Christopher S. Benton[b], Cheryl P. Andam[ac]

Article currently under review in *MEEGID*

[a] University of New Hampshire, Department of Molecular, Cellular and Biomedical Sciences, Durham, *New Hampshire,* USA

[b] *New Hampshire* Department of Health and Human Services, 29 Hazen Drive, Concord, *New Hampshire,* USA

[c] University at Albany, State University of New York, Department of Biological Sciences, Albany, *New York,* USA

ABSTRACT

Salmonella enterica, the causative agent of gastrointestinal diseases and typhoid fever, is a human and animal pathogen which causes significant mortality and morbidity worldwide. The use of whole genome sequencing in surveillance and monitoring of *Salmonella* infections creates tremendous opportunities to elucidate the genetic basis of antimicrobial resistance, virulence and diversity of *S. enterica* circulating in the community. In this study, we present our findings on the genomic diversity and phylogenetic relationships of 63 *S. enterica* isolates from human clinical specimens reported to the Department of Health and Human Services (DHHS) in the

state of New Hampshire, United States in 2017. We found a remarkably large genomic, phylogenetic and serotype variation among the *S. enterica* isolates co-circulating across the state, dominated by serotypes Enteritidis (sequence type [ST] 11), Heidelberg (ST 15) and Typhimurium (ST 19). We found that nearly all of the clinical *S. enterica* isolates carry numerous genetic determinants that confer resistance to multiple classes of antimicrobials, most notably aminoglycosides, fluoroquinolones and macrolides. Majority of the isolates (48 out of 63) carry at least four resistance determinants per genome. We also detected the genes *mdtK* and *mdsABC* that encode multidrug efflux pumps and the gene *sdiA* that encodes a regulator for a third multidrug resistance pump. Our results indicate rapid microevolution and geographical dissemination of multidrug resistant lineages over a short time span. These findings are critical to aid the DHHS and similar public health laboratories in the development of effective disease control measures, epidemiological studies and treatment options for serious *Salmonella* infections.

INTRODUCTION

Infections due to *Salmonella enterica* remains a major public health concern worldwide. The medical costs associated with surveillance, prevention and treatment further exacerbates the economic burden caused by *Salmonella* infections. Control of *Salmonella* infections is difficult partly because of the vast number and diversity of *Salmonella* serotypes. To date, more than 2,500 serotypes have been recognized (X. Zhang, Payne, and Lan 2019; Brenner et al. 2000), which display a broad range of epidemiological (e.g., virulence features, modes of transmission, disease outcomes, response to clinical interventions), ecological (e.g., host range, unique reservoirs, seasonal distribution) and evolutionary (e.g., rates of recombination) features (Gal-Mor, Boyle, and Grassl 2014; Judd et al. 2019; Jones et al. 2008). *S. enterica* serotypes are broadly divided into typhoidal and non-typhoidal based on the disease they cause (Balasubramanian et al. 2019; Darton, Blohmke, and Pollard 2014; Gal-Mor, Boyle, and Grassl 2014). The typhoidal serotypes (Typhi and Paratyphi) often cause severe systemic diseases in humans while the non-typhoidal serotypes mostly cause diarrhea, fever and abdominal cramps (Balasubramanian et al. 2019; Darton, Blohmke, and Pollard 2014; Gal-Mor, Boyle, and Grassl 2014).

Severe *Salmonella* infections can lead to death unless treated with antimicrobials (Crump et al. 2015; Stanaway et al. 2019). However, antimicrobial resistance in *S. enterica* has been increasing in the last four decades (Britto et al. 2018; Hawkey et al. 2019; Leekitcharoenphon et al. 2016). In the 2019 report of the United States Centers for Disease Control and Prevention (CDC) on antimicrobial resistance, both drug-resistant non-typhoidal and typhoidal *Salmonella* are considered to be serious level threats to human health (CDC 2019). In the United States, the

CDC estimates that non-typhoidal *Salmonella* causes an estimated 1.35 million infections, 26,500 hospitalizations, and 420 deaths annually, resulting in an estimated $400 million in direct medical costs (CDC 2019). Of these, 212,500 infections and 70 deaths every year are due to drug-resistant non-typhoidal *Salmonella* (CDC 2019). *Salmonella* serotype Typhi, which causes the potentially life-threatening typhoid fever (Britto et al. 2018; Darton, Blohmke, and Pollard 2014), causes an estimated 5,700 infections and 620 hospitalizations each year in the United States (CDC 2019). Of these, drug-resistant *Salmonella* Typhi results to 4,100 estimated infections and <5 deaths each year (CDC 2019). The emergence and geographic spread of multidrug resistance (i.e., resistant to three or more classes of antimicrobials) and resistance to currently antibiotics used for treatment such as ceftriaxone and ciprofloxacin (Klemm et al. 2018b; Mather et al. 2018; M. H. Y. Wong et al. 2014; Hawkey et al. 2019) is severely diminishing treatment options for *Salmonella* infections.

Salmonella infections are a nationally notifiable disease in the United States. Molecular surveillance activities based on pathogen subtyping and antimicrobial non-susceptibility testing are therefore important measures for tracking and controlling *Salmonella* infections by regional public health agencies. The use of whole genome sequencing by public health laboratories creates tremendous opportunities to elucidate the genetic basis of antimicrobial resistance, virulence and diversity of *S. enterica* circulating in the community. In this study, we present our findings on the genomic diversity and phylogenetic relationships of *S. enterica* isolates from human clinical specimens reported to the Department of Health and Human Services (DHHS) in the state of New Hampshire, United States in 2017. Our results reveal a genetically diverse assembly of multidrug resistant lineages and serotypes of *S. enterica* found across the state,

which can inform effective disease control, outbreak investigations and case studies of *Salmonella* infections in the region.

METHODS

Bacterial isolates

Isolates were submitted to the Public Health Laboratories, New Hampshire DHHS in Concord, New Hampshire, USA in 2017. These isolates were received from New Hampshire health care providers and were collected primarily from individuals diagnosed with *Salmonella* infection. Most of the *Salmonella* isolates were recovered from stool, while a few were obtained from bile, blood and urine. The state of New Hampshire considers *Salmonella* infections as a reportable disease and the DHHS strongly encourages isolate submission to the Public Health Laboratories. However, submission of isolates is not mandatory. No identifiable information is associated with the isolates submitted by the health care providers. In total, our data includes 63 isolates (Supplementary Table S1).

DNA extraction and whole genome sequencing

Sequencing of *Salmonella* isolates is part of the CDC-sponsored program PulseNet surveillance, a United States national laboratory network that connects foodborne illness cases to detect outbreaks (Tolar et al. 2019). DNA extraction, library preparation and whole genome sequencing were done following the PulseNet USA standard operating procedures (https://www.cdc.gov/pulsenet/index.html). Briefly, DNA extraction procedures were conducted using the DNeasy Blood & Tissue Kit (Qiagen, Valencia CA). DNA quality and concentration were measured using Qubit fluorometer and NanoDrop spectrophotometer. A total of 1 ng of

genomic DNA from each isolate was used to construct sequencing libraries using the Nextera XT DNA Library Preparation Kit (Illumina, Inc. San Diego, CA) per manufacturer's instructions. Samples were sequenced as multiplexed libraries on the Illumina MiSeq platform operated following the manufacturer's instructions for 500 cycles to produce paired end reads of 250 bp in length. The MiSeq sequencer is housed at the New Hampshire DHHS Public Health Laboratories.

De novo genome assembly, annotation, pan-genome and phylogenetic analyses

We used the Nullarbor pipeline v2.0 (https://github.com/tseemann/nullarbor) to perform read trimming, quality assessment, contig assembly, gene annotation, pan-genome, ST identification, sequence alignment and phylogenetic analysis of the entire dataset. The Nullarbor pipeline can be briefly described as follows: Adapters were trimmed using Trimmomatic v0.38 (Bolger, Lohse, and Usadel 2014). Trimmed reads were assembled into contigs using SKESA v2.3.0 (Souvorov, Agarwala, and Lipman 2018). using an *S. enterica* subsp. *enterica* serovar Typhimurium str. LT2 reference genome obtained from the RefSeq database (Accession ID: GCF_000006945.2) of the National Center for Biotechnology Information (NCBI). Quality of genome assemblies was assessed using Quast (Gurevich et al. 2013). Assembled genomes were annotated using Prokka v1.13.3 (T. Seemann 2014) with default parameters. To determine the degree of overall genomic relatedness between genomes, we calculated the genome-wide average nucleotide identity (ANI) for all possible pairs of genomes using the program FastANI (Jain et al. 2018). ANI estimates the average nucleotide identity of all orthologous genes shared between any two genomes (Jain et al. 2018). Pairwise ANI values and plots were generated and visualized using R (R Core Team 2019).

We used Roary v3.12.0 (Page et al. 2015) to characterize the pan-genome of the New Hampshire *S. enterica* dataset. The presence or absence of genes was visualized using post-processing scripts provided by the Roary program. Each orthologous gene family was aligned using MAFFT 7.407 (Kazutaka Katoh, Rozewicki, and Yamada 2017). The ST of each isolate was determined using the program mlst (multilocus sequence typing; https://github.com/tseemann/mlst), which extracts the sequences of seven housekeeping genes (*aroC, dnaN, hemD, hisD, purE, sucA, thrA*) from the Illumina raw sequences and compares them to the *S. enterica* MLST database (www.mlst.net) (Jolley, Chan, and Maiden 2004). Mobile genetic elements were identified from the assembled contigs using IslandViewer 4 (Bertelli et al. 2017), PlasmidFinder (Carattoli et al. 2014) and ViralRecall (https://github.com/faylward/viralrecall) to identify genomic islands, plasmids and prophages, respectively. Single nucleotide polymorphisms (SNPs) from the core genes were identified and aligned using Snippy v4.3.6 (https://github.com/tseemann/snippy) and were used to generate a maximum likelihood phylogeny with a general time reversible (GTR) nucleotide substitution model (Tavaré 1986) and four gamma categories for rate heterogeneity using the program IQ-TREE v1.6.9 (Nguyen, Lam-Tung, Schmidt, Heiko A., Haeseler, Arndt von, Minh 2015). Phylogenetic trees were visualized using iToL v5.5.1 (Letunic and Bork 2016). Statistical analysis of gene content differences between genomes was carried out using Mann-Whitney U pairwise tests (Mann and Whitney 1947) with Bonferroni adjusted p-values (Bonferroni 1936).

Serotype identification was carried out using both conventional phenotypic serotyping and genome-based methods. First, serotype was determined by agglutination of the bacterium with specific antisera to identify variants of the two surface structures O and H antigens based on

the WKL scheme (Grimont and Weill 2007). Second, we used a k-mer-based algorithm called SeqSero2 that uses raw reads to predict serotypes defined by the O and H antigens (S. Zhang et al. 2019). The k-mers were then compared to the serotype determinant database composed of the sequences of the *wzx* and *wzy* genes for the O antigen and the *fliC* and *fljB* genes for the H antigen (S. Zhang et al. 2019).

In silico identification of antimicrobial resistance genes

We screened all genomes for known resistance genes using a local assembly and contig mapping method called Antimicrobial Resistance Identification By Assembly (ARIBA) (Hunt et al., 2017). ARIBA identifies both horizontally acquired resistance genes and chromosomal mutations associated with resistance by mapping reads to a reference database. We used the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al., 2013) for comparison with the New Hampshire genomes. Sequence comparison was carried out by matching contigs to their closest reference sequence using MUMmer (Kurtz et al., 2004).

Data availability

All *S. enterica* genomic sequences generated under PulseNet USA surveillance (Tolar et al. 2019) are uploaded in real-time to the sequence read archive (SRA) hosted by NCBI. The genomes analyzed in this study are available in BioProject PRJNA230403. Accession numbers and Biosample IDs for the New Hampshire genomes are listed in Supplementary Table 1.

RESULTS

Genomic and phylogenetic characteristics of S. enterica in New Hampshire

We sequenced the genomes of 63 clinical *S. enterica* isolates collected from ten counties in New Hampshire, USA in 2017 (Fig. 1a and Supplementary Table 1). The isolates came from stool (n = 56 isolates), urine (n = 4), blood (n = 2) and bile (n = 1). The genome sequences contain between 21 – 126 contigs and N50 values range between 71,043 and 708,941bp (Supplementary Table 1). *De novo* genome assemblies generated sequences of sizes ranging from 4.51 – 5.03 Mb (mean = 4.74 Mb) (Supplementary Table 1). We used Roary to estimate the pan-genome (Page et al., 2015) of the entire New Hampshire *S. enterica* dataset. The pan-genome is defined as the totality of genes in a set of strains (Medini et al. 2005). Of the 9,850 gene families identified in the pan-genome, a total of 3,407 genes comprised the core genome (i.e., a core gene is present in 99% ≤ strains < 100%), which represents approximately 34.6% of the pan-genome. The maximum likelihood phylogenetic tree based on the alignment of 169,002 core SNPs revealed little overall population structure relative to the location of the healthcare provider (county) and date of collection (Fig 1a). Genome-wide ANI values (Jain et al. 2018) for every possible pair of *S. enterica* genomes ranged from 97.9 – 99.9% (mean = 98.8%) (Fig 1b and c). Together, the core genes (n = 3,407 genes) and the soft-core genes (n = 211 genes; defined as those genes present in 95% ≤ strains < 99%) constitute only 36.7% of the entire population's pan-genome. Accessory genes can be categorized into shell (n = 1,600 genes; defined as those genes present in 15% ≤ strains < 95%) and cloud genes (n = 4,632 genes; defined as those genes present in < 15% of strains). Together, both categories of accessory genes constitute 63.3% of the population's pan-genome. There was substantial strain-level variation in the New Hampshire population in terms of gene content. The number of protein-coding genes

per genome ranged from 4,190 – 4,758 (mean = 4446) (Fig 1d). The number of accessory genes per genome ranged from 427 – 898 (mean = 666) (Fig 1e). Many accessory genes were also unique to individual strains (2,519 genes representing 25.6% of the pan-genome), with 0 – 366 singleton genes identified per genome (Fig 1f).

Our results from the *in silico* MLST analysis showed that the *S. enterica* isolates belonged to 20 unique known STs (Fig. 1a). One novel ST found in a single strain had a MLST profile with no known match to the MLST database (Jolley, Chan, and Maiden 2004). We also identified a total of 18 serotypes using SeqSero2 (S. Zhang et al. 2019). The most common serotypes in the New Hampshire population were Enteritidis (ST 11), Heidelberg (ST 15) and Typhimurium (ST 19), which were represented by 15, 9, and 10 isolates respectively. Except for one genome, serotypes identified using the conventional phenotypic serotyping assay and the *in silico* method implemented in SeqSero2 were in concordance. Genome SRR6026010 was identified as serotype Panama by the former method but serotype Javiana by the latter. However, the core genome tree showed the isolate falling within the Javiana serotype cluster (Fig. 1a). We detected two typhoidal serotypes Typhi (ST 2) and Paratyphi B (ST 43), while the rest were all non-typhoidal.

Isolates from urine and blood were intermingled with the stool isolates and did not form source-specific clusters in the phylogenetic tree (Fig. 1a). The four urine isolates were represented by ST 1674 serotype Javiana, ST 32 serotype Reading, and two isolates of ST 11 serotype Enteriditis. The blood isolates were represented by ST 11 serotype Enteriditis and ST

19 serotype Typhimurium. Lastly, the single isolate from bile was represented by ST 23 serotype Oranienburg.

Genomic variation between closely related strains

The three most prominent STs were 11, 15 and 19, all of which correspond to the three common serotypes described above. We found genome content variation among members of each of the three STs. The core genome of each ST consisted of 4,311, 4,425 and 4,294 genes for STs 11, 15 and 19, respectively, while the accessory genome consisted of 624, 25 and 870 genes for STs 11, 15 and 19, respectively (Fig. 2a). Comparisons of these three STs revealed significant differences between their pan-genomes. Genomes within ST19 consistently demonstrated greater genomic diversity with higher counts of mobile genetic elements consisting of pathogenicity islands, plasmids and phages (Fig. 2b), accessory genes (Fig. 2c), protein coding genes (Fig. 2d) and singleton genes (i.e., genes unique to a single genome) (Fig. 2e) (Mann-Whitney U test). The presence of diverse mobile genetic elements that can rapidly disseminate genetic material between lineages may partly explain the large genomic variation between closely related *Salmonella* genomes (Emond-Rheault et al. 2020; Moreno Switt et al. 2012).

Notably, all ST15 genomes were sampled within a month of each other (August – September) and were primarily from the same county (Rockingham), which may explain the near absence of genomic diversity among individual genomes (Fig. 2bcde). Whether these Rockingham isolates are epidemiologically linked and/or comprise a local outbreak requires additional clinical data from the healthcare providers, which were not available to us. On the other hand, isolates of STs 11 and 19 were obtained from multiple counties and sampling months

throughout the year. Overall, we found a remarkably large genomic, phylogenetic and serotype variation among the *S. enterica* isolates co-circulating across the state of New Hampshire, which indicates rapid microevolution and geographical dissemination over a short time span.

Distribution of antimicrobial resistance genes

We used an *in silico* method implemented in the program ARIBA (Hunt et al. 2017) to determine the presence of horizontally acquired resistance genes and chromosomal mutations associated with resistance to a range of different classes of antimicrobials. We identified a total of 21 unique genes associated with resistance across ten different classes of antimicrobials (aminoglycosides, beta-lactams, cephalosporins, elfamycins, fosfomycins, fluoroquinolones, macrolides, phenicols, sulfonamides and tetracyclines) (Fig. 3a). We found that all 63 strains carry at least one resistance determinant. Three antimicrobial classes have the highest number of genomes carrying at least one resistance gene associated with it. These are aminoglycosides, fluoroquinolones and macrolides with 63/63 (100%), 57/63 (90.48%) and 46/63 (73.02%) genomes having at least one resistance gene for each class, respectively (Fig. 3b). We also detected *mdtK* gene, which encodes the multi-drug efflux pump and confers resistance against acriflavin, doxorubicin and norfloxacin (Nishino, Latifi, and Groisman 2006), in 92.06% (58/63) of the genomes. Only isolates with serotype Newport do not carry the *mdtK* gene. A remarkable 48 genomes, which constitute 76% of the dataset, carry four or more resistance determinants per genome (Fig 3c). Lastly, the *Salmonella*-specific multi-drug transporter efflux pump MdsABC and its promoter *golS* were found in all but one genome. MdsABC has been found to confer resistance to novobiocin and is required for full virulence to infect and colonize host cells (Nishino, Latifi, and Groisman 2006). The gene *sdiA*, which encodes a quorum-sensing

regulator that mediates the multi-drug resistance pump AcrAB (Rahmati et al. 2002), was also present in all genomes. It has been previously shown that overproduction of the gene product SdiA confers multidrug resistance and increased levels of AcrAB to the cell (Rahmati et al. 2002).

All genomes of STs 11, 15 and 19 carry resistance determinants associated with aminoglycosides, fluoroquinolones and the multidrug Mds efflux pump. Except for one genome, all ST11 genomes also carry resistance determinant for macrolides. All genomes of ST 15 also carry the *fosA7* gene which confers high-level resistance to fosfomycin (Rehman et al. 2017). The product of *fosA7* is a glutathione-S-transferase that binds to fosfomycin and ruptures its epoxide ring structure (Rehman et al. 2017). FosA7 was first detected in *S. enterica* serovar Heidelberg isolated from broiler chickens and has been shown to be transferrable via plasmid mobility (Rehman et al. 2017). Moreover, four of the eight genomes of ST 15 carry the resistance determinant for elfamycin, which targets the elongation factor-Tu (Prezioso, Brown, and Goldberg 2017). In ST 19, resistance determinants associated with cephalosporin, phenicol, sulfonamide and tetracycline are found in one, one, two and two genomes, respectively. Isolates that carry each of the genes that confer resistance to these four classes of antimicrobials were distributed in disparate parts of the phylogenetic tree and were not associated with any one ST or serotype, which may be indicative of horizontal transfer of resistance genes (Krauland et al. 2009; Cohen et al. 2020; Oladeinde et al. 2019; Park and Andam 2020). We also detected the presence of multiple resistance determinants per genome in other less common STs (i.e., those represented by only a single isolate), such as STs 2, 10, 132, 138 and 448.

DISCUSSION

S. enterica infections are a major public health concern in the United States and worldwide (Crump et al. 2015; Stanaway et al. 2019). The application of whole genome sequencing in infectious disease surveillance and epidemiological studies is a powerful tool for public health agencies and laboratories. Our study provides the initial genomic analysis of *S. enterica* isolates from clinical human specimens received by the New Hampshire DHHS and we show that nearly all the clinical isolates of this pathogen carry genetic determinants that confer resistance to multiple classes of antimicrobial compounds. We also show that although three lineages (STs 11, 15 and 19) are relatively common, numerous other STs and serotypes are also co-circulating in the clinical population.

There are three aspects of the New Hampshire *S. enterica* population worth highlighting. First, the distribution of many of the resistance genes in disparate parts of the phylogenetic tree reflects their rapid mobility between distinct lineages, including less common STs and serotypes (i.e., represented by one or two isolates). That the rarer lineages and serotypes were also multidrug resistant means that they can potentially increase in frequency in the population in the long term or act as a reservoir of horizontally transferrable resistance genes. Future work will help illuminate from which lineages they acquired the resistance genes from and whether these horizontally acquired resistance genes are maintained in in the population over many years. Horizontal gene transfer and recombination can also lead to the emergence of novel genetic variants with unique epidemiological characteristics, as has been reported in other bacterial pathogens (Sun et al. 2016; Chen et al. 2014) and *Salmonella* is not an exception (Brown et al. 2003; Criscuolo et al. 2019; Park and Andam 2020). Second, the remarkably diverse population

of *S. enterica* in the region within a short time span of one year highlights the need to implement a multi-year surveillance to understand the dynamics of these lineages and serotypes. It is possible that the long-term dynamics of the clinical *S. enterica* population in New Hampshire may be characterized by the persistence of the pre-existing dominant strains (STs 11, 15 and 19) and serotypes rather than through *de novo* adaptation (Andam et al. 2017). Genetic differences may accumulate in the three most common STs as they evolve and adapt over the long term and in response to environmental or host changes. The three STs are associated with the serotypes *S.* Typhimurium, Enteritidis and Heidelberg, which are major public health threats across the world. *S.* Typhimurium accounts for a quarter of total global infections and is exceptional in its wide host range (human, livestock, wildlife) and environmental distribution, global dissemination and multidrug resistance (Branchu, Bawn, and Kingsley 2018; Leekitcharoenphon et al. 2016). Major outbreaks worldwide due to contaminated food and animal sources have been attributed to S. Enteritidis (Vaughn et al. 2020; Dallman et al. 2016) and Heidelberg (Antony et al. 2018; Bearson et al. 2017). Future surveillance will be critical to documenting any of these scenarios, including whether these three serotypes will continue to persist in New Hampshire. On the other hand, less common lineages with unique features can have a selective advantage over their competitors as environmental conditions change (e.g., changes in host demography, clinical interventions implemented, food and animal sources) (X. Zhang, Payne, and Lan 2019) and replace the three most common STs or serotypes. Third, these multidrug resistant isolates are found all across the ten counties, suggesting the widespread geographical spread of multidrug resistance across the state. Such knowledge would be particularly relevant to public health officials to enable precise identification of priority regions (counties) and inform regulation strategies for antimicrobial compounds in the state.

This study has several limitations. First, sampling was limited to only clinical specimens received by DHHS and those collected from patients who went to see their health care providers. Hence, we were not able to determine the population structure and antimicrobial resistance profiles of *S. enterica* from the greater New Hampshire community, which may consist of a different suite of STs and serotypes or may exhibit dissimilar composition and distribution of resistance determinants. Since gastroenteritis is mainly a self-limiting disease (Gal-Mor, Boyle, and Grassl 2014), most people are likely to self-medicate and not see their healthcare provider. This means that a large subset of the New Hampshire population is not represented in the current study and only those STs and serotypes that result to more severe infections were included. A more comprehensive and systematic surveillance system is therefore needed to track the population structure of *S. enterica* isolates that cause less severe symptoms or those isolates from patients who experience rapid recovery. Gaps in disease surveillance means that isolates from these patients remain invisible from comparative analyses and may not be taken into account in clinical decision-making procedures. Second, we had limited metadata associated with the isolates because of current state policies regarding patient records. Such data would have been epidemiologically informative, especially in the case of the closely related ST 15 isolates from Rockingham county recovered in August - September. The small amount of genomic differences among them suggest a rapid emergence and/or spread, indicative of an unrecognized localized outbreak in the region. Cryptic transmission and outbreaks that might have otherwise gone unnoticed have been previously identified in *Salmonella* and other bacterial pathogens (Taylor et al. 1998; Roach et al. 2015; Turner et al. 2017) and our results suggest that this might have occurred in Rockingham. Moving forward, our results will prove useful as a basis to further

investigate this cluster of *Salmonella* cases and determine their relationships. Another limitation is the lack of isolates from food and environmental sources, where *S. enterica* is also known to inhabit and survive (Fernández, Guerra, and Rodicio 2018; Pornsukarom, Van Vliet, and Thakur 2018; Silva, Calva, and Maloy 2014) and to which we can compare our clinical isolates. Such information will be critical to ascertain the origin of the clinical isolates reported to DHHS and ensure safety in the food production and supply chains.

CONCLUSION

In summary, we found that nearly all of the clinical *S. enterica* isolates from the 2017 New Hampshire population carry numerous genetic determinants that confer resistance to multiple classes of antimicrobials. Our results suggest rapid microevolution and geographical dissemination of multidrug resistant lineages over a short time span. The disparate phylogenetic distribution of many of the resistance genes reflect their rapid mobility between phylogenetically distinct lineages and the potential threat of further geographical spread of multidrug resistance across the state. Future work should focus on implementing a multi-year genomic surveillance to help illuminate the population dynamics of clinical *S. enterica* in the state of New Hampshire.

FIGURES

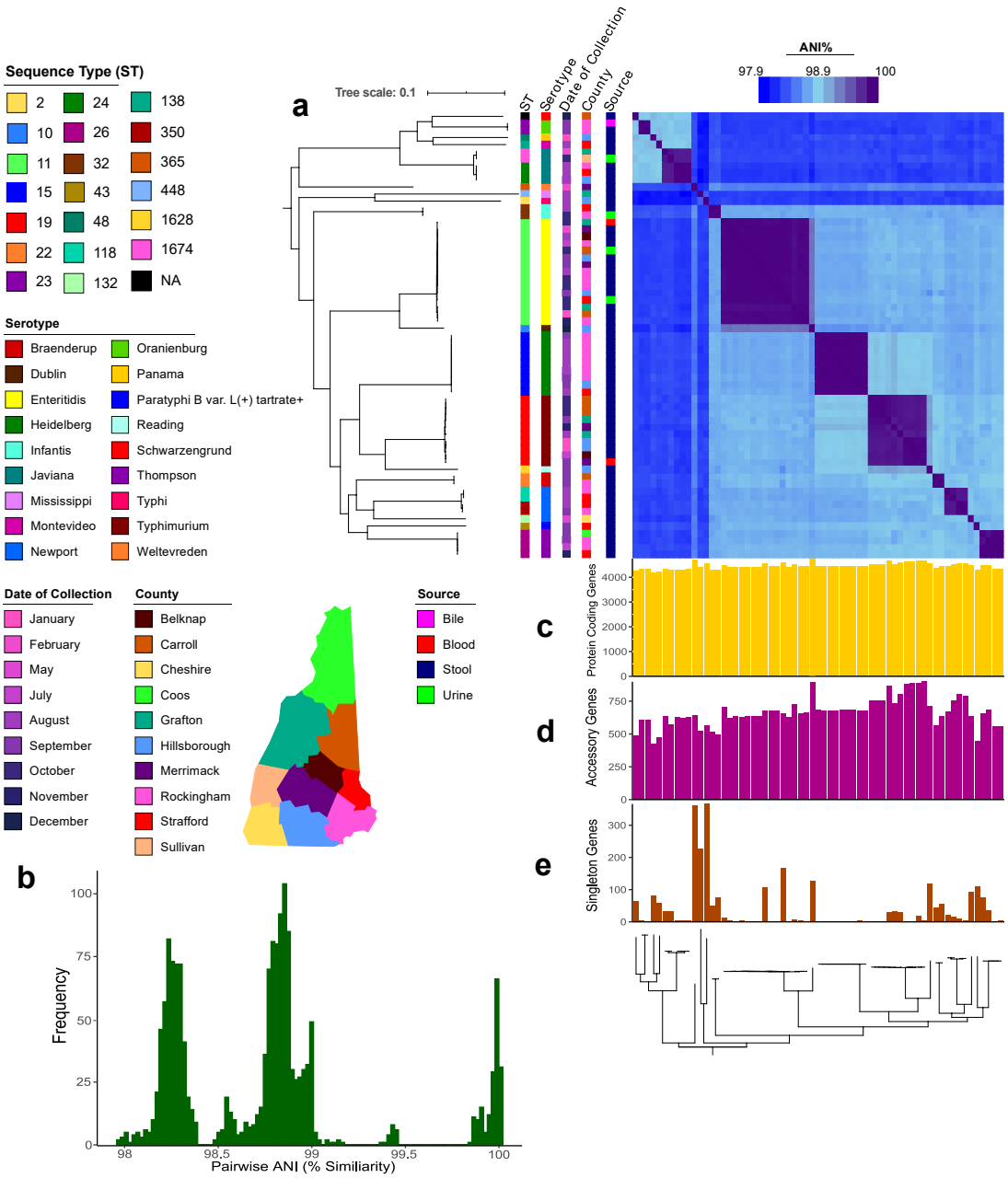

Figure 1. Phylogenetic relationship and genomic characteristics of the 63 clinical isolates of *S. enterica* from New Hampshire. (a) The phylogeny was reconstructed from 169,001 core SNPs using IQTree. The scale bar represents the number of nucleotide substitutions per site. The matrix on the right shows ANI values calculated for every pair of genomes in the entire data set. (b) Frequency distribution of all pairwise ANI values. (c) Number of protein coding genes per genome. (d) Number of accessory genes per genome. (e) Number of singleton genes per genome. Singleton genes are those that are unique to an individual genome.

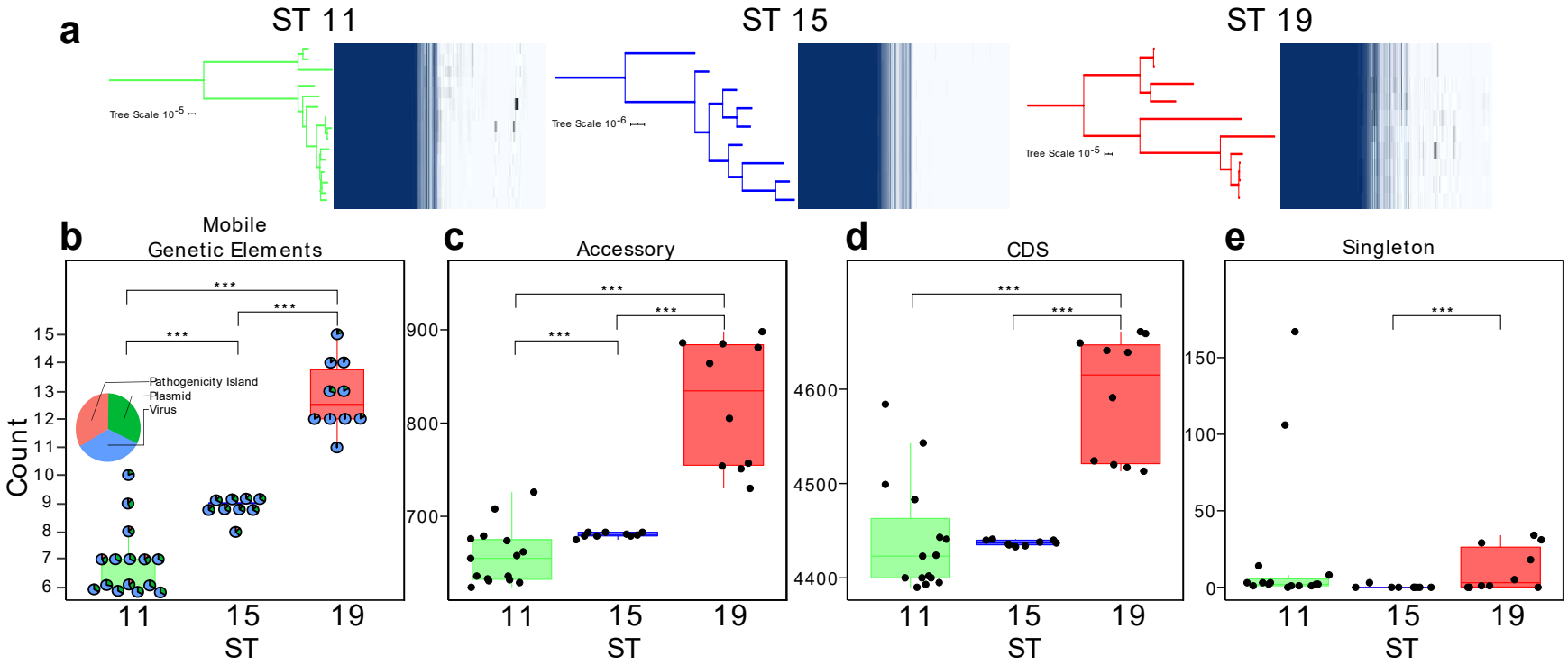

Figure 2. Genomic variation among strains of the same ST. (a) Phylogenetic trees of STs 11, 15 and 19 built from 870, 17, and 1,624, respectively core SNPs using IQTree. The matrix of the right of each tree shows the presence (dark blue) or absence (light blue) of gene families per genome. The scale bar represents the number of nucleotide substitutions per site. Comparison of mobile genetic elements (b), accessory genes (c), protein coding genes (d) and singleton genes (e) among the three STs. *** represents a p-value < 0.001 using a Mann-Whitney U pairwise test with a Bonferroni p-value adjustment.

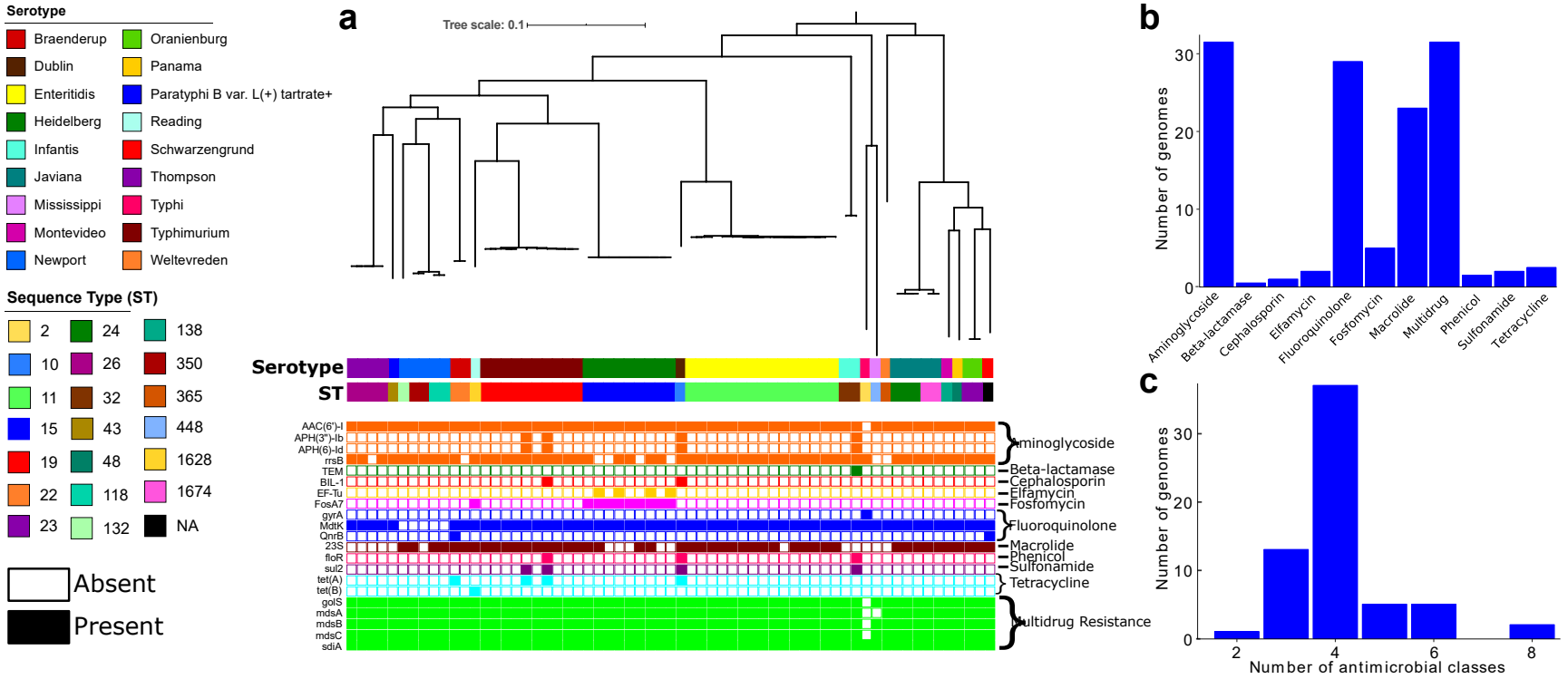

Figure 3. Antimicrobial resistance profiles of the 63 *S. enterica* isolates. (a) Names of specific resistance genes are listed on the left and the names of antimicrobial classes are listed on the right of the matrix. Solid blocks indicate the presence (≥95% sequence coverage) and empty boxes indicate the absence of the resistance determinant. The tree is identical to that in Fig. 1. A comprehensive list of all resistance genes identified in each strain is shown in Table S6. (b) Number of genomes carrying at least one resistance gene for each class of antimicrobial compound. (c) Number of genomes carrying multiple number of resistance genes.

ACKNOWLEDGEMENTS

We thank the UNH Resource Computing Center where all bioinformatics analyses were performed. We thank all the healthcare providers in New Hampshire who provided the bacterial samples to the New Hampshire DHHS. We acknowledge the start-up funds from the University of New Hampshire College of Life Sciences and Agriculture for the financial support to C.P.A.

APPENDIX 4

Supplementary Table 1. Accession numbers, metadata and genome characteristics of the 63 *S. enterica* isolates.

Genome	DOC	County	Sequence Type	CDC serotype	SeqSero_Serotype	Source	Contigs	bp	N50	CDS	rRNA	tRNA	tmRNA
PNUSAS25946	September	Rockingham	26	Thompson	Thompson	Fecal	22	4782193	708941	4478	7	0	0
PNUSAS26943	September	Rockingham	15	Heidelberg	Heidelberg	Fecal	35	4744532	235564	4440	6	0	0
PNUSAS26944	September	Rockingham	15	Heidelberg	Heidelberg	Fecal	32	4745684	270009	4440	6	0	0
PNUSAS26945	September	Coos	26	Thompson	Thompson	Fecal	49	4812295	202625	4490	8	0	0
PNUSAS26946	September	Carroll	22	Braenderup	Braenderup	Fecal	51	4684157	161515	4359	5	0	0
PNUSAS26947	September	Strafford	43	Paratyphi_B_var.L (+)tartrate+ 4,5	Paratyphi_B_var._L (+)_tartrate+	Fecal	44	4629689	258157	4298	3	0	0
PNUSAS26948	September	Rockingham	22	Braenderup	Braenderup	Fecal	27	4764080	386169	4438	5	0	0
PNUSAS26949	October	Carroll	19	Typhimurium	Typhimurium	Fecal	49	4819575	191426	4520	8	0	0
PNUSAS26951	October	Carroll	19	Typhimurium	Typhimurium	Fecal	55	4816957	185929	4517	8	0	0
PNUSAS26952	October	Grafton	11	Enteritidis	Enteritidis	Blood	45	4758090	244934	4483	6	0	0
PNUSAS26953	October	Grafton	11	Enteritidis	Enteritidis	Fecal	40	4761866	321396	4499	4	0	0
SRR5364221	January	Hillsborough	19	Typhimurium	Typhimurium	Fecal	48	4922423	200577	4639	5	0	0
SRR5364224	December	Hillsborough	10	Dublin	Dublin	Fecal	60	4950834	178709	4758	7	0	0
SRR5364225	January	Carroll	11	Enteritidis	Enteritidis	Fecal	21	4698759	444845	4423	4	0	0
SRR5364226	December	Carroll	0	Schwarzengrund	Schwarzengrund	Fecal	27	4565114	412740	4245	6	0	0
SRR5364227	January	Hillsborough	19	Typhimurium	Typhimurium	Fecal	49	4922449	225553	4641	5	0	0
SRR5364228	January	Hillsborough	48	Panama	Panama	Fecal	25	4514971	401124	4190	8	0	0
SRR5364229	February	Merrimack	365	Weltevreden	Weltevreden	Fecal	114	5032464	107850	4708	2	0	0
SRR5382665	February	Grafton	1674	Javiana	Javiana	Fecal	25	4610090	465452	4338	6	0	0
SRR5382673	February	Merrimack	11	Enteritidis	Enteritidis	Fecal	36	4689802	246231	4390	6	0	0
SRR6019672	July	Rockingham	11	Enteritidis	Enteritidis	Fecal	43	4688753	239013	4395	8	0	0
SRR6019677	July	Rockingham	26	Thompson	Thompson	Fecal	31	4668073	326092	4316	9	0	0
SRR6019679	July	Belknap	19	Typhimurium	Typhimurium	Fecal	59	4933506	164254	4659	8	0	0

SRR6026010	August	Rockingham	24	Panama	Javiana	Fecal	49	4604657	195247	4311	6	0	0
SRR6026018	August	Grafton	448	G	Mississippi	Fecal	60	4626896	153963	4402	5	0	0
SRR6026029	August	Strafford	24	Javiana	Javiana	Fecal	32	4608177	383767	4308	6	0	0
SRR6107297	September	Merrimack	11	Enteritidis	Enteritidis	Fecal	24	4692646	478724	4400	8	0	0
SRR6107309	August	Strafford	138	Montevideo	Montevideo	Fecal	33	4540052	243482	4203	7	0	0
SRR6107320	August	Rockingham	118	Newport	Newport	Fecal	33	4754786	405936	4444	6	0	0
SRR6107328	August	Hillsborough	24	Javiana	Javiana	Fecal	32	4608643	299062	4314	6	0	0
SRR6107330	August	Rockingham	15	Heidelberg	Heidelberg	Fecal	26	4746478	381308	4433	6	0	0
SRR6107334	August	Rockingham	15	Heidelberg	Heidelberg	Fecal	29	4745667	381280	4441	6	0	0
SRR6107340	August	Rockingham	15	Heidelberg	Heidelberg	Fecal	26	4746429	693625	4436	6	0	0
SRR6107350	August	Hillsborough	2	Typhi	Typhi	Fecal	54	4724761	185186	4564	3	0	0
SRR6107354	August	Strafford	350	Newport	Newport	Fecal	32	4836204	315961	4565	3	0	0
SRR6107359	August	Rockingham	15	Heidelberg	Heidelberg	Fecal	26	4746524	381295	4438	6	0	0
SRR6107360	August	Strafford	15	Heidelberg	Heidelberg	Fecal	31	4745286	298538	4437	6	0	0
SRR6107362	August	Strafford	32	Infantis	Infantis	Fecal	32	4641427	526636	4303	9	0	0
SRR6107371	August	Belknap	11	Enteritidis	Enteritidis	Fecal	25	4693061	478730	4402	7	0	0
SRR6158700	August	Strafford	118	Newport	Newport	Fecal	44	4808218	236852	4524	8	0	0
SRR6183270	August	Grafton	19	Typhimurium	Typhimurium	Fecal	77	4892969	114190	4591	7	0	0
SRR6183271	August	Rockingham	15	Heidelberg	Heidelberg	Fecal	34	4744849	288603	4435	6	0	0
SRR6183272	September	Hillsborough	15	Heidelberg	Heidelberg	Fecal	41	4745325	203449	4434	8	0	0
SRR6183274	September	Rockingham	23	Oranienburg	Oranienburg	Bile	59	4635557	122605	4324	5	0	0
SRR6183275	August	Rockingham	11	Enteritidis	Enteritidis	Fecal	32	4807134	433312	4543	6	0	0
SRR6183276	September	Merrimack	19	Typhimurium	Typhimurium	Blood	70	4927213	149645	4661	6	0	0
SRR6183302	September	Hillsborough	1628	Reading	Reading	Fecal	78	4871331	137427	4569	6	0	0
SRR6183303	September	Hillsborough	11	Enteritidis	Enteritidis	Fecal	30	4862126	479109	4584	10	0	0
SRR6183310	September	Grafton	19	Typhimurium	Typhimurium	Fecal	126	4922296	71043	4649	8	0	0
SRR6183315	September	Rockingham	350	Newport	Newport	Fecal	50	4823015	162766	4548	2	0	0
SRR6183316	September	Hillsborough	11	Enteritidis	Enteritidis	Fecal	26	4694442	478724	4400	8	0	0
SRR6183317	September	Rockingham	23	Oranienburg	Oranienburg	Fecal	62	4635842	134559	4320	7	0	0
SRR6366419	November	Rockingham	11	Enteritidis	Enteritidis	Fecal	31	4700781	328912	4424	5	0	0

SRR6366421	October	Strafford	11	Enteritidis	Enteritidis	Urine	50	4689049	225268	4393	6	0	0
SRR6366423	October	Carroll	11	Enteritidis	Enteritidis	Urine	41	4692651	348980	4400	8	0	0
SRR6366424	November	Merrimack	19	Typhimurium	Typhimurium	Fecal	65	4830191	162378	4524	6	0	0
SRR6366425	October	Rockingham	11	Enteritidis	Enteritidis	Fecal	25	4726093	464906	4441	6	0	0
SRR6366426	October	Rockingham	11	Enteritidis	Enteritidis	Fecal	23	4726788	433312	4443	6	0	0
SRR6366428	October	Rockingham	32	Infantis	Infantis	Urine	49	4647444	194908	4297	10	0	0
SRR6366430	October	Sullivan	1674	Javiana	Javiana	Urine	55	4565190	144059	4278	7	0	0
SRR6366432	November	Strafford	26	Thompson	Thompson	Fecal	29	4667339	285126	4317	8	0	0
SRR6366439	October	Carroll	19	Typhimurium	Typhimurium	Fecal	47	4816899	210210	4513	8	0	0
SRR6371549	May	Cheshire	132	Newport	Newport	Fecal	26	4761285	472041	4466	4	0	0

CHAPTER 5

HERO: Visualizing genome-wide patterns of recombination in microbial species and populations

Cooper J. Park[a], Pekka Marttinen[b] and Cheryl P. Andam[a,c]

Article in preparation for submission to *BMC Bioinformatics*

[a] University of New Hampshire, Department of Molecular, Cellular and Biomedical Sciences, Durham, New Hampshire 03824, USA

[b] Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland

[c] University at Albany, State University of New York, Department of Biological Sciences, Albany, New York 12222, USA

ABSTRACT

Background

Homologous recombination is known to influence a myriad of evolutionary and population processes within bacteria. However, growing evidence suggests that the frequency and distribution of recombination events can be influenced by genetic and ecological barriers between strains within a species. Despite the growing number of tools available to predict recombination events, no software provides the means to characterize donor-recipient relationships and other metrics of recombination heterogeneity within a population.

Results

We present HERO, a Python tool which uses the output of the recombination detection tool fastGEAR to identify donors and recipients in recombination events. HERO also maps recombination events to user-defined metadata categories to help elucidate potential drivers of biases in recombination partners and visualizes the results in publication-ready figures using Circos networks. It also reports and visualizes the variation in recombined DNA fragment size, and events per gene as additional measures of variation.

Conclusions

HERO is a freely available Python tool for measuring and visualizing heterogeneity within a bacterial population's history of recombination. The code and documentation are available to download from https://github.com/therealcooperpark/hero. An example of using the program can be found at https://github.com/therealcooperpark/hero_example.

INTRODUCTION

Genetic recombination allows a microbial cell to rapidly acquire novel traits through incorporation of DNA fragments from other strains or species into its own genome (Didelot and Maiden 2010). It often involves the non-reciprocal unidirectional transfer of a homologous or highly similar segment of DNA from a donor to a recipient (Didelot and Maiden 2010). The consequences of genetic recombination are vast. Homologous recombination is known to influence a myriad of evolutionary and population processes, including levels of standing diversity, niche expansion, spread of resistance and virulence determinants, and rapid adaptive changes in response to new or fluctuating environmental conditions (Levin and Cornejo 2009; William P. Hanage 2016). It can generate vaccine escape variants and the rapid diversification of surface antigens, allowing immune evasion (Croucher et al. 2017). Recombination of large DNA segments can also result to the emergence of novel genetic variants or hybrids with unique phenotypes such as multidrug resistance, hyper-virulence and increased transmissibility (Gabriel G. Perron et al. 2012; Spoor et al. 2015).

Although many studies have generated crucial insights into the nature and frequencies of recombination between bacterial species (González-Torres et al. 2019; Vos and Didelot 2009; Levin and Cornejo 2009), it is often assumed that all strains recombine at a uniform frequency and randomly across the entire species. Recombination rates between strains of the same species can vary along a continuum spanning several orders of magnitude. Some strains also donate or receive DNA more often than others (Rodríguez-Beltrán et al. 2015; Wyres et al. 2019), while some strains tend to preferentially recombine with specific partners (Chewapreecha et al. 2014; Park and Andam 2020). Such a pair of strains or lineages exchanging DNA more often between

them than with others is said to be linked by a highway of recombination (or biased recombination). Highways likely represent specific lineages that function as hubs of gene flow, facilitating the rapid spread of genes associated with antibiotic resistance, host adaptation and immune interactions (Chewapreecha et al. 2014). Within-species differences in recombination also suggest that lineages respond to selective pressures in different ways. Such variation also implies that recombination itself can evolve in response to natural selection (Lobkovsky, Wolf, and Koonin 2016; Peñalba and Wolf 2020) and can occur quickly on an evolutionary timescale (Cowley et al. 2018; Evans and Rozen 2013). Hence, the idea of a single effective recombination rate for a species does not provide a biologically realistic representation of microbial evolution. Equally problematic is when studies attempt to fit the data to evolutionary and population genetic models that assume a constant species-wide rate of recombination.

Rapid recombination detection programs have been developed that can be used to identify recombined DNA fragments in large-scale whole genome datasets. ClonalOrigin generates a clonal phylogenetic tree and considers recombination events as regions of DNA that create localized discrepancies to the clonal phylogeny (Didelot et al. 2010). BratNextGen clusters regions in a genome that may be more distinct from other taxa than expected by normal mutation-driven evolution and creates a proportion of shared ancestry tree to group genomes that have a greater proportion of shared DNA clusters (Marttinen et al. 2012). Gubbins identifies recombined sequences by iteratively scanning a sequence alignment and examining for elevated densities of nucleotide substitutions, and hence is more appropriate in investigations at the subspecies level (Croucher et al. 2015). FastGEAR uses a Hidden Markov Model to compare every nucleotide site in the target sequence to all remaining lineages and asks whether it is more

similar to something else compared to other strains in the same lineage (Mostowy et al. 2017). However, these programs do not provide a means to characterize population-wide patterns of donor-recipient relationships using the predicted recombination events. Here, we introduce the program HERO (Highways Enumerated by Recombination Observations), which uses the output of fastGEAR to identify donors and recipients in recombination events. HERO also maps recombination events to user-defined metadata categories to help elucidate potential drivers of biases in recombination partners.

IMPLEMENTATION

Identifying DNA donors and recipients

HERO is a Python-implemented tool that uses the results of fastGEAR as input to infer donor-recipient pairs in recombination events. Because fastGEAR identifies putative recombination events by predicting the origin of individual nucleotide sites from allelic patterns observed in different lineages, individual fastGEAR-defined lineages are reported as the potential donor for each recombination event rather than individual genomes. Additionally, because this process iterates over individual nucleotide sites, recombination events can be any length. For each putative recombined DNA segment that fastGEAR identifies, a Bayes factor (BF) is also computed based on the density of single nucleotide polymorphisms (SNP) that is compared between within the claimed recombination event and non-recombinant regions. FastGEAR uses a significance threshold of BF = 1 for recent recombination that represents a middle ground between false positive rate and power to detect recent recombination events.

HERO considers only the results of recent recombinations inferred by fastGEAR. It first filters predicted recent recombination events by their reported length in base pairs and their Bayes factor (BF) (Bernardo, JM, Smith 2001). The filtering criteria for both BF and length can be customized by the user, but the default minimum BF and fragment length are 10 and 0, respectively. Recombination detection methods rely on changes in the density of SNPs (between the donor and recipient) between the putative recombined segment and surrounding non-recombinant genome. However, recombined DNA is often very similar in sequence to the original recipient genome, especially when the event occurs within a species (Didelot and Maiden 2010). Therefore, filtering events by their length is an arbitrary cut-off when the short recombination events predicted by fastGEAR are likely to be only the divergent piece of larger DNA fragments. In order to remain conservative regarding the number of recombination events in the population, we increased the minimum BF from fastGEAR (BF > 1) to a more strict default in HERO (BF > 10).

HERO accepts associated metadata (e.g., clusters delineated in population structure analysis, environment, specimen source, human or animal host) for each genome in the dataset. HERO identifies a donor-recipient pair between the recipient's metadata group and the most likely donor metadata group. Because fastGEAR identifies a cluster of potential donor strains (i.e., lineage) rather than a single donor genome, HERO uses a simple distance matrix to compare the sequence similarity between the recombined DNA in the recipient to the same region from each genome in the donor lineage. Assuming the shared ecology facilitates recombination between closely related strains, the metadata group containing the genome with the highest similarity to the recipient is considered the donor group for that event.

Recombination events will be discarded from the analysis if donors from different metadata groups tie for the highest similarity to the recipient. Additionally, multiple recombination events with overlapping nucleotide ranges between the same donor-recipient metadata pair are considered to be a single recombination event.

Identifying highways of recombination

HERO identifies a highway of recombination as a pair of metadata groups with a number of recombination events greater than *3*IQR + Q3,* where *IQR* is the inter-quartile range and *Q3* is the third quartile of the distribution of recombination events per donor-recipient pair. Hence, the definition of a highway will vary based on the number of metadata groups and genomes included in the dataset being examined. Furthermore, the direction of a recombination event is considered when determining a highway, making it possible for a recombining pair to be a highway in one direction, but not the other.

Visualizing results

The primary output of HERO is a pair of network images generated using Circos (Krzywinski et al. 2009) . In the first figure "circos.svg", the fragments on the outer ring represent each metadata group involved in a recombination event (Fig. 1a). In this example, we used sequence clusters (SC) defined by a Bayesian hierarchical clustering method implemented in BAPS (Fig. 1a) (Corander et al. 2008). The length of the fragments in the outer ring (Fig. 1b) is proportional to the number of recombination events involving the group. The intertwining ribbons between groups represent donor-recipient pairs of recombination where the ribbon is colored to match the donor and the donor edge of the ribbon is indented towards the center of the

circle (Fig. 1b). The thickness of the ribbon is proportional to the number of recombination events between a pair of genomes. Because the direction of a recombination event is considered when visualizing these pairs, it is possible for two ribbons to exist between the same pair of metadata groups. There is an option to highlight highways of recombination as seen in the output "highway_circos.svg" (Fig. 1c). In addition to the circos networks, HERO generates frequency histograms showing the lengths of recombined DNA sequences, the number of recombination receipts per genome and the number of recombination events per gene (Fig. 2). HERO also provides supporting text files of the data for all figures.

RESULTS AND DISCUSSION

We next demonstrate the utility of HERO with the same collection of 616 whole-genome *Streptococcus pneumoniae* isolates sampled in Massachusetts, USA (Croucher et al. 2013) that was previously used to demonstrate the effectiveness of fastGEAR to detect recombination (Mostowy et al. 2017). The methods we used to prepare the dataset have been described in detail in Additional File 1 and Accession IDs for all genomes can be found in Additional File 2.

Exploring dynamics of recombination within Streptococcus pneumoniae

We first used Roary (Page et al. 2015) to characterize the pan-genome of the entire *S. pneumoniae* population. We identified 1,161 core genes (i.e., genes present in ≥ 99% of strains) and 6,133 shared accessory genes (i.e., genes present in at least 2 genomes, but less than 99% of the population) out of a total of 7,511 genes in the pan-genome. We identified 582 genes with evidence of recombination and 1,990 recent recombination events. We then used HERO to identify the distribution of these recombination events across the 16 SCs in which each genome

was assigned to in its original publication (Croucher et al. 2013) (Fig. 1a,b). Out of the 256 possible unidirectional pairs of SCs, 191 of them had evidence for recombination with between 1 and 191 recombination events in any one pair (mean ≈ 11 events). Using HERO's definition of a highway of recombination, we found 21 pairs that met the definition of a highway (i.e., pairs with ≥ 22 events) (Fig. 1c). Highways of recombination accounted for 1,052 of the total 1,990 (52.8%) recombination events inferred within the population.

All highways of recombination involved the only multiphyletic cluster SC16 as either a donor or recipient. Based on the phylogenetic tree for the population, SC16 is likely composed of multiple individual clusters too small to be detected independently by the BAPS clustering software (Corander et al. 2008). To improve the resolution of the analysis, we used HERO to re-calculate the distribution of events, but this time breaking SC16 into eight smaller SCs (labeled as SC16a-h) where each new SC is separated by at least one monophyletic SC (Fig. 3a,b). Using these newly assigned clusters, the number of possible unidirectional SC pairs in the population increased to 529. We found 347 of these pairs to have evidence of recombination, with between 1 and 58 events in any one pair (mean ≈ 5 events). These pairs shared a total of 2,230 recombination events across 558 different genes. In this new clustering scheme, the threshold for a highway of recombination decreased to 17 events per pair, yet only 18 pairs (5% of all recombining pairs) were identified as highways (Fig. 3c). These highways accounted for 466 (20%) out of the 2,230 recombination events. While 12 of these highways involved SC16c as either a donor or recipient, the remaining six highways were scattered between pairs involving SC16b, SC12, and SC6.

The number of recombination events per genome varied, with between 8 and 50 events in a single genome (mean ≈ 24 events) (Fig. 3a). The detected fragment size of recombination events varied from 1 - 4,447bp (mean ≈ 309bp) (Fig. 3b). Lastly, we detected variation in the number of recombination events per gene with between 1 and 33 events per gene (mean ≈ 4 events) (Fig. 3c).

Characterizing the properties of a recombination pair

Intra-species variation in recombination has been found to exist within multiple bacterial species and across broad ecological settings (Chewapreecha et al. 2014; Sheppard et al. 2014; Park and Andam 2020). However, the extent to which genetic and ecological factors drive this variation remains poorly understood. By combining results generated by HERO with other common measures of population diversity we sought to identify trends within the *S. pneumoniae* population that could be extrapolated to other species and populations.

One of the most significant challenges to predicting recombination pairs is the effect of sampling bias on donor identification. Under-sampling a population risks missing genomes with unique gene repertoires that could potentially be the source of a recombination event. In contrast, having one or a few well-sampled subpopulations may exaggerate the credit these larger groups get as a donor by being the dominant source of variation that suspected recombination events are compared against. To test the effect of sampling in our population we first compared the number of genomes in each cluster to the number of recombination events involving the cluster (Fig. 4a) and found a significant but weak positive correlation between the two (p-value < 0.05, R^2 = 0.14). We also compared the number of shared genes within a cluster to its number of

recombination events (Fig. 4b) and found a significant positive correlation (p-value < 0.001, R^2 = 0.72). Lastly, we calculated the Average Nucleotide Identity (ANI) for each SC using fastANI v1.0 (Jain et al. 2018). ANI estimates the average nucleotide identity of all orthologous genes shared between any two genomes, thus being analogous to a measure of their core-genome similarity (Jain et al. 2018). We calculated the SC-wide ANI, which refers to the mean of all possible pairwise ANI values between any two pairs in the SC. For each SC, we compared the number of recombination events with its SC-wide ANI (Fig. 4c) and found a statistically significant negative correlation (p-value < 0.001, R^2 = 0.47).

Collectively, these results indicate poor resolution of recombination events between closely related strains. While the size of a cluster can influence the amount of diversity within it, it is not clear that sample size alone is significantly influencing the distribution of predicted recombination events among the sequence clusters. Therefore, the primary limitations to HERO stem from the assigned metadata groups. Multiphyletic clades, such as SC16 from this *S. pneumoniae* population, are likely to distort findings from clusters derived from sequence data as the cumulative genetic diversity from multiple clades will contribute many more opportunities to find recombination than from within a single monophyletic clade. However, if multiphyletic clades are expected (e.g., in ecologically derived clusters), sufficient representation for each cluster will be crucial to accurately attributing recombination events to a donor cluster. Predicting whether a sampled population reflects the total genomic diversity of its natural population remains a challenging aspect of bacterial population genetics (L. M. Bobay and Ochman 2018). However, future advancements in metagenomic sequencing and genome assembly from metagenomic data are likely to improve the resolution of these analyses.

CONCLUSION

In summary, we present HERO, a user-friendly python program that uses the output from the popular recombination detection tool fastGEAR to identify and donor-recipient pairs in recombination events. We propose a definition of a “highway of recombination” which can capture unique trends in recombination frequencies within a population while maintaining flexibility across populations with different frequencies of recombination.. The simplicity of HERO’s usage combined with its informative visualizations provide a detailed look into a population’s history of recombination.

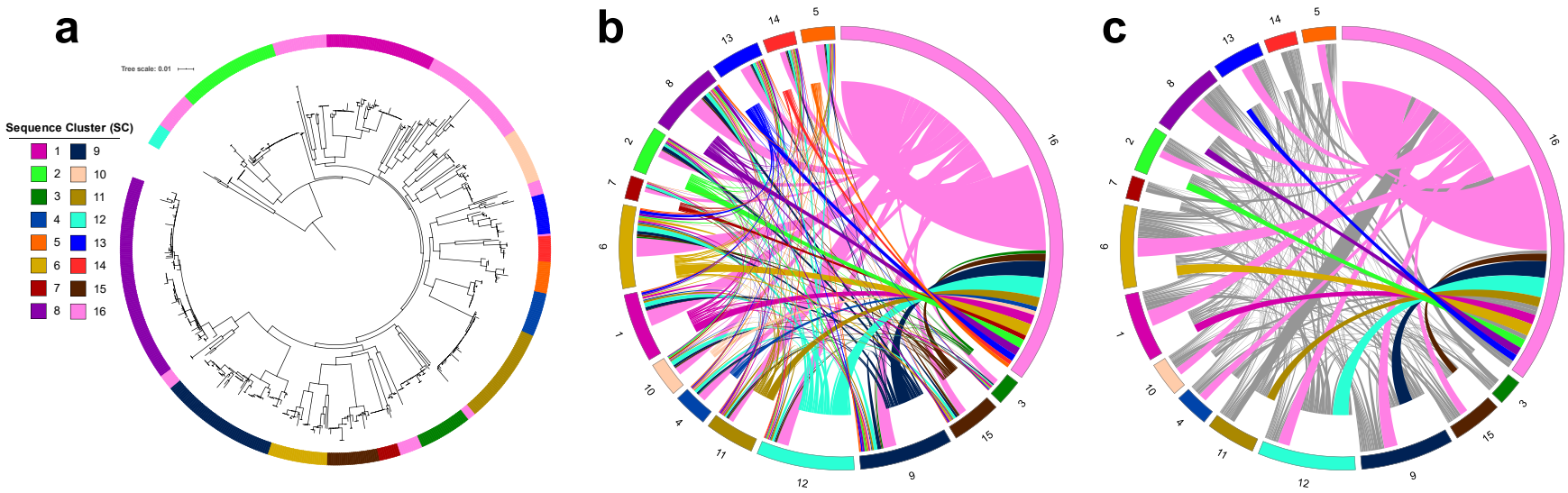

Figure 1. HERO recombination pairs compared to sequence cluster positions in a phylogeny. a) Core genome phylogeny of the *S. pneumoniae* population. The phylogeny was reconstructed using the concatenated alignment of 1,161 core genes. The scale bar represents substitutions per site. b) A recombination network generated by HERO. Outer ring fragments are individual BAPS-derived SCs. Length of each fragment is proportional to the number of recombination events affecting the SC. Ribbons connect clusters that share recombination events where the thickness of the ribbon is proportional to the number of shared events, the color of the ribbon matches the color of the donor cluster, and the donor edge of each ribbon is indented towards the middle of the circle. c) A recombination network (identical to panel b) highlighting the highways of recombination and non-highways are colored gray.

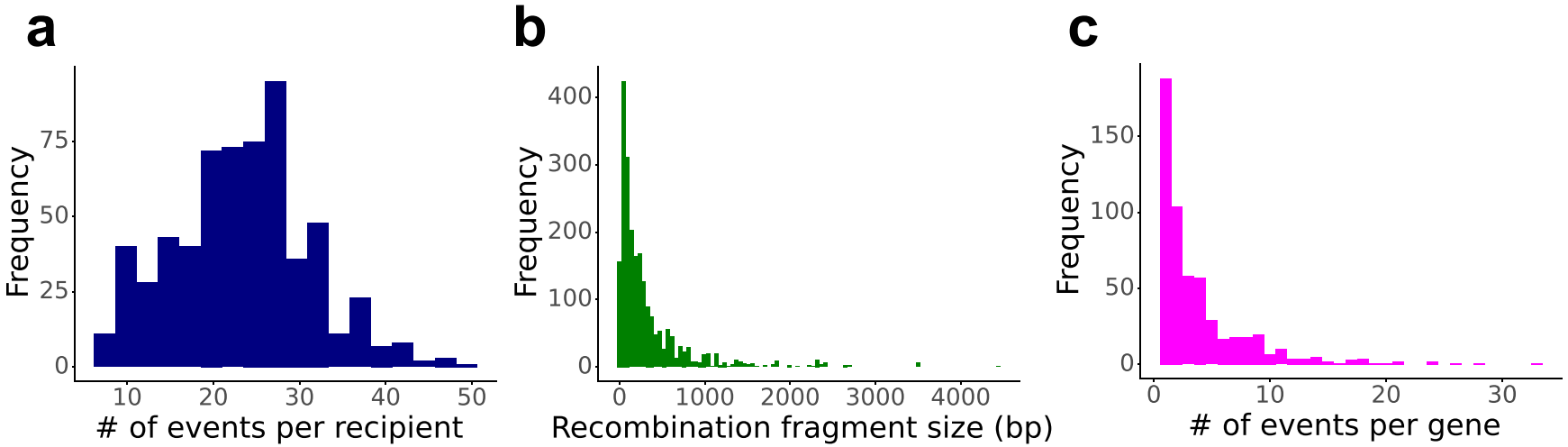

Figure 2. Measures of variability in recombination. a) Histogram showing the frequency distribution of events per recipient genome. b) Histogram showing the frequency distribution of recombination fragment size (bp). c) Histogram showing the frequency distribution of events per gene.

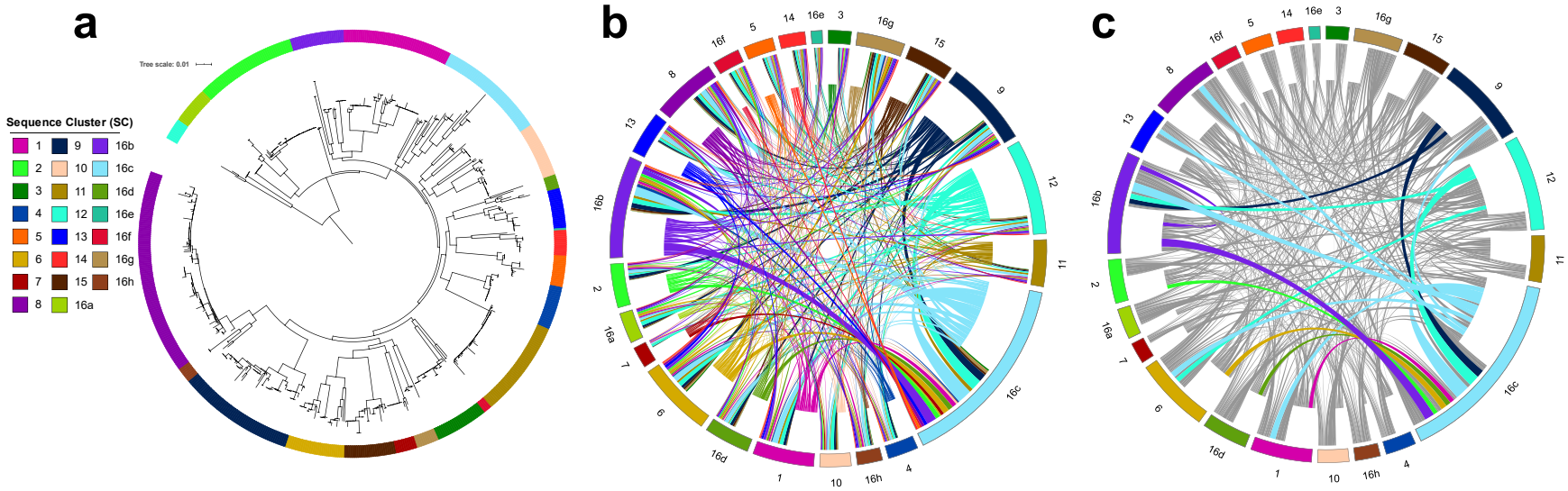

Figure 3. HERO recombination pairs compared to sequence cluster positions in a phylogeny with SC16 split into smaller clusters. a) Core genome phylogeny of the *S. pneumoniae* population. The phylogeny was reconstructed using the concatenated alignment of 1,161 core genes. The scale bar represents substitutions per site. b) A recombination network generated by HERO. Outer ring fragments are individual sequence clusters. Length of each fragment is proportional to number of recombination events affecting the cluster. Ribbons connect clusters that share recombination events where the thickness of the ribbon is proportional to the number of shared events, the color of the ribbon matches the color of the donor cluster, and the donor edge of each ribbon is indented towards the middle of the circle. c) A recombination network generated by HERO. Identical to figure 2, except only ribbons connecting highways of recombination are colored.

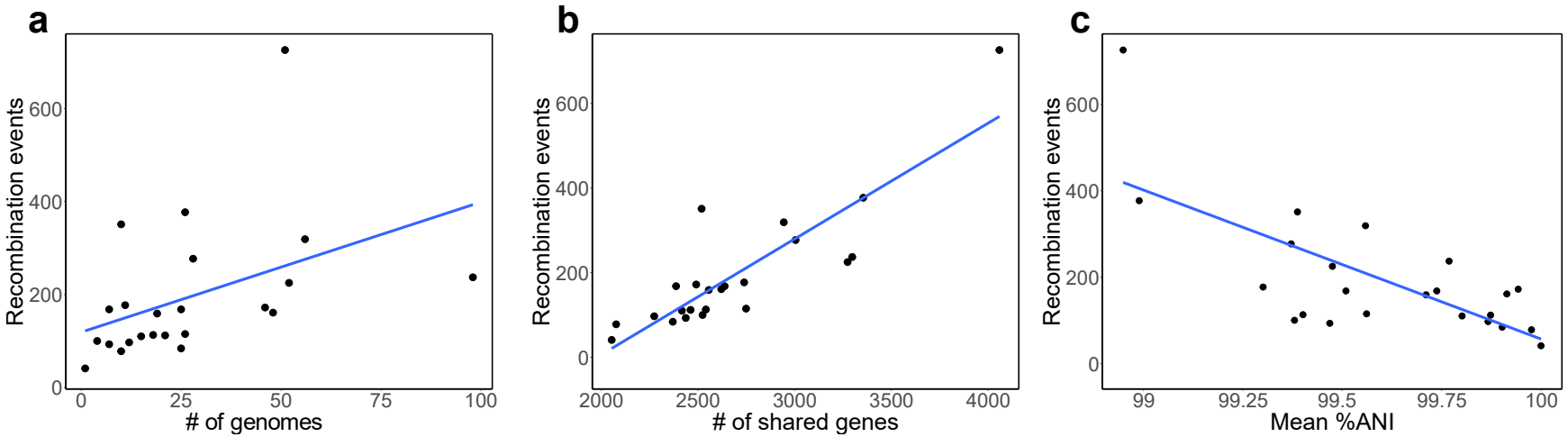

Figure 4. Characteristics of recombination pairs. a) Relationship between the number of genomes in a cluster and its number of predicted recombination events. b) Relationship between the number of shared genes in a cluster and its number of predicted recombination events. c) Relationship between SC-wide ANI and number of recombination events per SC.

DATA AVAILABILITY

Methods for preparing the dataset used here can be found in Additional File 1. Accession IDs for *S. pneumoniae* genomes can be found in Additional File 2. Additionally, a walkthrough of the sample dataset analysis including intermediate files for each step can be found at https://github.com/therealcooperpark/hero_example

AVAILABILITY AND REQUIREMENTS

Project name: HERO

Project home page: https://github.com/therealcooperpark/hero

Operating System(s): Linux

Programming language: Python 3.6

Other requirements: BioPython (Python), Pandas (Python), Plotnine (Python), fastGEAR, Circos, GNU Parallel

License: MIT

Any restrictions to use by non-academics: None

ACKNOWLEDGEMENTS

The authors thank the University of New Hampshire Resource Computing Center where all bioinformatics analyses were performed. The authors thank Anthony Westbrook for providing technical and bioinformatics assistance. We also thank Dr. Claire Chewapreecha for discussions on software implementation.

APPENDIX 5

Additional File 1 – Details about methods of *S. pneumoniae* analysis.

De novo genome assembly, annotation, pangenome, and phylogenetic analysis

Each of the 616 *S. pneumoniae* genomes was independently assembled using Spades v3.13.1 (Bankevich et al. 2012) with default parameters. Assembled genomes were annotated with Prokka v1.14.0 (T. Seemann 2014) and default parameters. We then used Roary v3.12.0 (Page et al. 2015) to characterize the pangenome of the population, including the '-z' parameter to generate alignments for each gene in the pangenome. Each gene alignment was aligned using MAFFT v7.407 (Kazutaka Katoh, Rozewicki, and Yamada 2017). A core genome phylogenetic tree was generated using the aligned concatenation of all core genes from the pangenome and the tool RAxML v8.2.11 (Stamatakis 2006) with a general time reversible (GTR) nucleotide substitution model (Tavaré 1986), four gamma categories for rate heterogeneity, and 100 bootstrap replicates. Phylogenies were visualized using the Interactive Tree of Life (Letunic and Bork 2016).

Detection of recombination

To identify recombination, we used fastGEAR (Mostowy et al. 2017) with default parameters on individual core and shared accessory genes identified by Roary. Prior to running fastGEAR, the protein specific headers in each FASTA gene alignment were replaced with a genome name using a custom script to make fastGEAR results comparable between genes. The custom script has been provided with HERO on its GitHub page (https://github.com/therealcooperpark/hero) as "sidekick.py" for reproducibility and convenience when using HERO in similar workflows.

Supplementary Table 1 – Accession IDs and metadata for 616 *S. pneumoniae* genomes used in analysis.

Accession	Strain_Cluster_(SC)
ERR069731	1
ERR069809	1
ERR129088	1
ERR129126	1
ERR129158	1
ERR129164	1
ERR129199	1
ERR129201	1
ERR124256	1
ERR124285	1
ERR065347	1
ERR065350	1
ERR068012	1
ERR068048	1
ERR067981	1
ERR069683	1
ERR069768	1
ERR065968	1
ERR129051	1
ERR129058	1
ERR129090	1
ERR129113	1
ERR129132	1
ERR124237	1
ERR124265	1
ERR124282	1
ERR124304	1
ERR068013	1
ERR068018	1
ERR068020	1
ERR067977	1
ERR069690	1
ERR069712	1
ERR069715	1
ERR069732	1
ERR069740	1
ERR069746	1
ERR069755	1
ERR069765	1
ERR069770	1
ERR069812	1
ERR069823	1
ERR069835	1
ERR065962	1
ERR065967	1
ERR124231	1
ERR065344	1
ERR065292	1
ERR065330	1
ERR067964	1
ERR069719	1
ERR069721	1
ERR069752	2
ERR069760	2
ERR069801	2
ERR069804	2
ERR069822	2
ERR069837	2
ERR069839	2
ERR065964	2
ERR129037	2
ERR129068	2
ERR129079	2
ERR129080	2
ERR129093	2
ERR129127	2
ERR129131	2
ERR129137	2
ERR129139	2
ERR129154	2
ERR129159	2
ERR129177	2
ERR129178	2
ERR129210	2
ERR129211	2
ERR129214	2
ERR129215	2
ERR124240	2
ERR124242	2
ERR124246	2
ERR124268	2
ERR124291	2
ERR124296	2
ERR124298	2
ERR124302	2
ERR065308	2
ERR065310	2
ERR065326	2
ERR067985	2
ERR067986	2
ERR068000	2
ERR068040	2
ERR068041	2
ERR068046	2
ERR068047	2
ERR067968	2
ERR069702	2
ERR069707	2
ERR069713	2
ERR069727	2
ERR129026	3
ERR129054	3
ERR129060	3
ERR129061	3
ERR129198	3
ERR124239	3
ERR124249	3
ERR124272	3
ERR124274	3
ERR124283	3
ERR124300	3
ERR065293	3
ERR065297	3
ERR065320	3
ERR065332	3
ERR068026	3
ERR068028	3
ERR068032	3
ERR068042	3
ERR068049	3
ERR067978	3
ERR069724	3
ERR069725	3
ERR124221	3
ERR069698	3
ERR069738	4
ERR069767	4
ERR069778	4
ERR069783	4
ERR069836	4
ERR129096	4
ERR129156	4
ERR129204	4
ERR124248	4
ERR124260	4
ERR124307	4
ERR124318	4
ERR065338	4
ERR065343	4
ERR067987	4
ERR067995	4
ERR067996	4
ERR068037	4
ERR067965	4
ERR069686	4
ERR069688	4
ERR129053	5

ERR129144	5
ERR129182	5
ERR129196	5
ERR068010	5
ERR129105	5
ERR129195	5
ERR067998	5
ERR069714	5
ERR069799	5
ERR069818	5
ERR124320	5
ERR065342	5
ERR129077	5
ERR065301	5
ERR124286	6
ERR124288	6
ERR124294	6
ERR124297	6
ERR124319	6
ERR129042	6
ERR129117	6
ERR129157	6
ERR129167	6
ERR129176	6
ERR129213	6
ERR124235	6
ERR124308	6
ERR065303	6
ERR068023	6
ERR069708	6
ERR129043	6
ERR124227	6
ERR068030	6
ERR069779	6
ERR069786	6
ERR069798	6
ERR065969	6
ERR065971	6
ERR129172	6
ERR069685	6

ERR069693	6
ERR129168	6
ERR129067	7
ERR129073	7
ERR124224	7
ERR124228	7
ERR124243	7
ERR124276	7
ERR124292	7
ERR124312	7
ERR067961	7
ERR069718	7
ERR069751	8
ERR069761	8
ERR069774	8
ERR069795	8
ERR069828	8
ERR069829	8
ERR069833	8
ERR065970	8
ERR065972	8
ERR065975	8
ERR129032	8
ERR129033	8
ERR129034	8
ERR129046	8
ERR129081	8
ERR129091	8
ERR129121	8
ERR129125	8
ERR129129	8
ERR129149	8
ERR129174	8
ERR129175	8
ERR129186	8
ERR129190	8
ERR124225	8
ERR124229	8
ERR124263	8
ERR124267	8

ERR124313	8
ERR065337	8
ERR065340	8
ERR065287	8
ERR065288	8
ERR065291	8
ERR068034	8
ERR068038	8
ERR068044	8
ERR068045	8
ERR068050	8
ERR067960	8
ERR067974	8
ERR069687	8
ERR069691	8
ERR069704	8
ERR069705	8
ERR069711	8
ERR129207	8
ERR124220	8
ERR124251	8
ERR065319	8
ERR069745	8
ERR069753	8
ERR069771	8
ERR069776	8
ERR069781	8
ERR069794	8
ERR069796	8
ERR069805	8
ERR069817	8
ERR069831	8
ERR129025	8
ERR129059	8
ERR129063	8
ERR129072	8
ERR129108	8
ERR129112	8
ERR129118	8
ERR129120	8

ERR129123	8
ERR129140	8
ERR129142	8
ERR129153	8
ERR129160	8
ERR129170	8
ERR129183	8
ERR129197	8
ERR129209	8
ERR124234	8
ERR124269	8
ERR124270	8
ERR124280	8
ERR124281	8
ERR124284	8
ERR124295	8
ERR124315	8
ERR065305	8
ERR065317	8
ERR065323	8
ERR065328	8
ERR067984	8
ERR068008	8
ERR068011	8
ERR068017	8
ERR068024	8
ERR069684	8
ERR069696	8
ERR069717	8
ERR069758	8
ERR069807	9
ERR069780	9
ERR069826	9
ERR129055	9
ERR129065	9
ERR129076	9
ERR129082	9
ERR129098	9
ERR129106	9
ERR129116	9

ERR129133	9
ERR129141	9
ERR129146	9
ERR129205	9
ERR124287	9
ERR065306	9
ERR065307	9
ERR065311	9
ERR065313	9
ERR065322	9
ERR068027	9
ERR067973	9
ERR067976	9
ERR069802	9
ERR129029	9
ERR129052	9
ERR129100	9
ERR129107	9
ERR129122	9
ERR129147	9
ERR129155	9
ERR129161	9
ERR129171	9
ERR129192	9
ERR129208	9
ERR124219	9
ERR124238	9
ERR124245	9
ERR065345	9
ERR065346	9
ERR065348	9
ERR068015	9
ERR067967	9
ERR069694	9
ERR069736	9
ERR069743	9
ERR069777	9
ERR069790	9
ERR069800	9
ERR069830	9

ERR065955	9
ERR065966	9
ERR065339	9
ERR065355	9
ERR065331	9
ERR067980	9
ERR069757	10
ERR069766	10
ERR069811	10
ERR129039	10
ERR129048	10
ERR129111	10
ERR129145	10
ERR129152	10
ERR129169	10
ERR129203	10
ERR124217	10
ERR124232	10
ERR124253	10
ERR124254	10
ERR124264	10
ERR124266	10
ERR124305	10
ERR124310	10
ERR124317	10
ERR065298	10
ERR065316	10
ERR065327	10
ERR067994	10
ERR067997	10
ERR068025	10
ERR065309	10
ERR069733	11
ERR069749	11
ERR069834	11
ERR129035	11
ERR129074	11
ERR129086	11
ERR129087	11
ERR129089	11

ERR129099	11
ERR129104	11
ERR129119	11
ERR129134	11
ERR129143	11
ERR129148	11
ERR129163	11
ERR129173	11
ERR129179	11
ERR129200	11
ERR124236	11
ERR124241	11
ERR124250	11
ERR124273	11
ERR124275	11
ERR124277	11
ERR124279	11
ERR124290	11
ERR124293	11
ERR124299	11
ERR124301	11
ERR124316	11
ERR065351	11
ERR065300	11
ERR065324	11
ERR065329	11
ERR067988	11
ERR067992	11
ERR068002	11
ERR068019	11
ERR068029	11
ERR068033	11
ERR067962	11
ERR067963	11
ERR067970	11
ERR069703	11
ERR069726	11
ERR069728	11
ERR069803	12
ERR069841	12

ERR065965	12
ERR065974	12
ERR129130	12
ERR124278	12
ERR065341	12
ERR065289	12
ERR065296	12
ERR067993	12
ERR069734	13
ERR069741	13
ERR069759	13
ERR069784	13
ERR069788	13
ERR069827	13
ERR065953	13
ERR129092	13
ERR065353	13
ERR068001	13
ERR068004	13
ERR068043	13
ERR067979	13
ERR069709	13
ERR069710	13
ERR068005	13
ERR129027	13
ERR129115	13
ERR124223	13
ERR069744	14
ERR069747	14
ERR069763	14
ERR069787	14
ERR069825	14
ERR065963	14
ERR129128	14
ERR124303	14
ERR124314	14
ERR067966	14
ERR067969	14
ERR069700	14
ERR129031	15

ERR129036	15
ERR129049	15
ERR129075	15
ERR129193	15
ERR124222	15
ERR124233	15
ERR124255	15
ERR069720	15
ERR069762	15
ERR069791	15
ERR069793	15
ERR069842	15
ERR069843	15
ERR065973	15
ERR129062	15
ERR129078	15
ERR065314	15
ERR067990	15
ERR067999	15
ERR068003	15
ERR068035	15
ERR068036	15
ERR069699	15
ERR069706	15
ERR069748	16A
ERR069750	16A
ERR069832	16A
ERR129180	16A
ERR129212	16A
ERR067983	16A
ERR067989	16A
ERR068006	16A
ERR069689	16A
ERR069701	16A
ERR069716	16A
ERR069772	16B
ERR069806	16B
ERR069816	16B

ERR069824	16B
ERR065956	16B
ERR069729	16B
ERR069730	16B
ERR129110	16C
ERR129057	16B
ERR129135	16B
ERR124311	16B
ERR065290	16B
ERR065959	16G
ERR129038	16G
ERR129066	16G
ERR065304	16G
ERR068016	16G
ERR067975	16E
ERR069815	16B
ERR129151	16B
ERR069775	16C
ERR129040	16C
ERR129056	16C
ERR129124	16C
ERR129188	16C
ERR065302	16C
ERR065321	16C
ERR068039	16C
ERR069808	16C
ERR069820	16C
ERR124244	16G
ERR129041	16C
ERR129181	16C
ERR129184	16C
ERR124247	16C
ERR124271	16C
ERR124289	16C
ERR124309	16C
ERR065318	16C
ERR124218	16C
ERR124252	16H

ERR124258	16C
ERR124306	16C
ERR065312	16C
ERR069739	16B
ERR069814	16A
ERR065315	16B
ERR068007	16A
ERR069695	16B
ERR129028	16B
ERR129030	16B
ERR129083	16B
ERR129084	16C
ERR129103	16B
ERR129136	16B
ERR129138	16B
ERR129202	16B
ERR129216	16B
ERR065294	16G
ERR067982	16H
ERR069737	16C
ERR069792	16C
ERR069810	16C
ERR065960	16C
ERR129045	16C
ERR129050	16C
ERR129094	16C
ERR129095	16C
ERR129206	16C
ERR065354	16C
ERR065299	16C
ERR065325	16C
ERR068014	16C
ERR068031	16C
ERR069692	16C
ERR069697	16C
ERR069723	16C
ERR069742	16G
ERR069785	16H

ERR069797	16H
ERR069821	16H
ERR069773	16C
ERR129185	16A
ERR069789	16A
ERR124257	16A
ERR068021	16A
ERR067972	16A
ERR065349	16C
ERR068022	16C
ERR069722	16C
ERR069764	16C
ERR065352	16G
ERR069735	16D
ERR069782	16D
ERR069813	16D
ERR069819	16D
ERR069838	16D
ERR069840	16F
ERR065957	16D
ERR065295	16D
ERR129044	16C
ERR129071	16B
ERR129194	16C
ERR124230	16G
ERR067991	16G
ERR129047	16C
ERR129187	16C
ERR129064	16F
ERR069769	16F
ERR065958	16H
ERR129162	16H
ERR124226	16F
ERR129102	16B
ERR068009	16C

CONCLUSIONS

Recombination and the pangenome as a reservoir for rapid ecological adaptation

The evolution of adaptive traits (*e.g.,* antibiotic resistance or virulence) is likely to happen more quickly when those traits are already present in the population instead of originating *de novo* (Andam et al. 2017). Therefore, an understanding of the baseline genetic diversity and the potential for genetic exchange within a population can be informative to public health endeavors. For example, the large pangenome and biosynthetic gene cluster diversity found in *Streptomyces rimosus* (Chapter 1) demonstrates the importance of drug discovery efforts testing multiple strains within the same species, as no individual genome is likely to be representative of the species' biochemical potential. Similarly, when searching for the origin of a new outbreak strain it will be imperative to understand the genomic variation of the strain's entire population. Species such as *Salmonella enterica* may utilize recombination as a metaphorical fishing rod to sample new traits from a pool of genetic diversity that includes individuals from different ecological backgrounds (Chapter 2). In fact, public health experts can preemptively sample the standing genomic diversity within their communities to identify potential mechanisms of infection that may be acquired by future outbreak isolates, as demonstrated in my analyses of *Campylobacter jejuni* and *Salmonella enterica* in New Hampshire (Chapter 3 & 4).

Barriers to recombination create bias in rates and patterns within a species

Despite the incredible diversity present within a species' pangenome, it is likely that not all of it is readily available to any one strain regardless of its affinity for recombined DNA.

Ecological and genetic barriers to recombination create implicit bias in the frequency and distribution of recombination events within a species (Andam and Gogarten 2011; Sheppard et al. 2014). However, lineages with the fewest barriers to recombination can act as "highways" of recombination that function as hubs of gene flow for adaptive alleles (Beiko, Harlow, and Ragan 2005; Chewapreecha et al. 2014). Understanding the specifics of any barrier to recombination and identifying extant highways of recombination will be crucial to refining our forecasts of public health. In my dissertation I have contributed a series of studies on several different pathogen and non-pathogen species which demonstrate a methodology for identifying specific events and overall biases of recombination in samples of global populations (Chapter 1 & 2). Additionally, I further demonstrate that similar studies can be conducted at the state level to assess potential biases in local populations (Chapter 3 & 4). These methods are anticipated to be especially beneficial as sampling strategies continue long-term and analyses can be conducted on regular intervals to elucidate changing dynamics in the population. Additionally, future work in identifying recombination highways will benefit tremendously from robust definitions that can be used repeatedly across different studies, species and populations. I attempt to establish one such definition and provide a convenient option for future researchers to use it with the implementation of HERO (Chapter 5). HERO also serves as a novel method to rapidly visualize trends of recombination within a population through its network graphics.

As the prevalence of whole genome sequencing continues to grow in both research and healthcare facilities, countless individual genomes will be sequenced from new environments including new countries, hosts, and environments. These studies will contribute vital knowledge to a broad range of public health challenges including the growth of antibiotic resistance,

potential pathogen reservoirs in the environment, changing transmission routes, discovery of novel therapeutics and host susceptibility to disease.

REFERENCES

Abi Khattar, Z., A. Rejasse, D. Destoumieux-Garzón, J. M. Escoubas, V. Sanchis, D. Lereclus, A. Givaudan, M. Kallassy, C. Nielsen-Leroux, and S. Gaudriault. 2009. "The Dlt Operon of Bacillus Cereus Is Required for Resistance to Cationic Antimicrobial Peptides and for Virulence in Insects." *Journal of Bacteriology* 191 (22): 7063–73. https://doi.org/10.1128/JB.00892-09.

Alikhan, Nabil Fareed, Zhemin Zhou, Martin J. Sergeant, and Mark Achtman. 2018. "A Genomic Overview of the Population Structure of Salmonella." *PLoS Genetics*. Public Library of Science. https://doi.org/10.1371/journal.pgen.1007261.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. http://www.ncbi.nlm.nih.gov/pubmed/2231712.

Andam, Cheryl P., Mallory J. Choudoir, Anh Vinh Nguyen, Han Sol Park, and Daniel H. Buckley. 2016. "Contributions of Ancestral Inter-Species Recombination to the Genetic Diversity of Extant Streptomyces Lineages." *ISME Journal* 10 (7): 1731–41. https://doi.org/10.1038/ismej.2015.230.

Andam, Cheryl P., and J. Peter Gogarten. 2011. "Biased Gene Transfer in Microbial Evolution." *Nature Reviews Microbiology* 9 (7): 543–55. https://doi.org/10.1038/nrmicro2593.

Andam, Cheryl P., Patrick K. Mitchell, Alanna Callendrello, Qiuzhi Chang, Jukka Corander, Chrispin Chaguza, Lesley McGee, Bernard W. Beall, and William P. Hanage. 2017. "Genomic Epidemiology of Penicillin- Nonsusceptible Pneumococci with Nonvaccine Serotypes Causing Invasive Disease in the United States." *Journal of Clinical Microbiology* 55 (4): 1104–15. https://doi.org/10.1128/JCM.02453-16.

Andino, A., and I. Hanning. 2015. "Salmonella Enterica: Survival, Colonization, and Virulence Differences among Serovars." *Scientific World Journal*. Hindawi Publishing Corporation. https://doi.org/10.1155/2015/520179.

Andreani, Nadia Andrea, Elze Hesse, and Michiel Vos. 2017. "Prokaryote Genome Fluidity Is Dependent on Effective Population Size." *ISME Journal* 11 (7): 1719–21. https://doi.org/10.1038/ismej.2017.36.

Antony-Babu, Sanjay, Didier Stien, Véronique Eparvier, Delphine Parrot, Sophie Tomasi, and Marcelino T. Suzuki. 2017. "Multiple Streptomyces Species with Distinct Secondary Metabolomes Have Identical 16S RRNA Gene Sequences." *Scientific Reports* 7 (1): 11089. https://doi.org/10.1038/s41598-017-11363-1.

Antony, Linto, Melissa Behr, Donald Sockett, Dale Miskimins, Nicole Aulik, Jane Christopher-Hennings, Eric Nelson, Marc W. Allard, and Joy Scaria. 2018. "Genome Divergence and Increased Virulence of Outbreak Associated Salmonella Enterica Subspecies Enterica Serovar Heidelberg." *Gut Pathogens* 10 (1): 53. https://doi.org/10.1186/s13099-018-0279-0.

Arber, Werner. 2000. “Genetic Variation: Molecular Mechanisms and Impact on Microbial Evolution.” *FEMS Microbiology Reviews* 24 (1): 1–7. https://doi.org/10.1111/j.1574-6976.2000.tb00529.x.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. *Gene Ontology: Tool for the Unification of Biology*. *Nature Genetics*. Vol. 25. https://doi.org/10.1038/75556.

Baig, Abiyad, Alan McNally, Steven Dunn, Konrad H. Paszkiewicz, Jukka Corander, and Georgina Manning. 2015. “Genetic Import and Phenotype Specific Alleles Associated with Hyper-Invasion in Campylobacter Jejuni.” *BMC Genomics* 16 (1): 852. https://doi.org/10.1186/s12864-015-2087-y.

Balasubramanian, Ruchita, Justin Im, Jung-Seok Lee, Hyon Jin Jeon, Ondari D. Mogeni, Jerome H. Kim, Raphaël Rakotozandrindrainy, Stephen Baker, and Florian Marks. 2019. “The Global Burden and Epidemiology of Invasive Non-Typhoidal *Salmonella* Infections.” *Human Vaccines & Immunotherapeutics* 15 (6): 1421–26. https://doi.org/10.1080/21645515.2018.1504717.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Bansal, Mukul S., Guy Banay, Timothy J. Harlow, J. Peter Gogarten, and Ron Shamir. 2013. “Systematic Inference of Highways of Horizontal Gene Transfer in Prokaryotes.” *Bioinformatics* 29 (5): 571–79. https://doi.org/10.1093/bioinformatics/btt021.

Baraúna, Rafael A., Rommel T.J. Ramos, Adonney A.O. Veras, Kenny C. Pinheiro, Leandro J. Benevides, Marcus V.C. Viana, Luís C. Guimarães, et al. 2017. “Assessing the Genotypic Differences between Strains of Corynebacterium Pseudotuberculosis Biovar Equi through Comparative Genomics.” *PLoS ONE* 12 (1). https://doi.org/10.1371/journal.pone.0170676.

Barka, Essaid Ait, Parul Vatsa, Lisa Sanchez, Nathalie Gaveau-Vaillant, Cedric Jacquard, Hans-Peter Klenk, Christophe Clément, et al. 2016. “Taxonomy, Physiology, and Natural Products of Actinobacteria.” *Microbiology and Molecular Biology Reviews* 80 (1): 1–43. https://doi.org/10.1128/mmbr.00019-15.

Bearson, Bradley L., Shawn M. D. Bearson, Torey Looft, Guohong Cai, and Daniel C. Shippy. 2017. “Characterization of a Multidrug-Resistant Salmonella Enterica Serovar Heidelberg Outbreak Strain in Commercial Turkeys: Colonization, Transmission, and Host Transcriptional Response.” *Frontiers in Veterinary Science* 4 (SEP): 25. https://doi.org/10.3389/fvets.2017.00156.

Beiko, Robert G., Timothy J. Harlow, and Mark A. Ragan. 2005. “Highways of Gene Sharing in Prokaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (40): 14332–37. https://doi.org/10.1073/pnas.0504068102.

Bentley, S. D., K. F. Chater, A. M. Cerdeño-Tárraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, et al. 2002. “Complete Genome Sequence of the Model Actinomycete

Streptomyces Coelicolor A3(2)." *Nature* 417 (6885): 141–47. https://doi.org/10.1038/417141a.

Bernardo, JM, Smith, AF. 2001. "Bayesian Theory | Wiley." IOP Publishing. 2001. https://www.wiley.com/en-us/Bayesian+Theory-p-9780471494645.

Bertelli, Claire, Matthew R. Laird, Kelly P. Williams, Britney Y. Lau, Gemma Hoad, Geoffrey L. Winsor, and Fiona S.L. Brinkman. 2017. "IslandViewer 4: Expanded Prediction of Genomic Islands for Larger-Scale Datasets." *Nucleic Acids Research* 45 (W1): W30–35. https://doi.org/10.1093/nar/gkx343.

Bezine, Elisabeth, Julien Vignard, and Gladys Mirey. 2014. "The Cytolethal Distending Toxin Effects on Mammalian Cells: A DNA Damage Perspective." *Cells* 3 (2): 592–615. https://doi.org/10.3390/cells3020592.

Bobay, Louis-Marie Marie, and Howard Ochman. 2018. "Factors Driving Effective Population Size and Pan-Genome Evolution in Bacteria." *BMC Evolutionary Biology* 18 (1): 153. https://doi.org/10.1186/s12862-018-1272-4.

Bobay, Louis Marie, and Howard Ochman. 2018. "Factors Driving Effective Population Size and Pan-Genome Evolution in Bacteria 06 Biological Sciences 0604 Genetics." *BMC Evolutionary Biology* 18 (1). https://doi.org/10.1186/s12862-018-1272-4.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Branchu, Priscilla, Matt Bawn, and Robert A. Kingsley. 2018. "Genome Variation and Molecular Epidemiology of Salmonella Enterica Serovar Typhimurium Pathovariants." *Infection and Immunity* 86 (8). https://doi.org/10.1128/IAI.00079-18.

Brenner, F. W., R. G. Villar, F. J. Angulo, R. Tauxe, and B. Swaminathan. 2000. "Salmonella Nomenclature." *Journal of Clinical Microbiology*. American Society for Microbiology. https://doi.org/10.1128/jcm.38.7.2465-2467.2000.

Brito, Patrícia H., Bastien Chevreux, Cláudia R. Serra, Ghislain Schyns, Adriano O. Henriques, and José B. Pereira-Leal. 2018. "Genetic Competence Drives Genome Diversity in Bacillus Subtilis." *Genome Biology and Evolution* 10 (1): 108–24. https://doi.org/10.1093/gbe/evx270.

Britto, Carl D., Vanessa K. Wong, Gordan Dougan, and Andrew J. Pollard. 2018. "A Systematic Review of Antimicrobial Resistance in Salmonella Enterica Serovar Typhi, the Etiological Agent of Typhoid." Edited by Gagandeep Kang. *PLOS Neglected Tropical Diseases* 12 (10): e0006779. https://doi.org/10.1371/journal.pntd.0006779.

Brown, Eric W., Mark K. Mammel, J. Eugene LeClerc, and Thomas A. Cebula. 2003. "Limited Boundaries for Extensive Horizontal Gene Transfer among Salmonella Pathogens." *Proceedings of the National Academy of Sciences of the United States of America* 100 (26): 15676–81. https://doi.org/10.1073/pnas.2634406100.

Bruen, Trevor C., Hervé Philippe, and David Bryant. 2006. "A Simple and Robust Statistical Test for Detecting the Presence of Recombination." *Genetics* 172 (4): 2665–81.

https://doi.org/10.1534/genetics.105.048975.

Brüggemann, Holger, Anders Jensen, Seven Nazipi, Hüsnü Aslan, Rikke Louise Meyer, Anja Poehlein, Elzbieta Brzuszkiewicz, Munir A. Al-Zeer, Volker Brinkmann, and Bo Söderquist. 2018. "Pan-Genome Analysis of the Genus Finegoldia Identifies Two Distinct Clades, Strain-Specific Heterogeneity, and Putative Virulence Factors." *Scientific Reports* 8 (1): 266. https://doi.org/10.1038/s41598-017-18661-8.

Carattoli, Alessandra, Ea Zankari, Aurora Garciá-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Mløler Aarestrup, and Henrik Hasman. 2014. "In Silico Detection and Typing of Plasmids Using Plasmidfinder and Plasmid Multilocus Sequence Typing." *Antimicrobial Agents and Chemotherapy* 58 (7): 3895–3903. https://doi.org/10.1128/AAC.02412-14.

Caro-Quintero, Alejandro, Gina P. Rodriguez-Castaño, and Konstantinos T. Konstantinidis. 2009. "Genomic Insights into the Convergence and Pathogenicity Factors of Campylobacter Jejuni and Campylobacter Coli Species." *Journal of Bacteriology* 191 (18): 5824–31. https://doi.org/10.1128/JB.00519-09.

Castillo-Ramírez, Santiago, Jukka Corander, Pekka Marttinen, Mona Aldeljawi, William P. Hanage, Henrik Westh, Kit Boye, et al. 2012. "Phylogeographic Variation in Recombination Rates within a Global Clone of Methicillin-Resistant Staphylococcus Aureus." *Genome Biology* 13 (1): R126. https://doi.org/10.1186/gb-2012-13-12-r126.

Chang, Qiuzhi, Izzeldin Abuelaish, Asaf Biber, Hanaa Jaber, Alanna Callendrello, Cheryl P. Andam, Gili Regev-Yochay, and William P. Hanage. 2018. "Genomic Epidemiology of Meticillin-Resistant Staphylococcus Aureus ST22 Widespread in Communities of the Gaza Strip, 2009." *Eurosurveillance* 23 (34). https://doi.org/10.2807/1560-7917.ES.2018.23.34.1700592.

Chaplin, Andrei V., Boris A. Efimov, Vladimir V. Smeianov, Lyudmila I. Kafarskaia, Alla P. Pikina, and Andrei N. Shkoporov. 2015. "Intraspecies Genomic Diversity and Long-Term Persistence of Bifidobacterium Longum." *PLoS ONE* 10 (8). https://doi.org/10.1371/journal.pone.0135658.

Chaudhry, Vasvi, and Prabhu B. Patil. 2020. "Evolutionary Insights into Adaptation of Staphylococcus Haemolyticus to Human and Non-Human Niches." *Genomics* 112 (2): 2052–62. https://doi.org/10.1016/j.ygeno.2019.11.018.

Chen, Liang, Barun Mathema, Johann D.D. Pitout, Frank R. DeLeo, and Barry N. Kreiswirth. 2014. "Epidemic Klebsiella Pneumoniae ST258 Is a Hybrid Strain." *MBio* 5 (3). https://doi.org/10.1128/mBio.01355-14.

Cheng, Kun, Xiaoying Rong, and Ying Huang. 2016. "Widespread Interspecies Homologous Recombination Reveals Reticulate Evolution within the Genus Streptomyces." *Molecular Phylogenetics and Evolution* 102 (September): 246–54. https://doi.org/10.1016/j.ympev.2016.06.004.

Chewapreecha, Claire, Simon R. Harris, Nicholas J. Croucher, Claudia Turner, Pekka Marttinen, Lu Cheng, Alberto Pessia, et al. 2014. "Dense Genomic Sampling Identifies Highways of Pneumococcal Recombination." *Nature Genetics* 46 (3): 305–9.

https://doi.org/10.1038/ng.2895.

Chidwick, Harriet S., and Martin A. Fascione. 2020. “Mechanistic and Structural Studies into the Biosynthesis of the Bacterial Sugar Pseudaminic Acid (Pse5Ac7Ac).” *Organic and Biomolecular Chemistry*. Royal Society of Chemistry. https://doi.org/10.1039/c9ob02433f.

Chisholm, Rebecca H., Patricia T. Campbell, Yue Wu, Steven Y.C. Tong, Jodie McVernon, and Nicholas Geard. 2018. “Implications of Asymptomatic Carriers for Infectious Disease Transmission and Control.” *Royal Society Open Science* 5 (2). https://doi.org/10.1098/rsos.172341.

Chopra, Ian, and Marilyn Roberts. 2001. “Tetracycline Antibiotics: Mode of Action, Applications, Molecular Biology, and Epidemiology of Bacterial Resistance.” *Microbiology and Molecular Biology Reviews* 65 (2): 232–60. https://doi.org/10.1128/mmbr.65.2.232-260.2001.

Choudoir, Mallory J., Charles Pepe-Ranney, and Daniel H. Buckley. 2018. “Diversification of Secondary Metabolite Biosynthetic Gene Clusters Coincides with Lineage Divergence in Streptomyces.” *Antibiotics* 7 (1): 1–15. https://doi.org/10.3390/antibiotics7010012.

Chung, Hattie, Tami D. Lieberman, Sara O. Vargas, Kelly B. Flett, Alexander J. McAdam, Gregory P. Priebe, and Roy Kishony. 2017. “Global and Local Selection Acting on the Pathogen Stenotrophomonas Maltophilia in the Human Lung.” *Nature Communications* 8 (January): 14078. https://doi.org/10.1038/ncomms14078.

Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, et al. 2014. “Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters.” *Cell* 158 (2): 412–21. https://doi.org/10.1016/j.cell.2014.06.034.

Cohen, Emiliano, Maya Davidovich, Assaf Rokney, Lea Valinsky, Galia Rahav, and Ohad Gal-Mor. 2020. “Emergence of New Variants of Antibiotic Resistance Genomic Islands among Multidrug-resistant *Salmonella Enterica* in Poultry.” *Environmental Microbiology* 22 (1): 413–32. https://doi.org/10.1111/1462-2920.14858.

Corander, Jukka, Pekka Marttinen, Jukka Sirén, and Jing Tang. 2008. “Enhanced Bayesian Modelling in BAPS Software for Learning Genetic Structures of Populations.” *BMC Bioinformatics* 9 (1): 539. https://doi.org/10.1186/1471-2105-9-539.

Cowley, Lauren A., Fernanda C. Petersen, Roger Junges, Med Jimson D. Jimenez, Donald A. Morrison, and William P. Hanage. 2018. “Evolution via Recombination: Cell-to-Cell Contact Facilitates Larger Recombination Events in Streptococcus Pneumoniae.” Edited by Ivan Matic. *PLOS Genetics* 14 (6): e1007410. https://doi.org/10.1371/journal.pgen.1007410.

Criscuolo, Alexis, Sylvie Issenhuth-Jeanjean, Xavier Didelot, Kaisa Thorell, James Hale, Julian Parkhill, Nicholas R. Thomson, François Xavier Weill, Daniel Falush, and Sylvain Brisse. 2019. “The Speciation and Hybridization History of the Genus Salmonella.” *Microbial Genomics* 5 (8): 1–11. https://doi.org/10.1099/mgen.0.000284.

Croucher, Nicholas J., Joseph J. Campo, Timothy Q. Le, Xiaowu Liang, Stephen D. Bentley,

William P. Hanage, and March Lipsitch. 2017. “Diverse Evolutionary Patterns of Pneumococcal Antigens Identified by Pangenome-Wide Immunological Screening.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (3): E357–66. https://doi.org/10.1073/pnas.1613937114.

Croucher, Nicholas J., Paul G. Coupland, Abbie E. Stevenson, Alanna Callendrello, Stephen D. Bentley, and William P. Hanage. 2014. “Diversification of Bacterial Genome Content through Distinct Mechanisms over Different Timescales.” *Nature Communications* 5 (1): 1–12. https://doi.org/10.1038/ncomms6471.

Croucher, Nicholas J., Jonathan A. Finkelstein, Stephen I. Pelton, Patrick K. Mitchell, Grace M. Lee, Julian Parkhill, Stephen D. Bentley, William P. Hanage, and Marc Lipsitch. 2013. “Population Genomics of Post-Vaccine Changes in Pneumococcal Epidemiology.” *Nature Genetics* 45 (6): 656–63. https://doi.org/10.1038/ng.2625.

Croucher, Nicholas J., Andrew J. Page, Thomas R. Connor, Aidan J. Delaney, Jacqueline A. Keane, Stephen D. Bentley, Julian Parkhill, and Simon R. Harris. 2015. “Rapid Phylogenetic Analysis of Large Samples of Recombinant Bacterial Whole Genome Sequences Using Gubbins.” *Nucleic Acids Research* 43 (3): e15. https://doi.org/10.1093/nar/gku1196.

Crump, John A., Maria Sjölund-Karlsson, Melita A. Gordon, and Christopher M. Parry. 2015. “Epidemiology, Clinical Presentation, Laboratory Diagnosis, Antimicrobial Resistance, and Antimicrobial Management of Invasive Salmonella Infections.” *Clinical Microbiology Reviews*. American Society for Microbiology. https://doi.org/10.1128/CMR.00002-15.

Cruz-Morales, Pablo, Johannes Florian Kopp, Christian Martínez-Guerrero, Luis Alfonso Yáñez-Guerra, Nelly Selem-Mojica, Hilda Ramos-Aboites, Jörg Feldmann, and Francisco Barona-Gómez. 2016. “Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes.” *Genome Biology and Evolution* 8 (6): 1906–16. https://doi.org/10.1093/gbe/evw125.

Dallman, Tim, Thomas Inns, Thibaut Jombart, Philip Ashton, Nicolas Loman, Carol Chatt, Ute Messelhaeusser, et al. 2016. “Phylogenetic Structure of European Salmonella Enteritidis Outbreak Correlates with National and International Egg Distribution Network.” *Microbial Genomics* 2 (8): e000070. https://doi.org/10.1099/mgen.0.000070.

Darton, Thomas C., Christoph J. Blohmke, and Andrew J. Pollard. 2014. “Typhoid Epidemiology, Diagnostics and the Human Challenge Model.” *Current Opinion in Gastroenterology* 30 (1): 7–17. https://doi.org/10.1097/MOG.0000000000000021.

David, Sophia, Leonor Sánchez-Busó, Simon R. Harris, Pekka Marttinen, Christophe Rusniok, Carmen Buchrieser, Timothy G. Harrison, and Julian Parkhill. 2017. “Dynamics and Impact of Homologous Recombination on the Evolution of Legionella Pneumophila.” *PLoS Genetics* 13 (6): e1006855. https://doi.org/10.1371/journal.pgen.1006855.

Davies, Mark R., Sarah E. Broadbent, Simon R. Harris, Nicholas R. Thomson, and Marjan W. van der Woude. 2013. “Horizontally Acquired Glycosyltransferase Operons Drive Salmonellae Lipopolysaccharide Diversity.” *PLoS Genetics* 9 (6): e1003568. https://doi.org/10.1371/journal.pgen.1003568.

Desai, Prerak T., Steffen Porwollik, Fred Long, Pui Cheng, Aye Wollam, Sandra W. Clifton, George M. Weinstock, and Michael McClelland. 2013. "Evolutionary Genomics of Salmonella Enterica Subspecies." *MBio* 4 (2). https://doi.org/10.1128/mBio.00579-12.

Destoumieux-Garzón, Delphine, Patrick Mavingui, Gilles Boetsch, Jérôme Boissier, Frédéric Darriet, Priscilla Duboz, Clémentine Fritsch, et al. 2018. "The One Health Concept: 10 Years Old and a Long Road Ahead." *Frontiers in Veterinary Science*. Frontiers Media S.A. https://doi.org/10.3389/fvets.2018.00014.

Didelot, Xavier, Rory Bowden, Teresa Street, Tanya Golubchik, Chris Spencer, Gil McVean, Vartul Sangal, et al. 2011. "Recombination and Population Structure in Salmonella Enterica." *PLoS Genetics* 7 (7): e1002191–e1002191. https://doi.org/10.1371/journal.pgen.1002191.

Didelot, Xavier, Daniel Lawson, Aaron Darling, and Daniel Falush. 2010. "Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences." *Genetics* 186 (4): 1435–49. https://doi.org/10.1534/genetics.110.120121.

Didelot, Xavier, and Martin C.J. Maiden. 2010. "Impact of Recombination on Bacterial Evolution." *Trends in Microbiology*. Elsevier. https://doi.org/10.1016/j.tim.2010.04.002.

Didelot, Xavier, Guillaume Méric, Daniel Falush, and Aaron E. Darling. 2012. "Impact of Homologous and Non-Homologous Recombination in the Genomic Evolution of Escherichia Coli." *BMC Genomics* 13 (1): 256. https://doi.org/10.1186/1471-2164-13-256.

Dixit, Purushottam D., Tin Yau Pang, and Sergei Maslov. 2017. "Recombination-Driven Genome Evolution and Stability of Bacterial Species." *Genetics* 207 (1): 281–95. https://doi.org/10.1534/genetics.117.300061.

Doroghazi, James R., and Daniel H. Buckley. 2010. "Widespread Homologous Recombination within and between Streptomyces Species." *ISME Journal* 4 (9): 1136–43. https://doi.org/10.1038/ismej.2010.45.

Doroghazi, James R., and William W. Metcalf. 2013. "Comparative Genomics of Actinomycetes with a Focus on Natural Product Biosynthetic Genes." *BMC Genomics* 14 (1): 611. https://doi.org/10.1186/1471-2164-14-611.

Dyken, J. David Van, Melanie J.I. Müller, Keenan M.L. MacK, and Michael M. Desai. 2013. "Spatial Population Expansion Promotes the Evolution of Cooperation in an Experimental Prisoner's Dilemma." *Current Biology* 23 (10): 919–23. https://doi.org/10.1016/j.cub.2013.04.026.

Elmberg, Johan, Charlotte Berg, Henrik Lerner, Jonas Waldenström, and Rebecca Hessel. 2017. "Potential Disease Transmission from Wild Geese and Swans to Livestock, Poultry and Humans: A Review of the Scientific Literature from a One Health Perspective." *Infection Ecology & Epidemiology* 7 (1): 1300450. https://doi.org/10.1080/20008686.2017.1300450.

Emond-Rheault, Jean-Guillaume, Jérémie Hamel, Julie Jeukens, Luca Freschi, Irena Kukavica-Ibrulj, Brian Boyle, Sandeep Tamber, et al. 2020. "The Salmonella Enterica Plasmidome as a Reservoir of Antibiotic Resistance." *Microorganisms* 8 (7): 1016. https://doi.org/10.3390/microorganisms8071016.

Eng, Shu Kee, Priyia Pusparajah, Nurul Syakima Ab Mutalib, Hooi Leng Ser, Kok Gan Chan, and Learn Han Lee. 2015. "Salmonella: A Review on Pathogenesis, Epidemiology and Antibiotic Resistance." *Frontiers in Life Science* 8 (3): 284–93. https://doi.org/10.1080/21553769.2015.1051243.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. "An Efficient Algorithm for Large-Scale Detection of Protein Families." *Nucleic Acids Research*. Oxford University Press. https://doi.org/10.1093/nar/30.7.1575.

Entwistle, Sarah, Xueqiong Li, and Yanbin Yin. 2019. "Orphan Genes Shared by Pathogenic Genomes Are More Associated with Bacterial Pathogenicity." *MSystems* 4 (1): e00290-18-e00290-18. https://doi.org/10.1128/msystems.00290-18.

Evans, Benjamin A., and Daniel E. Rozen. 2013. "Significant Variation in Transformation Frequency in Streptococcus Pneumoniae." *ISME Journal* 7 (4): 791–99. https://doi.org/10.1038/ismej.2012.170.

Fernández, Javier, Beatriz Guerra, and M. Rodicio. 2018. "Resistance to Carbapenems in Non-Typhoidal Salmonella Enterica Serovars from Humans, Animals and Food." *Veterinary Sciences* 5 (2): 40. https://doi.org/10.3390/vetsci5020040.

Flärdh, Klas, and Mark J. Buttner. 2009. "Streptomyces Morphogenetics: Dissecting Differentiation in a Filamentous Bacterium." *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/nrmicro1968.

Fondi, Marco, Antti Karkman, Manu V. Tamminen, Emanuele Bosi, Marko Virta, Renato Fani, Eric Alm, and James O. McInerney. 2016. "'Every Gene Is Everywhere but the Environment Selects': Global Geolocalization of Gene Sharing in Environmental Samples through Network Analysis." *Genome Biology and Evolution* 8 (5): 1388–1400. https://doi.org/10.1093/gbe/evw077.

Frazão, Miliane Rodrigues, Marta Inês Cazentini Medeiros, Sheila Da Silva Duque, and Juliana Pfrimer Falcão. 2017. "Pathogenic Potential and Genotypic Diversity of Campylobacter Jejuni: A Neglected Food-Borne Pathogen in Brazil." *Journal of Medical Microbiology* 66 (3): 350–59. https://doi.org/10.1099/jmm.0.000424.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data." *Bioinformatics* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

Fung, Connie, Shumin Tan, Mifuyu Nakajima, Emma C. Skoog, Luis Fernando Camarillo-Guerrero, Jessica A. Klein, Trevor D. Lawley, Jay V. Solnick, Tadashi Fukami, and Manuel R. Amieva. 2019. "High-Resolution Mapping Reveals That Microniches in the Gastric Glands Control Helicobacter Pylori Colonization of the Stomach." *PLoS Biology* 17 (5): e3000231. https://doi.org/10.1371/journal.pbio.3000231.

Gaiarsa, Stefano, Leone De Marco, Francesco Comandatore, Piero Marone, Claudio Bandi, and Davide Sassera. 2015. "Bacterial Genomic Epidemiology, from Local Outbreak Characterization to Species-History Reconstruction." *Pathogens and Global Health*. Maney Publishing. https://doi.org/10.1080/20477724.2015.1103503.

Gal-Mor, Ohad, Erin C. Boyle, and Guntram A. Grassl. 2014. “Same Species, Different Diseases: How and Why Typhoidal and Non-Typhoidal Salmonella Enterica Serovars Differ.” *Frontiers in Microbiology*. Frontiers Research Foundation. https://doi.org/10.3389/fmicb.2014.00391.

Giner-Lamia, Joaquín, Pablo Vinuesa, Laura Betancor, Claudia Silva, Julieta Bisio, Lorena Soleto, José A. Chabalgoity, et al. 2019. “Genome Analysis of Salmonella Enterica Subsp. Diarizonae Isolates from Invasive Human Infections Reveals Enrichment of Virulence-Related Functions in Lineage ST1256.” *BMC Genomics* 20 (1): 99. https://doi.org/10.1186/s12864-018-5352-z.

Golz, Julia C., Lennard Epping, Marie Theres Knüver, Maria Borowiak, Felix Hartkopf, Carlus Deneke, Burkhard Malorny, Torsten Semmler, and Kerstin Stingl. 2020. “Whole Genome Sequencing Reveals Extended Natural Transformation in Campylobacter Impacting Diagnostics and the Pathogens Adaptive Potential.” *Scientific Reports* 10 (1): 3686. https://doi.org/10.1038/s41598-020-60320-y.

González-Torres, Pedro, Francisco Rodríguez-Mateos, Josefa Antón, and Toni Gabaldón. 2019. “Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes.” *MBio* 10 (1). https://doi.org/10.1128/mBio.02494-18.

Grad, Yonatan H., Robert D. Kirkcaldy, David Trees, Janina Dordel, Simon R. Harris, Edward Goldstein, Hillard Weinstock, et al. 2014. “Genomic Epidemiology of Neisseria Gonorrhoeae with Reduced Susceptibility to Cefixime in the USA: A Retrospective Observational Study.” *The Lancet Infectious Diseases* 14 (3): 220–26. https://doi.org/10.1016/S1473-3099(13)70693-5.

Grad, Yonatan H., and Marc Lipsitch. 2014. “Epidemiologic Data and Pathogen Genome Sequences: A Powerful Synergy for Public Health.” *Genome Biology*. BioMed Central Ltd. https://doi.org/10.1186/s13059-014-0538-4.

Greenblum, Sharon, Rogan Carr, and Elhanan Borenstein. 2015. “Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species.” *Cell* 160 (4): 583–94. https://doi.org/10.1016/j.cell.2014.12.038.

Grinberg, Alex, Patrick J. Biggs, Ji Zhang, Stephen Ritchie, Zachary Oneroa, Charlotte O’Neill, Ali Karkaba, Niluka S. Velathanthiri, and Geoffrey W. Coombs. 2017. “Genomic Epidemiology of Methicillin-Susceptible Staphylococcus Aureus across Colonisation and Skin and Soft Tissue Infection.” *Journal of Infection* 75 (4): 326–35. https://doi.org/10.1016/j.jinf.2017.07.010.

Guerry, Patricia. 2007. “Campylobacter Flagella: Not Just for Motility.” *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2007.09.006.

Gupta, Vinod Kumar, Narendrakumar M. Chaudhari, Suchismitha Iskepalli, and Chitra Dutta. 2015. “Divergences in Gene Repertoire among the Reference Prevotella Genomes Derived from Distinct Body Sites of Human.” *BMC Genomics* 16 (1). https://doi.org/10.1186/s12864-015-1350-6.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75.

https://doi.org/10.1093/bioinformatics/btt086.

Guyomard-Rabenirina, Stéphanie, François Xavier Weill, Simon Le Hello, Sylvaine Bastian, Franck Berger, Séverine Ferdinand, Pierre Legreneur, et al. 2019. “Reptiles in Guadeloupe (French West Indies) Are a Reservoir of Major Human Salmonella Enterica Serovars.” *PLoS ONE* 14 (7): e0220145. https://doi.org/10.1371/journal.pone.0220145.

Hadfield, James, Nicholas J. Croucher, Richard J. Goater, Khalil Abudahab, David M. Aanensen, and Simon R. Harris. 2018. “Phandango: An Interactive Viewer for Bacterial Population Genomics.” *Bioinformatics* 34 (2): 292–93. https://doi.org/10.1093/bioinformatics/btx610.

Haegeman, Bart, and Joshua S. Weitz. 2012. “A Neutral Theory of Genome Evolution and the Frequency Distribution of Genes.” *BMC Genomics* 13 (1). https://doi.org/10.1186/1471-2164-13-196.

Halachev, Mihail R., Nicholas J. Loman, and Mark J. Pallen. 2011. “Calculating Orthologs in Bacteria and Archaea: A Divide and Conquer Approach.” *PLoS ONE* 6 (12). https://doi.org/10.1371/journal.pone.0028388.

Hanage, William P. 2016. “Not so Simple after All: Bacteria, Their Population Genetics, and Recombination.” *Cold Spring Harbor Perspectives in Biology* 8 (7): a018069. https://doi.org/10.1101/cshperspect.a018069.

Hanage, William P., Christophe Fraser, and Brian G. Spratt. 2005. “Fuzzy Species among Recombinogenic Bacteria.” *BMC Biology* 3 (March). https://doi.org/10.1186/1741-7007-3-6.

Hanage, William Paul, Christophe Fraser, Jing Tang, Thomas Richard Connor, and Jukka Corander. 2009. “Hyper-Recombination, Diversity, and Antibiotic Resistance in Pneumococcus.” *Science* 324 (5933): 1454–57. https://doi.org/10.1126/science.1171908.

Hawkey, Jane, Simon Le Hello, Benoît Doublet, Sophie A. Granier, Rene S. Hendriksen, W. Florian Fricke, Pieter Jan Ceyssens, et al. 2019. “Global Phylogenomics of Multidrug-Resistant Salmonella Enterica Serotype Kentucky ST198.” *Microbial Genomics* 5 (7). https://doi.org/10.1099/mgen.0.000269.

Hayes, Everett T., Jessica C. Wilks, Piero Sanfilippo, Elizabeth Yohannes, Daniel P. Tate, Brian D. Jones, Michael D. Radmacher, Sandra S. BonDurant, and Joan L. Slonczewski. 2006. “Oxygen Limitation Modulates PH Regulation of Catabolism and Hydrogenases, Multidrug Transporters, and Envelope Composition in Escherichia Coli K-12.” *BMC Microbiology* 6 (October): 89–89. https://doi.org/10.1186/1471-2180-6-89.

He, Zhen, Raad Z. Gharaibeh, Rachel C. Newsome, Jllian L. Pope, Michael W. Dougherty, Sarah Tomkovich, Benoit Pons, et al. 2019. “Campylobacter Jejuni Promotes Colorectal Tumorigenesis through the Action of Cytolethal Distending Toxin.” *Gut* 68 (2): 289–300. https://doi.org/10.1136/gutjnl-2018-317200.

Hertweck, Christian, Andriy Luzhetskyy, Yuri Rebets, and Andreas Bechthold. 2007. “Type II Polyketide Synthases: Gaining a Deeper Insight into Enzymatic Teamwork.” *Natural Product Reports*. Royal Society of Chemistry. https://doi.org/10.1039/b507395m.

Hickey, Thomas E., Annette L. McVeigh, Daniel A. Scott, Ronda E. Michielutti, Alyssa Bixby, Shannon A. Carroll, A. Louis Bourgeois, and Patricia Guerry. 2000. “Campylobacter Jejuni Cytolethal Distending Toxin Mediates Release of Interleukin-8 from Intestinal Epithelial Cells.” *Infection and Immunity* 68 (12): 6535–41. https://doi.org/10.1128/IAI.68.12.6535-6541.2000.

Hoelzer, Karin, Andrea Isabel Moreno Switt, and Martin Wiedmann. 2011. “Animal Contact as a Source of Human Non-Typhoidal Salmonellosis.” *Veterinary Research*. https://doi.org/10.1186/1297-9716-42-34.

Horstmann, Julia A., Erik Zschieschang, Theresa Truschel, Juana de Diego, Michele Lunelli, Manfred Rohde, Tobias May, et al. 2017. “Flagellin Phase-Dependent Swimming on Epithelial Cell Surfaces Contributes to Productive Salmonella Gut Colonisation.” *Cellular Microbiology* 19 (8): e12739. https://doi.org/10.1111/cmi.12739.

Huerta-Cepas, Jaime, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian Von Mering, and Peer Bork. 2017. “Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper.” *Molecular Biology and Evolution* 34 (8): 2115–22. https://doi.org/10.1093/molbev/msx148.

Huguet-Tapia, Jose C., Tristan Lefebure, Jonathan H. Badger, Dongli Guan, Gregg S. Pettis, Michael J. Stanhope, and Rosemary Loria. 2016. “Genome Content and Phylogenomics Reveal Both Ancestral and Lateral Evolutionary Pathways in Plant-Pathogenic Streptomyces Species.” *Applied and Environmental Microbiology* 82 (7): 2146–55. https://doi.org/10.1128/AEM.03504-15.

Hunt, Martin, Alison E. Mather, Leonor Sánchez-Busó, Andrew J. Page, Julian Parkhill, Jacqueline A. Keane, and Simon R. Harris. 2017. “ARIBA: Rapid Antimicrobial Resistance Genotyping Directly from Sequencing Reads.” *Microbial Genomics* 3 (10): e000131. https://doi.org/10.1099/mgen.0.000131.

Huson, Daniel H. 1998. “SplitsTree: Analyzing and Visualizing Evolutionary Data.” *Bioinformatics* 14 (1): 68–73. https://doi.org/10.1093/bioinformatics/14.1.68.

Hussein, Khetam, Ayelet Raz-Pasteur, Yael Shachor-Meyouhas, Yuval Geffen, Ilana Oren, Mical Paul, and Imad Kassis. 2016. “Campylobacter Bacteraemia: 16 Years of Experience in a Single Centre.” *Infectious Diseases* 48 (11–12): 796–99. https://doi.org/10.1080/23744235.2016.1195916.

Ichikawa, Natsuko, Machi Sasagawa, Mika Yamamoto, Hisayuki Komaki, Yumi Yoshida, Shuji Yamazaki, and Nobuyuki Fujita. 2013. “DoBISCUIT: A Database of Secondary Metabolite Biosynthetic Gene Clusters.” *Nucleic Acids Research* 41 (D1). https://doi.org/10.1093/nar/gks1177.

Ikeda, Haruo, Jun Ishikawa, Akiharu Hanamoto, Mayumi Shinose, Hisashi Kikuchi, Tadayoshi Shiba, Yoshiyuki Sakaki, Masahira Hattori, and Satoshi Omura. 2003. “Complete Genome Sequence and Comparative Analysis of the Industrial Microorganism Streptomyces Avermitilis.” *Nature Biotechnology* 21 (5): 526–31. https://doi.org/10.1038/nbt820.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes

Reveals Clear Species Boundaries." *Nature Communications* 9 (1): 5114. https://doi.org/10.1038/s41467-018-07641-9.

Jaspers, Elke, and Jörg Overmann. 2004. "Ecological Significance of Microdiversity: Identical 16S RRNA Gene Sequences Can Be Found in Bacteria with Highly Divergent Genomes and Ecophysiologies." *Applied and Environmental Microbiology* 70 (8): 4831–39. https://doi.org/10.1128/AEM.70.8.4831-4839.2004.

Jolley, Keith A., Man Suen Chan, and Martin C.J. Maiden. 2004. "MlstdbNet - Distributed Multi-Locus Sequence Typing (MLST) Databases." *BMC Bioinformatics* 5 (July): 86–86. https://doi.org/10.1186/1471-2105-5-86.

Jones, Timothy F., L. Amanda Ingram, Paul R. Cieslak, Duc J. Vugia, Melissa Tobin-D'Angelo, Sharon Hurd, Carlota Medus, Alicia Cronquist, and Frederick J. Angulo. 2008. "Salmonellosis Outcomes Differ Substantially by Serotype." *The Journal of Infectious Diseases* 198 (1): 109–14. https://doi.org/10.1086/588823.

Judd, M. C., R. M. Hoekstra, B. E. Mahon, P. I. Fields, and K. K. Wong. 2019. "Epidemiologic Patterns of Human Salmonella Serotype Diversity in the USA, 1996–2016." *Epidemiology and Infection* 147: 1–9. https://doi.org/10.1017/S0950268819000724.

Juhas, Mario. 2015. "Horizontal Gene Transfer in Human Pathogens." *Critical Reviews in Microbiology*. Informa Healthcare. https://doi.org/10.3109/1040841X.2013.804031.

Kaakoush, Nadeem O., Natalia Castaño-Rodríguez, Hazel M. Mitchell, and Si Ming Man. 2015. "Global Epidemiology of Campylobacter Infection." *Clinical Microbiology Reviews* 28 (3): 687–720. https://doi.org/10.1128/CMR.00006-15.

Kaltenpoth, Martin, Wolfgang Göttler, Gudrun Herzner, and Erhard Strohm. 2005. "Symbiotic Bacteria Protect Wasp Larvae from Fungal Infestation." *Current Biology* 15 (5): 475–79. https://doi.org/10.1016/j.cub.2004.12.084.

Katoh, K., and D. M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30 (14): 3059–66. http://www.ncbi.nlm.nih.gov/pubmed/12136088.

Katoh, Kazutaka, John Rozewicki, and Kazunori D. Yamada. 2017. "MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization." *Briefings in Bioinformatics*, September. https://doi.org/10.1093/bib/bbx108.

Kim, Ji Nu, Yeonbum Kim, Yujin Jeong, Jung Hye Roe, Byung Gee Kim, and Byung Kwan Cho. 2015. "Comparative Genomics Reveals the Core and Accessory Genomes of Streptomyces Speciesj." *Journal of Microbiology and Biotechnology* 25 (10): 1599–1605. https://doi.org/10.4014/jmb.1504.04008.

Kim, Jinki, and Gwan Su Yi. 2012. "PKMiner: A Database for Exploring Type II Polyketide Synthases." *BMC Microbiology* 12. https://doi.org/10.1186/1471-2180-12-169.

Kinashi, Haruyasu. 2011. “Giant Linear Plasmids in Streptomyces: A Treasure Trove of Antibiotic Biosynthetic Clusters.” *Journal of Antibiotics*. Nature Publishing Group. https://doi.org/10.1038/ja.2010.146.

Kirk, Martyn D., Sara M. Pires, Robert E. Black, Marisa Caipo, John A. Crump, Brecht Devleesschauwer, Dörte Döpfer, et al. 2015. “World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis.” *PLoS Medicine* 12 (12): e1001921. https://doi.org/10.1371/journal.pmed.1001921.

Kislyuk, Andrey O., Bart Haegeman, Nicholas H. Bergman, and Joshua S. Weitz. 2011. “Genomic Fluidity: An Integrative View of Gene Diversity within Microbial Populations.” *BMC Genomics* 12 (January). https://doi.org/10.1186/1471-2164-12-32.

Klemm, Elizabeth J., Sadia Shakoor, Andrew J. Page, Farah Naz Qamar, Kim Judge, Dania K. Saeed, Vanessa K. Wong, et al. 2018. “Emergence of an Extensively Drug-Resistant Salmonella Enterica Serovar Typhi Clone Harboring a Promiscuous Plasmid Encoding Resistance to Fluoroquinolones and Third-Generation Cephalosporins.” *MBio* 9 (1): e00105-19. https://doi.org/10.1128/mBio.00105-18.

Kominek, Jacek, Drew T. Doering, Dana A. Opulente, Xing Xing Shen, Xiaofan Zhou, Jeremy DeVirgilio, Amanda B. Hulfachor, et al. 2019. “Eukaryotic Acquisition of a Bacterial Operon.” *Cell* 176 (6): 1356-1366.e10. https://doi.org/10.1016/j.cell.2019.01.034.

Konstantinidis, Konstantinos T., Alban Ramette, and James M. Tiedje. 2006. “The Bacterial Species Definition in the Genomic Era.” In *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361:1929–40. Royal Society. https://doi.org/10.1098/rstb.2006.1920.

Kovács, Márta, Alexander Halfmann, Iris Fedtke, Manuel Heintz, Andreas Peschel, Waldemar Vollmer, Regine Hakenbeck, and Reinhold Brückner. 2006. “A Functional Dlt Operon, Encoding Proteins Required for Incorporation of D-Alanine in Teichoic Acids in Gram-Positive Bacteria, Confers Resistance to Cationic Antimicrobial Peptides in Streptococcus Pneumoniae.” *Journal of Bacteriology* 188 (16): 5797–5805. https://doi.org/10.1128/JB.00336-06.

Krauland, Mary G., Jane W. Marsh, David L. Paterson, and Lee H. Harrison. 2009. “Integron-Mediated Multidrug Resistance in a Global Collection of Nontyphoidal Salmonella Enterica Isolates.” *Emerging Infectious Diseases* 15 (3): 388–96. https://doi.org/10.3201/eid1503.081131.

Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. 2009. “Circos: An Information Aesthetic for Comparative Genomics.” *Genome Research* 19 (9): 1639–45. https://doi.org/10.1101/gr.092759.109.

Labeda, D. P., M. Goodfellow, R. Brown, A. C. Ward, B. Lanoot, M. Vanncanneyt, J. Swings, et al. 2012. “Phylogenetic Study of the Species within the Family Streptomycetaceae.” In *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 101:73–104. Springer. https://doi.org/10.1007/s10482-011-9656-0.

Lam, Margaret M.C., Kelly L. Wyres, Sebastian Duchêne, Ryan R. Wick, Louise M. Judd, Yunn Hwen Gan, Chu Han Hoh, et al. 2018. "Population Genomics of Hypervirulent Klebsiella Pneumoniae Clonal-Group 23 Reveals Early Emergence and Rapid Global Dissemination." *Nature Communications* 9 (1): 1–10. https://doi.org/10.1038/s41467-018-05114-7.

Lamas, Alexandre, José Manuel Miranda, Patricia Regal, Beatriz Vázquez, Carlos Manuel Franco, and Alberto Cepeda. 2018. "A Comprehensive Review of Non-Enterica Subspecies of Salmonella Enterica." *Microbiological Research*. Elsevier GmbH. https://doi.org/10.1016/j.micres.2017.09.010.

Lan, Ruiting, Peter R. Reeves, and Sophie Octavia. 2009. "Population Structure, Origins and Evolution of Major Salmonella Enterica Clones." *Infection, Genetics and Evolution*. Elsevier. https://doi.org/10.1016/j.meegid.2009.04.011.

Land, Miriam, Loren Hauser, Se-Ran Ran Jun, Intawat Nookaew, Michael R. Leuze, Tae-Hyuk Hyuk Ahn, Tatiana Karpinets, et al. 2015. "Insights from 20 Years of Bacterial Genome Sequencing." *Functional & Integrative Genomics* 15 (2): 141–61. https://doi.org/10.1007/s10142-015-0433-4.

Langridge, Gemma C., Maria Fookesa, Thomas R. Connora, Theresa Feltwella, Nicholas Feaseya, Bryony N. Parsonsb, Helena M.B. Seth-Smitha, et al. 2015. "Patterns of Genome Evolution That Have Accompanied Host Adaptation in Salmonella." *Proceedings of the National Academy of Sciences of the United States of America* 112 (3): 863–68. https://doi.org/10.1073/pnas.1416707112.

Lapierre, Pascal, and J. Peter Gogarten. 2009. *Estimating the Size of the Bacterial Pan-Genome*. *Trends in Genetics*. Vol. 25. https://www.sciencedirect.com/science/article/pii/S0168952509000055?via%3Dihub.

Lara-Tejero, M., and J. E. Galán. 2001. "CdtA, CdtB, and CdtC Form a Tripartite Complex That Is Required for Cytolethal Distending Toxin Activity." *Infection and Immunity* 69 (7): 4358–65. https://doi.org/10.1128/IAI.69.7.4358-4365.2001.

Lawrence, Jeffrey G., and John R. Roth. 1996. "Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters." *Genetics*. Genetics Society of America. /pmc/articles/PMC1207444/?report=abstract.

LeClerc, J. Eugene, Baoguang Li, William L. Payne, and Thomas A. Cebula. 1996. "High Mutation Frequencies among Escherichia Coli and Salmonella Pathogens." *Science* 274 (5290): 1208–11. https://doi.org/10.1126/science.274.5290.1208.

Leekitcharoenphon, Pimlapas, Rene S. Hendriksen, Simon Le Hello, François Xavier Weill, Dorte Lau Baggesen, Se Ran Jun, David W. Ussery, et al. 2016. "Global Genomic Epidemiology of Salmonella Enterica Serovar Typhimurium DT104." *Applied and Environmental Microbiology* 82 (8): 2516–26. https://doi.org/10.1128/AEM.03821-15.

Leonard, Susan R., Mark K. Mammel, David W. Lacher, and Christopher A. Elkins. 2016. "Strain-Level Discrimination of Shiga Toxin-Producing Escherichia Coli in Spinach Using Metagenomic Sequencing." *PLoS ONE* 11 (12). https://doi.org/10.1371/journal.pone.0167870.

Letunic, Ivica, and Peer Bork. 2016. “Interactive Tree of Life (ITOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees.” *Nucleic Acids Research* 44 (W1): W242–45. https://doi.org/10.1093/nar/gkw290.

Letzel, Anne Catrin, Jing Li, Gregory C.A. Amos, Natalie Millán-Aguiñaga, Joape Ginigini, Usama R. Abdelmohsen, Susana P. Gaudêncio, Nadine Ziemert, Bradley S. Moore, and Paul R. Jensen. 2017. “Genomic Insights into Specialized Metabolism in the Marine Actinomycete Salinispora.” *Environmental Microbiology* 19 (9): 3660–73. https://doi.org/10.1111/1462-2920.13867.

Levade, Inès, Yves Terrat, Jean Baptiste Leducq, Ana A. Weil, Leslie M. Mayo-Smith, Fahima Chowdhury, Ashraful I. Khan, et al. 2017. “Vibrio Cholerae Genomic Diversity within and between Patients.” *Microbial Genomics* 3 (12): e000142. https://doi.org/10.1099/mgen.0.000142.

Leventhal, Gabriel E., Carles Boix, Urs Kuechler, Tim N. Enke, Elzbieta Sliwerska, Christof Holliger, and Otto X. Cordero. 2018. “Strain-Level Diversity Drives Alternative Community Types in Millimetre-Scale Granular Biofilms.” *Nature Microbiology* 3 (11): 1295–1303. https://doi.org/10.1038/s41564-018-0242-3.

Levin, Bruce R., and Omar E. Cornejo. 2009. “The Population and Evolutionary Dynamics of Homologous Gene Recombination in Bacteria.” Edited by David S. Guttman. *PLoS Genetics* 5 (8): e1000601. https://doi.org/10.1371/journal.pgen.1000601.

Li, W., and A. Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–59. https://doi.org/10.1093/bioinformatics/btl158.

Liang, Hao, Aiyu Zhang, Yixin Gu, Yuanhai You, Jianzhong Zhang, and Maojun Zhang. 2016. “Genetic Characteristics and Multiple-PCR Development for Capsular Identification of Specific Serotypes of Campylobacter Jejuni.” Edited by Patrick Jon Biggs. *PLOS ONE* 11 (10): e0165159. https://doi.org/10.1371/journal.pone.0165159.

Lin, Mingzhi, and Edo Kussell. 2019. “Inferring Bacterial Recombination Rates from Large-Scale Sequencing Datasets.” *Nature Methods* 16 (2): 199–204. https://doi.org/10.1038/s41592-018-0293-7.

Linton, Dennis, Michel Gilbert, Paul G. Hitchen, Anne Dell, Howard R. Morris, Warren W. Wakarchuk, Norman A. Gregson, and Brendan W. Wren. 2000. “Phase Variation of a β-1,3 Galactosyltransferase Involved in Generation of the Ganglioside GM1-like Lipo-Oligosaccharide of Campylobacter Jejuni.” *Molecular Microbiology* 37 (3): 501–14. https://doi.org/10.1046/j.1365-2958.2000.02020.x.

Liu, Bo, Zheng, Dandan, Jin, Qi, Chen, Lihong, Yang, Jian. 2019. “VFDB 2019: A Comparative Pathogenomic Platform with an Interactive Web Interface.” *Nucleic Acids Research* 47 (D1): D687–92. https://academic.oup.com/nar/article/47/D1/D687/5160975.

Liu, Huanli, Chris A. Whitehouse, and Baoguang Li. 2018. “Presence and Persistence of Salmonella in Water: The Impact on Microbial Quality of Water and Food Safety.” *Frontiers in Public Health* 6 (May): 159. https://doi.org/10.3389/fpubh.2018.00159.

Liu, Yue, Dao Feng Zhang, Xiujuan Zhou, Li Xu, Lida Zhang, and Xianming Shi. 2017. "Comprehensive Analysis Reveals Two Distinct Evolution Patterns of Salmonella Flagellin Gene Clusters." *Frontiers in Microbiology* 8 (DEC): 2604. https://doi.org/10.3389/fmicb.2017.02604.

Lobkovsky, Alexander E., Yuri I. Wolf, and Eugene V. Koonin. 2016. "Evolvability of an Optimal Recombination Rate." *Genome Biology and Evolution* 8 (1): 70–77. https://doi.org/10.1093/gbe/evv249.

Long, Katherine S., Jacob Poehlsgaard, Corinna Kehrenberg, Stefan Schwarz, and Birte Vester. 2006. "The Cfr RRNA Methyltransferase Confers Resistance to Phenicols, Lincosamides, Oxazolidinones, Pleuromutilins, and Streptogramin A Antibiotics." *Antimicrobial Agents and Chemotherapy* 50 (7): 2500–2505. https://doi.org/10.1128/AAC.00131-06.

Lopes, Bruno S., Norval J.C. Strachan, Meenakshi Ramjee, Anne Thomson, Marion Macrae, Sophie Shaw, and Ken J. Forbes. 2019. "Nationwide Stepwise Emergence and Evolution of Multidrug-Resistant Campylobacter Jejuni Sequence Type 5136, United Kingdom." *Emerging Infectious Diseases* 25 (7): 1320–29. https://doi.org/10.3201/eid2507.181572.

Maayer, Pieter De, Wai Y. Chan, Enrico Rubagotti, Stephanus N. Venter, Ian K. Toth, Paul R.J. Birch, and Teresa A. Coutinho. 2014. "Analysis of the Pantoea Ananatis Pan-Genome Reveals Factors Underlying Its Ability to Colonize and Interact with Plant, Insect and Vertebrate Hosts." *BMC Genomics* 15 (1): 404. https://doi.org/10.1186/1471-2164-15-404.

Maayer, Pieter De, and Don A. Cowan. 2016. "Flashy Flagella: Flagellin Modification Is Relatively Common and Highly Versatile among the Enterobacteriaceae." *BMC Genomics* 17 (1): 377. https://doi.org/10.1186/s12864-016-2735-x.

Martinez-Gomez, Norma C., Matt Robers, and Diana M. Downs. 2004. "Mutational Analysis of ThiH, a Member of the Radical S-Adenosylmethionine (AdoMet) Protein Superfamily." *Journal of Biological Chemistry* 279 (39): 40505–10. https://doi.org/10.1074/jbc.M403985200.

Marttinen, Pekka, and William P. Hanage. 2017. "Speciation Trajectories in Recombining Bacterial Species." Edited by Mark M. Tanaka 13 (7): e1005640. https://doi.org/10.1371/journal.pcbi.1005640.

Marttinen, Pekka, William P. Hanage, Nicholas J. Croucher, Thomas R. Connor, Simon R. Harris, Stephen D. Bentley, and Jukka Corander. 2012. "Detection of Recombination Events in Bacterial Genomes from Large Population Samples." *Nucleic Acids Research* 40 (1): e6. https://doi.org/10.1093/nar/gkr928.

Mather, Alison E., Tu Le Thi Phuong, Yunfeng Gao, Simon Clare, Subhankar Mukhopadhyay, David A. Goulding, Nhu Tran Do Hoang, et al. 2018. "New Variant of Multidrug-Resistant Salmonella Enterica Serovar Typhimurium Associated with Invasive Disease in Immunocompromised Patients in Vietnam." *MBio* 9 (5). https://doi.org/10.1128/mBio.01056-18.

McDonald, Bradon R., and Cameron R. Curriea. 2017. "Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces." *MBio* 8 (3). https://doi.org/10.1128/mBio.00644-17.

McInerney, James O., Alan McNally, and Mary J. O'Connell. 2017. "Why Prokaryotes Have Pangenomes." *Nature Microbiology*. Nature Publishing Group. https://doi.org/10.1038/nmicrobiol.2017.40.

Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. 2005. *The Microbial Pan-Genome*. Vol. 15. http://www.ncbi.nlm.nih.gov/pubmed/16185861.

Meij, Anne van der, Sarah F. Worsley, Matthew I. Hutchings, and Gilles P. van Wezel. 2017. "Chemical Ecology of Antibiotic Production by Actinomycetes." *FEMS Microbiology Reviews*. Oxford University Press. https://doi.org/10.1093/femsre/fux005.

Mi, Huaiyu, Anushya Muruganujan, Xiaosong Huang, Dustin Ebert, Caitlin Mills, Xinyu Guo, and Paul D. Thomas. 2019. "Protocol Update for Large-Scale Genome and Gene Function Analysis with the PANTHER Classification System (v.14.0)." *Nature Protocols* 14 (3): 703–21. https://doi.org/10.1038/s41596-019-0128-8.

Montgomery, Martha P., Scott Robertson, Lia Koski, Ellen Salehi, Lauren M. Stevenson, Rachel Silver, Preethi Sundararaman, et al. 2018. " Multidrug-Resistant Campylobacter Jejuni Outbreak Linked to Puppy Exposure — United States, 2016–2018 ." *MMWR. Morbidity and Mortality Weekly Report* 67 (37): 1032–35. https://doi.org/10.15585/mmwr.mm6737a3.

Moreno Switt, Andrea I., Henk C. den Bakker, Craig A. Cummings, Lorraine D. Rodriguez-Rivera, Gregory Govoni, Matthew L. Raneiri, Lovorka Degoricija, et al. 2012. "Identification and Characterization of Novel Salmonella Mobile Elements Involved in the Dissemination of Genes Linked to Virulence and Transmission." Edited by Axel Cloeckaert. *PLoS ONE* 7 (7): e41247. https://doi.org/10.1371/journal.pone.0041247.

Mostowy, Rafal, Nicholas J. Croucher, Cheryl P. Andam, Jukka Corander, William P. Hanage, and Pekka Marttinen. 2017. "Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations." *Molecular Biology and Evolution* 34 (5): 1167–82. https://doi.org/10.1093/molbev/msx066.

Mostowy, Rafal, Nicholas J. Croucher, William P. Hanage, Simon R. Harris, Stephen Bentley, and Christophe Fraser. 2014. "Heterogeneity in the Frequency and Characteristics of Homologous Recombination in Pneumococcal Evolution." *PLoS Genetics* 10 (5): e1004300. https://doi.org/10.1371/journal.pgen.1004300.

Mourkas, Evangelos, Diego Florez-Cuadrado, Ben Pascoe, Jessica K. Calland, Sion C. Bayliss, Leonardos Mageiros, Guillaume Méric, et al. 2019. "Gene Pool Transmission of Multidrug Resistance among Campylobacter from Livestock, Sewage and Human Disease." *Environmental Microbiology* 21 (12): 4597–4613. https://doi.org/10.1111/1462-2920.14760.

Müller, Jens, Birgit Meyer, Ingrid Hänel, and Helmut Hotzel. 2007. "Comparison of Lipooligosaccharide Biosynthesis Genes of Campylobacter Jejuni Strains with Varying Abilities to Colonize the Chicken Gut and to Invade Caco-2 Cells." *Journal of Medical Microbiology* 56 (12): 1589–94. https://doi.org/10.1099/jmm.0.47305-0.

Neal-McKinney, Jason M., Kun C. Liu, Karen C. Jinneman, Wen Hsin Wu, and Daniel H. Rice. 2018. "Whole Genome Sequencing and Multiplex QPCR Methods to Identify Campylobacter Jejuni Encoding Cst-II or Cst-III Sialyltransferase." *Frontiers in*

Microbiology 9 (MAR): 408. https://doi.org/10.3389/fmicb.2018.00408.

Nguyen, Lam-Tung, Schmidt, Heiko A., Haeseler, Arndt von, Minh, Bui Quang. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution*, 268–74. https://academic.oup.com/mbe/article/32/1/268/2925592.

Nishino, Kunihiko, Tammy Latifi, and Eduardo A. Groisman. 2006. "Virulence and Drug Resistance Roles of Multidrug Efflux Systems of Salmonella Enterica Serovar Typhimurium." *Molecular Microbiology* 59 (1): 126–41. https://doi.org/10.1111/j.1365-2958.2005.04940.x.

Ohnishi, Yasuo, Jun Ishikawa, Hirofumi Hara, Hirokazu Suzuki, Miwa Ikenoya, Haruo Ikeda, Atsushi Yamashita, Masahira Hattori, and Sueharu Horinouchi. 2008. "Genome Sequence of the Streptomycin-Producing Microorganism Streptomyces Griseus IFO 13350." *Journal of Bacteriology* 190 (11): 4050–60. https://doi.org/10.1128/JB.00204-08.

Oladeinde, Adelumola, Kimberly Cook, Steven M. Lakin, Reed Woyda, Zaid Abdo, Torey Looft, Kyler Herrington, et al. 2019. " Horizontal Gene Transfer and Acquired Antibiotic Resistance in Salmonella Enterica Serovar Heidelberg Following In Vitro Incubation in Broiler Ceca ." *Applied and Environmental Microbiology* 85 (22). https://doi.org/10.1128/aem.01903-19.

Omelchenko, Marina V., Kira S. Makarova, Yuri I. Wolf, Igor B. Rogozin, and Eugene V. Koonin. 2003. "Evolution of Mosaic Operons by Horizontal Gene Transfer and Gene Displacement in Situ." *Genome Biology* 4 (9): R55–R55. https://doi.org/10.1186/gb-2003-4-9-r55.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. "Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis." *Bioinformatics* 31 (22): 3691–93. https://doi.org/10.1093/bioinformatics/btv421.

Papke, R. Thane, Olga Zhaxybayeva, Edward J. Feil, Katrin Sommerfeld, Denise Muise, and W. Ford Doolittle. 2007. "Searching for Species in Haloarchaea." *Proceedings of the National Academy of Sciences of the United States of America* 104 (35): 14092–97. https://doi.org/10.1073/pnas.0706358104.

Park, Cooper J., and Cheryl P. Andam. 2019. "Within-Species Genomic Variation and Variable Patterns of Recombination in the Tetracycline Producer Streptomyces Rimosus." *Frontiers in Microbiology* 10 (MAR): 552. https://doi.org/10.3389/fmicb.2019.00552.

Park, Cooper J., and Cheryl P. Andam. 2020. "Distinct but Intertwined Evolutionary Histories of Multiple Salmonella Enterica Subspecies." *MSystems* 5 (1): e00515-19-e00515-19. https://doi.org/10.1128/msystems.00515-19.

Paschos, Athanasios, Anette Bauer, Anja Zimmermann, Eva Zehelein, and August Böck. 2002. "HypF, a Carbamoyl Phosphate-Converting Enzyme Involved in [NiFe] Hydrogenase Maturation." *Journal of Biological Chemistry* 277 (51): 49945–51. https://doi.org/10.1074/jbc.M204601200.

Paul, Sandip, Elena V. Linardopoulou, Mariya Billig, Veronika Tchesnokova, Lance B. Price, James R. Johnson, Sujay Chattopadhyay, and Evgeni V. Sokurenko. 2013. "Role of Homologous Recombination in Adaptive Diversification of Extraintestinal Escherichia Coli." *Journal of Bacteriology* 195 (2): 231–42. https://doi.org/10.1128/JB.01524-12.

Peñalba, Joshua V., and Jochen B.W. Wolf. 2020. "From Molecules to Populations: Appreciating and Estimating Recombination Rate Variation." *Nature Reviews Genetics*. Nature Research. https://doi.org/10.1038/s41576-020-0240-1.

Perron, G. G., A. E. G. Lee, Y. Wang, W. E. Huang, and T. G. Barraclough. 2012. "Bacterial Recombination Promotes the Evolution of Multi-Drug-Resistance in Functionally Diverse Populations." *Proceedings of the Royal Society B: Biological Sciences* 279 (1733): 1477–84. https://doi.org/10.1098/rspb.2011.1933.

Perron, Gabriel G., Alexander E. G. Lee, Yun Wang, Wei E. Huang, and Timothy G. Barraclough. 2012. "Bacterial Recombination Promotes the Evolution of Multi-Drug-Resistance in Functionally Diverse Populations." *Proceedings of the Royal Society B: Biological Sciences* 279 (1733): 1477–84. https://doi.org/10.1098/rspb.2011.1933.

Petković, Hrvoje, John Cullum, Daslav Hranueli, Iain S. Hunter, Nataša Perić-Concha, Jasenka Pigac, Arinthip Thamchaipenet, Dušica Vujaklija, and Paul F. Long. 2006. "Genetics of Streptomyces Rimosus, the Oxytetracycline Producer." *Microbiology and Molecular Biology Reviews* 70 (3): 704–28. https://doi.org/10.1128/mmbr.00004-06.

Pilla, Giulia, and Christoph M. Tang. 2018. "Going around in Circles: Virulence Plasmids in Enteric Pathogens." *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/s41579-018-0031-2.

Pinto, A. Viviana, Aurélie Mathieu, Stéphanie Marsin, Xavier Veaute, Luis Ielpi, Agnès Labigne, and J. Pablo Radicella. 2005. "Suppression of Homologous and Homeologous Recombination by the Bacterial MutS2 Protein." *Molecular Cell* 17 (1): 113–20. https://doi.org/10.1016/j.molcel.2004.11.035.

Plener, Laure, Pierre Boistard, Adriana González, Christian Boucher, and Stéphane Genin. 2012. "Metabolic Adaptation of Ralstonia Solanacearum during Plant Infection: A Methionine Biosynthesis Case Study." *PLoS ONE* 7 (5): e36877–e36877. https://doi.org/10.1371/journal.pone.0036877.

Pohl, Sarah, Jens Klockgether, Denitsa Eckweiler, Ariane Khaledi, Monika Schniederjans, Philippe Chouvarine, Burkhard Tümmler, and Susanne Häussler. 2014. "The Extensive Set of Accessory Pseudomonas Aeruginosa Genomic Components." *FEMS Microbiology Letters* 356 (2): 235–41. https://doi.org/10.1111/1574-6968.12445.

Pornsukarom, Suchawan, Arnoud H.M. Van Vliet, and Siddhartha Thakur. 2018. "Whole Genome Sequencing Analysis of Multiple Salmonella Serovars Provides Insights into Phylogenetic Relatedness, Antimicrobial Resistance, and Virulence Markers across Humans, Food Animals and Agriculture Environmental Sources." *BMC Genomics* 19 (1): 801. https://doi.org/10.1186/s12864-018-5137-4.

Prezioso, Samantha M., Nicole E. Brown, and Joanna B. Goldberg. 2017. "Elfamycins: Inhibitors of Elongation Factor-Tu." *Molecular Microbiology* 106 (1): 22–34.

https://doi.org/10.1111/mmi.13750.

Pritchard, J K, M Stephens, and P Donnelly. 2000. “Inference of Population Structure Using Multilocus Genotype Data.” *Genetics* 155 (2): 945–59. http://www.ncbi.nlm.nih.gov/pubmed/10835412.

Pulford, Caisey V., Nicolas Wenner, Martha L. Redway, Ella V. Rodwell, Hermione J. Webster, Roberta Escudero, Carsten Kröger, et al. 2019. “The Diversity, Evolution and Ecology of Salmonella in Venomous Snakes.” *PLoS Neglected Tropical Diseases* 13 (6): e0007169. https://doi.org/10.1371/journal.pntd.0007169.

R Core Team. 2019. “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria. 2019. http://www.r-project.org/.

Rahmati, Sonia, Shirley Yang, Amy L. Davidson, and E. Lynn Zechiedrich. 2002. “Control of the AcrAB Multidrug Efflux Pump by Quorum-Sensing Regulator SdiA.” *Molecular Microbiology* 43 (3): 677–85. https://doi.org/10.1046/j.1365-2958.2002.02773.x.

Rauch, Joseph, Jane Kondev, and Alvaro Sanchez. 2017. “Cooperators Trade off Ecological Resilience and Evolutionary Stability in Public Goods Games.” *Journal of the Royal Society Interface* 14 (127). https://doi.org/10.1098/rsif.2016.0967.

Rehman, Muhammad A., Xianhua Yin, Marissa G. Persaud-Lachhman, and Moussa S. Diarra. 2017. “First Detection of a Fosfomycin Resistance Gene, FosA7, in Salmonella Enterica Serovar Heidelberg Isolated from Broiler Chickens.” *Antimicrobial Agents and Chemotherapy* 61 (8). https://doi.org/10.1128/AAC.00410-17.

Reid, Anne N., Reenu Pandey, Kiran Palyada, Hemant Naikare, and Alain Stintzi. 2008. “Identification of Campylobacter Jejuni Genes Involved in the Response to Acidic PH and Stomach Transit.” *Applied and Environmental Microbiology* 74 (5): 1583–97. https://doi.org/10.1128/AEM.01507-07.

Roach, David J., Joshua N. Burton, Choli Lee, Bethany Stackhouse, Susan M. Butler-Wu, Brad T. Cookson, Jay Shendure, and Stephen J. Salipante. 2015. “A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota.” Edited by Diarmaid Hughes. *PLOS Genetics* 11 (7): e1005413. https://doi.org/10.1371/journal.pgen.1005413.

Rodríguez-Beltrán, Jerónimo, Jérôme Tourret, Olivier Tenaillon, Elena López, Emmanuelle Bourdelier, Coloma Costas, Ivan Matic, Erick Denamur, and Jesús Blázquez. 2015. “High Recombinant Frequency in Extraintestinal Pathogenic Escherichia Coli Strains.” *Molecular Biology and Evolution* 32 (7): 1708–16. https://doi.org/10.1093/molbev/msv072.

Rouli, L., V. Merhej, P. E. Fournier, and D. Raoult. 2015. “The Bacterial Pangenome as a New Tool for Analysing Pathogenic Bacteria.” *New Microbes and New Infections* 7 (September): 72–85. https://doi.org/10.1016/j.nmni.2015.06.005.

Schauer, Kristina, Jürgen Stolz, Siegfried Scherer, and Thilo M. Fuchs. 2009. “Both Thiamine Uptake and Biosynthesis of Thiamine Precursors Are Required for Intracellular Replication of Listeria Monocytogenes.” *Journal of Bacteriology* 191 (7): 2218–27. https://doi.org/10.1128/JB.01636-08.

Schnappinger, Dirk, and Wolfgang Hillen. 1996. “Tetracyclines: Antibiotic Action, Uptake, and Resistance Mechanisms.” *Archives of Microbiology*. Springer Verlag. https://doi.org/10.1007/s002030050339.

Seco, Elena M., Trinidad Cuesta, Serge Fotso, Hartmut Laatsch, and Francisco Malpartida. 2005. “Two Polyene Amides Produced by Genetically Modified Streptomyces Diastaticus Var. 108.” *Chemistry and Biology* 12 (5): 535–43. https://doi.org/10.1016/j.chembiol.2005.02.015.

Seco, Elena M., Francisco J. Pérez-Zúñiga, Miriam S. Rolón, and Francisco Malpartida. 2004. “Starter Unit Choice Determines the Production of Two Tetraene Macrolides, Rimocidin and CE-108, in Streptomyces Diastaticus Var. 108.” *Chemistry and Biology* 11 (3): 357–66. https://doi.org/10.1016/j.chembiol.2004.02.017.

Seemann, T. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.

Segerman, B. 2012. “The Genetic Integrity of Bacterial Species: The Core Genome and the Accessory Genome, Two Different Stories.” *Frontiers in Cellular and Infection Microbiology* 2: 116. https://doi.org/10.3389/fcimb.2012.00116.

Seipke, Ryan F. 2015. “Strain-Level Diversity of Secondary Metabolism in Streptomyces Albus.” *PLoS ONE* 10 (1). https://doi.org/10.1371/journal.pone.0116457.

Seipke, Ryan F., Jörg Barke, Charles Brearley, Lionel Hill, Douglas W. Yu, Rebecca J.M. Goss, and Matthew I. Hutchings. 2011. “A Single Streptomyces Symbiont Makes Multiple Antifungals to Support the Fungus Farming Ant Acromyrmex Octospinosus.” *PLoS ONE* 6 (8). https://doi.org/10.1371/journal.pone.0022028.

Seipke, Ryan F., Martin Kaltenpoth, and Matthew I. Hutchings. 2012. “Streptomyces as Symbionts: An Emerging and Widespread Theme?” *FEMS Microbiology Reviews*. Oxford Academic. https://doi.org/10.1111/j.1574-6976.2011.00313.x.

Sela, Uri, Chad W. Euler, Joel Correa da Rosa, and Vincent A. Fischetti. 2018. “Strains of Bacterial Species Induce a Greatly Varied Acute Adaptive Immune Response: The Contribution of the Accessory Genome.” Edited by Vance G Fowler. *PLoS Pathogens* 14 (1): e1006726. https://doi.org/10.1371/journal.ppat.1006726.

Sheppard, Samuel K., Lu Cheng, Guillaume Méric, Caroline P.A. A. De Haan, Ann-Katrin Katrin Llarena, Pekka Marttinen, Ana Vidal, et al. 2014. “Cryptic Ecology among Host Generalist Campylobacter Jejuni in Domestic Animals.” *Molecular Ecology* 23 (10): 2442–51. https://doi.org/10.1111/mec.12742.

Sheppard, Samuel K., and Martin C.J. Maiden. 2015. “The Evolution of Campylobacter Jejuni and Campylobacter Coli.” *Cold Spring Harbor Perspectives in Biology* 7 (8). https://doi.org/10.1101/cshperspect.a018119.

Silby, Mark W., Craig Winstanley, Scott A.C. Godfrey, Stuart B. Levy, and Robert W. Jackson. 2011. “Pseudomonas Genomes: Diverse and Adaptable.” *FEMS Microbiology Reviews* 35

(4): 652–80. https://doi.org/10.1111/j.1574-6976.2011.00269.x.

Silva, Claudia, Edmundo Calva, and Stanley Maloy. 2014. “One Health and Food-Borne Disease: Salmonella Transmission between Humans, Animals, and Plants.” In *One Health*, 137–48. American Society of Microbiology. https://doi.org/10.1128/microbiolspec.oh-0020-2013.

Skarp, C. P.A., O. Akinrinade, R. Kaden, C. Johansson, and H. Rautelin. 2017. “Accessory Genetic Content in Campylobacter Jejuni ST21CC Isolates from Feces and Blood.” *International Journal of Medical Microbiology* 307 (4–5): 233–40. https://doi.org/10.1016/j.ijmm.2017.04.001.

Skippington, Elizabeth, and Mark A. Ragan. 2012. “Phylogeny Rather than Ecology or Lifestyle Biases the Construction of Escherichia Coli-Shigella Genetic Exchange Communities.” *Open Biology* 2 (9): 120112. https://doi.org/10.1098/rsob.120112.

Smillie, Chris S., Mark B. Smith, Jonathan Friedman, Otto X. Cordero, Lawrence A. David, and Eric J. Alm. 2011. “Ecology Drives a Global Network of Gene Exchange Connecting the Human Microbiome.” *Nature* 480 (7376): 241–44. https://doi.org/10.1038/nature10571.

Smith, N. H., P. Beltran, and R. K. Selander. 1990. “Recombination of Salmonella Phase 1 Flagellin Genes Generates New Serovars.” *Journal of Bacteriology* 172 (5): 2209–16. https://doi.org/10.1128/jb.172.5.2209-2216.1990.

Snipen, Lars, and Kristian Hovde Liland. 2015. “Micropan: An R-Package for Microbial Pan-Genomics.” *BMC Bioinformatics* 16 (1): 1–8. https://doi.org/10.1186/s12859-015-0517-0.

Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten. 2015. “Horizontal Gene Transfer: Building the Web of Life.” *Nature Reviews Genetics* 16 (8): 472–82. https://doi.org/10.1038/nrg3962.

Souvorov, Alexandre, Richa Agarwala, and David J. Lipman. 2018. “SKESA: Strategic k-Mer Extension for Scrupulous Assemblies.” *Genome Biology* 19 (1): 153. https://doi.org/10.1186/s13059-018-1540-z.

Spoor, Laura E., Emily Richardson, Amy C. Richards, Gillian J. Wilson, Chriselle Mendonca, Ravi Kr Gupta, Paul R. McAdam, et al. 2015. “Recombination-Mediated Remodelling of Host–Pathogen Interactions during Staphylococcus Aureus Niche Adaptation.” *Microbial Genomics* 1 (4): 1–14. https://doi.org/10.1099/mgen.0.000036.

Stamatakis, Alexandros. 2006. “RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.” *Bioinformatics* 22 (21): 2688–90. https://doi.org/10.1093/bioinformatics/btl446.

Stanaway, Jeffrey D., Andrea Parisi, Kaushik Sarkar, Brigette F. Blacker, Robert C. Reiner, Simon I. Hay, Molly R. Nixon, et al. 2019. “The Global Burden of Non-Typhoidal Salmonella Invasive Disease: A Systematic Analysis for the Global Burden of Disease Study 2017.” *The Lancet Infectious Diseases* 19 (12): 1312–24. https://doi.org/10.1016/S1473-3099(19)30418-9.

Stoesser, N., A. E. Sheppard, C. E. Moore, T. Golubchik, C. M. Parry, P. Nget, M. Saroeun, et al. 2015. “Extensive Within-Host Diversity in Fecally Carried Extended-Spectrum-Beta-

Lactamase-Producing Escherichia Coli Isolates: Implications for Transmission Analyses." *Journal of Clinical Microbiology* 53 (7): 2122–31. https://doi.org/10.1128/JCM.00378-15.

Sun, Jian, Run Shi Yang, Qijing Zhang, Youjun Feng, Liang Xing Fang, Jing Xia, Liang Li, et al. 2016. "Co-Transfer of BlaNDM-5 and Mcr-1 by an IncX3-X4 Hybrid Plasmid in Escherichia Coli." *Nature Microbiology* 1 (12): 1–4. https://doi.org/10.1038/nmicrobiol.2016.176.

Taboada, Blanca, Karel Estrada, Ricardo Ciria, and Enrique Merino. 2018. "Operon-Mapper: A Web Server for Precise Operon Identification in Bacterial and Archaeal Genomes." *Bioinformatics* 34 (23): 4118–20. https://doi.org/10.1093/bioinformatics/bty496.

Tack, Danielle M., Ellyn P. Marder, Patricia M. Griffin, Paul R. Cieslak, John Dunn, Sharon Hurd, Elaine Scallan, et al. 2019. "Preliminary Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food — Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2015–2018." *MMWR. Morbidity and Mortality Weekly Report* 68 (16): 369–73. https://doi.org/10.15585/mmwr.mm6816a2.

Talukder, Kaisar A., Mohammad Aslam, Zhahirul Islam, Ishrat J. Azmi, Dilip K. Dutta, Sabir Hossain, Alam Nur-E-Kamal, et al. 2008. "Prevalence of Virulence Genes and Cytolethal Distending Toxin Production in Campylobacter Jejuni Isolates from Diarrheal Patients in Bangladesh." *Journal of Clinical Microbiology* 46 (4): 1485–88. https://doi.org/10.1128/JCM.01912-07.

Tang, Yizhi, Lei Dai, Orhan Sahin, Zuowei Wu, Mingyuan Liu, and Qijing Zhang. 2017. "Emergence of a Plasmid-Borne Multidrug Resistance Gene Cfr(C) in Foodborne Pathogen Campylobacter." *Journal of Antimicrobial Chemotherapy* 72 (6): 1581–88. https://doi.org/10.1093/jac/dkx023.

Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. "The Evolutionary Origin of Orphan Genes." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3053.

Taylor, Jeffery P., Ben J. Barnett, Lemuel Del Rosario, Karen Williams, and Suzanne S. Barth. 1998. "Prospective Investigation of Cryptic Outbreaks of Salmonella Agona Salmonellosis." *Journal of Clinical Microbiology* 36 (10): 2861–64. https://doi.org/10.1128/jcm.36.10.2861-2864.1998.

Tettelin, Hervé, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, et al. 2005. "Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial 'Pan-Genome.'" *Proceedings of the National Academy of Sciences of the United States of America* 102 (39). https://doi.org/10.1073/pnas.0506758102.

Tettelin, Hervé, David Riley, Ciro Cattuto, and Duccio Medini. 2008. "Comparative Genomics: The Bacterial Pan-Genome." *Current Opinion in Microbiology*. https://doi.org/10.1016/j.mib.2008.09.006.

Tolar, Beth, Lavin A. Joseph, Morgan N. Schroeder, Steven Stroika, Efrain M. Ribot, Kelley B. Hise, and Peter Gerner-Smidt. 2019. "An Overview of PulseNet USA Databases." *Foodborne Pathogens and Disease*. Mary Ann Liebert Inc. https://doi.org/10.1089/fpd.2019.2637.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. "Microbial Strain-Level Population Structure & Genetic Diversity from Metagenomes." *Genome Research* 27 (4): 626–38. https://doi.org/10.1101/gr.216242.116.

Turner, Claire E., Luke Bedford, Nicholas M. Brown, Kim Judge, M. Estée Török, Julian Parkhill, and Sharon J. Peacock. 2017. "Community Outbreaks of Group A Streptococcus Revealed by Genome Sequencing." *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-08914-x.

Udwary, Daniel W., Lisa Zeigler, Ratnakar N. Asolkar, Vasanth Singan, Alla Lapidus, William Fenical, Paul R. Jensen, and Bradley S. Moore. 2007. "Genome Sequencing Reveals Complex Secondary Metabolome in the Marine Actinomycete Salinispora Tropica." *Proceedings of the National Academy of Sciences of the United States of America* 104 (25): 10376–81. https://doi.org/10.1073/pnas.0700962104.

Underthun, Kristina, Jaysankar De, Alan Gutierrez, Rachael Silverberg, and Keith R. Schneider. 2018. "Survival of Salmonella and Escherichia Coli in Two Different Soil Types at Various Moisture Levels and Temperatures." *Journal of Food Protection* 81 (1): 150–57. https://doi.org/10.4315/0362-028X.JFP-17-226.

Vaughn, Eric L., Quynh T. Vo, Johanna Vostok, Tracy Stiles, Andrew Lang, Catherine M. Brown, R. Monina Klevens, and Lawrence Madoff. 2020. "Linking Epidemiology and Whole-Genome Sequencing to Investigate Salmonella Outbreak, Massachusetts, USA, 2018." *Emerging Infectious Diseases* 26 (7): 1538–41. https://doi.org/10.3201/eid2607.200048.

Vegge, Christina S., Lone Brøndsted, Małgorzata Ligowska-Marzeta, and Hanne Ingmer. 2012. "Natural Transformation of Campylobacter Jejuni Occurs Beyond Limits of Growth." *PLoS ONE* 7 (9): e45467. https://doi.org/10.1371/journal.pone.0045467.

Vicente, Cláudia M., Annabelle Thibessard, Jean Noël Lorenzi, Mabrouka Benhadj, Laurence Hôtel, Djamila Gacemi-Kirane, Olivier Lespinet, Pierre Leblond, and Bertrand Aigle. 2018. "Comparative Genomics among Closely Related Streptomyces Strains Revealed Specialized Metabolite Biosynthetic Gene Cluster Diversity." *Antibiotics* 7 (4). https://doi.org/10.3390/antibiotics7040086.

Vos, Michiel, and Xavier Didelot. 2009. "A Comparison of Homologous Recombination Rates in Bacteria and Archaea." *ISME Journal* 3 (2): 199–208. https://doi.org/10.1038/ismej.2008.93.

Vos, Michiel, and Adam Eyre-Walker. 2017. "Are Pangenomes Adaptive or Not?" *Nature Microbiology*. Nature Publishing Group. https://doi.org/10.1038/s41564-017-0067-5.

Vos, Michiel, Matthijn C. Hesselman, Tim A. te Beek, Mark W.J. van Passel, and Adam Eyre-Walker. 2015. "Rates of Lateral Gene Transfer in Prokaryotes: High but Why?" *Trends in Microbiology*. Elsevier Ltd. https://doi.org/10.1016/j.tim.2015.07.006.

Weber, Tilmann, Kai Blin, Srikanth Duddela, Daniel Krug, Hyun Uk Kim, Robert Bruccoleri, Sang Yup Lee, et al. 2015. "AntiSMASH 3.0—a Comprehensive Resource for the Genome Mining of Biosynthetic Gene Clusters." *Nucleic Acids Research* 43 (W1): W237–43. https://doi.org/10.1093/nar/gkv437.

Weimer, Bart C. 2017. “100K Pathogen Genome Project.” *Genome Announcements* 5 (28): e00594-17. https://doi.org/10.1128/genomeA.00594-17.

Weissbach, H., and N. Brot. 1991. “Regulation of Methionine Synthesis in Escherichia Coli.” *Molecular Microbiology*. https://doi.org/10.1111/j.1365-2958.1991.tb01905.x.

Whitehouse, Chris A., Shaohua Zhao, and Heather Tate. 2018. “Antimicrobial Resistance in Campylobacter Species: Mechanisms and Genomic Epidemiology.” In *Advances in Applied Microbiology*, 103:1–47. Academic Press Inc. https://doi.org/10.1016/bs.aambs.2018.01.001.

Wilson, Daniel J., Edith Gabriel, Andrew J.H. Leatherbarrow, John Cheesbrough, Steven Gee, Eric Bolton, Andrew Fox, C. Anthony Hart, Peter J. Diggle, and Paul Fearnhead. 2009. “Rapid Evolution and the Importance of Recombination to the Gastroenteric Pathogen Campylobacter Jejuni.” *Molecular Biology and Evolution* 26 (2): 385–97. https://doi.org/10.1093/molbev/msn264.

Wong, Anthony, Dirk Lange, Sebastien Houle, Nikolay P. Arbatsky, Miguel A. Valvano, Yuriy A. Knirel, Charles M. Dozois, and Carole Creuzenet. 2015. “Role of Capsular Modified Heptose in the Virulence of Campylobacter Jejuni.” *Molecular Microbiology* 96 (6): 1136–58. https://doi.org/10.1111/mmi.12995.

Wong, Marcus Ho Yin, Meiying Yan, Edward Wai Chi Chan, Kan Biao, and Sheng Chen. 2014. “Emergence of Clinical Salmonella Enterica Serovar Typhimurium Isolates with Concurrent Resistance to Ciprofloxacin, Ceftriaxone, and Azithromycin.” *Antimicrobial Agents and Chemotherapy* 58 (7): 3752–56. https://doi.org/10.1128/AAC.02770-13.

Woodcock, Dan J., Peter Krusche, Norval J. C. Strachan, Ken J. Forbes, Frederick M. Cohan, Guillaume Méric, and Samuel K. Sheppard. 2017. “Genomic Plasticity and Rapid Host Switching Can Promote the Evolution of Generalism: A Case Study in the Zoonotic Pathogen Campylobacter.” *Scientific Reports* 7 (1): 9650. https://doi.org/10.1038/s41598-017-09483-9.

Wotanis, Caitlin K., William P. Brennan, Anthony D. Angotti, Elizabeth A. Villa, Josiah P. Zayner, Alexandra N. Mozina, Alexandria C. Rutkovsky, Richard C. Sobe, Whitney G. Bond, and Ece Karatan. 2017. “Relative Contributions of Norspermidine Synthesis and Signaling Pathways to the Regulation of Vibrio Cholerae Biofilm Formation.” *PLoS ONE* 12 (10): e0186291. https://doi.org/10.1371/journal.pone.0186291.

Wyres, Kelly L., Ryan R. Wick, Louise M. Judd, Roni Froumine, Alex Tokolyi, Claire L. Gorrie, Margaret M.C. Lam, Sebastián Duchêne, Adam Jenney, and Kathryn E. Holt. 2019. “Distinct Evolutionary Dynamics of Horizontal Gene Transfer in Drug Resistant and Virulent Clones of Klebsiella Pneumoniae.” *PLoS Genetics* 15 (4): e1008114. https://doi.org/10.1371/journal.pgen.1008114.

Xu, Min, Yemin Wang, Zhilong Zhao, Guixi Gao, Sheng Xiong Huang, Qianjin Kang, Xinyi He, et al. 2016. “Functional Genome Mining for Metabolites Encoded by Large Gene Clusters through Heterologous Expression of a Whole-Genome Bacterial Artificial Chromosome Library in Streptomyces Spp.” *Applied and Environmental Microbiology* 82 (19): 5795–5805. https://doi.org/10.1128/AEM.01383-16.

Yang, Yichao, Kristina M. Feye, Zhaohao Shi, Hilary O. Pavlidis, Michael Kogut, Amanda J. Ashworth, and Steven C. Ricke. 2019. "A Historical Review on Antibiotic Resistance of Foodborne Campylobacter." *Frontiers in Microbiology*. Frontiers Media S.A. https://doi.org/10.3389/fmicb.2019.01509.

Youngblut, Nicholas D., Joseph S. Wirth, James R. Henriksen, Maria Smith, Holly Simon, William W. Metcalf, and Rachel J. Whitaker. 2015. "Genomic and Phenotypic Differentiation among Methanosarcina Mazei Populations from Columbia River Sediment." *ISME Journal* 9 (10): 2191–2205. https://doi.org/10.1038/ismej.2015.31.

Yu, Robert K., Seigo Usuki, and Toshio Ariga. 2006. "Ganglioside Molecular Mimicry and Its Pathological Roles in Guillain-Barré Syndrome and Related Diseases." *Infection and Immunity*. https://doi.org/10.1128/IAI.00967-06.

Zahrt, Thomas C., and Stanley Maloy. 1997. "Barriers to Recombination between Closely Related Bacteria: MutS and RecBCD Inhibit Recombination between Salmonella Typhimurium and Salmonella Typhi." *Proceedings of the National Academy of Sciences of the United States of America* 94 (18): 9786–91. https://doi.org/10.1073/pnas.94.18.9786.

Zankari, Ea, Hasman, Henrik, Cosentino, Salvatore, Vestergaard, Martin, Rasmussen, Simon, Lund, Ole, Aarestrup, Frank M., Larsen, Mette Voldby. 2012. "Identification of Acquired Antimicrobial Resistance Genes." *Journal of Antimicrobial Chemotherapy* 67 (11): 2640–44. https://academic.oup.com/jac/article/67/11/2640/707208.

Zautner, Andreas E., Carolin Ohk, Abdul Malik Tareen, Raimond Lugert, and Uwe Groß. 2012. "Epidemiological Association of Campylobacter Jejuni Groups with Pathogenicity-Associated Genetic Markers." *BMC Microbiology* 12: 171–171. https://doi.org/10.1186/1471-2180-12-171.

Zhang, Shaokang, Hendrik C. den Bakker, Shaoting Li, Jessica Chen, Blake A. Dinsmore, Charlotte Lane, A. C. Lauer, Patricia I. Fields, and Xiangyu Deng. 2019. "SeqSero2: Rapid and Improved Salmonella Serotype Determination Using Whole-Genome Sequencing Data." *Applied and Environmental Microbiology* 85 (23). https://doi.org/10.1128/AEM.01746-19.

Zhang, Wenjun, Brian D. Ames, Shiou Chuan Tsai, and Yi Tang. 2006. "Engineered Biosynthesis of a Novel Amidated Polyketide, Using the Malonamyl-Specific Initiation Module from the Oxytetracycline Polyketide Synthase." *Applied and Environmental Microbiology* 72 (4): 2573–80. https://doi.org/10.1128/AEM.72.4.2573-2580.2006.

Zhang, Xiaomei, Michael Payne, and Ruiting Lan. 2019. "In Silico Identification of Serovar-Specific Genes for Salmonella Serotyping." *Frontiers in Microbiology* 10 (APR): 835. https://doi.org/10.3389/fmicb.2019.00835.

Zhaxybayeva, Olga, W. Ford Doolittle, R. Thane Papke, and J. Peter Gogarten. 2009. "Intertwined Evolutionary Histories of Marine Synechococcus and Prochlorococcus Marinus." *Genome Biology and Evolution* 1 (January): 325–39. https://doi.org/10.1093/gbe/evp032.

Zhou, Zhan, Jianying Gu, Yong Quan Li, and Yufeng Wang. 2012. "Genome Plasticity and Systems Evolution in Streptomyces." *BMC Bioinformatics* 13 Suppl 10.

https://doi.org/10.1186/1471-2105-13-S10-S8.

Zhou, Zhemin, Nabil Fareed Alikhan, Khaled Mohamed, Yulei Fan, and Mark Achtman. 2020. "The EnteroBase User's Guide, with Case Studies on Salmonella Transmissions, Yersinia Pestis Phylogeny, and Escherichia Core Genomic Diversity." *Genome Research* 30 (1): 138–52. https://doi.org/10.1101/gr.251678.119.

Zhu, Ana, Shinichi Sunagawa, Daniel R. Mende, and Peer Bork. 2015. "Inter-Individual Differences in the Gene Content of Human Gut Bacterial Species." *Genome Biology* 16 (1): 82. https://doi.org/10.1186/s13059-015-0646-9.

Zhu, Bo, Muhammad Ibrahim, Zhouqi Cui, Guanlin Xie, Gulei Jin, Michael Kube, Bin Li, and Xueping Zhou. 2016. "Multi-Omics Analysis of Niche Specificity Provides New Insights into Ecological Adaptation in Bacteria." *ISME Journal* 10 (8): 2072–75. https://doi.org/10.1038/ismej.2015.251.

Zomorrodi, Ali R., and Daniel Segrè. 2017. "Genome-Driven Evolutionary Game Theory Helps Understand the Rise of Metabolic Interdependencies in Microbial Communities." *Nature Communications* 8 (1). https://doi.org/10.1038/s41467-017-01407-5.

Zotchev, Sergey B. 2014. "Genomics-Based Insights into the Evolution of Secondary Metabolite Biosynthesis in Actinomycete Bacteria." In *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, 35–45. Springer International Publishing. https://doi.org/10.1007/978-3-319-07623-2_2.