Universidad de **Cantabria**

2025

Programa de Doctorado en Biología Molecular y Biomedicina

Autora:
Arancha Peñil Celis

Directores:
María del Pilar Garcillán Barcia
Fernando de la Cruz Calahorra

Tesis Doctoral

INCLUSIÓN DEL PANGENOMA PARA LA VIGILANCIA GENÓMICA DE LOS SEROVARES TYPHI Y HADAR DE *SALMONELLA ENTERICA*

PhD Thesis

HARNESSING THE PANGENOME FOR THE GENOMIC SURVEILLANCE OF *SALMONELLA ENTERICA* SEROVARS TYPHI AND HADAR

UC | Escuela de **Doctorado**

UNIVERSIDAD DE CANTABRIA

PROGRAMA DE DOCTORADO EN

BIOLOGÍA MOLECULAR Y BIOMEDICINA

Tesis doctoral

Inclusión del pangenoma para la vigilancia genómica de los serovares Typhi y Hadar de *Salmonella enterica*

PhD Thesis

Harnessing the pangenome for the genomic surveillance of *Salmonella enterica* serovars Typhi and Hadar

Autora:

Arancha Peñil Celis

Directores:

María del Pilar Garcillán Barcia

Fernando de la Cruz Calahorra

Escuela de Doctorado de la Universidad de Cantabria, Santander, 14 de abril de 2025

This work has been carried out at “Instituto de Biomedicina y Biotecnología de Cantabria”, funded by the following projects:

- Structure and dynamics of *Salmonella* plasmids and their involvement in the dissemination of antibiotic resistance. Funded by Centers for Disease Control and Prevention (CDC). Contract No. 75D30119C06679.
- Regional and global risk estimates for the emergence and persistence of MDR and XDR strains of enterobacterial human pathogens. Funded by Centers for Disease Control and Prevention (CDC). Contract No. 75D30121C11978.
- MAPMAR: Marine plasmids driving the spread of antibiotic resistances. Project PCI2021-121978. Funded by MCIN/AEI/10.13039/501100011033 and by the European Union NEXTGENERATIONEU/PRTR.
- Supremacía de los plásmidos. Project PID2020-117923GB-I00. Funded by MICIU/AEI /10.13039/501100011033.

A mi hermana,

Irene

Agradecimientos

Con la escritura de esta tesis culmina una etapa muy significativa de mi vida, marcada por el aprendizaje y el crecimiento personal y profesional, en la que he estado rodeada de personas excepcionales.

En primer lugar, quiero agradecer a mis supervisores, Fernando y Mapi. Gracias por darme siempre vuestro extraordinario apoyo, por vuestra guía a lo largo de este proyecto y por haber estado siempre dispuestos a compartir conmigo vuestro tiempo y conocimiento.

Mapi, tengo mucho que agradecerte, pero quiero empezar por la confianza que depositaste en mi desde el primer momento al abrirme las puertas de tu laboratorio. Eso ha significado mucho para mí, de verdad. Gracias por haber estado siempre ahí, dispuesta a ayudar, a hablar, a apoyar y, sobre todo, a enseñarme. Es una verdadera suerte haber tenido de supervisora a una de las personas más generosas y bondadosas, algo que ha hecho que este camino sea, sin duda, infinitamente mejor y más fácil.

Fernando, gracias por todo lo que me has enseñado, y por todas nuestras reuniones y viajes que nos han llevado desde profundas conversaciones científicas hasta explorar temas tan diversos como la llegada de trabajadores japoneses a Canadá. Guardo con cariño incontables horas de reflexión y charlas enriquecedoras, que reflejan tu pasión por la enseñanza, de la que tanto he disfrutado y aprendido.

Santi, gracias por enseñarme este mundo de la bioinformática, he aprendido muchísimo gracias a ti. Agradezco tu infinita paciencia a lo largo de este proceso, así como tu constante disposición a enseñarme, explicarme y supervisarme. Has sido una pieza fundamental en el desarrollo de esta tesis.

Kaitlin and Hattie, this thesis would never have come together without you. You both know how essential your contributions have been. I will always be grateful for your trust, your support, your understanding and encouragement when I needed it most. Thank you also for generously sharing your knowledge, which helped shape this thesis into something I am truly proud of. Whenever I think about this thesis, I will always think fondly of both of you as indispensable parts of the journey.

Y cuando se empieza una tesis doctoral, una puede ser más o menos consciente del papel fundamental que van a tener sus supervisores y el proyecto en sí en el desarrollo de la tesis. Sin embargo, de lo que se es menos consciente al principio y que resulta ser uno de los

pilares más importantes son los compañeros/amigos/familia que haces en el laboratorio, y yo he tenido la suerte de encontrarme con los mejores. Siempre seréis unas personas súper especiales para mí.

Mariadel, me siento muy afortunada de haberte tenido al lado estos años. Hemos vivido todas las etapas del doctorado a la vez, incluso trabajando a veces en cosas parecidas, lo cual nos ha hecho ayudarnos y necesitarnos mucho. Realmente no tengo palabras para agradecerte todo el cariño, el apoyo y los ánimos que me has dado durante todo este camino. Todo ha sido más fácil gracias a ti, especialmente en esta última etapa. **Dani**, tú nos cambiaste el laboratorio, ha sido una auténtica suerte tenerte al lado todos estos años. Gracias por tu apoyo, por estar siempre para animar, por hacer el día a día más bonito y divertido. **Irene**, siempre serás una persona muy especial en esta etapa del IBBTEC que empezamos a la vez. Muchas gracias por estar siempre que lo he necesitado, tu apoyo siempre ha sido muy importante para mí y esa alegría y energía que te caracterizan me han alegrado muchos días.

A todos mis amigos de "Las dos perfumadas" a quienes admiro profundamente, tanto en lo personal como profesional: **Lolu, David, Antonio, Alfonso, Pablo, Aurora, Marina, Carlos, Andrea, Luisa, Celia** y **Patri**. Cuántas horas hemos compartido de cafés, celebraciones, bingos, tortillas y comidas que realmente me han dado la vida. Cada uno de vosotros ha formado una parte esencial de este viaje. Gracias de corazón por estar ahí, por compartir tanto, y por hacer que cada día en el laboratorio sea más feliz.

Gracias al resto de personas del laboratorio por ser también partícipes de esta etapa: Yelina, Marta, Miguel; y a los primos: María, Iván y Arturo. En especial, quiero agradecer a Juanma por estar siempre tan dispuesto a ayudar, es una suerte saber que siempre puedo contar contigo. Raúl y Elena, gracias por vuestros consejos en cada evaluación que han sido fundamentales para la versión final de esta tesis.

Un agradecimiento muy especial para los pilares del laboratorio: Sheila, Ana y Raquel. Gracias a vuestra gran experiencia, fruto de haber acompañado tan de cerca a tantos doctorandos, vuestros consejos siempre han sido acertados y fundamentales. En particular, quiero agradecer a Sheila (y, por supuesto, a Carol) por darme la mano siempre que lo necesité durante mis primeros pasos en el IBBTEC y por enseñarme todo en el wet lab; a Ana por estar siempre dispuesta a ayudar y por todo el apoyo cuando fuimos compis de lab; y a Raquel por tu alegría y la increíble energía que transmites cada día.

Cuando entras en el mundo de la ciencia, especialmente en el camino de una tesis doctoral, muchas veces esta te absorbe completamente. Por eso, tener siempre cerca a mis amigas de Sanvi me ha ayudado a mantener la perspectiva. Siempre estaré profundamente agradecida por vuestro cariño y por estar siempre ahí, en cada etapa. Gracias Ángela, Laura C., Elsa, Mara, Rocío, Laura G., María; a mis niñas Sofía, Mara, Candela y Camila, y a mi ahijado Romeo.

La familia siempre es el gran pilar en todas las etapas de la vida. Irene, porque todo lo que haga yo en la vida, también será tuyo. Y todo lo que haga yo en la vida estará muy influido por todo lo que tú me enseñas sobre ella. Gracias a ello, esta tesis también ha salido adelante. Un agradecimiento infinito y especial a mi abuela, por estar siempre a mi lado, con ese amor que todo lo sostiene. Y, por supuesto, he llegado hasta aquí gracias a mis padres. Sin ellos, esto no hubiese sido posible.

Finalmente, agradecer a mi compañero de piso, mi amigo y mi todo, Miguel. Gracias por apoyarme siempre, por esperarme en casa, llegue a la hora que llegue, por no entender muchas cosas de este trabajo, pero seguir ahí, haciendo los días difíciles un poco más fáciles y los días buenos, aún mejores. A ti y, por supuesto, a tu familia, Nando, Cristi, María y Marcos, gracias siempre por estar. Y a D.

De corazón, gracias.

Arancha

TABLE OF CONTENTS

Abbreviations

AMR	Antimicrobial Resistance
ANI	Average Nucleotide Identity
APHIS	Animal and Plant Health Inspection Service
ARS	Agricultural Research Service
BGMM	Beta-Gaussian Mixture Modeling
BIC	Bayesian Information Criterion
bp	base pairs
BYPAS	Backyard Poultry-Associated Salmonellosis
CDC	Centers for Disease Control and Prevention
CDS	Coding Sequences
CFSAN	Center for Food Safety and Applied Nutrition
cgMLST	core-genome Multi-locus Sequence Typing
cgST	core genome Sequence Type
CI	Confidence Intervals
CipR	Ciprofloxacin-resistant
CipNS	Ciprofloxacin non-susceptibility
COVID-19	Coronavirus disease 2019
CVM	Center for Veterinary Medicine
eBGs	eBurst Groups
FDA	Food and Drug Administration
GBD	Global Burden of Diseases, Injuries, and Risk Factors Study
GLD	Genome Length Distance
HCC	Hierarchical Clustering of cgST
HGT	Horizontal Gene Transfer
HPC	Homologous Protein Clusters
ICE	Integrative Conjugative Element
IME	Integrative and Mobilizable Elements
Indel	Insertion/deletion
iNTS	invasive Non-Typhoidal *Salmonella*
IQR	Interquartile Range
IS	Insertion Sequence

JI	Jaccard Index
JI-group	Jaccard Index group
JINA	Jaccard Index Network Analysis
Kb	Kilobase
Mb	Megabase
MDR	Multi Drug Resistance
MGE	Mobile Genetic Element
ML	Maximum Likelihood
MLST	Multi-locus Sequence Typing
MLVA	Multilocus Variable-Number Tandem-Repeat Analysis
MPF	Mating Pair Formation
MSTree	Minimum Spanning Tree
NAHLN	National Animal Health Laboratory Network
NARMS	National Antimicrobial Resistance Monitoring System
NCBI	National Center for Biotechnology Information
NTS	Non-Typhoidal *Salmonella*
NVSL	National Veterinary Services Laboratories
OMS	Outer Membrane Vesicles
OR	Odds Ratios
ORA	Office of Regulatory Affairs
oriT	origin of Transfer
PATO	Pangenome Analysis Toolkit
PCR	Polymerase Chain Reaction
PFGE	Pulse-field Gel Electrophoresis
PHLs	Public Health Laboratories
PMNs	Polymorphonuclear leukocytes
PMQR	Plasmid-Mediated Quinolone Resistance
PopPUNK	Population Partitioning Using Nucleotide K-mers
PTU	Plasmid Taxonomic Unit
QRDR	Quinolone Resistance Determining Region
REP	Reoccurring, Emerging, or Persisting
rMLST	ribosomal Multi-locus Sequence Typing
SCV	*Salmonella*-Containing Vacuole

SNP	Single Nucleotide Polymorphism
SPI	*Salmonella* Pathogenicity Island
ST	Sequence Type
T3SS	Type III Secretion Systems
T4CP	Type IV Coupling Protein
T4SS	Type IV Secretion System
U.K.	United Kingdom
U.S.	United States
UI	Uncertainty interval
USDA-FSIS	The United States Department of Agriculture's Food Safety and Inspection Service
Vet-LIRN	Veterinary Laboratory Investigation and Response Network
ViCPS	Vi Capsular Polysaccharide
wgMLST	whole genome Multi-locus Sequence Typing
WGS	Whole-Genome Sequencing
Zot	Zonular occludens toxin protein

List of tables

Supplementary online tables ***(link available in Section 3.9)***

List of figures

Abstract

Bacterial relatedness measured using selected chromosomal loci forms the basis of public health genomic surveillance. While approximating vertical evolution through this approach has proven exceptionally valuable for understanding pathogen dynamics, it excludes a fundamental dimension of bacterial evolution, horizontal gene transfer. Incorporating the accessory genome is the logical remediation and has recently shown promise in expanding epidemiological resolution for enteric pathogens. Employing *k*-mer-based Jaccard Index (JI) analysis, optionally complemented by a novel genome length distance metric (GLD), we computed pangenome relatedness for two serovars of *Salmonella,* Typhi and Hadar. This method simultaneously captures both vertical (homology-by-descent) and horizontal (homology-by-admixture) evolutionary relationships in a reticulate network. This dual perspective provided high-resolution stratification of closely related genomes and highlighted epidemiologically relevant subgroups that traditional core genome or gene-by-gene methods may have overlooked. The addition of GLD further enhanced the detection of insertions and deletions in populations that harbor core genome variation.

To investigate the pangenome *of Salmonella enterica* serotype Typhi, a pathogen of significant global public health concern, we analyzed the largest United States (U.S.) dataset to date, containing over 2,200 genomes. This analysis revealed a non-random structure in the Typhi pangenome primarily driven by the gain and loss of mobile genetic elements (MGEs). It confimed and expanded upon known epidemiological patterns, uncovered novel plasmid dynamics, and identified avenues for further genomic epidemiological exploration. Notably, our analysis elucidated the existence of two distinct IncY plasmids containing the $bla_{\text{CTX-M-15}}$ gene that belong to different plasmid taxonomic units (PTUs). This finding challenges previous classifications that treated these plasmids as a single type and emphasizes the need for more detailed analyses that go beyond plasmid replicons and known antimicrobial resistance (AMR) genes. Further, JI-grouping robustly linked specific clusters to globally significant Typhi lineages, such as the 4.3.1 multi-drug resistant (MDR) and extensively drug-resistant (XDR) strains. Genomes harboring the $bla_{\text{CTX-M-15}}$ gene integrated into the chromosome were separated from those in which the gene remained plasmid-borne. We further documented novel chromosomal integrations of $bla_{\text{CTX-M-15}}$, highlighting the dynamic nature of resistance evolution in Typhi. Our results also confirmed known epidemiological patterns, including the emergence of XDR Typhi strains and their

geographic associations with travel-related cases in regions like Pakistan, and extended these findings by demonstrating that the overall pangenome structure observed in U.S. isolates is generalizable to global populations. Moreover, analysis of pre-antibiotic era isolates showed that certain Typhi lineages were established before the advent of antibiotics and have circulated with minimal genetic change, underscoring the long-term persistence and evolutionary stability of these lineages.

With public health applications in mind, this work highlights the diversity of the accessory genome and provides a valuable complement to traditional GenoTyphi typing by supplying an additional layer of genomic information that enhances strain discrimination and deepens our understanding of Typhi's evolutionary dynamics.

In the case of *Salmonella enterica* serotype Hadar, an emerging zoonotic pathogen in the U.S. linked to both commercial and backyard poultry, we explored the population structure and epidemiology in the U.S. using over 3,300 genomes. Between 2019 and 2020, U.S. Hadar populations experienced substantial shifts driven by the expansion of a lineage carrying a previously uncommon prophage-like element, which was detected in both backyard and commercial poultry. This emergent lineage likely gained a selective advantage through the acquisition of this novel prophage. While no distinct pangenomic differences were detected between strains isolated from backyard versus commercial poultry, we did observe a division within this lineage involving the presence or absence of a PTU-I1 plasmid. The JI-based clustering allowed for finer discrimination of closely related isolates, thereby facilitating more accurate linkage between human cases and putative exposure sources. Examination of non-U.S. Hadar populations revealed geographical variations in pangenome diversity, suggesting that factors such as poultry trade, farming practices, and regional ecology play significant roles in shaping global population structures. Additionally, our study highlighted that PTU-I1 plasmids are the most frequent plasmid type among Hadar genomes. Although resistance genes have not yet posed a major threat in Hadar, the presence of new AMR determinants on these broad-host range plasmids signals the potential for future horizontal gene transfer events that could lead to more problematic resistance profiles.

Moreover, this work emphasizes that short-term population shifts in both Typhi and Hadar are frequently driven by the gain or loss of MGEs. As these elements continue to evolve, there is a clear need for consistent and updated genomic surveillance using flexible bioinformatic workflows capable of detecting newly emerging MGEs. Such efforts are

particularly important for tracking the evolution of antimicrobial resistance and the spread of novel plasmids and prophages across different environments and geographic regions. Incorporating accessory genome data into public health genomics deepens our understanding of pathogen adaptation and specialization. The demonstrated utility of the JI-based approach for both Typhi and Hadar underscores its potential for application in other pathogens.

Resumen

El estudio de la relación genética entre bacterias a través de loci cromosómicos específicos constituye la base de la vigilancia genómica en salud pública. Aunque este enfoque basado en la evolución vertical ha sido sumamente valioso para descifrar la dinámica de los patógenos, deja de lado un aspecto crucial de la evolución bacteriana: la transferencia genética horizontal. La integración del genoma accesorio en el análisis genómico permite evaluar esta dimensión de la evolución, mostrando un notable potencial para aumentar la resolución epidemiológica en patógenos entéricos.

Empleando un análisis basado en *k*-meros, el Índice de Jaccard (JI), complementado opcionalmente con una nueva métrica de distancia basada en la longitud del genoma (GLD), se ha calculado la relación del pangenoma para dos serovares de *Salmonella*, Typhi y Hadar. Este método permite capturar de manera simultánea las relaciones evolutivas verticales (homología por descendencia) y horizontales (homología por mezcla) en una red reticulada. Este enfoque dual ha permitido una estratificación de alta resolución de genomas estrechamente relacionados, destacando subgrupos epidemiológicamente relevantes que los métodos tradicionales basados en el genoma central o el análisis gen a gen habrían pasado por alto. La incorporación de GLD optimizó aún más la detección de inserciones y deleciones en poblaciones que presentaban variación en el genoma central.

Para explorar el pangenoma de *Salmonella enterica* serotipo Typhi, un patógeno muy relevante para la salud pública global, se analizó el mayor conjunto de genomas aislados en Estados Unidos (EE. UU.) hasta la fecha, que comprende más de 2,200 genomas. Este análisis reveló una estructura no aleatoria en el pangenoma de Typhi, impulsada predominantemente por la ganancia y pérdida de elementos genéticos móviles (MGEs), lo cual confirma y amplía los patrones epidemiológicos conocidos, desvela nuevas dinámicas plasmídicas e identifica vías para futuras investigaciones en epidemiología genómica. Concretamente, se identificaron dos tipos de plásmidos IncY distintos que contienen el gen $bla_{\text{CTX-M-15}}$ y pertenecen a diferentes unidades taxonómicas plasmídicas (PTUs). Este hallazgo desafía las clasificaciones previas que consideraban estos plásmidos como un único tipo y subraya la necesidad de análisis más detallados que trasciendan los replicones plasmídicos y los genes de resistencia a antibióticos conocidos. Asimismo, la agrupación basada en el JI vinculó grupos específicos a linajes globalmente significativos de Typhi, tales

como las cepas 4.3.1 multirresistentes (MDR) y extensamente resistentes (XDR). Los grupos de genomas que portan el gen *bla*$_{\text{CTX-M-15}}$ integrado en el cromosoma se distinguieron de aquellos en los que el gen estaba codificado en un plásmido. Además, se documentaron nuevas integraciones cromosómicas de *bla*$_{\text{CTX-M-15}}$, subrayando la naturaleza dinámica de la evolución de la resistencia en Typhi. Los resultados de este estudio también confirmaron patrones epidemiológicos conocidos, incluida la vinculación de cepas XDR con viajes a regiones como Pakistán. Este trabajo demuestra que la estructura del pangenoma observada en los aislados de EE.UU. es representativa de las poblaciones a nivel mundial. Por otro lado, el análisis de aislados de la era pre-antibiótica evidenció que ciertos linajes de Typhi se establecieron antes de la introducción de los antibióticos y han circulado con mínimas modificaciones genéticas, destacando su persistencia a largo plazo y estabilidad evolutiva.

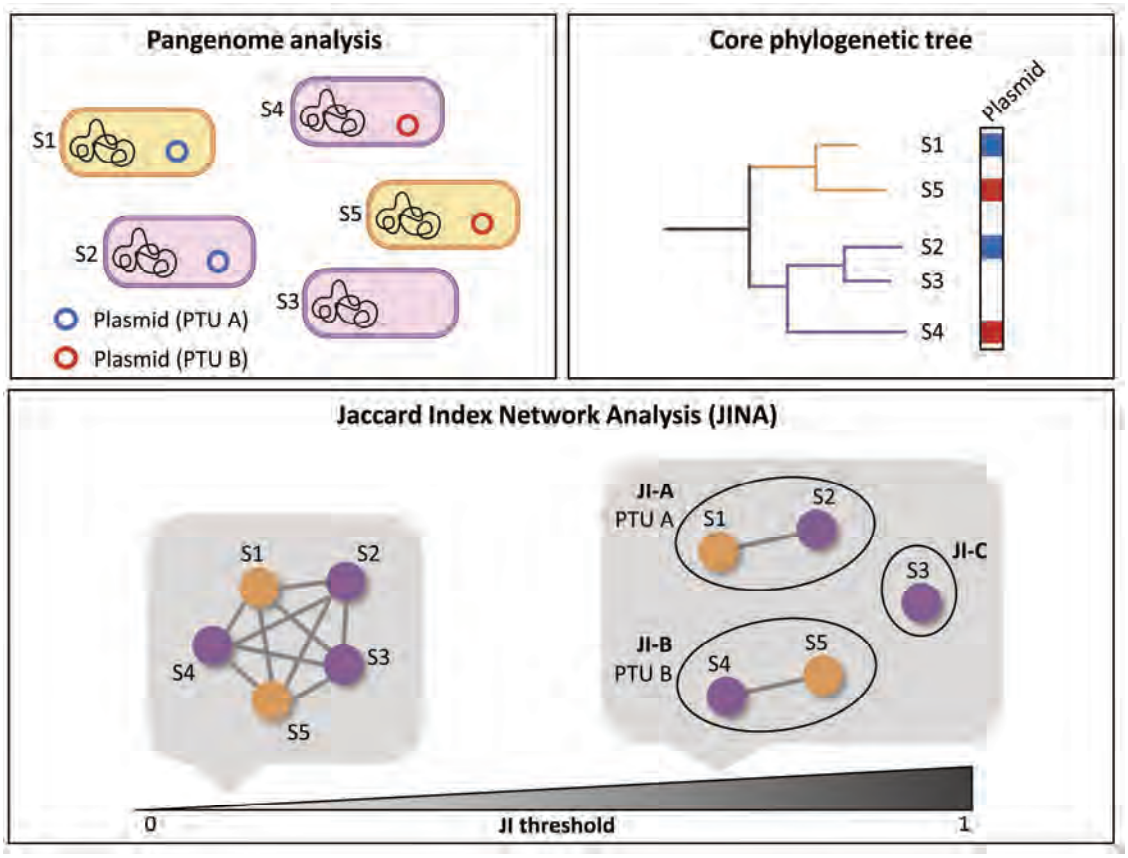
Enfocado en aplicaciones en salud pública, este trabajo demuestra la diversidad del genoma accesorio y complementa la tipificación tradicional de GenoTyphi, aportando una capa adicional de información genómica que mejora la discriminación de cepas y profundiza el entendimiento de la dinámica evolutiva de Typhi.

En el caso de *Salmonella enterica* serotipo Hadar, un patógeno zoonótico emergente en EE.UU. vinculado tanto a la avicultura comercial como a la doméstica, se exploró la estructura poblacional y la epidemiología en EE.UU., utilizando más de 3,300 genomas. Durante el período 2019-2020, las poblaciones de Hadar experimentaron cambios significativos, impulsados por la expansión de un linaje portador de un profago previamente poco común, identificado tanto en brotes relacionados con aves comerciales como con domésticas. Este linaje emergente podría haber adquirido una ventaja selectiva a través de la incorporación de este nuevo profago. Aunque no se detectaron diferencias pangenómicas distintivas entre cepas aisladas de aves domésticas y de entornos comerciales, se observó una diferenciación dentro de este linaje en función de la presencia o ausencia de un plásmido PTU-I1. Asimismo, la agrupación mediante el JI permitió una discriminación de aislados estrechamente relacionados, facilitando la vinculación entre casos humanos y las posibles fuentes de exposición. El análisis de poblaciones de Hadar fuera de EE. UU. ha revelado variaciones geográficas en la diversidad del pangenoma, sugiriendo que factores como el comercio avícola, las prácticas agrícolas y la ecología regional influyen en la configuración de las estructuras poblacionales globales. Además, el estudio destacó que los plásmidos PTU-I1 son el tipo de plásmido más frecuente entre los genomas de Hadar. Aunque los genes de resistencia aún no constituyen una amenaza en Hadar, la presencia de nuevos

determinantes de resistencia antimicrobiana en estos plásmidos de amplio rango de hospedador indica el potencial para futuros eventos de transferencia horizontal de genes, los cuales podrían derivar en perfiles de resistencia más problemáticos.

Finalmente, este trabajo enfatiza que los cambios poblacionales a corto plazo en Typhi y Hadar se deben con frecuencia a la ganancia o pérdida de MGEs. Estos elementos continúan evolucionando y movilizándose, por lo que se necesita una vigilancia genómica constante y actualizada, apoyada en herramientas bioinformáticas flexibles capaces de detectar MGEs emergentes. Dichos esfuerzos son especialmente cruciales para el seguimiento de la evolución de la resistencia a los antimicrobianos y la diseminación de nuevos plásmidos y profagos a través de diversos entornos y regiones geográficas. En conclusión, la integración de datos del genoma accesorio en la genómica aplicada a la salud pública enriquece nuestra comprensión de la adaptación y especialización de los patógenos. La utilidad demostrada del enfoque basado en el JI tanto para Typhi como para Hadar resalta su potencial de aplicación en otros patógenos.

Graphical abstract



The core genome reflects chromosomal phylogeny but does not account for accessory genome components. In contrast, JINA integrates all genomic information, revealing similarities that are invisible to traditional phylogenetic methods. The combination of these complementary approaches enabled the detailed stratification of two serovars of *Salmonella enterica*, Typhi and Hadar, revealing epidemiologically relevant subgroups that traditional core genome methods alone may overlook.

CHAPTER 1: INTRODUCTION

1.1 Bacterial pathogens population

Pathogenic bacteria pose a significant and growing threat to global public health, causing widespread disease outbreaks that demand effective monitoring and control strategies [1]. Their rapid evolution, coupled with increased global connectivity, complicates efforts to predict and mitigate infections. Consequently, accurately identifying and distinguishing pathogenic strains is essential, not only for epidemiological surveillance but also for the design of targeted public health interventions.

An estimated 60% of all human pathogens have originated from other animal species [2]. Conversely, human-to-animal transmissions threaten sustainable livestock production [3,4], while emerging plant pathogens increasingly jeopardize crop yields and global food security. Therefore, a critical aspect of bacterial evolution is their ability to adapt to new host species, which represents a considerable risk to both human health and global ecosystems. Investigating these host transitions provides valuable insights into the evolutionary mechanisms enabling bacteria to establish themselves in new niches and may uncover novel targets for infection control and guide strategies to limit the emergence of new pathogens.

Bacteria exist as complex communities or populations of organisms shaped by various selective pressures, including interactions with hosts (e.g. host immunity), environmental and ecological factors, and human interventions such as drug treatments. These factors drive changes in the genetic structure and the evolutionary trajectory of bacterial populations, leading to differences within and between species [5–7]. Studying the genetic diversity of bacterial pathogen populations can provide valuable insights into their evolutionary history, population dynamics, host interactions, and adaptive strategies. In turn, these insights help elucidate clinically important traits such as virulence, drug resistance, and antigenic variation, which are essential for designing effective medical and public health interventions [8].

1.1.1 Evolution and variation in pathogen populations

Bacterial diversity is a product of adaptation to specific ecological niches. Each host species (animal and plant kingdom) represents a unique collection of sub-niches that bacteria may colonize. When bacteria transfer to a new host, they frequently face new challenges, including interactions with the host and its microbiota. Such interactions require a balance

of survival, proliferation, and competition. The speed of adaptation depends on factors such as the frequency of beneficial mutations, the fitness advantage conferred by these mutations, the effective population size, or the capacity to acquire new genetic material (encoding beneficial traits) through horizontal gene transfer (HGT) [6,8]. Higher values for these factors can accelerate adaptation, giving certain strains a selective advantage and enabling them to establish themselves in the new host. This establishment can disrupt the balance of the resident microbiota, potentially leading to disease in the new host.

The lifestyle of a pathogen, whether it is obligate, opportunistic, or accidental, also shapes its evolutionary trajectory and population structure [9]. Obligate pathogens rely on host infection for survival, while opportunistic pathogens can survive outside the host and may cause disease in some individuals but not others. Accidental pathogens, on the other hand, cause disease by chance without benefiting from it in terms of transmission.

A crucial aspect of bacterial pathogens is the dynamic evolution of their genomes, which allows microorganisms to adapt to new environments and ecological niches. These evolutionary processes include the acquisition of new genes, homologous recombination, gene deletion, and point mutations [10,11].

Gene gain may occur via duplication (duplications can emerge from errors in DNA replication or asymmetric recombination) or through HGT [11]. By contrast, gene loss involves the shedding of unnecessary genes to reduce metabolic costs [12]. Point mutations, which can arise spontaneously during DNA replication, constitute a fundamental source of genetic variation. Most of these mutations are neutral or slightly deleterious, but a small fraction may confer a selective advantage, promoting adaptation to new niches or enhancing resistance to external pressures [10,13]. When such mutations arise, they are passed on through vertical inheritance. Finally, homologous and non-homologous recombination events can rearrange or replace existing genetic material, further contributing to bacterial diversity (**Figure 1.1**) [11,14].

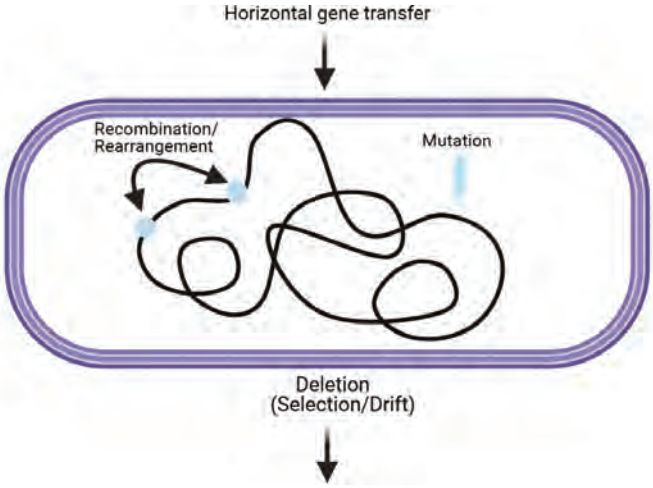


Figure 1.1 | Scheme of the main evolutionary mechanisms in bacteria. The diagram illustrates key evolutionary processes in bacteria, including HGT, genetic recombination/rearrangement, mutations, and gene deletion through selection or genetic drift.

1.1.1.1 Horizontal gene transfer

Horizontal gene transfer is a major driver of bacterial evolution, enabling the acquisition of new genetic material, often in the form of mobile genetic elements (MGEs) like bacteriophages, pathogenicity islands, transposons, insertion sequences (ISs), and plasmids. MGEs can integrate into the bacterial chromosome or replicate autonomously as extra-chromosomal elements, facilitating the rapid acquisition of adaptive traits such as antibiotic resistance, virulence factors, and new metabolic capabilities, which greatly accelerate bacterial adaptation and diversification [15]. HGT occurs through several mechanisms: transformation, conjugation and transduction. In transformation, a bacterium takes up extracellular DNA release into the environment by a donor organism. During conjugation, direct contact between cells enables the transfer of plasmid DNA [16]. In transduction, DNA from a donor cell is delivered via a bacteriophage and integrated into the recipient genome as a prophage [17]. Prophage DNA may contain transposable elements that facilitate movement to other chromosomal locations [18]. Transformation and transduction events result in either homologous recombination, where DNA from another lineage replaces homologous sequences, or non-homologous recombination, which involves the insertion of novel genetic material [19,20].

Conjugation is arguably the most common mechanism of HGT [21], enabling the movement of large DNA fragments containing diverse adaptive traits [22]. These adaptive

traits are encoded either in autonomously replicating conjugative plasmids or in integrative conjugative elements (ICEs) inserted in the bacterial chromosome. Indeed, they are major vehicles for the spread of antimicrobial resistance (AMR) genes [23,24]. This is a process renowned for its promiscuity in transferring genetic material across diverse bacterial species, including those separated by significant phylogenetic distances [25].

Conjugative systems typically carry two modules: a mobility (MOB) module responsible for conjugative DNA processing and a mating pair formation (MPF) module that facilitates DNA delivery through the membranes of donor and recipient bacteria [22]. The MOB module includes an origin of transfer (*oriT*), a short DNA sequence required in *cis* for plasmid mobility, a relaxase to initiate conjugation and a type IV coupling protein (T4CP) to interconnect DNA processing with DNA transport. Meanwhile, MPF module encodes for a complex of proteins that build the type IV secretion system (T4SS) and the conjugation pilus. According to mobility, plasmids can be classified into three categories: conjugative, mobilizable and non-mobilizable. A conjugative plasmid contains the two sets of genes necessary for their own transfer, whereas a mobilizable plasmid lacks *MPF* genes, and generally the T4CP, and instead relies on these components from a co-resident, self-transmissible element. Moreover, some mobilizable plasmids are even more streamlined, containing only the *oriT*, which is sufficient to initiate the DNA processing necessary for conjugation [26]. Non-mobilizable plasmids cannot be transferred by conjugation (**Figure 1.2**).

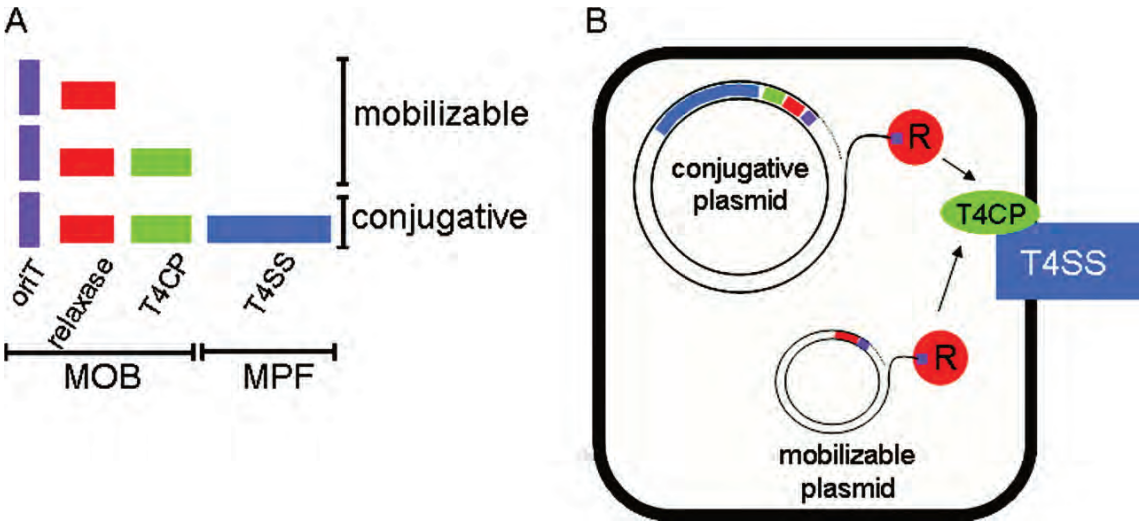


Figure 1.2 | Genetic organization of plasmids and key steps in the conjugation process. **(A)** Schematic view of the genetic elements involved in conjugation. **(B)** Scheme of interactions in the process of conjugation. The relaxase cleaves a specific site within *oriT*, and this step starts

conjugation. The DNA strand that contains the relaxase protein covalently bound to its 5′ end is displaced. The relaxase interacts with the T4CP and then with other components of the T4SS. As a result, it is transported to the recipient cell. Subsequently, the DNA is pumped into the recipient by the ATPase activity of the T4CP. Text adapted from the original legend and figure taken from [22].

Relaxases are enzymes that play a central role in the transfer of plasmids and, as they are common to most conjugative elements, serve as excellent molecular markers for classification of plasmids [27]. Based on MOB sequences, mobilizable plasmids are classified into nine MOB families: MOB_F, MOB_H, MOB_Q, MOB_C, MOB_P, MOB_T, MOB_B, MOB_M, and MOB_V [28]. The tool MOBscan enables the identification and classification of MOBs genes, which has provided crucial evolutionary insights in understanding the diversity of plasmid transfer processes across different bacterial populations. Another traditional approach for plasmid typing focuses on replicons, which are sequences within the plasmid that contain the origin of replication (*ori*), which is essential for the autonomous replication of the plasmid inside the host cell [29]. They determine the plasmid's ability to replicate independently of the bacterial chromosome and dictate its copy number, stability, and compatibility with other plasmids present within the same host. PlasmidFinder is a tool that uses replicon sequence data to identify and classify plasmids [30]. Both relaxase-based and replicon-based methods rely on detecting a single gene (the relaxase or the replicon) and therefore have inherent limitations [31]. However, recent advances in plasmid classification, as seen in systems like MOB-suite [32] and Plasmid Taxonomic Units (PTUs) [25], leverage the complete plasmid sequence to provide a more comprehensive and accurate grouping. These approaches offer a more robust framework for understanding plasmid evolution across diverse bacterial populations.

1.1.1.2 Gene loss

Although bacteria frequently acquire new genes, they also undergo gene loss to maintain optimal genome size and reduce metabolic burdens. Genes that confer no clear advantage in a stable niche may be lost over time. This “use it or lose it” phenomenon can be viewed as an adaptive strategy; by discarding unnecessary DNA, cells conserve resources and improve fitness [11,12].

1.1.1.3 Vertical inheritance

In addition to HGT, vertical inheritance also plays an important role in bacterial evolution [6,8,20,33]. Through vertical inheritance, genetic material is passed from parent to offspring, allowing evolutionary changes, such as mutations, to accumulate over generations. Point mutations involve the substitution of a few nucleotides or small insertions and deletions (indels). Mutations within protein-coding sequences can influence host adaptation by altering protein function and driving new evolutionary trajectories or by causing gene function loss. The majority of mutations in bacterial genomes are neutral, while some are detrimental and eliminated by purifying selection, and only a few increase fitness, spreading through the population via positive selection.

The dynamics of recombination and mutation also influence the population structure of bacterial pathogens. Single nucleotide polymorphisms (SNPs) contribute to genome divergence, but recombination can reduce this divergence. Recombination rates vary significantly between lineages, complicating phylogenetic analysis [34]. The relative importance of SNP accumulation versus recombination in bacterial evolution is debated [20]. In clonal populations, mutations accumulate vertically, leading to the propagation of fixed lineages. In contrast, in populations with frequent recombination, genetic material is constantly reshuffled, resulting in higher genetic diversity and a non-clonal population structure [35].

While vertical inheritance promotes gradual genetic changes, HGT introduces genetic material from unrelated individuals, facilitating the rapid acquisition of new traits. Therefore, HGT is a fundamental driver of bacterial evolution, introducing substantial genetic variability even within a single species. Unlike in eukaryotes, where genetic variation primarily arises from allelic differences, bacterial strains of the same species can possess entirely distinct genes and large genomic regions, acquired through HGT. This dual mechanism of vertical inheritance and HGT shapes the remarkable plasticity of bacterial genomes.

1.1.2 Bacterial pangenomes

Bacterial genomic plasticity is reflected in the structure of their genomes, which are often divided into distinct components [36]. The core genome comprises genes shared by all strains of a species. The accessory genome includes genes found in only some strains, often

encoding traits that confer ecological or evolutionary advantages, such as antibiotic resistance. Together, the core and accessory genomes form the pangenome, which represents the full genetic repertoire of a species (**Figure 1.3**).

Bacterial pangenomes can be classified as either open or closed [36,37]. Species with open pangenomes exhibit extensive gene content variability, largely driven by frequent HGT events and differential gene loss, with gene duplications playing a smaller role. In contrast, species with closed pangenomes show limited gene content variation (**Figure 1.3**). The size and nature of the pangenome of a species depend on the genetic diversity sampled and the number of genomes sequenced from that diversity.

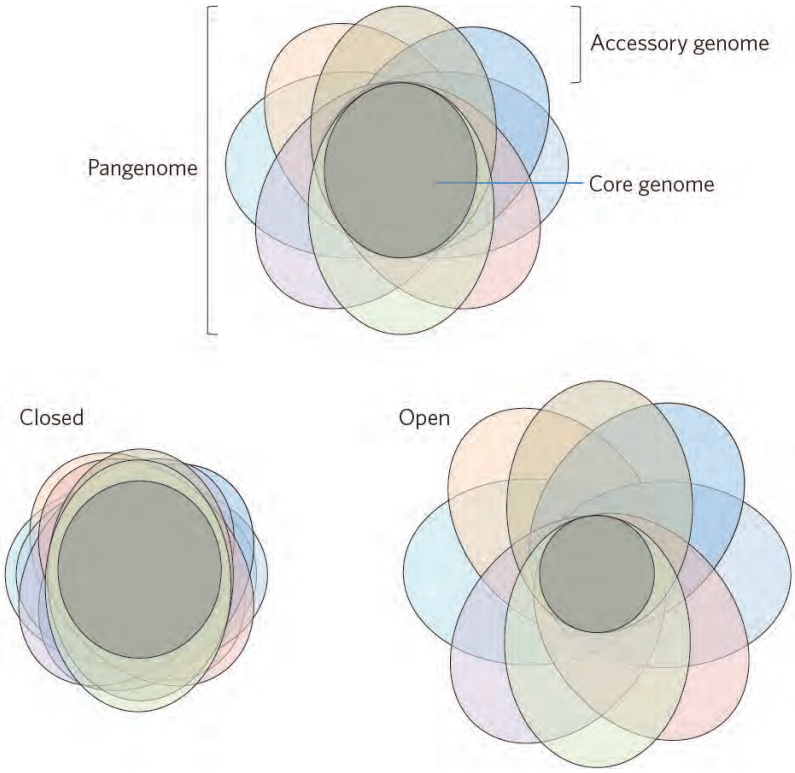


Figure 1.3 | Schematic representation of pangenomes as Venn diagrams. Each colored circle represents a bacterial genome, while the gray circle indicates the core genome. Text adapted from the original legend and figure taken from [36].

Different evolutionary forces shape the core and accessory genomes. The core genome evolves primarily through vertical inheritance, with mutations playing a major role in driving gradual genetic changes [20]. In contrast, the accessory genome is largely influenced by HGT [37]. This distinction highlights the complementary roles of vertical and horizontal evolutionary mechanisms in shaping bacterial genomes, with mutations

maintaining core functionality and HGT driving the dynamic plasticity of the accessory genome.

1.2 Methods for studying bacterial pathogen populations

Accurately identifying and distinguishing pathogen strains is essential for understanding their population structure, tracking outbreaks, monitoring disease spread, and designing targeted interventions. Several methods have been developed to analyze bacterial population structures [38,39]. These include traditional techniques based on discriminatory markers, such as serotyping, phage typing, and pulsed-field gel electrophoresis (PFGE), which do not rely on DNA sequencing. More advanced genome methods rely on DNA sequencing and include multilocus sequence typing (MLST), core genome MLST (cgMLST), and whole genome MLST (wgMLST), as well as techniques focusing on SNPs and *k*-mer analysis.

1.2.1 Allele calling methods before Whole-Genome Sequencing (WGS)

Despite the development of DNA sequencing methods in 1977, PFGE remained the standard for bacterial strain typing for many years [40]. PFGE involves digesting genomic DNA with restriction enzymes and separating fragments on a pulsed-field gel. Fragment sizes vary due to mutations, DNA gain/loss, and genomic rearrangements. However, PFGE may offer limited resolution.

DNA sequencing gained popularity in the late 1980s with the advent of polymerase chain reaction (PCR), revolutionizing sequencing and enabling methods like MLST in 1998 [41]. MLST typing involves the amplification of seven loci of housekeeping genes by PCR, followed by DNA sequencing of the PCR products. A single nucleotide variation at any of these loci defines a distinct allele, which is used to determine the sequence type (ST).

However, MLST sometimes struggles to distinguish closely related strains during outbreaks. This limitation led to the development of multilocus variable-number tandem-repeat analysis (MLVA) [42], which provides greater resolution by analyzing variations in tandem DNA repeats, offering a faster and more precise tool for tracking outbreaks. While methods like PFGE and MLVA transformed bacterial epidemiology, they still lacked the necessary resolution for evolutionary and spatiotemporal studies [38].

1.2.2 The Post-WGS era

Advances in next-generation sequencing technologies and the accumulation of large, diverse datasets have led to a significant transformation in the field of bacterial genomics. WGS can now be performed rapidly, offering a high resolution for differentiating closely related strains and enabling in-depth analysis like phylogenetics, outbreak tracing, and phenotype prediction [38]. However, this surge in data production has created new challenges: extracting biologically meaningful insights from large datasets demands sophisticated bioinformatics approaches and considerable computational power.

One key benefit of WGS is its application in comparative genomics, where diverse isolates can be compared to trace outbreak sources and identify clonal strains. The process typically involves assessing genome similarity through various methods, followed by clustering to infer phylogenetic relationships. These methods include gene-by-gene approaches (e.g., cgMLST) and SNP-based analyses.

1.2.3 Allele calling methods in the post-WGS era

Within the WGS framework, gene-by-gene approaches have evolved and expanded significantly. A key factor in designing typing schemes is determining the required resolution for distinguishing isolates, which depends on the specific question being addressed [43]. For example, high resolution is needed for outbreak detection and within-patient variation studies, while lower resolution is sufficient for identifying the species causing an infection.

One of the earliest gene-by-gene method is MLST. Although MLST was initially developed as a PCR-based method, WGS allows researchers to derive MLST data directly from assemblies and determine the ST without the need for traditional PCR-based workflows [43]. MLST is widely used in molecular epidemiology and has been adapted for numerous bacterial species, each supported by extensive online databases for easy querying. These resources facilitate the identification of population clusters, often by grouping related STs into minimum-spanning trees (MSTree), such as those produced by eBurst [44]. However, while MLST has greatly advanced our ability to track bacterial diversity, its low resolution (based on 7 housekeeping genes) limits its use in some cases (**Figure 1.4**).

To address MLST`s limitation, other MLST-based approaches have been developed, each utilizing different numbers of loci, such as cgMLST and wgMLST [43]. Numerous schemes for both cgMLST and wgMLST are available in EnteroBase [39,45], a software environment for identifying global population structures across multiple bacterial genera.

cgMLST schemes analyze many more genes common to all isolates in a sample. This method combines the speed and ease of assigning indices to alleles with the increased resolution gained from using larger portions of the core genome. This method has proven useful in outbreak investigation [46,47]. However, cgMLST relies on predefined loci, which limits its ability to capture the full genetic diversity of bacterial populations. In *Salmonella*, for example, cgMLST involves 3,002 core genes that were found to be present in >=98%, intact in >=94%, of 3,144 representative *Salmonella* genomes (**Figure 1.4**) [39].

To further enhance the resolution of cgMLST, wgMLST schemes have been developed to incorporate accessory genes in addition to core loci. The wgMLST approach is particularly useful for differentiating single-clone pathogens with closed pangenomes or for closely related variants within more diverse organisms [43]. For example, EnteroBase implements a wgMLST scheme for *Salmonella* based on 21,065 orthologs from 537 complete genomes (**Figure 1.4**). However, the use of wgMLST for fine-scale epidemiological analysis has been questioned for several reasons [39]. Firstly, *Salmonella* has an open pangenome, and the number of orthologs in an open pangenome grows with the number of sequenced genomes. This growth necessitates continuous annotation of new loci, resulting in ever-evolving wgMLST schemes, a process that becomes computationally intensive for large databases. Alternatively, it might be sufficient to call genotypes on the basis of a frozen scheme, as is the case with EnteroBase. However, a frozen wgMLST scheme would lack any newly imported genes, such as those encoding AMR. Excluding these emerging loci reduces the scheme's ability to detect critical genetic changes introduced by HGT, a key aspect of pathogen surveillance.

The literature presents numerous conflicting claims about the best approaches for studying fine-scale epidemiology and tracing transmission chains. PulseNet International [48], a network of public health laboratories dedicated to tracking foodborne infections world-wide, now recommends wgMLST as the preferred tool for fine-scale epidemiology, replacing PFGE as its primary method.

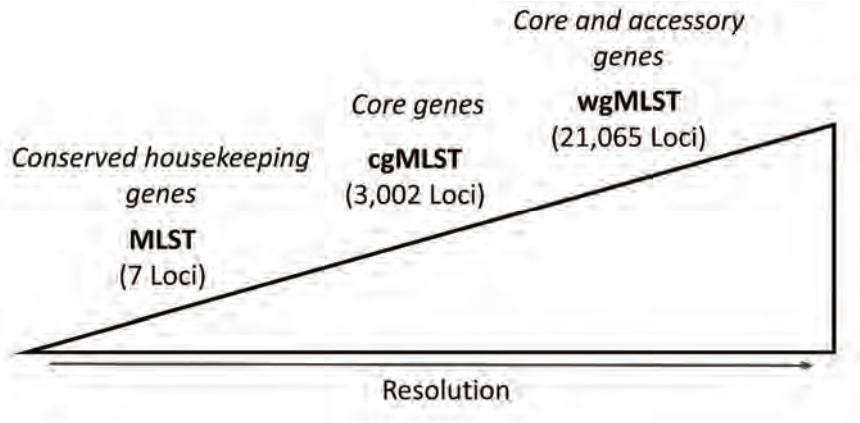


Figure 1.4 | Schematic representation of the MLST approaches for *Salmonella*. The resolution increases from left to right, progressing from traditional MLST (7 loci of conserved housekeeping genes) to cgMLST (3,002 core gene loci) and finally to wgMLST (21,065 loci including core and accessory genes).

1.2.4 Single Nucleotide Polymorphisms (SNPs) for bacterial typing

The identification of SNPs is another key approach for studying strain variation. SNPs are detected by mapping sequence reads or assemblies to a closely related reference genome and recording nucleotide differences [38]. A set of core SNPs, identified from positions covered by all query genomes, can be used to generate an SNP distance matrix for phylogenetic analysis, such as with neighbor-joining or maximum likelihood (ML) trees. This method has been successfully applied in resolving large-scale outbreaks [49,50].

The choice of reference genome is critical, as a high-quality, closed reference provides more accurate SNP calling, while a distant reference or unrelated isolates may reduce the number of core SNPs. Specialized tools, like SAMtools [51] and Snippy (https://github.com/tseemann/snippy), facilitate the process of SNP calling.

Phylogenetic trees can also be constructed by aligning the core genes of the bacterial genomes under study. This approach involves identifying the set of core genes shared among all query genomes and performing multiple sequence alignments on these conserved regions. Once the core genes are aligned, the SNPs identified within these alignments are extracted and used to generate a comprehensive SNP dataset. This dataset serves as the foundation for constructing phylogenetic trees.

Both MLST methods and SNP-based approaches generate distance matrices that are used to analyze genetic relatedness. Methods such as neighbor-joining, ML trees, MSTree, and hierarchical clustering are commonly employed to construct phylogenetic trees or visualize relationships between isolates (dendograms). The resulting groupings (e.g., clusters derived from SNP analysis or MLST methods) facilitate the interpretation of how closely related samples are [38,44].

1.2.5 *K*-mer based methods

In addition to the commonly used methods described in the previous sections, new computationally faster strategies have emerged to avoid the need for a reference genome or predefined schemes. Computational speed is achieved through alignment-free approaches [52], which quantify sequence similarity without generating alignments. These methods rely on breaking the genome into short nucleotide subsequences of defined length, called *k*-mers. Because they use the entire genome assembly, *k*-mer-based methods can capture both core and accessory genome variation. In practice, *k*-mer-based tools divide each genome sequence into all possible overlapping nucleotide segments of a specified length, *k*. The presence or absence of each *k*-mer in the sequences is recorded, generating a vector for each genome. Pairwise comparisons are then performed using these vectors, with dissimilarity between sequences quantified through the application of a distance function [38,52,53].

1.2.5.1 Split-k-mer methods

One approach to inferring differences in SNPs involves the use of odd-length *k*-mers which can be divided into two smaller fragments around a variable middle base, forming a structure called a “split *k*-mer” [38]. This structure allows the left and right parts of the *k*-mer to serve as local reference points for the position of the middle base. If the same left-right *k*-mer combination appears in another strain with a different middle base, homology between the middle bases can be hypothesized, and the difference can be interpreted as a SNP. This approach enables alignment-free, reference-free comparisons between strains while maintaining the ability to map to a reference sequence for interpretation. However, in highly variable samples, closely spaced SNPs may disrupt *k*-mer matching, reducing accuracy.

The split *k*-mer approach was first introduced in the kSNP program [54] and later in SKA1 [55], offering improvements in efficiency and flexibility. kSNP is an open-source tool, currently in its fourth version (released in 2023) [56], designed to detect SNPs between strains and construct phylogenetic trees from those SNPs. It allows the user to adjust various parameters and customize outputs. For example, users can choose to analyze core SNPs or SNPs from the entire genome. It also provides options to generate different types of phylogenetic trees, including parsimony, neighbor-joining, and ML. While other programs either detect SNPs or build phylogenetic trees, kSNP is unique in that it does both [56].

Another tool that employs the "split *k*-mer" approach is SKA1 [55], which has recently been updated to SKA2 [57]. It is a tool designed for fast and alignment-free SNP detection and comparison across genomic datasets. By leveraging split *k*-mers, SKA enables reference-free alignments, mapping to reference genomes, and the calculation of SNP distances, facilitating clustering and phylogenetic analysis [57]. Unlike kSNP, which directly generates a phylogenetic tree based on the core or pangenome SNPs, SKA2 provides detailed SNP data and distances but does not generate the tree itself. Instead, it outputs the necessary data for users to construct the phylogenetic tree or visualize the results using other tools.

Although the split-*k*-mer method is effective for SNP detection, the current implementation is limited to detect SNPs only, and cannot identify indels or other structural variants. However, it is suggested that such variants could be inferred by performing the split-*k*-mer analysis at various split sizes and matching flanking bases across regions of different lengths [57].

1.2.5.2 Full k-mer similarity method: Jaccard Index

Another widely used metric in *k*-mer-based approaches is the Jaccard Index (JI), which measures the fraction of shared complete *k*-mers between two datasets, rather than individual fragments of them as "split-*k*-mers" methods do. JI is a common proximity measurement used to compute the similarity between two objects, with wide use in numerous domains, such as ecology [58,59], text mining [60,61], and genome comparison [53,62–64]. The JI is defined as:

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are the sets of k k-mers from two sequences.

To further reduce computational resources, Mash [53], one of the most popular tools for *k*-mer based distance estimation, uses the MinHash technique. This technique allows approximation of the JI using a reduced set of *k*-mers called "sketches" which enables a significant reduction in memory usage and runtime, while preserving a reliable estimate of genomic similarity. More recent tools, such as BinDash [62] incorporate advanced MinHash variants that further reduce root-mean-square error, increase compression, lower memory consumption, and improve runtime efficiency.

A tool that utilizes a *k*-mer based method for large-scale population analysis and bacterial strain clustering is PopPUNK (Population Partitioning Using Nucleotide *K*-mers) [65]. PopPUNK employs the Mash algorithm to calculate genome distances based on *k*-mer sketches, while specifically separating core and accessory genome *k*-mers. Core *k*-mers vary due to SNPs, while accessory *k*-mers reflect gene presence-absence variations. These two distances are plotted against one another and a machine learning algorithm (two-dimensional Gaussian Mixture Model (BGMM) or HDBSCAN) identifies components. The cluster closest to the origin (indicating minimal differences in both the core and accessory components) is interpreted as representing within-strain relationships, and pairwise distances within this cluster are used as thresholds to link samples and form the final PopPUNK cluster. The threshold for within-strain links is refined using a network score (transitivity and density) ensuring that the resulting network is sparse but highly clustered (**Figure 1.5**). The tool is optimized for certain species and provides reference databases (https://www.bacpop.org/poppunk/) with predefined models that include thresholds for cluster definition, allowing users to directly assign strains to new sets of isolates.

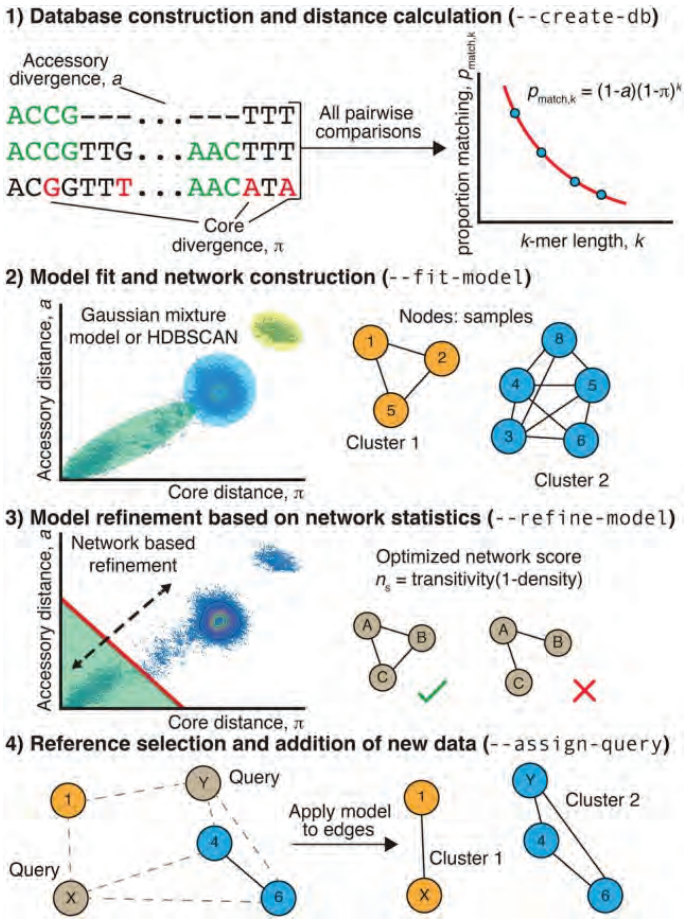


Figure 1.5 | Summary of the PopPUNK algorithm. Original legend **"**(Step 1) For each pairwise comparison of sequences, the proportion of shared *k*-mers of different lengths is used to calculate a core and accessory distance. Differences in gene content cause *k*-mers (examples highlighted in green) to mismatch irrespective of length, whereas point mutations distinguishing orthologous sequences cause longer *k*-mers to mismatch more frequently than shorter *k*-mers. (Step 2) The scatterplot of these core and accessory distances is clustered to identify the set of distances representing "within-strain" comparisons between closely related isolates. A network is then constructed from nodes, corresponding to isolates, linked by short genetic distances, corresponding to within-strain comparisons. Connected components of this network define clusters. (Step 3) The threshold defining within-strain links is then refined using a network score, ns, in order to generate a sparse but highly clustered network. (Step 4) Finally, the network is pruned by taking one sample from each clique. The distances between new query sequences and references are calculated, and within-strain distances used to add new edges. The clusters are then reevaluated as in Step 3, with the nomenclature being kept consistent with the original reference cluster names." Taken from [65].

Despite its advantages, PopPUNK has limitations, particularly for species with low genetic diversity, where cluster boundaries may connect in non-transitive ways [66]. The

calculation of core and accessory distances will in theory work to any resolution. However, if there is no clear within-strain versus between-strain separation in the distances and instead just a cloud of points, the spatial clustering methods are not likely to converge on a good solution. Network-based model refinement is needed in this case, though it is likely to split the strain into many sub-strains.

A recent version, called Iterative-PopPUNK [66] employs a multi-level clustering approach by applying several thresholds and saving the resulting clusters at each level. At higher thresholds, strains are grouped into parent clusters, while at lower thresholds, these same strains split into multiple subclusters (child clusters). All clusters are then integrated into a genealogical tree, where each cluster is nested as a child under its corresponding parent based on strain membership. For every node in the tree, the average core genome distance among the strains it contains is calculated, and its branch length relative to the root is set accordingly. This approach provides users with a tree representation of clustering results across multiple thresholds based on core distances, ultimately allowing them to select the most appropriate threshold for defining clusters according to the core distance between strains.

1.2.6 Pangenome analysis tools

In the preceding sections, a variety of methods for typing and differentiating bacterial genomes has been presented. Specifically, these methods focus on differentiating strains based on specific markers or nucleotide sequences, typing or clustering them into sets of closely related isolates. However, they do not address the gene families contained within each strain, nor the shared or distinguishing genes across different populations. To fill this gap and provide a more comprehensive view of bacterial diversity, pangenome analysis tools identify, classify, and compare gene families on a broader scale. This section introduces various programs and approaches developed for this purpose. Some popular tools are Roary [67], PanACoTA [68], Panaroo [69], or PanX [70]. These pipelines typically start by identifying and annotating genes from the input genomes, producing a set of protein sequences for each isolate. Next, they use homology search methods to estimate the similarity between proteins using tools like DIAMOND [71], USEARCH [72], or CD-HIT [73], followed by clustering algorithms such as Markov clustering. The clustering step is computationally intensive, as it groups genes into families that are common across a set of

genomes. By mapping which gene families are common or unique, these tools offer a detailed view of pangenome structure and reveal key differences among strains.

Programs like Roary [67], and PanX [70] remain among the most widely used tools for clustering gene families and constructing pangenome profiles. More recent tools, such as PATO (Pangenome Analysis Toolkit) [74] have introduced more efficient methods, using tools like MMseqs2 [75], MASH [53] or Minimap2 [76] to define core and accessory genome components. Moreover, it provides advanced visualization tools, and all its functionalities are integrated in an R package. Another recent tool, PanACoTA [68], provides a comprehensive pipeline for pangenome analysis, integrating steps such as genome gathering, quality control, alignment, core genome identification, and phylogenetic analysis.

In contrast to the gene-based pipelines described above, PanGraph [77] takes a gene-agnostic approach, relying solely on sequence homology. It uses sequence aligners such as Minimap2 [76] or MMseqs2 [75] to align genome sequences and extract blocks from homologous regions called “pancontigs”. These continuous segments of homologous sequence can be common to all genomes or exclusive to specific ones, enabling a detailed analysis of the pangenome structure.

Once pancontigs are identified, PanGraph can create a pangenome graph, which is a compressed representation of the genomes. In this graph, each genome is represented as a path, that is, a list of pancontigs. This visualization highlights how genomes overlap or differ, with pancontigs serving as the fundamental unit of comparison (**Figure 1.6**). The output of PanGraph can be exported into various formats, making it suitable for further downstream analysis.

One of the features of PanGraph is the ability to adjust two parameters (α and β) to modify the sensitivity of the alignments and the maximum sequence divergence allowed between homologous sequences in the same pancontig. These parameters allow users to fine-tune the tool’s performance depending on the specific dataset and research goals [77].

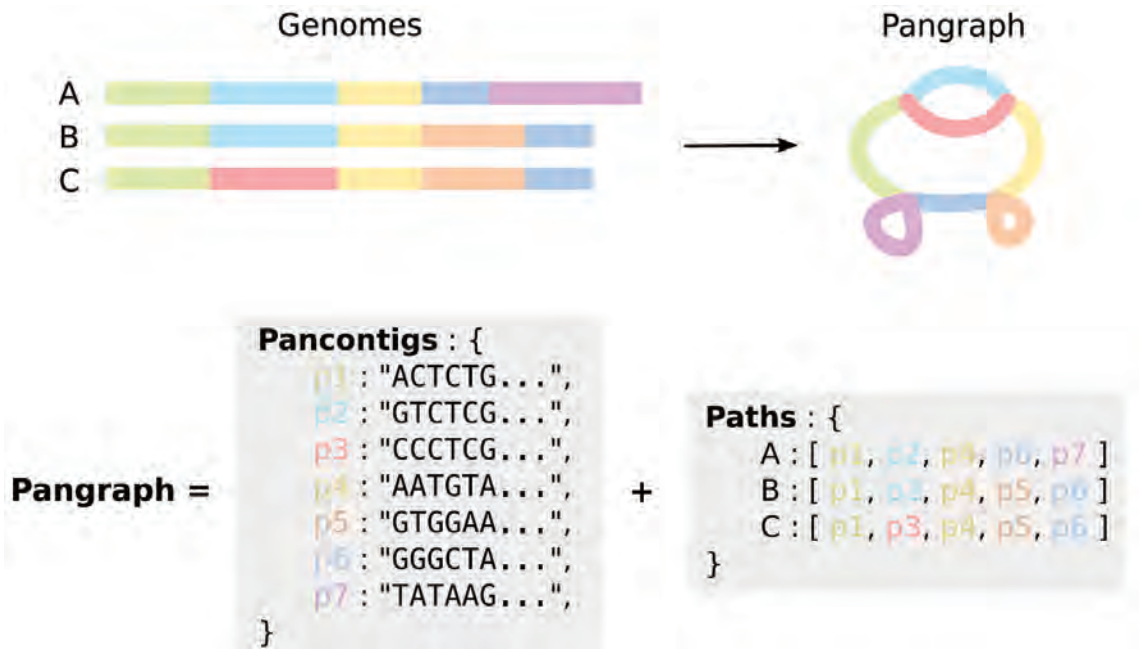


Figure 1.6 | Schematic pangenome graph representation generated by PanGraph. Given a set of genomes (A, B and C), PanGraph will construct a pangenome graph by condensing all homologous sequences in pancontigs (colored the same), containing the alignment of these sequences. From these pancontigs, a consensus sequence is generated for each one, making it easily accessible for downstream analysis. Genomes can then be represented as paths through the graph, i.e., sequences of pancontigs. In this representation, a pangenome graph is essentially a collection of pancontigs and paths. Taken from [77].

1.3 A prototype bacterial pathogen: *Salmonella*

Salmonella is a bacterial genus that belongs to the family *Enterobacteriaceae* and includes many pathogens that can cause disease in humans and other animals. These bacteria are commonly found in the intestinal tract of many animals including reptiles, amphibians, and poultry [78–80]. Although human salmonellosis is primarily foodborne, it can also be acquired through direct or indirect contact with infected animals in settings such as homes, veterinary clinics, zoological gardens, and farms. Approximately 11% of *Salmonella* infections are attributed to animal exposure annually, making it important for healthcare providers to be aware of this zoonosis. Both clinically affected and healthy animals can shed *Salmonella* over extended periods, with the prevalence of shedding often being higher in sick animals [81,82].

The public health risk varies based on factors such as age and health status, with certain human populations at higher risk due to biological or behavioral factors. Efforts to control *Salmonella* are further complicated by environmental contamination and indirect transmission through contaminated food and water. *Salmonella* exemplifies the One Health paradigm, as it can be transmitted through multiple food products and environmental sources, with reservoirs in humans, animals, plants, and in the environment [83,84]. Several initiatives, such as those intended to eradicate poultry-specific *Salmonella,* highlight the importance of ecological principles in understanding pathogen niches, and emphasize the need for an integrated approach to address S*almonella* infections [82].

1.3.1 Classification and taxonomy

The genus *Salmonella* is divided into two species, *S. bongori* and *S. enterica* [85]. *Salmonella enterica* is further divided into six subspecies: *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae* and *indica*, which contain over 2,500 serovars or serotypes (**Figure 1.7**) [85,86]. These subspecies divisions were initially classified by their biochemical properties and nucleotide similarity [85,87,88] and are supported by more recent sequence data [89]. The majority of the diseases associated with *Salmonella* are caused by serovars of *S. enterica* subspecies *enterica* [86].

Serovars were originally classified based on their O (somatic) and H (flagellar) antigens, with antigenic formulae represented by O antigens and H (phase 1, phase 2) antigens [85]. Specific antibodies target the cell wall (O) and phase 1 and phase 2 flagella (H) antigens, and each unique combination of O:H1:H2 is designated as a serovar (or serotype) name. The official list of serovars, known as the Kauffmann-White scheme [86], is maintained and regularly updated by the WHO Collaborating Centre for Reference and Research on *Salmonella*. Most serovars are named after the geographic location from which they were first isolated. These names are written without italics and start with a capital letter [85]. The formal name includes the full taxonomy, such as *Salmonella enterica* subspecies *enterica* serovar Typhi. However, they are often shortened to *S. enterica* serovar Typhi, S. Typhi, or just Typhi. In this thesis, serovars of *S. enterica* subspecies *enterica* will be introduced with their full names but referred to by their serovar name for simplicity.

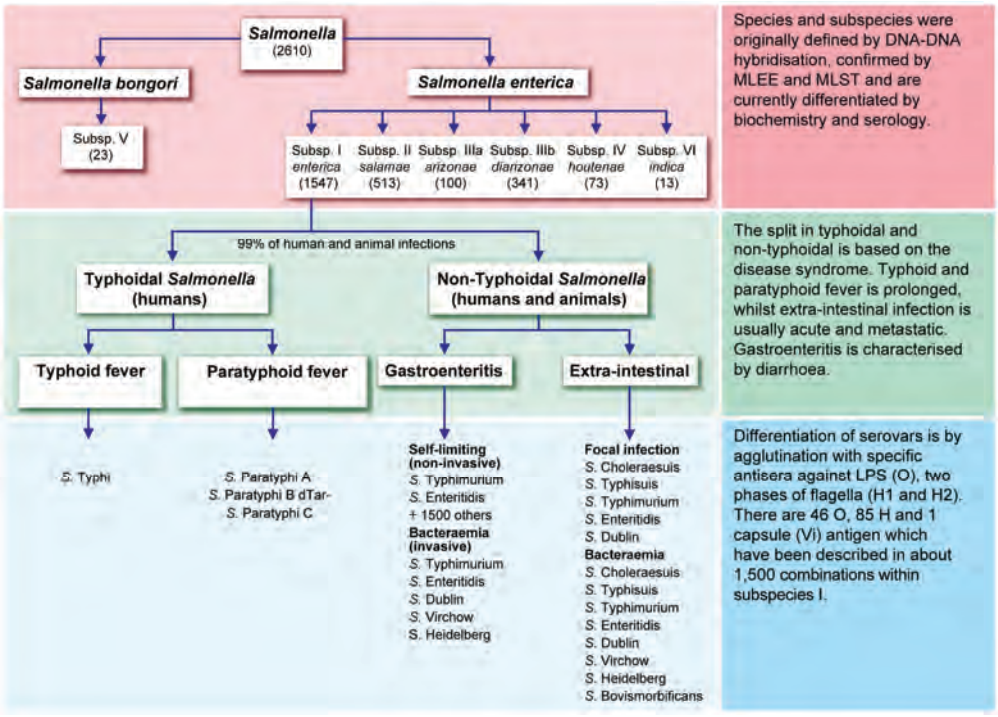


Figure 1.7 | General overview of the current classification of *Salmonella* [90]**.**

However, genomic studies [90] suggests that serovar classification, despite its epidemiological utility, does not always reflect the true evolutionary relationships among *S. enterica* isolates. Analysis based on MLST, which assigns each isolate to a ST, have demonstrated that clusters of closely related ST, referred to as eBurst Groups (eBGs), represent clonal complexes derived from a common ancestor (**Figure 1.8**). For example, most Typhimurium isolates belong to eBG1, but certain monophasic variants and related serovars, such as Farsta and Hato, also fall within this group despite antigenic differences. These observations suggest that traditional serotyping often confounds genetically unrelated isolates and fails to capture natural evolutionary groupings. Building on these findings, Achtman et al. (2012) [90] argued that, because MLST provides a clearer view of evolutionary relationships than traditional serotyping, *Salmonella* classification should transition from serovar-based designations to MLST or equivalent methodologies.

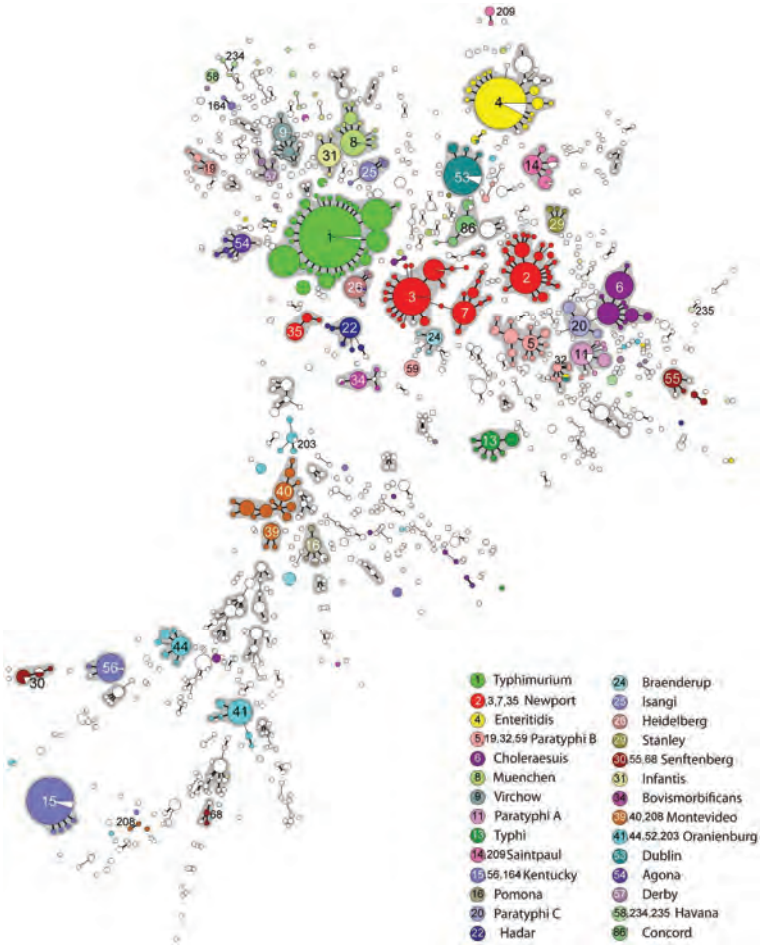


Figure 1.8 | Minimum spanning tree (MSTree) of MLST data on 4257 isolates of *S. enterica* subsp. *enterica*. Each circle corresponds to one STs, with its size proportional to the number of isolates. The topological arrangement within the MSTree is dictated by its graphic algorithm, which uses an iterative network approach to identify sequential links of increasing distance (fewer shared alleles), beginning with central STs that contain the largest numbers of isolates. As a result, singleton STs are scattered throughout the MSTree proximal to the first node that was encountered with shared alleles. The serovar associated with most of the isolates in each eBG or singleton ST is indicated by color coding for the 28 most frequent serovars. Text adapted from the original legend and figure taken from [90].

To further confirm that MLST provides a more accurate framework for classifying *Salmonella* than serotyping, complementary methods have been employed [39]. For instance, ribosomal MLST (rMLST) clusters isolates into rBGs based on sequences from 51 ribosomal protein genes, and these rBGs have largely confirmed the genetic groupings suggested by conventional MLST (**Figure 1.9**).

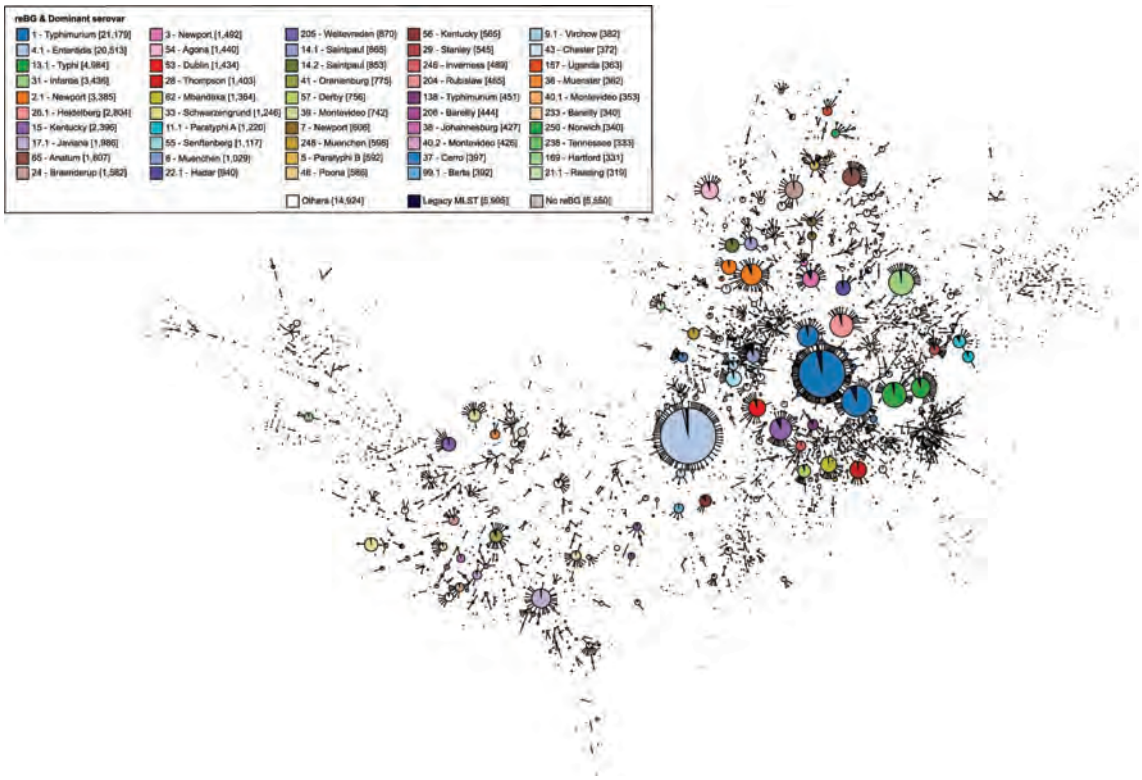


Figure 1.9 | Correspondence between eBGs from MLST and reBGs from rMLST. The figure shows a MSTree [91] from 118,391 *Salmonella* strains in EnteroBase. Each node corresponds to a single ST, with its size proportional to the number of isolates. The colours, reBG designations, dominant serovar, and numbers of genomic assemblies are indicated in the key (top left). Lines connect nodes that are single-locus variants. Text adapted from the original legend and figure taken from [39].

Additionally, another study [92] used cgMLST to generate core genome Sequence Types (cgSTs). Hierarchical clustering of these cgSTs (HCC) further refines this approach by grouping isolates into clusters. It was found that clusters that differed at less than 900 genes were correlated with single serotypes and were called HC900 clusters (**Figure 1.10**). Thereby bridging traditional serotyping with robust genomic frameworks. HC900 clusters provide higher resolution than eBGs,

In summary, while serovar nomenclature will continue to be used for initial identification and epidemiological reference, genomic classifications, such as those based on eBG, and HierCC clusters offer a more precise representation of the evolutionary relationships and genetic diversity within *Salmonella enterica.*

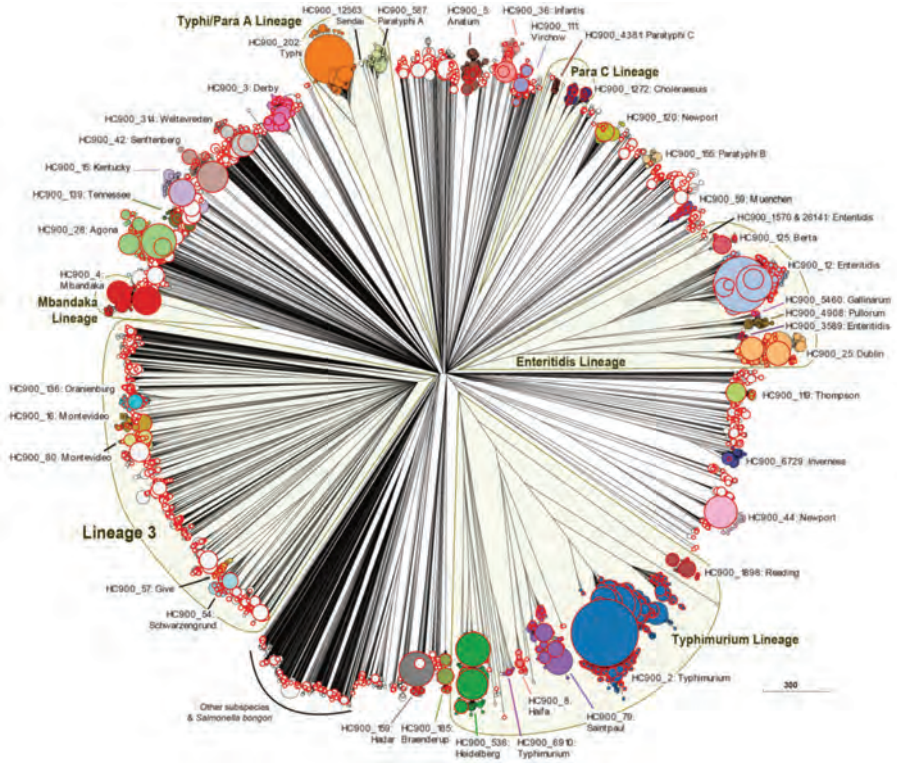


Figure 1.10 | Genomic diversity of *Salmonella* using HCC. The figure shows a neighbor-joining tree of the numbers of different alleles between cgSTs as generated within EnteroBase using GrapeTree [91]. Nodes from 41 common HC900 clusters are indicated by distinct colors, HC900 designations and predominant serovars. Lineages of HC900 clusters are indicated in yellow. Node sizes are proportional to the numbers of genomes they include. Scale bar: 300 alleles. Text adapted from the original legend and figure taken from [92].

1.3.2 Host range

The *Salmonella* serovars differ widely in their ability to infect different hosts and cause distinct disease syndromes [93]. They are classified into three groups based on their host range:

- Host-generalist serovars, such as Typhimurium and Enteritidis, cause infections in both humans and animals. The primary symptoms they produce include acute and self-limiting gastroenteritis [94].
- Host-adapted serovars are those that are prevalent in a particular host but can also colonize and potentially cause disease in other hosts. This group primarily causes systemic infections, such as Dublin in cattle and Choleraesuis in pigs but can also cause infections in humans and other animals [93,95,96].

- Host-specific serovars cause severe systemic infections in a specific host. Examples include Typhi and Paratyphi, which cause typhoid fever exclusively in humans and other primates [97]. In animals, notable examples include Typhisuis, which induces paratyphoid fever in pigs; Gallinarum, responsible for typhoid fever in poultry; and Abortusovis, which causes abortions in sheep [98].

The genetic differences between host-generalist, host-adapted, and specific *Salmonella* serovars provide insights into the factors that determine their host range [84]. Broad-host-range pathogens face selective constraints because they must survive in a variety of host environments, each with different physiological needs. Even minor fitness impacts can limit their ability to compete with other bacteria. In contrast, host-specific pathogens are adapted to a narrow niche, allowing them to grow slower but more effectively in a controlled environment. These pathogens have fewer selective pressures, enabling them to accumulate more loss-of-function mutations (pseudogenes) than broad-host-range serovars. Some of these pseudogenes help host-specific *Salmonella* strains evade immune responses and survive in specific hosts [99].

1.3.3 The disease and its global burden

Once transmitted to humans, the fate of *Salmonella* largely depends on the host's immune status and the serovar involved. The infection can be cleared by the immune system without causing disease, resulting in an asymptomatic carrier state, or it can lead to various types of illness [100]. Among these, we can differentiate between enteric fevers caused by typhoidal serovars and intestinal infections (enteritis and enterocolitis) caused by non-typhoidal serovars.

Non-typhoidal *Salmonella* (NTS) typically causes localized intestinal infections, including enteritis and enterocolitis [100,101]. These infections are characterized by acute inflammation of the gastrointestinal tract, often leading to symptoms such as nausea, vomiting, abdominal cramps, and diarrhea, which usually appear within 6-48 hours of ingestion. In healthy adults, the disease is typically self-limiting and resolves without the need for antibiotics, as the host's immune system can control and eliminate the infection. However, in vulnerable populations such as children, the elderly, and immunocompromised individuals, the infection may progress to extraintestinal or focal infections, requiring antibiotic treatment.

Beyond diarrheal disease, NTS can invade sterile body sites, leading to severe conditions such as bacteremia, meningitis, and localized infections [102]. These more serious forms, collectively known as invasive non-typhoidal *Salmonella* disease (iNTS), are typically characterized by nonspecific febrile symptoms rather than diarrhea and are associated with significantly higher fatality rates [100]. Vulnerable groups are at an elevated risk of developing these invasive infections [102]. According to the Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) 2017, NTS enterocolitis caused an estimated 95.1 million cases (95% uncertainty interval [UI]: 41.6-184.8 million), and 50,771 deaths (95% UI: 2,824-129,736) in 2017 [103]. In Europe, NTS was the second most commonly reported foodborne gastrointestinal infection in 2023, with 77,486 confirmed human cases and 88 reported deaths [104]. In the United States (U.S.), NTS are estimated to cause 1.35 million infections and 420 deaths each year [105].

Typhoidal *Salmonella* infections, caused by *S. enterica* serovars Typhi and Paratyphi A, B, and C, pose significant public health challenges, particularly in regions with poor water supply and sanitation [106]. The transmission of *Salmonella* occurs via the fecal-oral route, where infected individuals excrete bacteria in their feces and urine. In unsanitary conditions, these pathogens can contaminate food or water, which may be ingested by others. Upon ingestion, the bacteria pass through the acidic environment of the stomach and reach the lower small intestine (ileum), where they invade the intestinal epithelium. This leads to an incubation period of 7-14 days, during which the bacteria are released into the bloodstream, causing bacteremia and the onset of symptoms like headache, cough, weight loss, and abdominal pain. If untreated, the infection can spread to other organs, such as the liver, spleen, bone marrow, and gallbladder, potentially resulting in serious complications like gastrointestinal bleeding, intestinal perforation, septic shock, and death [100,101,107]. These infections are particularly prevalent in South Asia, Southeast Asia, and sub-Saharan Africa, where they remain major contributors to mortality and disability, especially among children [106,108]. In 2017, an estimated 14.3 million cases of enteric fever occurred globally (95% UI: 12.5-16.3 million) and 135,000 (76.9-218.9) deaths. Of these cases, Typhi was responsible for 76.3% (95% UI: 71.8-80.5) [106]. Typhoid and paratyphoid fevers are relatively rare in Europe and in the U.S. and they are mainly acquired during travel to other countries, particularly South Asia. In 2020, there were 315 reported cases of Typhi infection in Europe [109], while in the U.S., Typhi causes an estimated 5,700 infections and 620 hospitalizations each year [105].

1.3.4 Mechanisms of invasion and pathogenesis

The pathogenicity of *Salmonella* is determined by the presence of virulence factors, many of which are encoded in specific genomic regions called *Salmonella* Pathogenic Islands (SPIs). These islands harbor genes for critical determinants, including Type 3 Secretion Systems (T3SS), toxins, flagella, and capsules, that collectively enable the bacterium to invade and manipulate host cells [13,101,110] .

The differences in disease outcomes between typhoidal and non-typhoidal *Salmonella* are largely due to variations in their invasion mechanisms and interactions with the host immune system. Both types of *Salmonella* use similar strategies to invade the host, but their ability to evade immune responses and the resulting inflammatory reaction differ significantly.

After surviving the acidic environment of the stomach, *Salmonella* reaches the small intestine, where it invades the epithelial cells of the intestinal mucosa. The invasion process is mediated by the T3SS-1, encoded by SPI-1, which induces cytoskeletal rearrangements to facilitate bacterial uptake into enterocytes. The persistence of the bacteria inside these cells is due to *Salmonella*'s ability to create a vacuole known as the *Salmonella*-containing vacuole (SCV), which allows it to survive and replicate within these cells. Following invasion, the bacterium expresses a second secretion system, T3SS-2, encoded by SPI-2. This system modulates the trafficking and maduration of the SCV, further enhancing intracellular survival and replication, and facilitating the systemic phase of infection. Thus, SPI-1 is mainly active when *Salmonella* is extracellular, enabling the invasion of non-phagocytic cells, while SPI-2 T3SS is active once internalized and promotes the development of the SCV inside the macrophages (**Figure 1.11**) [13,101,110].

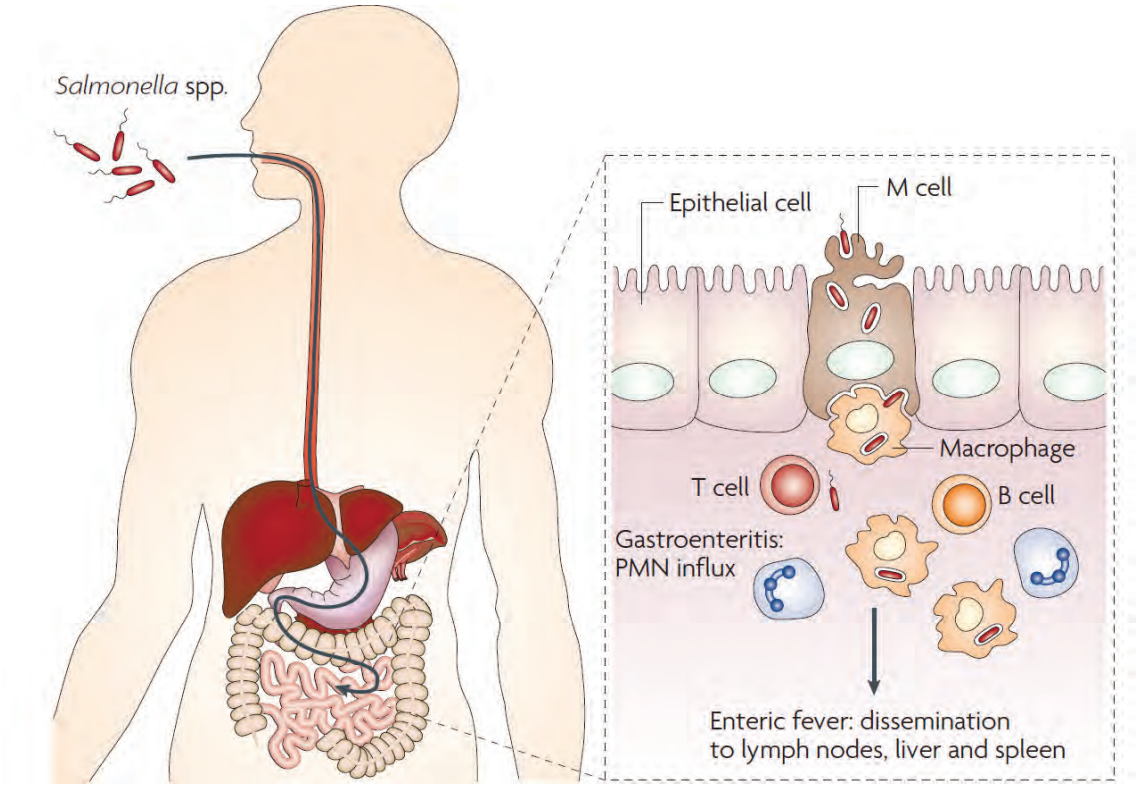


Figure 1.11 | Biology of *Salmonella* infection. Original legend: "Orally ingested *Salmonella* survive at the low pH of the stomach and evade the multiple defences of the small intestine in order to gain access to the epithelium. *Salmonella* preferentially enter M cells, which transport them to the lymphoid cells (T and B) in the underlying Peyer's patches. Once across the epithelium, *Salmonella* serotypes that are associated with systemic illness enter intestinal macrophages and disseminate throughout the reticuloendothelial system. By contrast, non-typhoidal *Salmonella* strains induce an early local inflammatory response, which results in the infiltration of PMNs (polymorphonuclear leukocytes) into the intestinal lumen and diarrhoea." Taken from [101].

The genetic and molecular differences between NTS and typhoidal *Salmonella,* which lead to different disease outcomes, remain an ongoing area of research. Non-typhoidal *Salmonella* infects many hosts, and therefore it is well-studied, with a clear understanding of the infection and invasion process in cells. In contrast, Typhi only infects humans, and there is no optimal animal model for studying its pathogenesis, leaving many aspects still to be elucidated. Much of the current research has focused on the differences between Typhi and Typhimurium, two of the most well-known serovars in each group. Despite sharing over 96% DNA sequence identity in their genomes, Typhi and Typhimurium result in vastly different clinical manifestations and immune responses in humans [100,111]. Key differences that have been identified are the following.

Typhoid *Salmonella* is adapted to systemic infection and has evolved to evade the local immune responses in the intestines. *In vitro* studies showed that Typhi induces significantly lower levels of the neutrophil chemoattractant IL-8, preventing inflammatory diarrhea [112]. It is hypothesized that one factor that helps Typhi evade the immune system is the Vi capsular polysaccharide (ViCPS) encoded by SPI-7. ViCPS prevents phagocytosis and limits immune recognition by masking bacterial surface patterns, thus enhancing its survival in the host [112]. However, the role of ViCPS in pathogenesis is not the sole explanation for the differences in clinical presentation, as Paratyphi A does not contain it and can still cause typhoid-like illnesses in humans [113]. Furthermore, Typhi has approximately 200 pseudogenes compared to Typhimurium [114]. Many of these disrupted genes are involved in motility, chemotaxis, T3SS effectors, fimbriae, and adhesins (factors crucial for the bacterium's pathogenicity). Additionally, Typhi also harbors other pathogenicity islands, such as SPI-15, SPI-17, and SPI-18, which are absent in Typhimurium. On the other hand, Typhi lacks SPI-14, which is specific to Typhimurium [107].

In contrast, NTS serovars, such as Typhimurium, typically trigger a much stronger local inflammatory response in the intestines [101]. This inflammation helps the recruitment of immune cells like neutrophils to the site of infection, contributing to symptoms like diarrhea and abdominal pain. While NTS strains can survive within the SCV, they rely on the inflammation in the gut to facilitate their growth by using compounds like nitrate and tetrathionate, which are abundant during inflammatory conditions [115]. This preference for the inflamed gut contrasts with Typhi, which avoids such localized inflammation and instead spreads throughout the body [101].

1.3.5 Treatment of salmonellosis, antibiotic resistance and vaccines

The discovery of antibiotics initiated a period of innovation and the implementation of drugs in both human and animal health, as well as in agriculture. However, these discoveries were quickly overshadowed by the emergence of resistant microbes [116]. Infectious pathogens are capable of evolving rapidly, and many have developed resistance to currently prescribed antibiotics. When a pathogen becomes resistant, the antibiotic loses its effectiveness in combating the infection. As a result, the infected host does not receive the help of the treatment to fight the disease. Antibiotic resistance is an old, dynamic and growing problem. Some of the factors that contribute to this situation are overpopulation,

increased global migration, the growing use of antibiotics in clinics and animal production, selection pressure, poor sanitary conditions, the spread of wildlife, and deficiencies in wastewater disposal [117].

Antibiotic resistance in *Salmonella* Typhi

The antibiotic chloramphenicol was introduced for the treatment of enteric fever and other bacterial diseases in 1948 (**Figure 1.12**) [118]. Chloramphenicol-resistant typhoid fever was reported two years later but was not common until the early 1970s, when a number of chloramphenicol-resistant typhoid outbreaks swept through Central and South America and Asia [119], even leading to a widespread epidemic in Mexico in 1972 [120]. This led to the replacement of chloramphenicol with a combination of ampicillin and co-trimoxazole from that point onward [121]. However, their prescription was also compromised by the emergence of multidrug-resistant (MDR) strains globally in the late 1980s (**Figures 1.12 and 1.13**). Consequently, due to these complications, the possibility of using first-line antibiotics against this bacterium diminished [122].

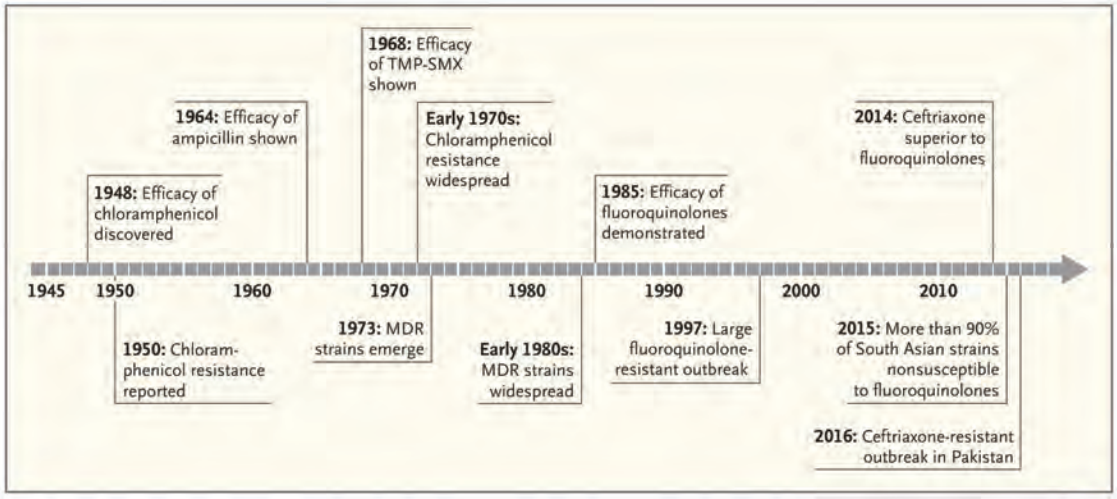


Figure 1.12 | History of antibiotic efficacy studies and the emergence of antimicrobial resistance in *Salmonella* Typhi. MDR denotes multidrug-resistant, and TMP-SMX trimethoprim–sulfamethoxazole. Strains noted to be "nonsusceptible" are intermediately or fully resistant. Taken from [123].

Following the emergence of MDR strains, fluoroquinolones (mainly ciprofloxacin) became the primary treatment for enteric fever for two decades. However, ciprofloxacin non-susceptibility (CipNS), defined by a minimum inhibitory concentration (MIC) ≥0.06 mg/L, soon emerged and become widespread, now representing the majority of typhoid cases in

South Asia (**Figure 1.13**) [123,124]. This phenotype is primarily attributable to substitutions in the quinolone-resistance-determining region (QRDR) of the *gyrA* and *parC* genes or via plasmid-mediated quinolone resistance (PMQR) genes (e.g., *qnrB*, *qnrD*, *qnrS*), all of which impair fluoroquinolone binding. The addition of further QRDR mutations or the co-occurrence of PMQR genes in strains already carrying a QRDR mutation can lead to clinically significant ciprofloxacin resistance (CipR), generally marked by MIC >1 mg/L [124,125].

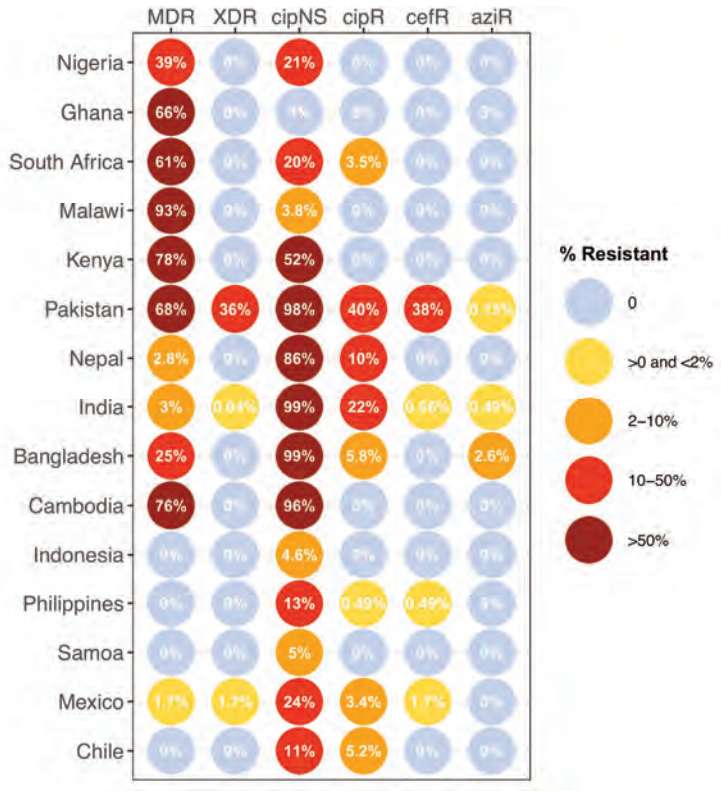


Figure 1.13 | Prevalence of key antimicrobial resistance profiles by typhoid-endemic countries from 2010 to 2020. The percentage resistance values are shown for each country-drug combination and color-coded by categorical ranges to reflect escalating levels of concern for empirical therapy: (i) 0%, no resistance detected; (ii) >0% to ≤2%, resistance present but rare; (iii) 2-10%, emerging resistance; (iv) 10-50%, resistance is common; (v) >50%, resistance is well established. MDR, multidrug resistant; XDR, extensively drug resistant; CipNS, ciprofloxacin non-susceptible; CipR, ciprofloxacin resistant; CefR, ceftriaxone resistant; AziR, azithromycin resistant. Text adapted from the original legend and figure taken from [124].

Finally, and up to the present day, both Typhi and Paratyphi are treated with third-generation cephalosporins (ceftriaxone) or azithromycin [121]. However, strains with resistances to these drugs have already emerged (**Figures 1.12 and 1.13**). In particular, an

outbreak in Pakistan in 2016 was caused by an extremely drug-resistant (XDR) Typhi strain (**Figure 1.12**), which subsequently spread through travelers to the United Kingdom (U.K.) and to the U.S. in 2018 [123,126]. This strain exhibited resistance to nearly all antimicrobials: ampicillin, chloramphenicol, co-trimoxazole, fluoroquinolones (ciprofloxacin), and third-generation cephalosporins, leaving azithromycin as one of the few remaining oral treatment options (**Figure 1.12**). Although CipNS is prevalent worldwide, strains exhibiting CipR, azithromycin resistance, or XDR remain predominantly confined to South Asia (**Figure 1.13**).

Therefore, today, as treatment options for enteric fever become increasingly limited, the evaluation and monitoring of the re-emergence of strains susceptible to any of these agents should be controlled to assess their potential therapeutic applications.

Antimicrobial resistance in Non-Typhoidal *Salmonella*

Most cases of NTS infection are self-limiting and do not require antibiotics for resolution. However, when the infection progresses to more severe forms, typically in vulnerable population, antibiotic treatment becomes necessary. Historically, antibiotics used to treat NTS have included ampicillin, chloramphenicol, and co-trimoxazole. However, since the late 1980s, the emergence of antimicrobial resistance has also affected these serovars, primarily in sub-Saharan African countries by Typhimurium and Enteritidis [127]. Therefore, the distribution of MDR strains poses a major challenge for the treatment and management of the disease in Africa, as MDR is also associated with higher mortality rates and disease transmission.

The MDR region is often encoded in *Salmonella* Genomic Island 1 (SGI1) which is integrated into the bacterial chromosome and can be transmitted vertically, remaining stably maintained even in the absence of selective pressure [128,129]. SGI1 and its variants have been described in a wide range of *S. enterica* serovars [130,131], and in other bacterial species such as in *Proteus mirabilis* [132], *Acinetobacter baumannii* [133], and in *Escherichia coli* (*E. coli*) [134]. In addition, several studies have demonstrated SGI1's capacity to excise, circularize, and transfer horizontally with the assistance of an IncA/C conjugative plasmid [135,136]. Therefore, SGI1 is an important vehicle for disseminating this resistant phenotype.

Consequently, alternative drugs such as ciprofloxacin and ceftriaxone have been increasingly used. However, these drugs are more expensive and less available in many

resource-limited settings across the continent [137], which presents a serious problem, especially considering that sub-Saharan Africa accounted for more than 79% of the global NTS cases diagnosed in 2017 [138].

Vaccines

Typhoid vaccination is an important component in the prevention and control of typhoid fever, and it is recommended for public health programs in both endemic and outbreak settings [139]. Vaccines for Typhi have been in use for many years, with two main types: the live attenuated Ty21a vaccine and the injectable Vi polysaccharide vaccine. These vaccines have successfully reduced the incidence of typhoid fever, but their effectiveness is limited by factors such as serotype-specific protection. For example, the Ty21a vaccine provides protection primarily against the Typhi serotype, while the Vi polysaccharide vaccine is not as effective in young children.

Recent developments in vaccine research have led to the introduction of new candidates, such as the conjugate Vi-TT vaccine [140]. This new vaccine demonstrated superior efficacy compared to older vaccines and offers broader protection, including in children under two years of age [141]. As a result, the conjugate Vi-TT vaccine is being increasingly used in endemic regions, particularly in countries with high rates of typhoid transmission and a signifiant disease burden.

For NTS, no widely approved vaccines are currently available for humans. However, as they constitute a critical group of zoonotic pathogens, several vaccines have been developed for use in animals such as poultry, swine, and cattle. Vaccination not only decreases susceptibility and alleviates the clinical manifestations of salmonellosis but also reduces fecal shedding and subsequent environmental contamination [142]. Multiple vaccines, including inactivated, live attenuated or live recombinant, have been developed over the years [143,144]. Nevertheless, more efforts are still needed to develop more effective vaccines against NTS. In recent years, outer membrane vesicles (OMVs) have emerged as promising vaccine candidates [145]. Naturally released by Gram-negative bacteria, OMVs display multiple bacterial antigens and possess intrinsic properties, such as immunogenicity and self-adjuvant activity. These features make them attractive for developing vaccines not only for a broad range of pathogenic bacteria but also specifically for *Salmonella* [146,147].

1.3.6 *Salmonella* genetics and evolution

Salmonella diverged from *E. coli* approximately 100 million years ago [148]. Both *Salmonella* and *E. coli* have circular chromosomes that are generally 4.5-5 Mb in size and encode around 4,500 genes. These genomes are highly similar, sharing approximately 70% of their genes, with 80% identity at the nucleotide level [149]. The differentiation between *E. coli* and *Salmonella* can be attributed to a variety of evolutionary, ecological, and genetic factors.

E. coli evolved primarily as a commensal bacterium that thrives in the intestines of mammals, including humans. In this role, *E. coli* has developed a mutually beneficial relationship with its host, aiding in digestion, producing vitamins, and outcompeting harmful bacteria [150]. This commensal lifestyle does not require the bacterium to be pathogenic because its survival is supported by its presence in a stable and nutrient-rich environment (the intestines). However, some strains of *E. coli*, such as *E. coli* O157:H7, have developed pathogenic characteristics [150,151].

In contrast, *Salmonella* evolved with a more aggressive approach to survival. While some *Salmonella* species can exist in the intestines of animals, they also evolved mechanisms to invade host cells and cause diseases such as gastroenteritis and systemic infections. A significant evolutionary event was the acquisition of the SPIs, which provide the genetic tools necessary for its survival and virulence in hosts [152]. For example, SPI1 is found in both *Salmonella bongori* and *Salmonella enterica*, but it is absent in *E. coli*. This suggests that SPI1 was acquired by the common ancestor of all modern *Salmonella* species after their divergence from *E. coli*. Conversely, SPI2, which is involved in intracellular survival and virulence, is present only in *S. enterica* and not in *S. bongori*, indicating that it was acquired after the two species diverged (**Figure 1.14**).

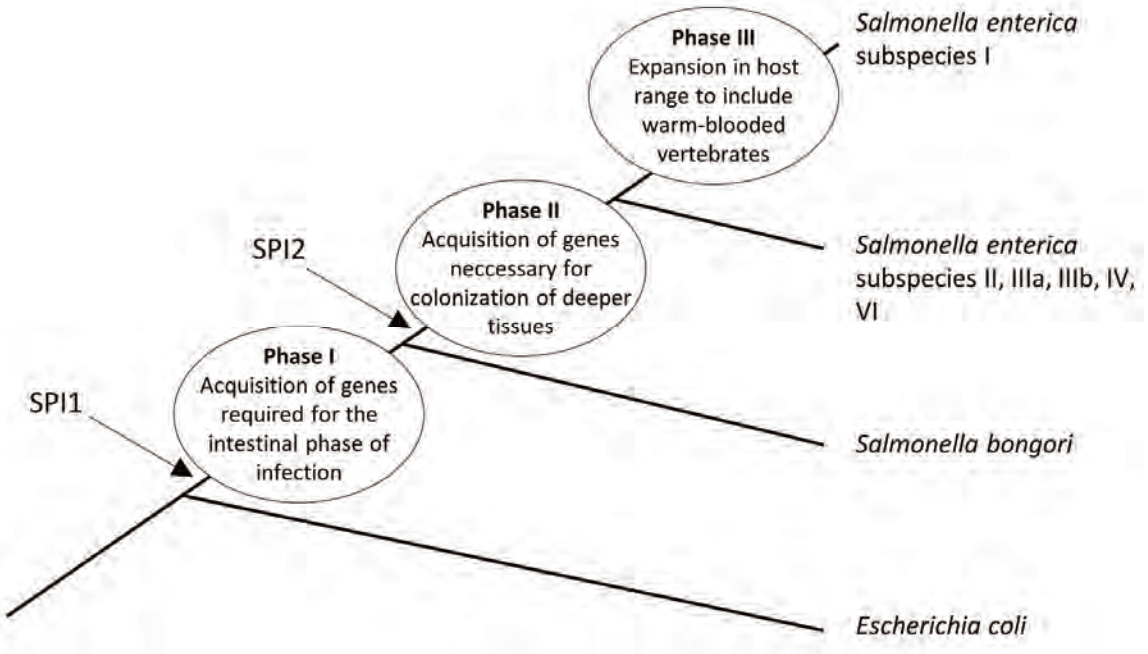


Figure 1.14 | Model for the evolution of virulence in the genus *Salmonella*. The three phases in which virulence evolved since divergence from the *Escherichia coli* lineage are depicted. The subspecies of *Salmonella enterica* are designated as I-VI. *Salmonella enterica* subspecies I is an alternative designation for *S. enterica* subspecies *enterica.* The phylogenetic tree is not drawn to scale. Modified from [153].

Transmission routes significantly influence *Salmonella* evolution. *Salmonella* is often transmitted through contaminated food, water, or contact with infected animals [13,84]. This mode of transmission may have driven the bacterium to evolve enhanced mechanisms for survival outside of the host (e.g., in water or food). In contrast, *E. coli* typically remains within its host and is adapted to a more stable and confined ecological niche in the gut [150]. This ecological difference may have contributed to the diverging evolutionary paths.

The genetic diversification within *Salmonella* is another important aspect and is driven by several complementary mechanisms. One contributing factor is the variation in antigen biosynthesis genes, which underlies the formation of distinct serovars [86]. On average, serovars share approximately 90% of their genes at >98% nucleotide identity [111]. Over time, these serovars have diversified further through point mutations, as well as through HGT [90,154]. These processes allow genes to be shared between different strains, blurring the boundaries defined by antigen synthesis gene variation and leading to further diversification of genetic material.

Consequently, differences in gene content and allelic variations influence both virulence and host range [130], leading to the emergence of highly pathogenic, and, in some cases, MDR strains. This continuous genetic evolution poses a significant public health challenge, particularly in regions with poor sanitation and limited access to medical care.

As discussed in the *1.3.1* section, traditional serovar classifications, based primarily on surface antigen variation, may therefore conflate genetically unrelated isolates and fail to capture natural evolutionary groupings [90].

1.3.7 The pangenome of *Salmonella*

The growing integration of WGS into public health surveillance has created unprecedented opportunities to analyze the complete pangenome of *Salmonella*. As a result, the number of *Salmonella* genomic sequences has been increasing exponentially since 2015 (**Figure 1.15**). Most of these genomic sequences have been deposited as short read archives [155].

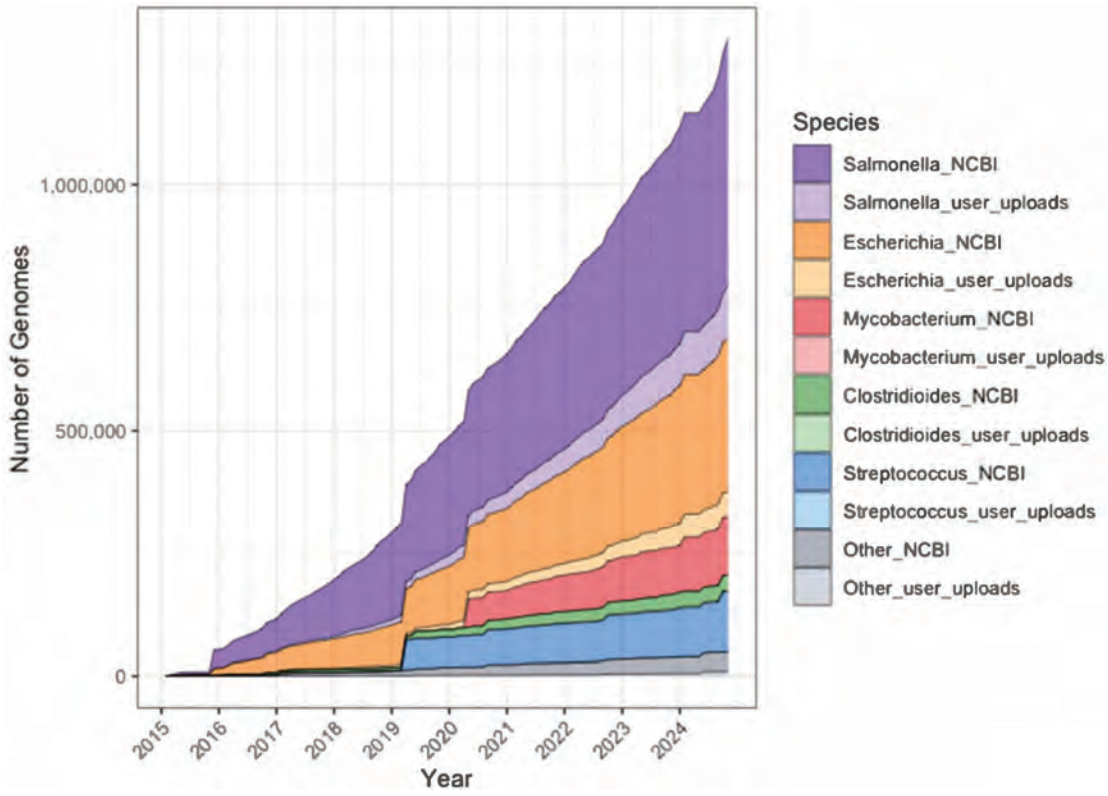


Figure 1.15 | Exponential growth in bacterial genome sequences in EnteroBase databases. The figure illustrates the increasing number of bacterial genome sequences, categorized by species and data source (public sequence read archives and user uploads). The 'other' sections include the smaller databases (*Vibrio*, *Helicobacter*, *Moraxella* and *Yersinia*). Taken from [155].

Despite this abundance, the global genetic diversity of this genus remains underexplored. This is primarily due to two factors. First, sequencing efforts have predominantly focused on human infections and domesticated animals. Second, analytical methods typically focus on the core genome, using SNPs analysis to build phylogenetic trees, while only overlaying elements of the accessory genome already cataloged in databases (e.g., plasmid replicons or AMR genes) [156,157]. For instance, in the analysis shown in **Figure 1.16** of *Salmonella enterica* serovar Reading genomes [156], a phylogenetic tree was constructed using core genome SNPs, and known accessory elements were then mapped onto the resulting tree. This revealed that strains closely related at the core genome level can exhibit substantial differences in accessory gene content. However, a limitation of this approach is that it can only detect accessory elements already present in known databases.

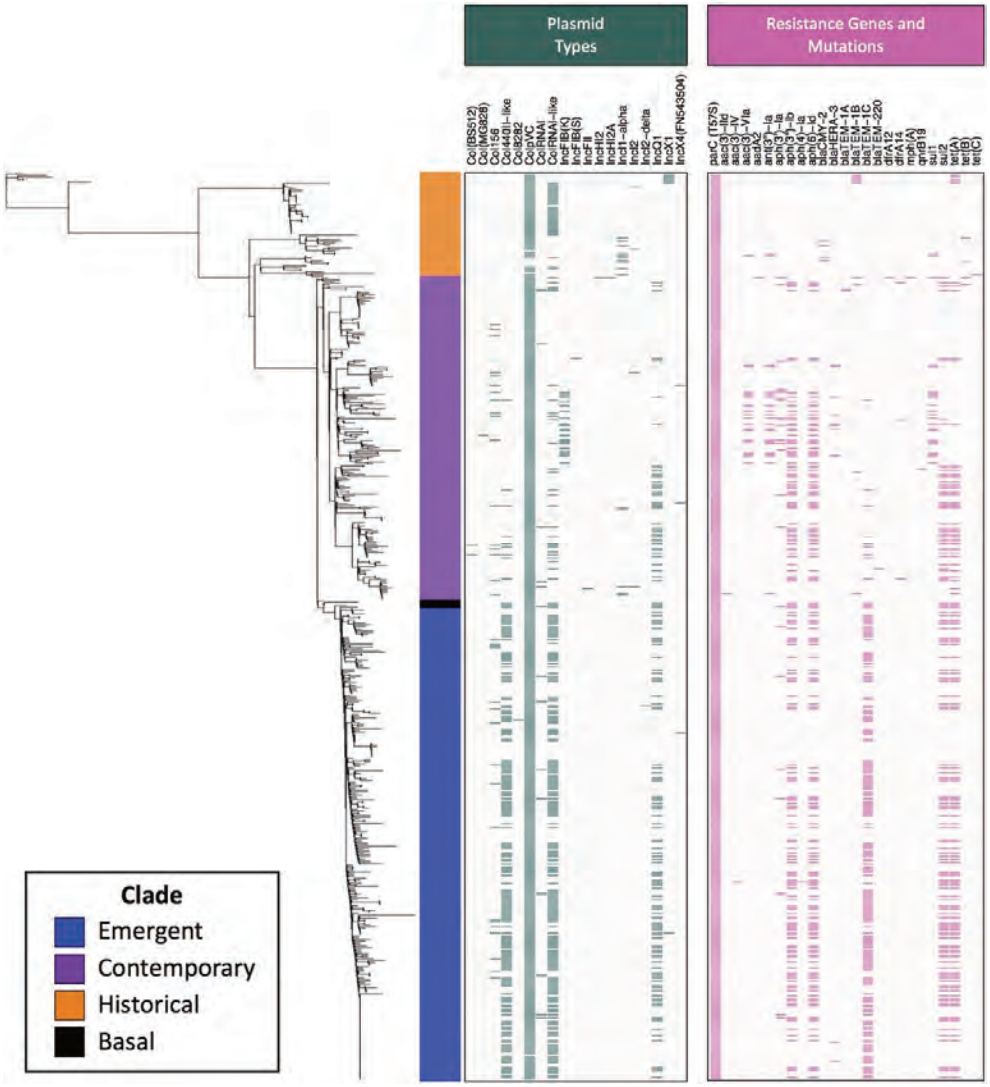


Figure 1.16 | SNP-based phylogenetic tree of *Salmonella* Reading with accessory genome features. The figure depicts a SNP-based phylogenetic tree of *Salmonella* Reading isolates, grouped

into distinct clades (Emergent, Contemporary, Historical, and Basal) as indicated by the color-coded sidebar. The accessory genome features are displayed as heatmaps to the right of the tree. The green heatmap shows the presence of plasmid replicons, while the pink heatmap represents AMR genes and point mutations associated with AMR. Each row corresponds to an isolate, and each column represents a specific plasmid type or resistance gene/mutation. Taken from [156].

Comparative genomic analyses have repeatedly demonstrated the fundamental role of HGT in the diversification of *S. enterica*. For example, studies of isolates of serovar Typhimurium [158] have reported core genome mutation rates ranging from 1.9×10^{-7} to 1.49×10^{-6} substitutions per site per year. Similarly, studies of serovar Typhi [159] have reported a mutation rate of 1.42×10^{-7} substitutions per site per year. In contrast, the composition of accessory elements vary considerably even within the same clone. Such insights highlight the interplay between relatively slow core genome evolution (vertical inheritance) and faster-paced changes driven by MGEs (horizontal acquisition).

Recognizing the challenges, several pangenomic studies have begun to shed light on the structure and variability of the *Salmonella* genome beyond the core. One of the largest such analyses examined 4,893 *Salmonella* genomes [160], identifying a total pangenome of 25.3 Mb. A "strict core" of 1.5 Mb was shared by all isolates, while a "conserved core" of 3.2 Mb was found in at least 96% of them, indicating that a substantial fraction of the *Salmonella* genome is accessory. In addition, several smaller-scale pangenomic studies have provided insights into the accessory genome. For instance, an investigation analyzing the pangenome of environmental *Salmonella* using a limited dataset (25 genomes) demonstrated that the accessory genome is highly abundant [161]. Similar studies of serovars Concord [162] and Reading [156] have highlighted that differences in accessory genetic elements, including plasmid acquisitions conferring AMR, phage-like sequences, virulence factors, correlate with the success of emergent subclades. The importance of including accessory genetic elements in genomic surveillance frameworks was also exemplified by the 2020 outbreak of *Salmonella enterica* serovar Newport linked to red onions [163]. Although epidemiological investigations identified two farms as the likely contamination source, SNPs analysis of WGS data failed to conclusively link clinical isolates to environmental samples from these locations. Instead, analysis of the plasmid content provided evidence to connect the clinical isolates to the implicated farms, as the same plasmid was found in some clinical and environmental samples.

Further reinforcing the importance of the accessory genome, a recent study [164] applied machine learning models to evaluate the role of accessory genome variations in clustering epidemiologically related *Salmonella* cases. By analyzing genomes from 24 outbreaks of food, animal, or environmental origin, their models found that polymorphisms and gain/loss events in MGEs were highly informative in defining outbreak clusters, particularly when core genome-based methods lacked resolution.

Among the various MGEs, plasmids and prophages play a pivotal role in shaping *Salmonella* population structure and dynamics. A large-scale study using the MOB-suite toolset analyzed 150,767 publicly available *Salmonella* genomes across 1,204 serovars and reconstructed 183,017 plasmids, revealing that 22% of these carried at least one resistance gene [32]. The study also found that plasmid carriage varies widely among serotypes, with prevalence ranging from 17% to 99% in serotypes with more than 100 isolates and an overall average of 65%. Additionally, the mean number of plasmids per isolate varied from 1.0 to 3.77 (average 1.88) and certain *Salmonella* serovars appear to be associated with specific types of plasmids [165,166].

Prophages, the integrated forms of bacteriophage genomes within bacterial chromosomes, are also significant contributors to the evolution and adaptability of *Salmonella*. Their diversity is driven by both homologous and non-homologous recombination events, and they facilitate the transfer and acquisition of adaptive genes. Prophages are very frequent in *Salmonella* [167]. For example, a study of 21 genomes of *Salmonella* [168] estimated that each genome contain in average 5.29 prophages representing around 3.52% of the total gene content and nearly 30% of the accessory genome. In a more extensive analysis of nearly 300 *Salmonella* genomes from 254 unique serovars [169], prophage regions were found to account for an average of 3.7% of the total genomic content (ranging from 0.1% to 8.8%) in each isolate.

Prophages are well known to encode virulence factors [170], but recent studies suggest that prophages also play broader roles in bacterial metabolism, regulation, resistance to heavy metals and modifications of cell surface structures [167,169]. Some prophages remain dormant and are transmitted vertically during bacterial replication, while others can be induced under stress, such as in the animal gut or during DNA damage. Interestingly, even defective prophages, which cannot form infectious particles, persist in *Salmonella* genomes, suggesting they may still provide evolutionary or physiological advantages to the

host [167]. This suggest a complex interplay between host and prophage which will be better understood as more bacterial genomes are sequenced and comprehensively analyzed.

In parallel with these genomic advances (including pangenome analyses, plasmid characterization, and prophage studies), source attribution has emerged as a critical application of WGS data for public health [13,171,172]. The combination of high-resolution WGS and detailed metadata has revolutionized *Salmonella* epidemiology by enabling precise identification of outbreak clusters and transmission pathways. However, source attribution remains challenging, particularly in cases involving complex transmission routes or limited metadata availability. Efforts to overcome these challenges are increasingly leveraging computational approaches such as machine learning that integrate phylogeographical signals and genomic features to improve the accuracy of contamination source predictions.

Several studies have explored novel genomic-based source attribution models for *Salmonella*. Zhang et al. [173] employed a random forest classifier, identifying 50 key genetic features, primarily accessory genes, sufficient for distinguishing livestock sources in Typhimurium strains. This highlights the role of the accessory genome in driving niche-specific adaptations. Another recent approach [174] leveraged accessory genome data in a multinomial logistic regression classifier, demonstrating its potential to accurately predict sources of bacterial contamination. As WGS and subtyping methods continue to advance, source attribution models are becoming increasingly precise, ultimately contributing to improved public health interventions.

Overall, *Salmonella*'s evolutionary diversification and niche adaptation are profoundly shaped by the accessory genome, which is highly variable [175]. While inclusion of accessory content is often thought to confound genomic epidemiological analyses [13,176], the accessory genome is not randomly structured, nor is it under neutral selection [36,175,177,178]. In fact, accessory genome have offered deeper epidemiological resolution for foodborne pathogen investigations [13,163,179,180]. However, the value of accessory genome data for differentiating related strains and pinpointing sources likely depends on the unique ecology of each *Salmonella* serovar and should be assessed within the context of serovar-specific population analyses. Such high-resolution analyses are particularly important for clonal *Salmonella* lineages that exhibit limited genomic variability or that are linked to multiple sources and transmission pathways. Moving forward, systematically

incorporating accessory genetic material into genomic surveillance frameworks holds considerable promise for deepening our understanding of *Salmonella* evolution, epidemiology, and control. This comprehensive perspective will be central to addressing the questions and objectives posed in this thesis.

CHAPTER 2: OBJECTIVES

This thesis presents a comprehensive pangenome-wide analysis of two *Salmonella enterica* serovars (Typhi and Hadar) using a Jaccard Index-based approach. It researches how vertical (core genome) and horizontal (accessory genome) processes shape population structure and drive short-term changes. The *Results* section is structured into two chapters, each focusing on either Typhi or Hadar and addressing distinct research questions relevant to their specific epidemiological and evolutionary challenges.

1. **Evaluate a Jaccard Index-based strategy** to capture both core and accessory genome variations.
2. **Pangenome characterization**: Characterize the genetic diversity within the largest U.S. datasets of Typhi and Hadar to date, and compare these findings with datasets from other geographic regions.
3. **Accessory genome dynamics**: Demonstrate how the success or short-term shifts of pathogen populations can be largely attributed to changes in the accessory genome.
4. **Refinement of bacterial stratification**: Refine bacterial stratification through pangenome analysis and evaluate how incorporating both known and uncharacterized MGEs enhances the resolution for distinguishing closely related strains.
5. **Epidemiological insights**: Identify key epidemiological patterns and evolutionary dynamics that may be overlooked by core genome methods alone.
6. **Public health applications:** Illustrate how pangenomic approaches can inform outbreak detection, source attribution, and targeted public health responses, highlighting the benefits of real-time surveillance that integrates accessory genome data.

CHAPTER 3: MATERIALS AND METHODS

3.1 Data collection

Salmonellosis is a nationally notifiable disease in the U.S., and isolates obtained from patients are routinely submitted to public health laboratories (PHLs) as part of the national enteric disease surveillance network, PulseNet USA, coordinated by the Centers for Disease Control and Prevention (CDC) [181]. Since 2019, PHLs have performed WGS on all *Salmonella* isolates they receive and upload sequence data to a centralized national database for genetic analysis, including computed serotyping [181,182], and to the National Center for Biotechnology Information (NCBI) under the BioProject PRJNA230403. Additionally, public health departments routinely collect demographic information for all laboratory-confirmed cases of salmonellosis. For cases included in multistate outbreak investigations, public health officials conduct additional patient interviews, whenever possible, with supplementary standardized questionnaires to obtain further details about foods eaten and animal contact before illness onset [183]. Approximately 5% of isolates detected by PHL also fall within the CDC arm of the National Antimicrobial Resistance Monitoring System (NARMS), a structured collection of enteric isolates from all 50 U.S. states used to monitor temporal trends in AMR (https://www.cdc.gov/narms/index.html). CDC NARMS has been routinely generating WGS data for this smaller subset of *Salmonella* isolates since 2016.

All datasets from the U.S. used in this study were collected by our project collaborators, Kaitlin Tagg and Hattie Webb (CDC NARMS, U.S.). The collected datasets encompass Typhi and Hadar genomes and the following sections detail the data sources for each pathogen.

3.1.1 *Salmonella enterica* serovar Typhi

An overview of all *Salmonella* Typhi genomes analyzed in this study is provided in **Table 3.1**, summarizing the data sources and the number of genomes included. The subsections (3.1.1.1 and 3.1.1.3) describe each dataset in more detail.

Table 3.1: Summary of *Salmonella* Typhi data collection

Dataset	Nº genomes
U.S. NARMS and PulseNet	2,272
RefSeq200	120
Indian Subcontinent	1,606
Globally representative genomes	1,804
Murray Collection	38

3.1.1.1 CDC and PulseNet dataset

Since 2016, NARMS and PulseNet USA have routinely performed WGS on Typhi isolates [182]. CDC's National Typhoid and Paratyphoid Fever Surveillance system collects metadata on all Typhi cases reported to PHL, including history of international travel in the 30 days before illness onset (https://www.cdc.gov/typhoid-fever/surveillance.html).

A total of 2,272 Typhi isolates were collected from January 1st, 2008, through September 30th, 2021 (**Table S1**). The dataset is divided as follows:

- 2008-2015 (prior to routine WGS): All Typhi isolates in the PulseNet national database with WGS data available were included (n=68).
- 2016–2018: All Typhi isolates sent to NARMS for WGS were included (n=1,343), representing U.S. Typhi cases reported to CDC for these years.
- 2019-2021: Due to logistics and delays in shipping for NARMS surveillance isolates, this period is represented by Typhi isolates in PulseNet with WGS data available (n=861), with expected underreporting due to Coronavirus disease 2019 (COVID-19) pandemic-related factors.

3.1.1.2 Sequencing methods

All genomes were sequenced using WGS through NARMS and PulseNet, followed standard operating procedures for the Illumina Miseq platform (https://www.aphl.org/programs/global_health/Documents/PNL38_WGS%20on%20MiSeq%20SOP_v4.pdf). Reads with a base call quality score ≥28 and coverage ≥40 x were assembled using shovill v.1.0.9 (https://github.com/tseemann/shovill), and contigs with coverage below 10% average genome coverage were excluded from the final assemblies.

During this study, long-read sequencing was required to confirm specific genomic features identified in short-read sequencing data. Consequently, long-read sequencing followed by hybrid assembly was performed by NARMS for selected isolates: PNUSAS224101, PNUSAS195139, and PNUSAS198714. The first two isolates were selected to confirm the absence of *Salmonella* Genomic Island 11 (SGI11) and the disruption of the *yidA* gene, while the third was chosen to detect the integration of $bla_{\text{CTX-M-15}}$ in SG11. Corresponding Illumina short reads were generated from the same DNA extraction; libraries were prepared using the Illumina DNA Flex preparation kit following the PulseNet protocol (https://www.aphl.org/programs/global_health/Documents/PNL35%20Illumina%20DNA%20Prep%20SOP_v5.pdf) and sequenced on the Illumina MiSeq platform as described above. Hybrid assemblies were uploaded to the NCBI.

3.1.1.3 Additional genomes

3.1.1.3.1 RefSeq200 genomes

One-hundred twenty Typhi genomes from NCBI RefSeq200 database (accessed on May 14th, 2020) were included in the analysis as reference genomes. These genomes were collected between 1916 and 2019 (**Table S1**).

3.1.1.3.2 Indian subcontinent genomes

A dataset comprising all Typhi genomes isolated in the Indian subcontinent available in Pathogenwatch (n=1,606, accessed on March 22nd, 2021) was generated for comparative analysis against the U.S. dataset. Specifically, this dataset included genomes linked to Bangladesh (n=637), India (n=487), Nepal (n=318), Pakistan (n=158), and Sri Lanka (n=3), or a combination of these countries (n=3) (**Table S2**).

3.1.1.3.3 Globally representative genomes

A collection of 1,804 globally representative Typhi genomes, used to develop the GenoTyphi typing scheme [184] (**Table S3**), was included for comparative analysis against the U.S. dataset. These genomes are available at Pathogenwatch (https://pathogen.watch/genomes/all?collection=nti046ubbs7t-wong-et-al-2015&genusId=590).

3.1.1.3.4 Murray collection genomes

Thirty-eight Typhi genomes from the Murray collection [185] were included to compare contemporary genomes with those isolated in the pre-antibiotic era. These genomes

are available at the European Nucleotide Archive at (https://www.ebi.ac.uk/ena/browser/view/PRJEB3255) (**Table S4**).

3.1.2 *Salmonella enterica* serovar Hadar

A total of 3,384 U.S. Hadar genomes were included in this analysis (**Table S5**), collected between 1990 and 2023 (August 30th), from U.S. surveillance systems and *ad hoc* sampling. Hadar genomes from ill humans with exposure information available were categorized as follows:

- Backyard poultry contact: when contact occurred within seven days of illness onset. Contact is defined as direct interaction with chickens, ducks, turkeys, geese, guinea fowl, or quail; direct exposure to the environment where backyard poultry live and roam; consumption of eggs or meat obtained from backyard poultry; or living with a household member who directly interacted with backyard poultry [186].
- Turkey consumption: when ground turkey was consumed within seven days before illness onset.
- Unknown: when exposure information was unavailable, or when neither backyard poultry contact nor turkey consumption was reported.

Genomes from non-human sources were categorized according to the commodity from which they were sampled, for example, “commercial poultry” or “swine”. The category “Other” was used for samples from unknown food, animal, or environmental sources.

An overview of all the Hadar genomes analyzed in this study is provided in **Table 3.2**, summarizing the data sources and the number of genomes included. The subsections (3.1.2.1, 3.1.2.2 and, 3.1.2.4) describe each dataset in more detail.

Table 3.2: Summary of *Salmonella* Hadar data collection

Dataset	Sample System	Source of data	Nº genomes
United States	U.S. surveillance system	CDC NARMS	2,494
		PulseNet	55
		FDA	300
		USDA-FSIS	367
		USDA-FSIS NARMS	102
	ad hoc	ORA, CFSAN, CVM	20
		Vet-LIRN	9
		APHIS	32
		ARS	3
		National Wildlife Health Center	2
No United States	Unknown	EnteroBase	1,145

3.1.2.1 United States surveillance systems

3.1.2.1.1 CDC NAMRS and PulseNet

WGS data for 2,494 Hadar isolates collected through the CDC NARMS from patients between January 1st, 2016, and August 30th, 2023, were included in this analysis (**Table S5**). For years prior to routine WGS (2005–2015), all Hadar isolates in PulseNet USA's national database with WGS data available were included (n=55); these represent isolates that were sequenced for various special interest projects.

3.1.2.1.2 FDA NARMS retail meats

The U.S. Food and Drug Administration (FDA) arm of NARMS routinely collects WGS data on *Salmonella* isolated from retail meats (chicken, ground turkey, ground beef, pork) purchased from U.S. grocery stores (https://www.fda.gov/animal-veterinary/national-antimicrobial-resistance-monitoring-system/about-narms). This data, along with source information, is uploaded to the NCBI under the BioProject PRJNA292661. As of August 30th, 2023, an NCBI Pathogen Detection query identified 300 Hadar genomes (**Table S5**).

3.1.2.1.3 USDA-FSIS

The U.S. Department of Agriculture's Food Safety and Inspection Service (USDA-FSIS) routinely collects WGS data on *Salmonella* isolated from regulated food and animal products within U.S. food processing facilities (https://www.fsis.usda.gov/science-data/sampling-program/sampling-results-fsis-regulated-products). Sequencing data and

source information are uploaded under NCBI BioProject PRJNA242847. On August 30th 2023, NCBI Pathogen Detection query identified 367 Hadar genomes (**Table S5**).

3.1.2.1.4 USDA-FSIS NARMS

Additionally, the USDA-FSIS arm of NARMS routinely collects WGS data from *Salmonella* isolated from the intestinal content of food animals at slaughter (https://www.fsis.usda.gov/science-data/national-antimicrobial-resistance-monitoring-system-narms). These data are uploaded under NCBI BioProject PRJNA292666. The above NCBI Pathogen Detection query (August 30th, 2023) identified 102 Hadar genomes (**Table S5**).

3.1.2.2 Ad hoc sampling systems

To expand source type representation along the farm-to-fork continuum, Hadar genomes isolated from North America were included from *ad hoc* sampling systems. The FDA's Office of Regulatory Affairs (ORA), Center for Food Safety and Applied Nutrition (CFSAN), and Center for Veterinary Medicine (CVM) perform *ad hoc* WGS on human food and animal food (including imported) product sampling and upload sequencing data to the GenomeTrakr project at NCBI (BioProject PRJNA186035). Twenty genomes (**Table S5**) collected between 2003 and 2022 were selected and included in this analysis. An additional nine isolates representing all sequenced Hadar collected from sick animals as part of FDA-CVM's Veterinary Laboratory Investigation and Response Network (Vet-LIRN) AMR monitoring program were also included.

USDA's Animal and Plant Health Inspection Service (APHIS) provides ongoing animal disease surveillance and animal disease diagnostic services through the National Veterinary Services Laboratories (NVSL; https://www.aphis.usda.gov/labs/about-nvsl) and the National Animal Health Laboratory Network (NAHLN; https://www.aphis.usda.gov/labs/nahln). Thirty-two Hadar genomes (**Table S5**) collected from chickens or turkeys from 2018 until 2023 as part of on farm monitoring or for diagnostic purposes were included in this analysis. Three Hadar genomes previously sequenced and published by USDA's Agricultural Research Service (ARS) [187], and two Hadar genomes collected from wild ducks by the National Wildlife Health Center were also included (**Table S5**). Additional Hadar genomes were available on NCBI, but source

information availability (through NCBI or personal communication) was a requirement for inclusion in this analysis.

3.1.2.3 Sequencing methods

All genomes were sequenced using WGS, performed through NARMS and PulseNet, as indicated in section *3.1.1.2*.

During this study, long-read sequencing was necessary to obtain reference genomes from each Jaccard Index group (JI-group) (see *3.3* section*)*. Long-read sequencing was requested in this study to be performed by NARMS on 53 selected isolates from each JI-group, chosen strategically to maximize connectivity to other internal nodes and to best achieve JI-group representation [188]. Thirty-six Hadar isolates from people or food products were sequenced on the Oxford Nanopore GridION sequencing platform; reads were assembled using a pipeline previously described [188,189]. Seventeen isolates collected from food or animal samples were sequenced on PacBio Sequel (Pacific BioSciences, CA), using the 10-kb SMRTLink template preparation protocol, as previously described [190]. Long-read data were uploaded under BioSample numbers listed in **Table S6**.

3.1.2.4 Additional genomes

A dataset of global non-U.S. Hadar genomes was obtained from EnteroBase [45] for comparative analysis against the pangenome of the U.S. collection. All genomes with predicted serotype "Hadar" (EnteroBase employs SISTR1 [191] and SeqSero2 [192]) isolated in any country other than the U.S. were downloaded (n=1,145) (accessed on December 21st, 2023) (**Table S7**).

Antother dataset of 259 PTU-I1 plasmids from RefSeq200 was used to compare with the PTU-I1 plasmids identified in the Hadar genomes of this study (**Table S8**).

3.2 Genomic characterization

3.2.1 *Salmonella* serotyping

The serotype of *Salmonella* isolates was determinad in silico using SeqSero 2.0 v1.2.1 [192]. This tool employs a database containing the *rfb* cluster genes involved in the synthesis of O antigens, as well as the *fliC* and *fliB* genes, which encode the first- and second-phase flagelar antigens.

3.2.2 Allele-based typing

Allele-based typing involves clustering isolates by identifying the different alleles present in a population for a given set of genes. ST were assigned using classical MLST (https://github.com/tseemann/mlst), which analyzes seven conserved housekeeping genes [90].

For higher resolution analysis, cgMLST [39,43] comprising 3,002 loci was employed to generate phylogenetic relationships and allele codes. The cgMLST allele codes were collapsed to the third digit (e.g., allele codes SALM1.0-6771.1.1.30.1.21 and SALM1.0-6771.1.1.30.1.44 would be collapsed into SALM1.0–6771.1.1), to simplify representation of allele codes. Genomes of the same condensed allele code are expected to differ by less than ~15 allele loci.

For *Salmonella* Typhi, typing was performed using the updated GenoTyphi scheme v1.9.1 [193] (https://github.com/katholt/genotyphi), which provides a phylogenetically informative nomenclature specific for lineages of *Salmonella* Typhi.

3.2.3 Characterization of known accessory genomic elements

Known accessory genome elements were characterized using a combination of bioinformatics tools. PlasmidFinder [30] was used to identify plasmid replicons with databases downloaded on July 31st, 2019 for Typhi analysis and on May 27th, 2023 for Hadar. This analysis was conducted using thresholds of 90% identity and 60% gene coverage. MOBscan [28] identified conjugative relaxases using as thresholds a HMM coverage >60%, and e-value < 0.01, while CONJScan [194] was employed to detect complete conjugative systems with databases downloaded on May 30th, 2019 for Typhi analysis and on February 24th, 2023 for Hadar. Bakta v1.9.1 [195] was used for gene annotation. Phage-related

elements were detected and annotated using PhageScope [196] or PHASTEST [197]. Easyfig [198] was used to perform linear comparisons between prophage sequences, further supporting the analysis of genomic organization and sequence similarities.

3.2.4 Antimicrobial resistance (AMR) determinants

AMR determinants, including acquired genes and chromosomal mutations, were detected using staramr, v.0.4.0 [199], which employs the ResFinder database (updated on 30th July, 2020; 90% identity, 50% gene coverage) and the *Salmonella* spp. PointFinder scheme (https://github.com/phac-nml/staramr). Predicted AMR was determined by staramr according to ResFinder and PointFinder results.

Typhi genomes were defined as MDR if they contain genes conferring resistance to ampicillin, chloramphenicol and co-trimoxazole. These resistance genes are typically acquired within an IS*1*-mediated composite transposon encoded in SGI11 [122,200]. XDR Typhi was defined as MDR with the addition of a ciprofloxacin resistance mechanism (quinolone resistance determining region [QRDR] mutation and/or plasmid-mediated quinolone resistance [PMQR] gene), and a ceftriaxone resistance gene (typically $bla_{\text{CTX-M-15}}$) [126,201].

Chromosomal integration events of AMR-containing regions were detected using the typing mode of ISMapper v.2.0.2 [202] to identify acquisition of an IS relative to a reference chromosome. To detect the integration sites of $bla_{\text{CTX-M-15}}$, its mobilizer, IS*Ecp1*, was used as a bait against a reference chromosome (Typhi 311189_291186, NZ_CP029894.1). Integration of SGI11 was detected using IS*1* as a bait element, and Typhi CT18 (NC_003198) as the reference chromosome.

3.2.5 Detection of plasmid-associated contigs in short read sequencing data

Draft Illumina contigs were assigned to plasmids or chromosomes using either MOB-suite v3.1.9 [32] or PLACNETw [203]. PLACNETw integrates assembly and plasmid-specific analysis. The tool performs BLAST searches to identify proteins within each contig and generates scaffold graphs. Based on the results and user expertise, contigs

are manually pruned to distinguish plasmids from chromosomes according to their protein-content.

MOB-suite is a toolkit that includes MOB-recon, a tool designated to assign Illumina contigs to plasmids or chromosomes [32]. Specifically, MOB-recon identifies plasmid contigs by searching for known plasmid replicons and relaxases using BLAST, detecting circularity, and comparing contigs to a curated database of complete plasmids. Contigs meeting any of these criteria are classified as plasmid-related, while the remaining contigs are assigned to the chromosome.

3.2.6 Plasmid analysis and classification

Plasmids were classified into PTUs [25] using COPLA [204]. This classification facilitates the categorization of plasmids based on their genetic content.

AcCNET [205] was used to build the plasmid ORFeome network. Homologous protein clusters (HPCs) were generated using kClust [206], with thresholds of >=80% protein identity, >=80% alignment coverage and clustering e-value < 1E-14. All edges were assigned equal weights. The network layout was visualized in Gephi v10 [207] using the force-directed continuous algorithm ForceAtlas2 [208]. The network consists of two types of nodes: HPCs and their corresponding plasmids. Edges connect both types whenever a plasmid contains a member in a given protein cluster (**Figure 3.1**).

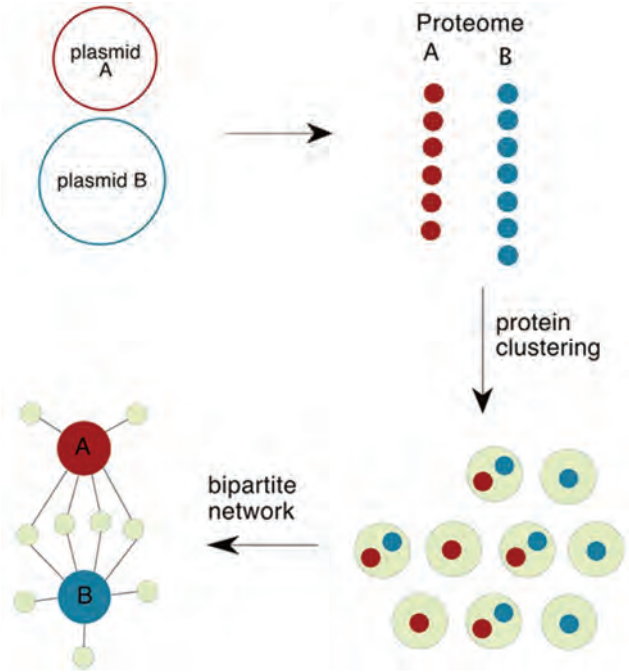


Figure 3.1 | Graphical representation of AcCNET. The protein set of all plasmids to be compared is clustered based on sequence similarity. The HPCs and their corresponding plasmids are the two kinds of nodes represented in the network and edges connect both types whenever a plasmid contains a member in a given protein cluster. Figure modified from [209].

The comparative sequence analysis of the different plasmids was carried out with BRIG [210] using default parameters. BRIG visualizes sequence similarity between each query with the chosen reference. Nucleotide identity ≥50% is indicated by color-coded bands, with the darker shading corresponding to higher sequence similarity. Blank regions indicate <50% nucleotide identity. Predicted coding sequences (CDS) of the reference genome are displayed in the outermost black ring.

3.3 Jaccard Index Network Analysis

Jaccard Index Network Analysis (JINA) was developed to explore genomic relationships through an integrated network-based approach. This workflow combines existing tools, including JI [62], Gephi v10 [207], BLAST v2.6 [211], and PanGraph v0.7.3 [77], alongside the newly introduced Genome Length Distance (GLD) metric. By integrating these methodologies into a unified framework, JINA enables efficient visualization and stratification of genomic data, facilitating the identification of meaningful patterns, groups, and associations within bacterial populations. The use of JI ensures precision in capturing genomic variation including SNPs, insertions and deletions. While JINA does not implement Gephi, Blast, and PanGraph directly in a single software, it guides their coordinated use to analyze and interpret genomic data effectively.

The JINA workflow proceeds through the following steps, which are detailed in subsequent sections. The process begins with the collection of genomes, followed by the calculation of the JI (and GLD if needed) and the construction of an adjacency matrix to represent pairwise genome similarities. The next step involves visualizing the adjacency matrix as an undirected network. To identify groups of genomes with high sequence similarity, an optimized JI threshold is applied to filter the network. A clustering algorithm is then used to detect these groups, referred to as “JI-groups”. The final threshold-filtered network can then be mapped with relevant metadata. Lastly, the differences between JI-groups are further examined to detect indels (**Figure 3.2**).

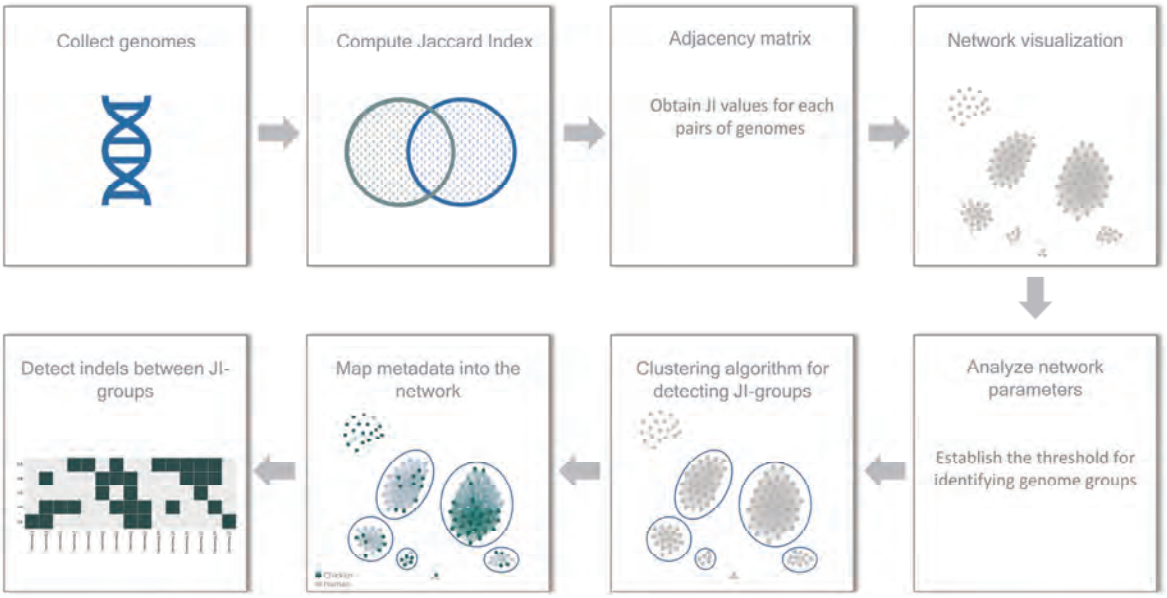


Figure 3.2 | Overview of the Jaccard Index Network Analysis workflow. This schematic summarizes the key steps of the workflow, from genome collection and pairwise similarity computation to network visualization, threshold filtering, clustering, metadata mapping, and subsequent analysis of genomic differences.

3.3.1 Jaccard Index calculation

I would like to acknowledge Santiago Redondo-Salvo for his work in developing the equivalences of the Jaccard Index (JI) in terms of SNPs, indels, and replacements described in this section. It is included in this thesis to facilitate the reader's understanding of the JI.

The exact JI was used as a measure of similarity between all genome pairs. First, the complete assembly of each genome was converted into a set of *k*-mers. JI was calculated as the ratio of shared *k*-mers over the total number of different *k*-mers between the two sets (including shared *k*-mers, SNP *k*-mers differing by a single base pair, and indel *k*-mers differing between the datasets and excluding duplicated *k*-mers). JI considers only once those identical *k*-mer between both sets disregarding genome differences due to sequence repetitions. The JI value ranges from 0 to 1, where 1 indicates 100% *k*-mer similarity and 0 indicates no *k*-mers shared. The formula to calculate JI between genomes A and B is shown in Equation 1.

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Eq. 1

BinDash v1 [62] was employed to calculate JI, using parameters minhashtype=-1, ensuring exact JI computation by considering the complete set of *k*-mers instead of an estimated JI based on a *k*-mer subset. The *k*-mer length (*k*=21) was selected as previously determined to be optimal in [53].

JI is a symmetrical measure, that is, $JI_{(A,B)} = JI_{(B,A)}$. A closely related notion, the Jaccard Distance, measures the dissimilarity of two datasets, as defined by Equation 2:

$$d_{JI(A,B)} = 1 - JI_{(A,B)} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} = \frac{|A \,\Delta\, B|}{|A \cup B|}$$

Eq. 2

At first glance, applying JI to compare genomic sequences may seem like a straightforward exercise of counting shared *k*-mers between sequences. However, two key characteristics of DNA can influence JI estimation. First, the molecular topology of DNA sequences, whether linear or circular, affects *k*-mer computation at contig edges (as shown in Equation 3):

$$\#\text{ of } k\text{ mers} = \begin{cases} N, & if\ circular\ contig\ of\ size\ N \\ N - k + 1, & if\ linear\ contig\ of\ size\ N \end{cases}$$

Eq. 3

Second, the double-stranded nature of DNA complicates JI estimation, as it is impossible to determine the strand orientation in a draft genome. To overcome this, *k*-mers must be extracted from both the original contig sequence and its reverse complement. An alternative approach is to use canonical *k*-mers, which are defined as the lexicographically smaller sequence between a *k*-mer and its reverse complement. By using canonical *k*-mers, the strand direction of the assembled contigs can simply be ignored, ensuring the validity of Eq. 3. However, a drawback of this method is that it effectively halves the *k*-mer space and, thus doubling the probability of random *k*-mer duplication for a given *k* value. To mitigate this issue, an adequately large *k* value must be chosen.

JI captures both SNPs and gene content differences that arise as the result of gain and loss of genetic material (indels) (**Figure 3.3**). When two genomes differ by *L* nucleotides (e.g., SNPs), the number of different *k*-mers will be approximately *Lk*, assuming SNPs are infrequent enough that their co-occurrence within the same *k*-mer is negligible. The insertion of a DNA sequence in one genome contributes $L + k - 1$ new *k*-mers, while a deletion results

in $k - 1$ new k-mers in the deleted genome and $L + k - 1$ in the non-deleted genome. Additionally, the acquisition of a plasmid of size L introduces L new k-mers.

3.3.1.1 Influence of SNPs in the Jaccard Index

Let A and B be two genomes of equal size N. If a mutation is introduced in a position of sequence B, the number of different canonical k-mers between the two sequences will be k. Therefore, the JI between both sequences is obtained from Equation 4:

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{N - k}{N + k}$$

Eq. 4

In the case of multiple SNPs, if the number of SNPs (L) is very small relative to the genome size (N), and assuming that SNPs follow a Poisson distribution [53], the probability of two SNPs occurring within the same k-mer (i.e. closer than k base pairs) can be disregarded. Under these conditions, JI can be calculated using the formula:

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{N - Lk}{N + Lk}$$

Eq. 5

Equation 5 can be illustrated with a simple example: genomes A and B are identical circular sequences of length N=22 (where length is defined as the total number of k-mers and k=5). Three SNPs (L=3) were introduced in the sequence of genome B (**Figure 3.3a**). The number of different k-mers between the two sequences will be Lk. Therefore, the JI between both sequences is:

$$JI_{(A,B)} = \frac{N - Lk}{N + Lk} = 0.19$$

In the case of SNP hotspots, each SNP influences fewer than k k-mers. **Figure 3.3b** illustrates this scenario.

As the number of SNPs gets larger the probability of two SNPs occurring in the same k-mer cannot be dismissed, thus Eq. 5 is not a good JI estimator. In general, with the sole assumption of a Poisson distribution of SNPs, equation 4 from the MASH paper [53] can be applied to estimate JI as a function of the sequence dissimilarity ($D = L / N$):

$$JI_{(A,B)} = 1/(2e^{kD} - 1) = 1/\left(2e^{kL/N} - 1\right)$$

Eq. 6

3.3.1.2 Influence of Indels in the Jaccard Index

The acquisition of a plasmid of size *L* by a genome A of size *N* is simply a particular case of sequence insertion not affected by the *k*-mer length, and its contribution to the JI is defined in Equation 7:

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{N}{N+L} = 1 - \frac{L}{N+L}$$

Eq. 7

Sequence insertion in a genome is slightly more complex. Let A be a genome with a sequence of length *N*, and B a genome identical to A, except for an inserted sequence of length *L* (making the total size of genome B = *N* + *L*). Assuming no duplicated *k*-mers due to random repetitions or duplication events, genome B will contain all *k*-mers shared with A plus $L + k - 1$ new *k*-mers (see **Figure 3.3c**). Meanwhile, in genome A, $k - 1$ *k*-mers surrounding the insertion site will be unique to this genome. Therefore, JI between sequences A and B can be calculated as follows:

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{N-(k-1)}{N+L+k-1} = 1 - \frac{L+2k-2}{N+L+k-1} \approx 1 - \frac{L}{N+L}$$

Eq. 8

When $k << L$ *k*-mer size effect on Equation 8 can be neglected, simplifying it into Eq. 7. This can be illustrated with a simple example: let genome A be a circular genome of length *N*=22, and B a genome identical to A, except for a 7 bp insertion (*L*=7). JI between sequences A and B, using *k*-mers of length *k*=5, is calculated as follows:

$$JI_{(A,B)} = \frac{N-k+1}{N+L+k-1} = 0.55$$

For the case of plasmid loss, the formula (Equation 9) is slightly different from the plasmid gain scenario described in Equation 7:

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{N-L}{N} = 1 - \frac{L}{N}$$

Eq. 9

Equation 10 allows JI calculation when a sequence is deleted in one of the genomes (**Figure 3.3d**). When $k << L$, the result of Eq. 10 converges with that of Eq. 9.

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{N-(L+k-1)}{N+k-1} = 1 - \frac{L+2k-2}{N+k-1} \approx 1 - \frac{L}{N}$$

Eq. 10

Equation 9 can be illustrated with the following example: let genome A be a circular sequence of length N=22 and genome B be identical to A except for a 7 bp deletion (L=7). The JI, using k-mers of length k=5, between both sequences can be estimated as follows:

$$JI_{(A,B)} = \frac{N-L-k+1}{N+k-1} = 0.42$$

Table 3.3 shows the impact of increasing SNP counts on the JI, using typical values observed in real cases of bacterial genome comparisons and it also presents the equivalent insertion lengths (in bp) corresponding to these JI values.

Table 3.3: Discrimination of SNPs by using the Jaccard Index with k=21 and k=31 and equivalent insertion lengths (bp) in a sequence of length N=5x10^6.

SNPS	SNPS/MB	*INSERTION* [A]	*%ID* = 1 - *D* [B]	JI_{21} [C]	JI_{31} [D]
50	10	2060	0.99999	0.99958	0.99938
100	20	4161	0.99998	0.99916	0.99876
500	100	20982	0.9999	0.99581	0.99383
1000	200	42048	0.9998	0.99165	0.98771
1440	288	60623	0.99971	0.98801	0.98238
2050	410	86431	0.99959	0.98300	0.97506
5000	1000	212180	0.999	0.95928	0.94076
5880	1176	250000	0.99882	0.95237	0.93088
10000	2000	428903	0.998	0.92099	0.88658

[A] | Estimated length (bp) of an insertion computed with JI using k = 21. Data obtained with Equation 8.

[B] | Percentage of sequence identity.

[C] | JI computed with k = 21.

[D] | JI computed with k = 31.

In orange, the row corresponding to JI=0.988, the threshold used to sparsify the Hadar network. In green, the row corresponding to JI=0.983, the threshold used to sparsify the Typhi network. In blue, the row corresponding to 5,880 SNPs, which results in the same JI as a 250 kb insertion in a 5 Mb genome (see **Figure 3.3f**)

3.3.1.3 Influence of sequence replacement in JI

A special case arises when a DNA stretch of length *L* bp is substituted (**Figure 3.3e**). This scenario includes clusters of closely located SNPs, where the distance between each successive pair is less than *k* nucleotides. The formula to calculate JI in this case is given by:

$$JI_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{N - \big(L + (k-1)\big)}{N - (L+k-1) + (L+k-1) + (L+k-1)} = \frac{N - L + k - 1}{N + L + k - 1} \approx \frac{N - L}{N + L}$$

Eq. 11

Equation 11 can be illustrated with the following example: let A and B be circular genomes of length N=22 that differ in a sequence stretch of 7 bp (L=7). JI between sequences A and B is calculated as follows:

$$JI_{(A,B)} = \frac{N - L + k - 1}{N + L + k - 1} = 0.58$$

3.3.1.4 JI vs sequence identity

A graphical representation of the influence of a variety of genome differences in the JI is shown in **Figure 3.3f**.

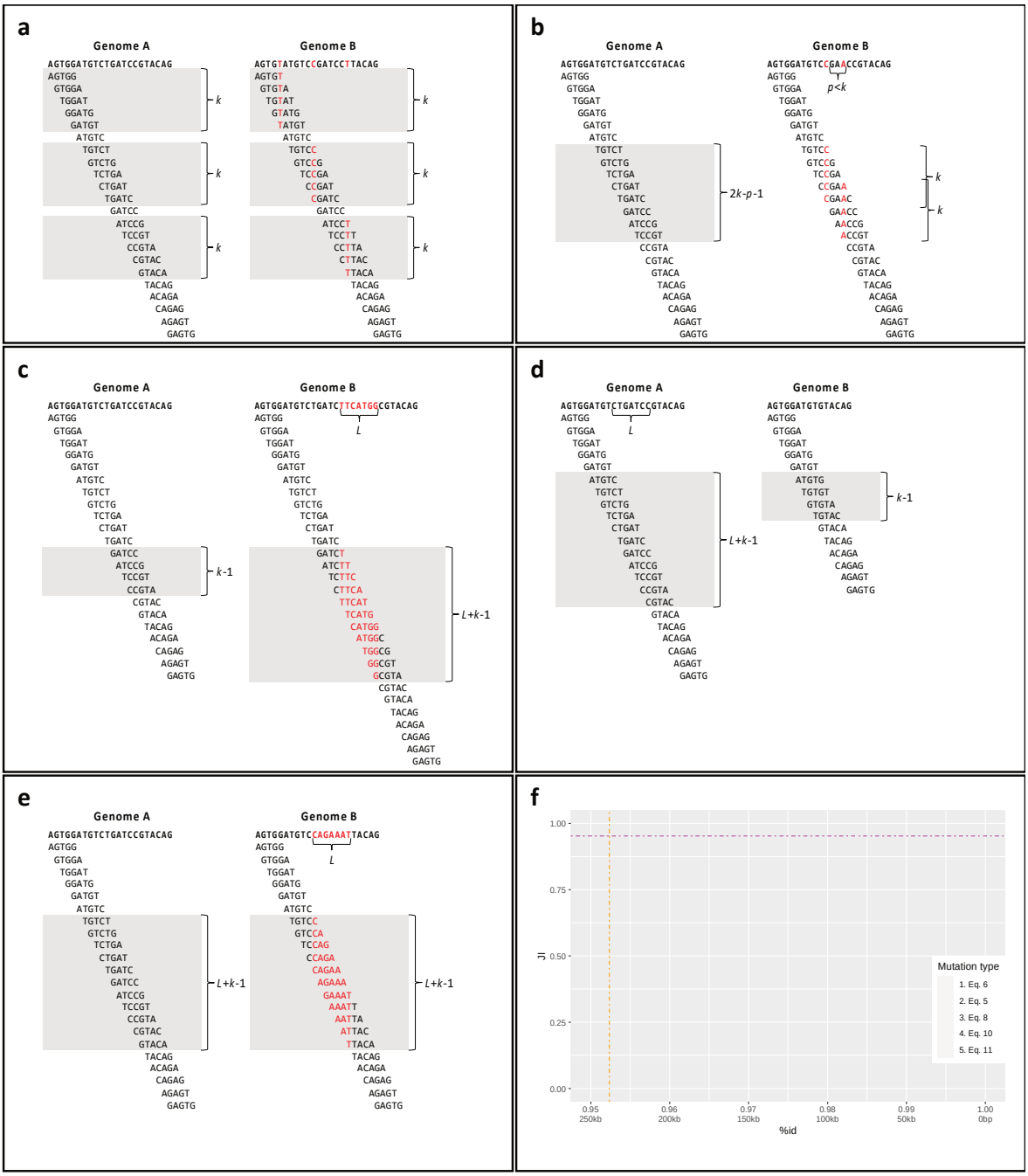


Figure 3.3 | Influence of SNPs and indels in JI. (a) Influence of SNPs between two sequences. Each genome sequence is split into *k*-mers of length *k*=5. Genomes A and B differ in *L*=3 SNPs (colored in red in genome B). These SNPs alter the *k*-mers in which they appear, leading to differences in shared *k*-mers between the two genomes. *k*-mers not shared between both genomes (*Lk* for each genome) are shaded in gray. **(b)** Influence of neighboring SNPs. The sequence of each genome is split into *k*-mers of length *k*=5. Both genomes differ in two SNPs (colored in red in genome B) located within *k* base pairs. *p* means the distance between SNPs, and it is calculated as the difference between the nucleotide position of each SNP (in this case, *p*=3). Because of their proximity, these SNPs affect fewer unique *k*-mers than if they were spread further apart. *k*-mers not shared between both genomes are shaded and they correspond to 2*k*-*p*-1 *k*-mers for each genome. **(c)**

Influence of sequence insertion. The sequence of each genome is split into *k*-mers of length *k*=5. Genome B is equal to genome A, except for an *L*=7 bp insertion (highlighted in red). *k*-mers that are not shared by both genomes are shaded in gray, and they correspond to *k*-1 unique *k*-mers in genome A and *L*+*k*-1 new *k*-mers in genome B. **(d)** Influence of sequence deletion. The sequence of each genome is split into *k*-mers of length *k*=5. Genome B is identical to genome A, except for a *L*=7 bp deletion. The *k*-mers that spanned across the deleted region are lost in genome B. *k*-mers that are not shared between both genomes are shaded in gray, and they correspond to *L*+*k*-1 unique *k*-mers in genome A and *k*-1 unique *k*-mers in genome B. **(e)** Influence of sequence replacement. Each genome sequenceis split into *k*-mers of length *k*=5. Genome B differs from genome A by a substitution of *L*=7 bp, which is indicated in red letters. *k*-mers that are not shared between both genomes are shaded in gray, and they correspond to *L*+*k*-1 *k*-mers for each genome. Since the entire replaced sequence introduces new *k*-mers without preserving any from the original, the same number of *k*-mers is unique to each genome. **(f)** Relationship between JI and % id. Each curve represents the comparison of a 5Mb reference genome with a second genome. *x*-axis represents both the percent identity (*%id*) between the genomes and the corresponding number of different base pairs between them. *y*-axis represents JI, calculated using *k*-mers of 21 bp. The cases analyzed were the following: 1. Randomly distributed SNPs: The second genome differs from the reference only by SNPs allocated using a Poisson model, according to Eq. 6 (JI=0.21208 for id=0.95%). SNPs scattered throughout the genome significantly impact JI; 2. SNPs with a Minimum Distance of *k* Base Pairs: The second genome differs only by SNPs. When SNPs are spaced exactly *k* base pairs apart, no *k*-mers are shared (Eq. 5 results in JI=0 for id=0.95%). A discontinuous vertical orange line indicates the *% id* threshold beyond which SNPs no longer satisfy the spacing condition (id=0.95238%=1-(1/*k*)); 3. The second genome contains insertions, Eq. 8 (JI=0.95237 for id=0.95%). Since the inserted sequence adds new *k*-mers but does not remove any, the JI remains relatively high. A discontinuous horizontal magenta line indicates the JI value for which 5,880 random SNPs (id=0.99882%, case 1) are equivalent to a 250 kb sequence insertion (id=0.95%); 4. The second genome differs by sequence deletions, Eq. 10 (JI=0.94999 for id=0.95%). The effect is similar to insertions since most of the original *k*-mers are retained; and 5. The second genome differs by sequence substitutions, and this replacement produces entirely new *k*-mers, Eq. 11 (JI=0.90475 for id=0.95%). This has a greater impact on JI compared to insertions or deletions because the original *k*-mers are lost.

3.3.2 Genome length distance calculation

Genome length was estimated based on the number of unique *k*-mers in a genome (*S*). The upper *k*-mer length limit in Jellyfish v.2.2.6 62 (*k*=27) [212] was used to generate *k*-mers from each genome sequence, as a greater *k*-mer length reduces the probability of

random *k*-mer repetition and improves the accuracy of genome length estimation. S was computed by counting the occurrences of identical *k*-mers only once, that is, unique *k*-mers. To obtain a relative measure of genome size, the unique *k*-mer count is divided by one million base pairs (S /1000000). For each genome pair (A and B), the difference between their unique *k*-mer counts is recorded as the GLD value (Equation 12).

$$GLD_{(A,B)} = \frac{|SA - SB|}{1000000}$$

Eq. 12

Taking into account that contig ends affect *k*-mer count, a correction described by Equation 13 was applied in draft genomes, considering S, *k*, and the number of contigs of the assembly (C).

$$GLD_{(A,B)} = \frac{|SA + (k-1)CA - (SB + (k-1)CB|}{1000000}$$

Eq. 13

The difference between the unique *k*-mer counts of two given genomes is used as a proxy for genome size variation. The lowest GLD value is 0, meaning the genomes compared do not differ in size. The theoretical upper limit to GLD depends on the size difference between the smallest and largest genomes analyzed. For example, two genomes with GLD=0.05 would differ in 50 kb (0.05 x 1,000,000 bp), whereas GLD=2.0 would correspond to a difference of at least 2 Mb between them (excluding sequence duplications). Thus, the GLD value is not determined by the absolute genome sizes but rather by the difference in size between them.

Pairwise GLD values can be used on top of a given JI threshold to emphasize differences in genome size as a proxy for indels. Without the GLD filter, the difference in genome size is not considered for clustering, which is equivalent to setting the GLD threshold at its upper limit. On the other extreme, when GLD=0, only genomes fulfilling the JI threshold and having equal size will be linked. As SNPs do not alter the total genome length, whereas indels change genome size, the GLD filter ensures that any pair of genomes exceeding a given threshold of indel-derived size difference is also separated, even if they meet the JI requirement. For instance, applying only a JI threshold of ≥0.983 could explain between-group differences as either 86 kb (from indels), 2,050 SNPs, or a combination of both. However, by applying both the JI ≥0.983 threshold and the GLD filter (GLD ≤0.05), any pair of genomes differing by more than 50 kb of indels will also be separated, regardless

of whether it fulfills the JI threshold. This approach enhances genome separation based on size variations while minimizing the influence of SNPs.

3.3.3 Network visualization

This step involves visualizing genomic relationships as an undirected network, which is constructed using the adjacency matrix. Gephi v10 [207] was employed for network visualization, applying the ForceAtlas2 algorithm for layout.

This network visualization enables the exploration of clusters and relationships based on genetic similarities and genome size differences. Pairwise genome similarities can be represented in an undirected network, where nodes (genomes) are connected if the pairwise JI equals or exceeds the specified JI threshold (and GLD threshold, if applied). At the initial network stage, genomes sharing any JI value greater than 0 will be linked by an edge, resulting in most genomes forming a single connected component. By increasing the stringency of the JI threshold, separate connected components emerge (**Figure 3.4**).

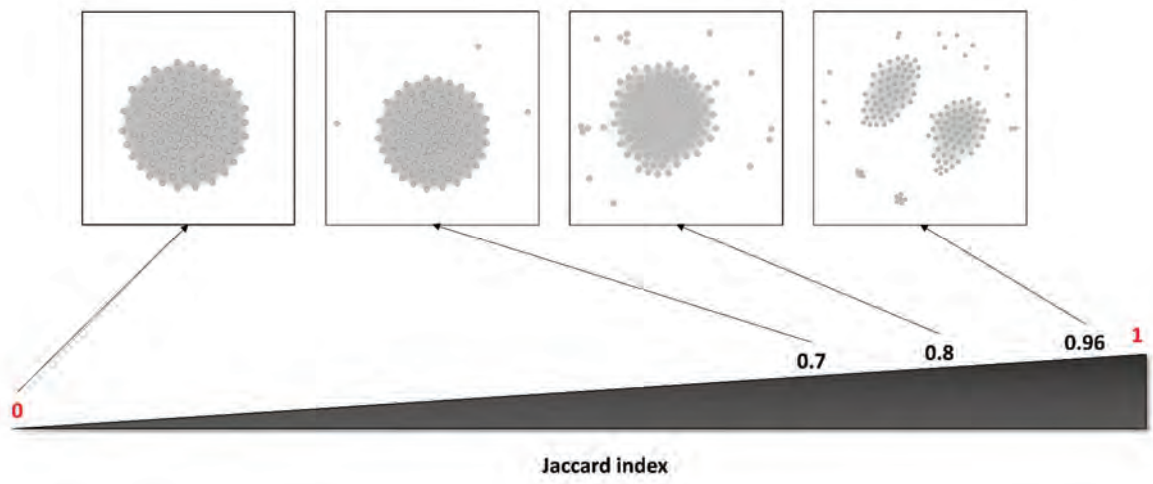


Figure 3.4 | Example of a network representation based on pairwise genome similarities. Nodes represent genomes, and edges are drawn when the pairwise JI value meets or exceeds the defined threshold. The initial network includes all genomes connected by any JI value above 0, forming a single component, while increasing the JI threshold (0.7, 0.8 and 0.96) results in the emergence of separate connected components.

3.3.4 Analysis of network parameters

To define the final components to study, referred to as JI-groups, a range of JI thresholds must be assessed for a given application. In the studied datasets, no distinct valley was observed in the distribution of JI values. We thus evaluated multiple statistical measures to optimize the network sparsification. Specifically, we considered transitivity, which indicates groups of nodes with strong internal connections, quantifying how likely it is that two neighbors of a node are also neighbors of each other. We also evaluated the number of communities (with sizes either smaller or larger than five members), and the proportion of genomes within these two types of communities.

The optimal threshold depends on the specific study population and research objectives. To reduce complexity in this study, we recommend setting a threshold that results in a manageable number of JI-groups. Ideally, the number of clusters should not exceed the natural logarithm of the total genomes, ensuring the largest possible grouping while maintaining high transitivity and alignment with relevant genetic determinants, if available. The different network properties (transitivity, number of communities and proportion of clustered genomes) were calculated using the igraph v2.0.1 package in R (https://r.igraph.org/articles/igraph.html).

3.3.5 Clustering algorithm for detecting JI-groups

The Louvain method was used to define JI-groups. This algorithm optimizes modularity to detect communities with dense internal connections while minimizing links between groups. The resolution parameter is key to controlling the granularity of the clusters. The Louvain method implemented in Gephi, was used with a resolution of 1.5. A similar implementation of this method is available in the igraph package in R (https://r.igraph.org/articles/igraph.html); however, due to the inverse relationship of the resolution parameter in igraph, a resolution of 0.55 was used to obtain the same grouping results. Once the main JI-groups are defined (each containing at least five genomes), they can be further subdivided into several subgroups within the network using a more stringent JI and the same community detection algorithm. The resulting JI-groups were named alphabetically. For JI-subgroups, the parent group's letter is followed by a number, with the largest subgroup labeled as JI-A1.

3.3.6 Mapping metadata into the network

The network nodes, representing genomes, can be colored based on metadata and genetic determinants of interest. This mapping allows for the visual evaluation of associations between the identified groups and various epidemiological data, genetic factors, or other relevant variables. By integrating these layers of information, the network visualization can reveal potential correlations and patterns, highlighting relationships that may guide further analyses. Moreover, metadata can assist in selecting the optimal JI threshold.

3.3.7 Detection of insertions and deletions between JI-groups

The identification of indels, including MGEs and other accessory genome regions, between JI-groups was carried out using two distinct approaches. The first approach, applied to Typhi, was initially employed due to unavailability of the second method at the time of analysis. The second approach, introduced later, was specifically designed for pangenome analysis.

3.3.7.1 BLASTn-based indel detection

The first approach involves BLAST v. 2.6 [211] to compare reference genomes from each JI-group. Reference genomes refer to complete genomes retrieved from a public databases or genomes sequenced using long reads. For those JI-groups without an available reference genome, a representative genome was reconstructed using PLACNETw [203] or MOB-suite [32]. The BLASTn searches using a cutoff e-value of 10^{-5} were conducted between all possible reference pairs from different JI-groups to detect regions present in one genome but absent in the other. Following this step, the identified regions were mapped against all genomes using another round of BLASTn searches using a cutoff e-value of 10^{-5}. This step ensures that the detected regions were present in >=90% of the genomes within the specific JI-group from which they originated (**Figure 3.5**).

Figure 3.5 | Detection of JI-group specific indels using BLASTn. (A) BLASTn was performed between the reference genomes of each JI-group. **(B)** BLASTn searches between all possible reference genome pairs enabled the identification of regions present in one genome but absent in the other. (C) The identified regions were further validated through a second round of BLASTn searches, where they were confirmed to be present in at least 90% of the genomes within the corresponding JI-group.

3.3.7.2 PanGraph-based indel detection

PanGraph v0.7.3 [77] identifies blocks of homologous sequence and was used to detect indels specific to each JI-group. PanGraph was run on all genomes using parameters `α=20` and `β=20`. The α parameter controls the cost of splitting a block into smaller units, with a value of 20 chosen to minimize excessive fragmentation of the graph. The β parameter, which regulates the diversity cost, was set to 20, allowing a sequence diversity threshold of 20%. Only homologous sequences (pancontigs), larger than 250 bp, present in ≥85% of the members of a given JI-group but absent from all JI-groups, were retained as "core" pancontigs. Core pancontigs of each JI-group were then mapped with BLASTn against a reference genome (preferably sequenced with long-read technology) from their respective JI-group. This step ensured proper ordering of the pancontigs and the identification of the regions they form. This is particularly important because accessory genome elements often consist of multiple consecutive pancontigs. The continuity provided by long-read sequencing ensures accurate reconstruction and ordering of these regions (**Figure 3.6**).

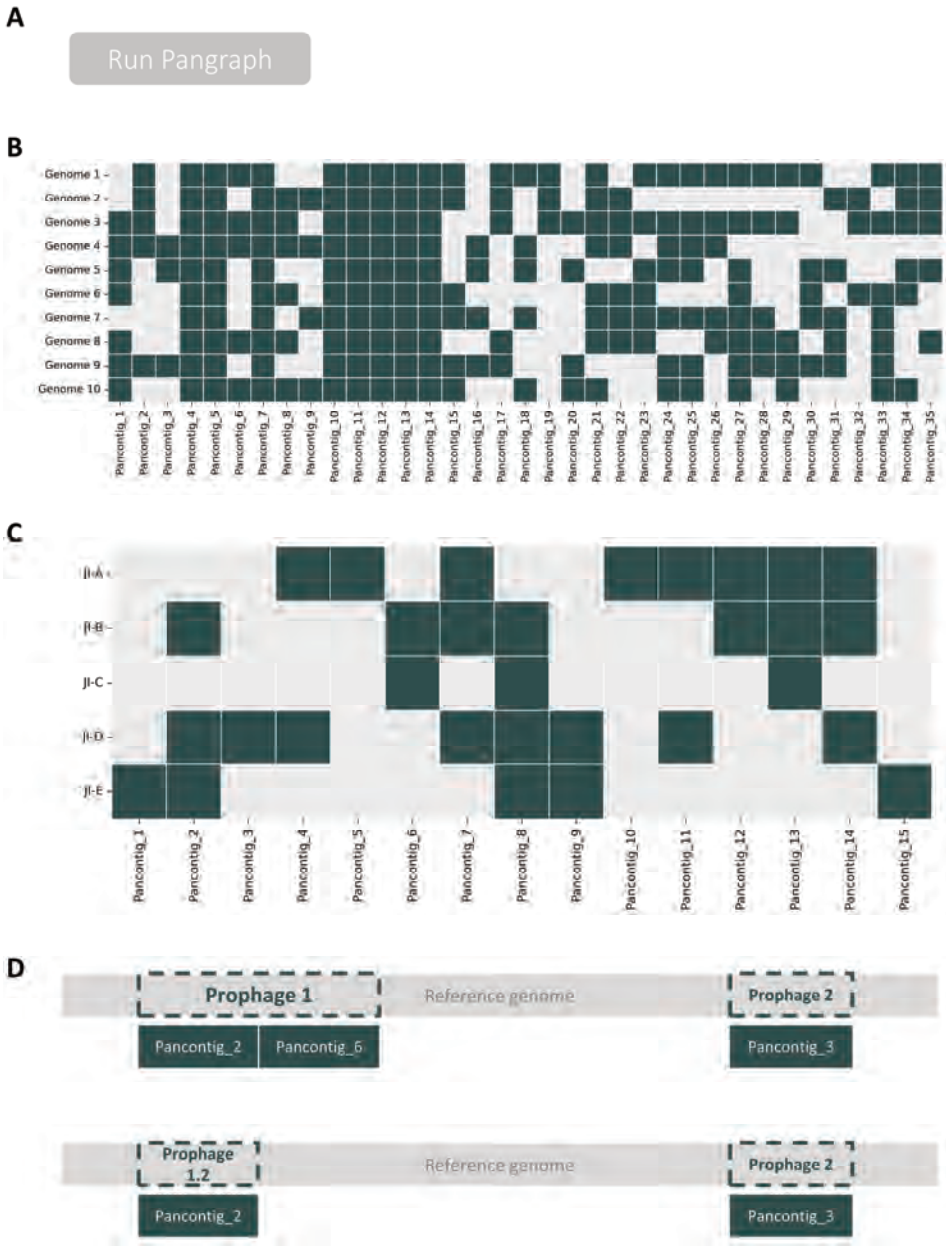


Figure 3.6 | Detection of JI-group specific indels using PanGraph. (A) Run PanGraph. **(B)** Presence/Absence matrix: A matrix is generated to display the presence (green cells) or absence (gray cells) of pancontigs across all genomes. **(C)** Classification into JI-groups: Genomes are classified by the JI-groups they belong to. Pancontigs that meet the threshold for defining "core" pancontigs in each JI-group are retained for further analysis. This process generates a matrix, where the core pancontigs of each JI-group are indicated (green cells). **(D)** Mapping core pancontigs to a reference genome: The core pancontigs are mapped with BLASTn against a reference genome to determine their order and identify the regions they form. For example, pancontig_2 and pancontig_6 are found to be consecutive, so they are considered to form an accessory element, in this case, prophage 1.

3.3.7.3 Classification of the detected regions

The indels detected when analyzing the differences between the JI-groups were classified as follows. Chromosomally-integrated region containing at least five phage-related genes according to PhageScope [196] or PHASTEST [197] were classified as a prophage. Regions of unknown function (also dentoted as “Region” or “RUF” for simplicity) were designated for sequences with no clearly defined function; however, if they contained a MPF system, this was noted in the name. Regions with multiple genes associated with specific traits were labeled accordingly, such as AMR-encoding transposons. Integrative and Mobilizable Elements (IME) were identified when they contained relaxase genes (MOB genes), while ICEs were classified when they carried both MPF system genes and MOB genes. Regions classified as plasmids were assigned to the corresponding PTU when available, or labeled "NA" (not assigned) if the PTU could not be determined.

3.4 Complementary genomic similarity metrics

3.4.1 FastANI

FastANI v1.34 [213] uses alignment-free approximate sequence mapping to calculate Average Nucleotide Identity (ANI) between pairs of genomes, providing a rapid and accurate estimation of genetic relatedness. For this study, pairwise ANI values were computed for all genome pairs within each JI-group to estimate and compare the results obtained from the JI analysis, using default parameters.

3.4.2 PopPUNK

PopPUNK v2.7 [66] was employed to generate a *k*-mer sketch database of the analyzed genomes (*poppunk --create-db*), with a sketch size of 100,000 to improve the accuracy of JI calculations. This approach enabled an independent assessment of core and accessory genome variation, offering a comprehensive overview of genomic diversity in both compartments.

3.5 Pangenome comparison and gene prediction

For pangenome comparison between U.S. and non-U.S. Hadar genomes, gene prediction of the assembled genomes was performed with Prokka v1.14.5 [214]. Annotated assemblies in GFF3 format were used as input for pangenome calculation using Roary v3.13 [67], with 80% minimum percent identity and coverage. Pangenome gene categories were defined as: core genes (shared by 80–100% of the genomes); shell genes (15–79%); and cloud genes (0–14%). Heaps' law was used to evaluate pangenome openness and closeness, using the script available at https://github.com/SethCommichaux/Heap_Law_for_Roary.

3.6 Phylogenetic analysis

The methods used for phylogenetic reconstruction were guided by the type of genetic data and the specific objectives of each analysis. WGS and plasmid sequences require different approaches due to differences in genome structure and variation patterns. The strategies employed focused either on SNPs derived from the core genome or leveraged existing phylogenetic frameworks such as cgMLST. All the phylogenetic trees generated in this study were visualized with iTol v5-6 [215,216].

3.6.1 SNP-based phylogenetic analysis

The detection of SNPs is essential for reconstructing accurate phylogenies. These SNPs can be identified in different regions of the genome, but for the purpose of this analysis, we focused on SNPs from the core genome. Two approaches were used for SNP detection:

3.6.1.1 K-mer-based method

SNPs were identified by analyzing the *k*-mer patterns from the WGS using kSNP 3.0 [217]. The optimal *k*-mer size was determined using the Kchooser tool implemented within kSNP3. The resulting phylogenetic trees were generated using the maximum parsimony method, which is the default tree type in kSNP3. For this analysis, we specifically used the "-core" option to extract SNPs from the core genome.

3.6.1.2 Core genome alignment-based methods

This approach focuses on generating core genome alignments and extracting the SNPs for phylogenetic reconstruction. Two strategies were employed to obtain these alignments.

(1) **Reference-based mapping and variant calling:** assemblies were mapped to a reference genome, and SNPs were identified using the variant-calling tool Snippy v4.6 (https://github.com/tseemann/snippy). When SNPs were called for multiple isolates against the same reference genome, a "core SNP" alignment was generated. This alignment represents genome positions (core sites) present across samples, which may either be identical in all samples (monomorphic) or show variation across samples (polymorphic or variant). By focusing on polymorphic core sites and excluding complex variations like insertions and deletions, a core SNP genome alignment was produced.

(2) **Multi-sequence alignment and SNP-sites:** core genome alignments were generated using PanACoTA v1.3.1 [68]. In this study, the *pangenome* and *corepers* modules were used to identify and align the core genes. Genes were considered as core if present in at least 80% of the genomes. Briefly, the process began by constructing the pangenome through clustering all protein sequences using MMseqs2 [75], applying a protein identity threshold of >80%. From the resulting pangenome, the core genome was subsequently retrieved. Multiple sequence alignments of the core genes families were performed using the *align* module, which uses MAFFT v7.467 [218]. These amino acid alignments were then back-translated to nucleotide alignments. Finally, the nucleotide alignments were concatenated to create a unified core genome alignment.

In both cases, the alignments were used to reconstruct Maximum Likelihood (ML) phylogenies using IQ-TREE2 v2.2.2.1 and v2.2.2.3 [219] with the ultra-fast bootstrap option (-bb 1000 bootstraps) [220]. The best fitting model estimated using ModelFinder Plus (-MFP) [221] and selected based on the Bayesian Information Criterion (BIC). All these phylogenetic trees were midpoint-rooted.

3.6.2 Multi-locus sequence typing scheme

cgMLST of *Salmonella* is a well-established framework that analyzes 3,002 loci within the core genome [39]. *The cgMLST-based phylogenetic tree was generously generated by our collaborator, Kaitlin Tagg, using BioNumerics v7.6.3* [222].

3.7 MGE removal to assess the contribution of MGEs in the JI-groups

To explore the contribution of MGEs to JI-group clustering, MGE (plasmids and/or SGI11) sequences were manually removed from the nucleotide fasta files of selected genomes. The resulting "cured" sequences were then used to compute JI and generate networks, following the methodology explained in the *3.3* section.

3.8 Statistical analysis

I acknowledge Kaitlin Tagg for computing all statistical analysis detailed in this section. Each dataset required a different approach based on its specific aims, and the details of these analyses are described in the following subsections.

3.8.1 Statistical analysis of the *Salmonella* Typhi dataset

The varpart function implemented in the vegan Community Ecology R package (https://search.r-project.org/CRAN/refmans/vegan/html/varpart.html) was used to partition the variance in JI-groups with respect to GenoTyphi lineages and MOB relaxase genes using an adjusted R^2 value. Chi-squared tests of independence were performed to examine geographic signals associated with JI-groups.

3.8.2 Statistical analysis of the *Salmonella* Hadar dataset

Statistical analyses were performed using genomes collected through NARMS (CDC, FDA, FSIS), PulseNet (CDC), and FSIS U.S. surveillance systems from years 2016 through 2023, aligning with the introduction of routine sequencing for NARMS, PulseNet, and FSIS surveillance isolates. Corrected Cramer's *V* was used to measure the strength of

associations between all categorical variables [223]; chi-squared tests of independence were used to test associations between specific epidemiological and genomic variables (Bonferroni adjusted significance value: $p < 0.005$). Odds ratios (OR) (95% confidence intervals (CI)) were used to quantify the strength and direction of significant associations. For statistical tests involving a specific JI-group, the comparison group was always "all other JI-groups". All tests were conducted using the stats subpackage of SciPy v1.14.1 implemented in Python v3.11.7 (https://docs.scipy.org/doc/scipy/reference/stats.html). JI-groups with less than 20 genomes were not analyzed for statistical associations. Only NARMS surveillance data collected by CDC, FDA, and FSIS (cecal sampling) were used to assess shifts in pangenome group abundance over time. The NARMS dataset was systematically collected and more resistant to large outbreaks and regulatory testing changes than the PulseNet and FSIS product sampling datasets.

3.9 Data availability

All supplementary tables (Table S1 to Table S8) are available at the following Google Drive link:

https://drive.google.com/drive/folders/1zgdx6FQfu2mLsS1J0n9DZ5BcDt8XsB5x?usp=sharing. These tables contain all genomes, accession numbers, and associated metadata for the genomes analyzed in this thesis.

Script for the Jaccard Index calculation using BinDash are available in the public repository: https://github.com/PenilCelis/Salmonella_Typhi_JINA

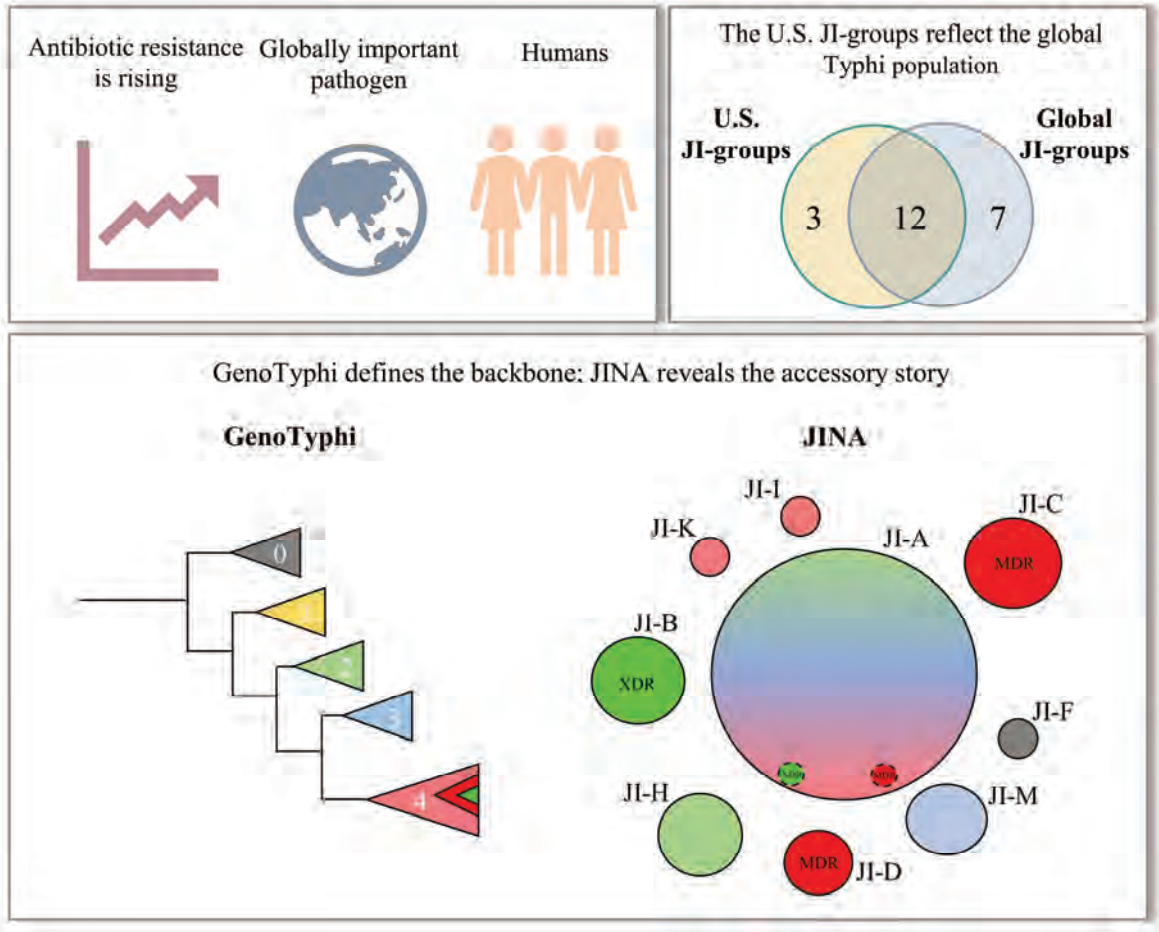
The scripts used to generate most of the figures, created using R or Python, are available at the following Google Drive link:

https://drive.google.com/drive/folders/1qvA52_AzAo3lkGPTrQldlHcLBbJK02P0?usp=sharing. Each file corresponds to a script used to generate a figure, and the filename matches the figure number as referenced in this thesis. Whenever possible, the scripts include data loading, analysis, and plotting in a self-contained manner. Additional information or data sources (if applicable) are described in comments within each script.

CHAPTER 4: RESULTS I

SALMONELLA ENTERICA SEROVAR TYPHI

Graphical abstract



Salmonella enterica serovar Typhi (Typhi) is a human-restricted pathogen and the causative agent of typhoid fever, a major global public health concern. Antimicrobial treatment is essential for disease management, but the rise of antimicrobial resistance in Typhi threatens effective control efforts. This study integrates two complementary genomic frameworks, GenoTyphi and Jaccard Index Network Analysis (JINA) to resolve population structure and uncover hidden diversity among Typhi isolates from the United States (U.S.). GenoTyphi defines the phylogenetic lineage structure based on core genome SNPs, while JINA provides additional resolution by distinguishing strains based on mobile genetic elements (MGEs). Notably, strains within the same GenoTyphi type may belong to different JINA groups if they carry distinct MGEs. All MDR and XDR strains fall within GenoTyphi clade 4, yet JINA uncovers finer substructure based on mobile elements. These two complementary perspectives reveal distinct layers of genomic diversity. The U.S. Typhi isolates reflect broad global diversity, with most JINA groups overlapping between local and international datasets, reinforcing the global relevance of these findings.

4.1 Background and specific objectives

Salmonella enterica subsp. *enterica* serovar Typhi (Typhi) is a human-restricted bacterial pathogen that causes typhoid fever, a serious systemic illness associated with high morbidity and mortality [108]. In 2017 [106], approximately 14 million cases of enteric fever were reported globally, resulting in an estimated 135,000 deaths. It is transmitted primarily through the fecal-oral route, typically through ingestion of contaminated food or water. Typhi is prevalent in regions with poor sanitation and hygiene, including urban slums in South Asia and sub-Saharan Africa.

Antimicrobials are essential for the effective treatment of typhoid fever. However, the emergence and spread of AMR in Typhi pose a significant threat to disease control [122,200]. As noted in the *Introduction* chapter, the widespread emergence of MDR strains rendered first-line antibiotics ineffective. Subsequently, fluoroquinolones (particularly ciprofloxacin) became the primary oral treatment for typhoid fever. However, their use led to the emergence and dissemination of fluoroquinolone-resistant strains. More recently, XDR Typhi (defined by resistance to fluoroquinolones, third-generation cephalosporins, and MDR phenotypes) has emerged, leaving azithromycin as one of the few remaining oral treatment options (**Figure 4.1**) [123,126].

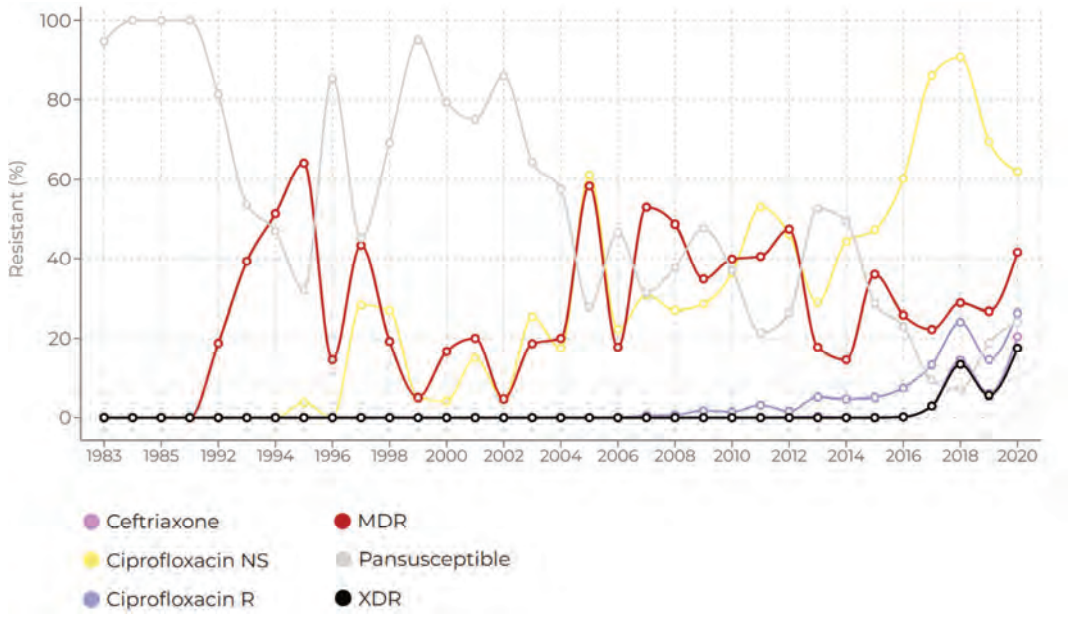


Figure 4.1 | Trends in antimicrobial resistance among *Salmonella* Typhi isolates, 1983–2020. Percentage of Typhi isolates resistant to selected antibiotics over time, based on data from TyphiNET, which included 11,836 genomes at the time the figure was generated. Resistance is

shown for ceftriaxone (purple), ciprofloxacin non-susceptible (NS, yellow), ciprofloxacin-resistant (R, blue), multidrug-resistant (MDR, red), extensively drug-resistant (XDR, black), and pansusceptible (gray) strains. Figure taken from https://www.typhi.net/.

Moreover, azithromycin-resistant Typhi has been documented in Bangladesh [224], Pakistan [225], Nepal [226], India [227], and the U.S. [189]. A recent report described a clinical case in Pakistan involving XDR Typhi resistant to both carbapenems and azithromycin [225], highlighting the potential emergence and expansion of untreatable typhoid with current oral antimicrobials. Such a scenario would impose significant costs on healthcare systems and exacerbate challenges in resource-limited settings.

Typhi is a pathogen with low genetic variability, a slow mutation rate, and infrequent recombination events [159,228]. Despite its limited genetic diversity, Typhi is phylogenetically informative for tracking antimicrobial resistance and understanding transmission dynamics [124,229–231]. An early SNP-based genotyping scheme (based on a limited number of genes) [228] identified 85 haplotypes, providing a foundational framework for epidemiological studies. Notably, this study was the first to report the expansion of the highly clonal MDR haplotype H58. With the advent of WGS, a more comprehensive genotyping approach, the GenoTyphi scheme [184,193], was developed in 2016. Derived from an analysis of nearly 2,000 Typhi genomes spanning more than 60 countries, this scheme employed 68 marker SNPs to classify Typhi into 4 primary clades, 16 clades, and 49 subclades. Primary clade 1 is subdivided into clades 1.1 and 1.2; clade 1.1 is further subdivided into subclades 1.1.1, 1.1.2, 1.1.3, and so on. The median divergence between genomes decreases from primary clades to subclades: 243 SNPs for genomes contained in the same primary clade, 109 within the clades, and 25 within the subclades. This framework has played a crucial role in identifying key epidemiological trends, including the global dissemination of the MDR-associated H58 clade, designated as genotype subclade 4.3.1 [159,232].

GenoTyphi was expanded in 2021 [193] to include new genotypes that address regional diversity and emerging AMR. For instance, subclade 4.3.1 was divided into three lineages (4.3.1.1, 4.3.1.2, and 4.3.1.3), with additional designations for epidemiologically significant populations, such as 4.3.1.1.P1 (the XDR Typhi strain from Pakistan) and 4.3.1.3.Bdq (a fluoroquinolone-resistant lineage from Bangladesh) sublineages (**Figure 4.2**). Notably, these new sublineages (4.3.1.1.P1, 4.3.1.3.Bdq) were defined based on specific AMR

profiles or geographic distribution rather than differences in the core genome. By refining genotype definitions and increasing resolution, GenoTyphi has improved the discrimination of lineages, thereby enabling precise tracking of regional and global transmission patterns.

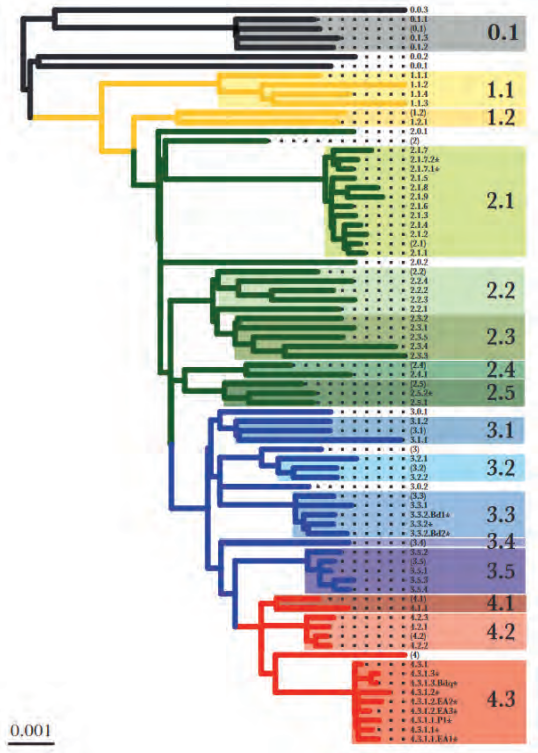


Figure 4.2 | GenoTyphi scheme. Phylogenetic tree showing the relationships among 16 clades and 63 subclades, lineages, and sublineages. Tree tips represent unique genotypes as labeled, and background shading highlights clades (labeled in larger font). An asterisk (*) indicates genotypes added to the scheme after its initial publication, and brackets indicate undifferentiated clades and primary clades. Modified from [193].

Nevertheless, a purely core-genome or SNP-based perspective may overlook the short-term evolutionary changes driven by the acquisition, loss, or rearrangement of MGEs in the accessory genome. These MGEs often carry AMR and virulence genes, rapidly altering the pathogen's resistance and virulence profiles. Understanding the accessory genome's contribution to Typhi evolution is particularly relevant given the rise of resistant phenotypes that are frequently associated with plasmid-borne resistance determinants. For example, MDR Typhi strains initially acquired resistance determinants via plasmids [200], which subsequently integrated into the chromosome [233]. This pattern was also observed in XDR Typhi, where an MDR strain further acquired resistance to fluoroquinolones and third-generation cephalosporins through plasmid-borne genes [126]. Over time, these resistance

genes have become chromosomally stabilized [234], reducing fitness costs linked to plasmid carriage and increasing the likelihood that resistance will persist.

The emergence of azithromycin resistance in Typhi further underscores the importance of monitoring these MGEs. For instance, in the azithromycin-resistant Typhi strain isolated in the U.S. [189], the resistance-conferring gene was encoded in a plasmid. Additionally, the emergence of azithromycin-resistant XDR Typhi in Pakistan is hypothesized to result from an HGT event, in which an azithromycin-resistance plasmid entered an existing XDR background. Such events highlight the dynamic nature of the accessory genome and the necessity of pangenomic approaches to monitor the spread of highly resistant strains.

Therefore, this chapter aims to evaluate the benefits of extending beyond core genome typing, where Typhi is already well-characterized, by incorporating accessory genome analysis to highlight how MGEs drive resistance and genomic diversity, particularly in the U.S. context. To address this, we propose the following specific objectives:

1. Assess the value of the Jaccard Index in capturing both core and accessory genome relationships, and integrate the Genome Length Distance metric to emphasize the impact of insertions and deletions.
2. Conduct a comprehensive pangenome analysis of the largest U.S. Typhi dataset to date.
3. Compare U.S. Typhi isolates with a global reference dataset to examine the genomic diversity and regional variations.
4. Investigate the specific role of MGEs in driving rapid genomic shifts, such as the emergence of novel lineages and resistance phenotypes that are not captured by core genome analyses alone.
5. Demonstrate how integrating pangenome data with GenoTyphi enhances resolution for lineage tracing, outbreak detection, and epidemiological surveillance, particularly in the context of emerging resistance elements.

4.2 Pangenome analysis of U.S. Typhi population

JI was used as a similarity measure between all genome pairs and was calculated with BinDash [62]. Specifically, exact JI values were obtained from pairwise comparisons within a 2,392 Typhi genome dataset, comprising 2,272 genomes isolated in the U.S. (2008-

2021) and 120 RefSeq reference genomes (**Table S1**). Their JI value distribution showed that most comparisons (99.84%) yielded JI values above 0.90 (**Figure 4.3**).

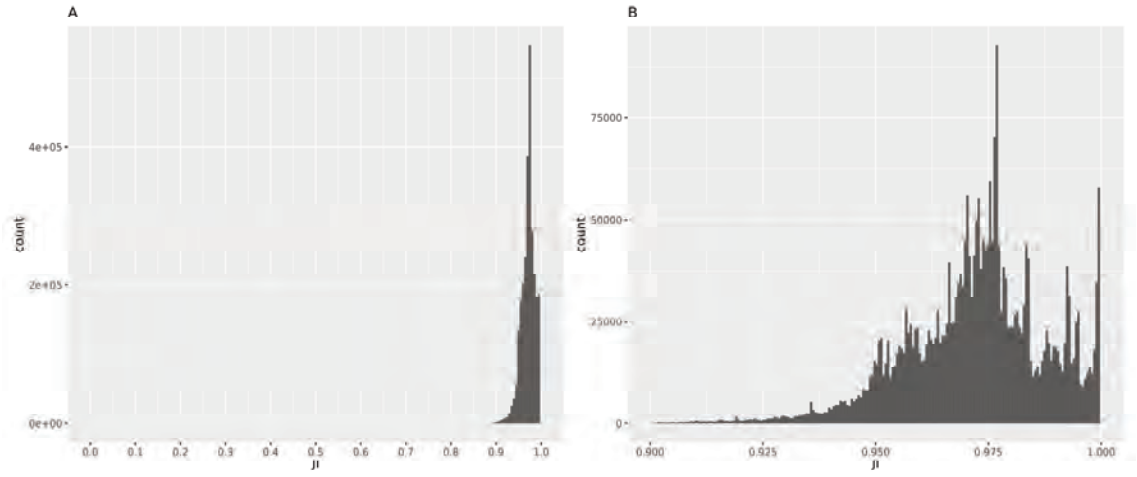


Figure 4.3 | JI distribution obtained from the pairwise comparison of Typhi genomes. (A) Histogram displaying the distribution of JI values ranging from 0 to 1. **(B)** Zoom in on JI values between 0.9 and 1.

These pairwise JI comparisons were visualized as a network, where each node corresponds to a genome and edges represent similarity based on the JI values. To identify clusters of highly similar genomes, the network was filtered by applying a threshold, removing connections between genomes that did not meet the similarity requirement (as explained in the *Materials and Methods* chapter).

For robust cluster definition, networks should exhibit a community structure characterized by subgraphs with highly interconnected members and sparser connections between subgraphs. Furthermore, to provide meaningful insights, most genomes should belong to non-singleton communities (**Figure 4.4**). The threshold used for this analysis was determined through a detailed assessment of various network metrics across different cutoff values (**Figure 4.4**). For JI values above 0.97, transitivity (an indicator of the likelihood that two neighbors of a node are also connected) increased, reflecting tightly knit subgroups of nodes within the network. The number of small communities increased exponentially for JI>=0.975. Within the JI range between 0.979 and 0.983, transitivity remained relatively stable. Beyond 0.983, transitivity changed significantly, indicating a notable shift in the network´s structure. However, at JI=0.983 and 0.984 the number of communities remained largely unchanged, but the percentage of clustered genomes slightly decreased at JI=0.984.

Given this, the optimal threshold for analyzing Typhi genomes was set at JI=0.983, as higher values caused excessive fragmentation, while lower values resulted in fewer but larger genome clusters due to excessive merging. This empirically derived threshold provides a foundation for downstream comparative genomic analyses.

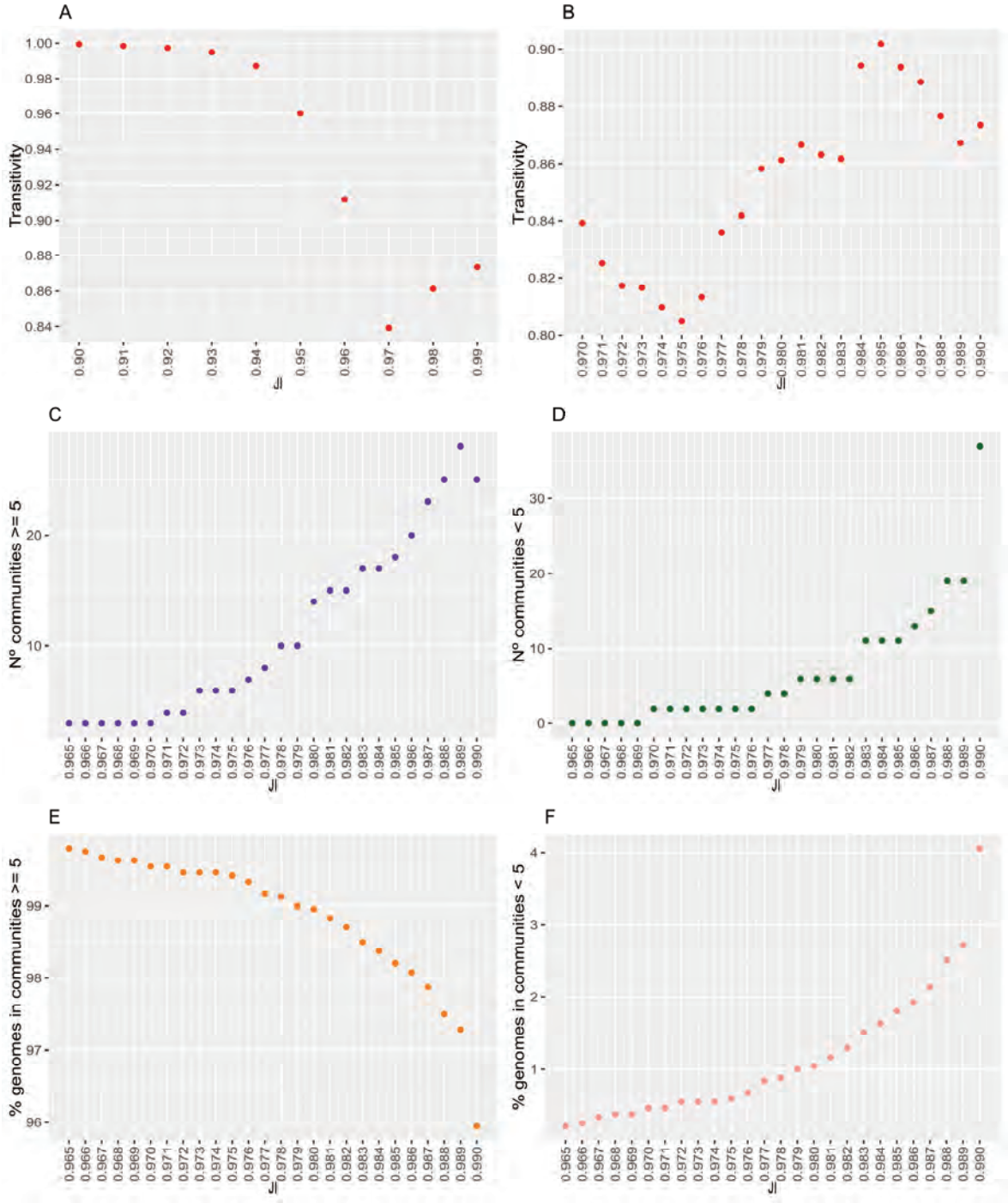


Figure 4.4 | Analysis of different networks parameters in the Typhi dataset. A range of JI thresholds was applied to the original network and several criteria were explored. **(A)** Transitivity for JI values in the range from 0.9 to 1. **(B)** Transitivity for JI values between 0.97 and 0.99. **(C)**

Number of communities containing at least five members. **(D)** Number of communities containing fewer than five members. **(E)** Percentage of genomes contained in communities with at least five members. **(F)** Percentage of genomes contained in communities with fewer than five members.

While JI captures both core and accessory genome variation, it does not distinguish their individual contributions. To better understand how Typhi genomes differ in these two genomic compartments, PopPUNK [65] was used (**Figure 4.5**). This tool calculates core and accessory Jaccard distances independently, where Jaccard distance is defined as 1-JI. In this dataset, core distances were low, consistent with the narrow genetic diversity reported for this serovar [184]. However, accessory distances exhibited a significant broader range, suggesting that gains and losses of genomic regions account for a significant portion of the observed diversity. A small subset of genomes displayed elevated core and accessory distances, indicating more divergent lineages. Although the majority of Typhi isolates shared a very tight core-genome distance, some variation remained, underscoring the importance of phylogenetic analyses in tracking evolutionary changes [184].

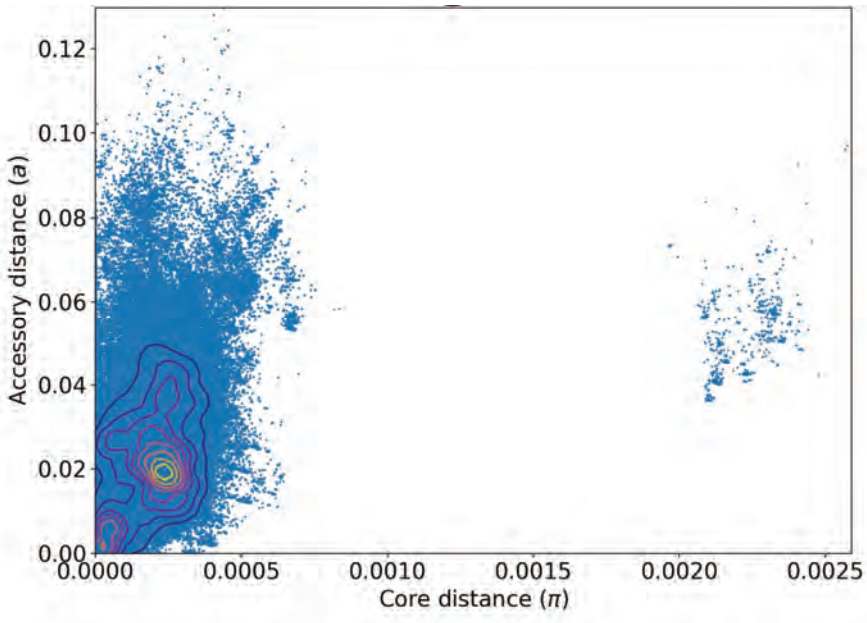


Figure 4.5 | PopPUNK analysis of Typhi isolates showing core versus accessory genome distances. Each dot represents a pairwise comparison between two genomes, with the x-axis indicating the core distance (primarily driven by SNP differences) and the y-axis indicating the accessory distance (reflecting variation in gene content). Contour lines highlight density regions, illustrating that most isolates cluster at very low core distances while displaying a broader range of accessory distances.

Therefore, in Typhi, both core and accessory genomes contribute to overall diversity, but their relative contributions differ significantly. To better capture indels differences beyond what the JI measures, an additional metric was introduced: GLD. This new metric serves a proxy for genome size variation, calculated by the difference in unique *k*-mer counts between two genomes.

On top of a given JI threshold, pairwise GLD values can be used as a proxy of indels, as SNPs do not alter genome length. This allows for emphasizing and further exploring differences in genome size. At threshold JI=0.983, between-group differences can be explained by insertions larger than 86 kb in size, or by >=2,050 SNPs across the entire genome, or a mix of both (see *Materials and Methods* chapter). By applying a conservative GLD filter of 0.05, any pair of genomes differing by more than 50 kb is separated, even if they meet the JI requirement. The chosen threshold (0.05) was further supported by the plasmid size distribution in our dataset, as the majority of plasmids found in Typhi exceed 50 kb. In our dataset, the maximum GLD value was 0.845, meaning that the largest size difference between two Typhi genomes was 845 kb. GLD thus provides an extra layer that can be conditionally applied, depending on the dataset and user requirements.

One might consider raising the JI threshold to 0.991 to approximate a 50 kb difference between groups; this threshold corresponds to around 41 kb in indels or roughly 1,000 SNPs. However, using such a high JI cutoff alone would excessively fragment the network, since any pair of genomes differing by about 41 kb or 1,000 SNPs or a combination of both would also be separated. Therefore, using a lower JI threshold combined with a GLD filter reduces the influence of SNPs. This approach achieves the same minimum 50 kb difference while emphasizing indel-driven differences over SNP-based differences.

At JI=0.983 and GLD=0.05, the 2,392 Typhi genomes self-organized into 17 distinct clusters according to the Louvain method, named JI-groups A-Q, with only 38/2,392 (1.6%) nodes not assigned (singletons or JI-clusters with less than 5 members) (**Figure 4.6**).

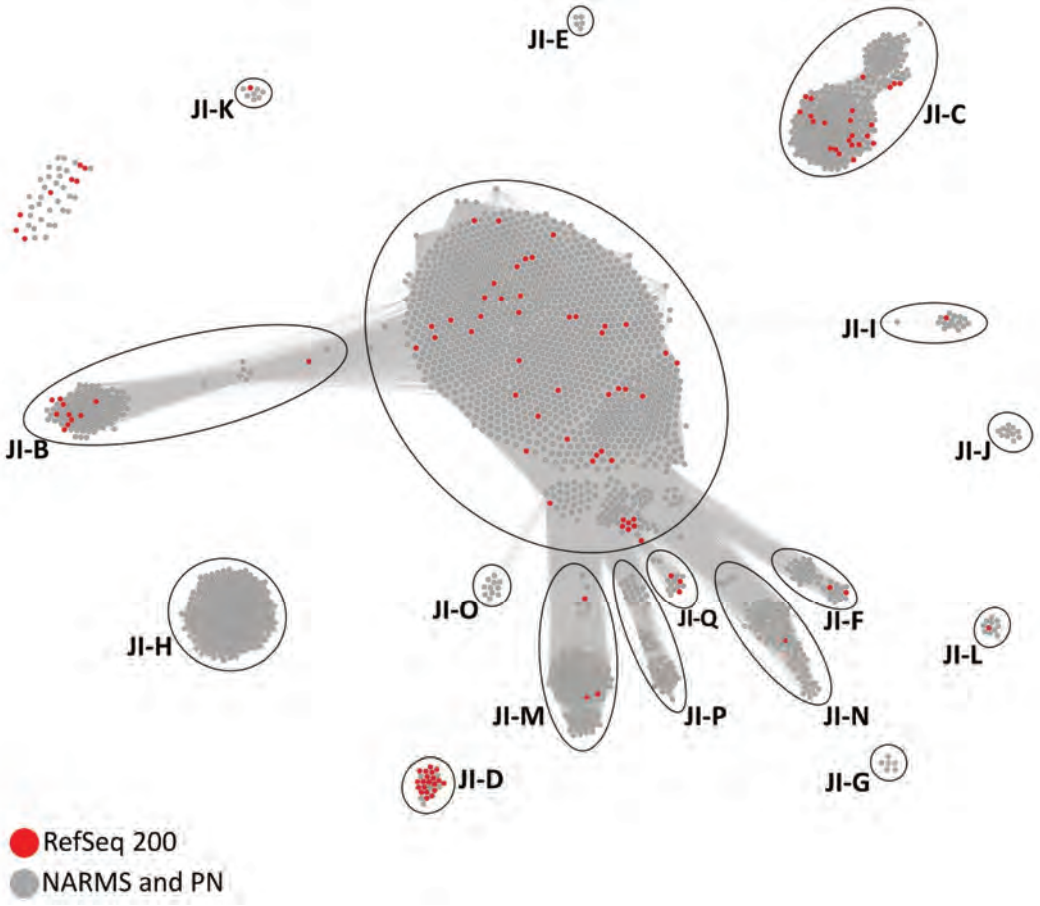


Figure 4.6 | Distribution of Typhi genomes by JI. The network contains 2,392 nodes, connected when JI>=0.983 and GLD<=0.05. Seventeen clusters (named JI-A to JI-Q) are indicated by circles. Nodes depicted in red represent RefSeq200 genomes (references) and those in grey represent NARMS and PN (PulseNet) genomes.

The relatedness of genomes within each group was > 0.988 JI and > 99.95% ANI, reflecting high genomic similarity within each group (**Figure 4.7**). However, while most groups displayed interquartile ranges (IQRs) entirely above 0.985 JI, groups A, B, C, N, and P were exceptions, with the lower edge of their boxplots (first quartile) extending below this threshold. This reveals that a significant fraction of genomes (approximately 25%) within these 5 groups have genomic relatedness values below 0.985, potentially indicating distinct genetic subgroups or less closely related genomes within these otherwise cohesive clusters. Additionally, the presence of whiskers and outliers across several groups further highlights individual genomes with lower relatedness values, reflecting isolated genomic diversity or genetic variants distinct from the main cluster of genomes in each group (**Figure 4.7**).

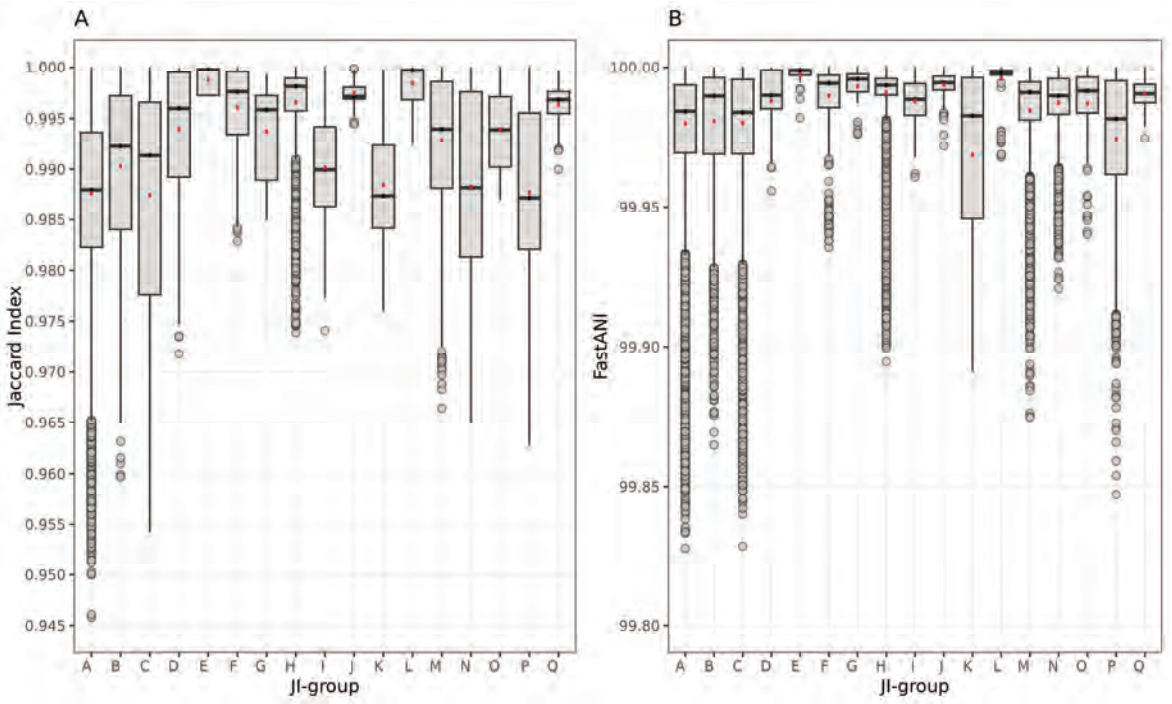


Figure 4.7 | Relatedness of Typhi genomes within each JI-group. Boxplot illustrating the distribution of **(A)** Jaccard Index and **(B)** FastANI values across different JI-groups. The boxplot displays the IQR of JI or ANI values within each JI-group, with the lower and upper edges of the box indicating the first quartile and third quartile, respectively. Within each boxplot, horizontal lines represent the median (black) and the average (red) values. The 'whiskers' of the boxplot extend to the most extreme values within 1.5 times the IQR from the edges of the box, while outliers are depicted as individual points beyond the whiskers.

JI-group A was the largest (n=1,320/2,392), with all other JI-groups represented by at least five genomes (**Table 4.1**).

Table 4.1: Summary of Typhi JI-group information for 2,272 U.S. CDC and 120 RefSeq200 genomes.

JI group	Count[a]	%[b]	GenoTyphi primary cluster
A	1320	55.1	0, 1, 2, 3, 4
B	114	4.8	4
C	265	11.1	2, 3, 4
D	26	1.1	3, 4
E	5	0.2	4
F	39	1.6	0
G	6	0.3	2
H	225	9.4	2
I	17	0.7	2, 3, 4
J	11	0.5	3
K	8	0.3	4
L	11	0.5	2
M	133	5.6	3
N	90	3.8	2, 4
O	11	0.5	3
P	58	2.4	2
Q	15	0.6	2
Singletons	38	1.6	0, 2, 3, 4

[a] | Number of genomes present in each JI group.
[b] | Percentage of genomes from the total data set that belong to each JI group.

Three of the largest JI-groups (A, B, and C) were further divided into JI-subgroups using an increased JI threshold. Subgroups were identified using the Louvain method, with different minimum size criteria based on the overall structure of each JI-group network. JI-A subgroups A1 to A17 were defined at JI=0.995 and included only those with more than five members; JI-B subgroups B1 to B3 at JI=0.986 with a minimum of two members per subgroup; and JI-C subgroups C1 to C6 at JI=0.997, including only those with more than three members (**Figure 4.8**). This higher resolution was implemented to capture genetic variations and structural patterns that were not apparent at the broader JI threshold, thus, enabling a more detailed exploration of the genomic relationships within these major groups. The chosen thresholds were determined through visual inspection of the network, balancing the number of groups and the transitivity of the network.

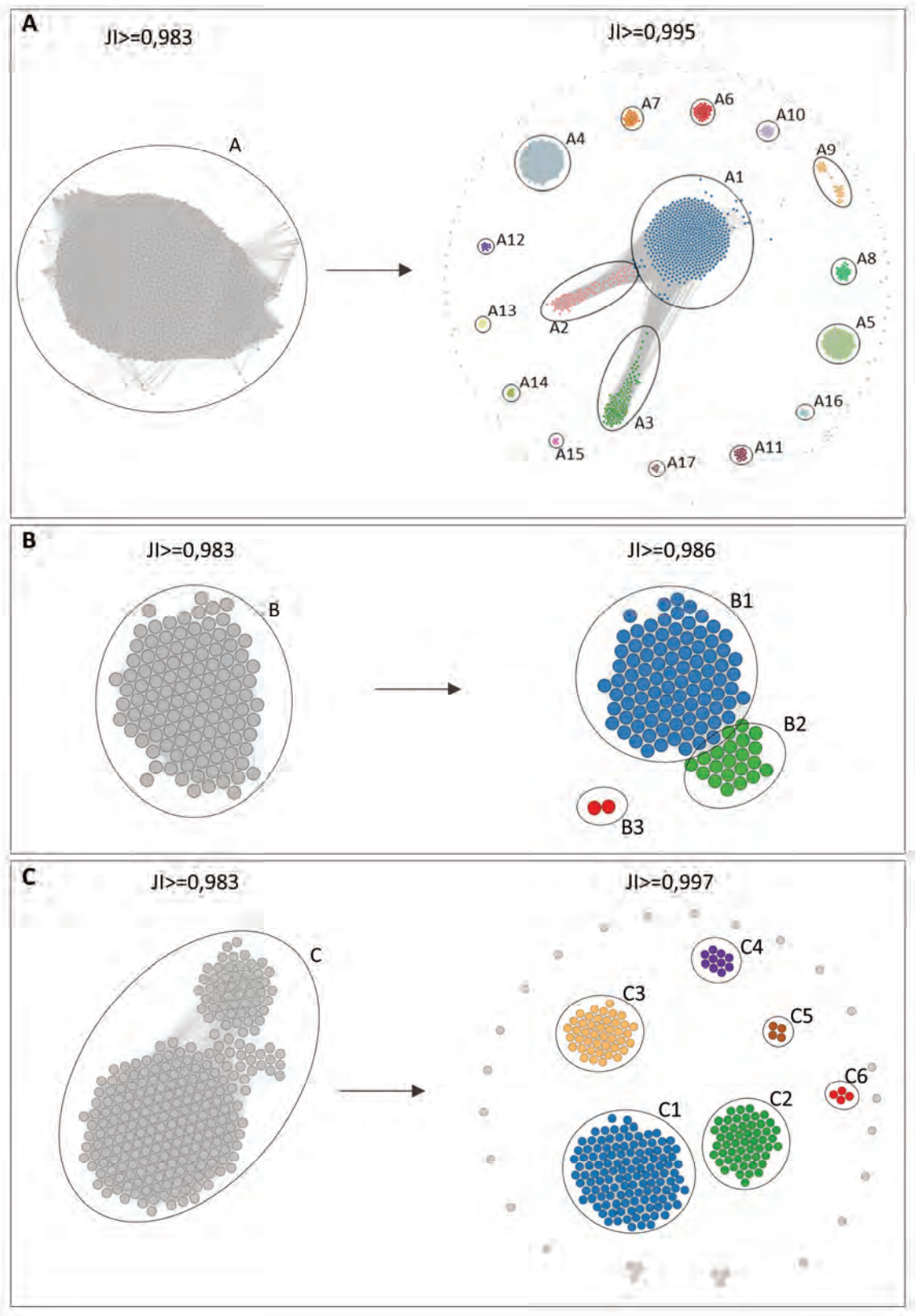


Figure 4.8 | Subclustering analysis of Typhi JI-groups A, B, and C. Each panel contains two networks, both filtered at GLD ≤ 0.05, but at different JI thresholds. In the left network, a lower JI threshold displays the entire group. In the right network, a higher JI threshold reveals subgroups, which are highlighted with circles and assigned distinct colors. **(A)** Subclustering analysis of 1,320

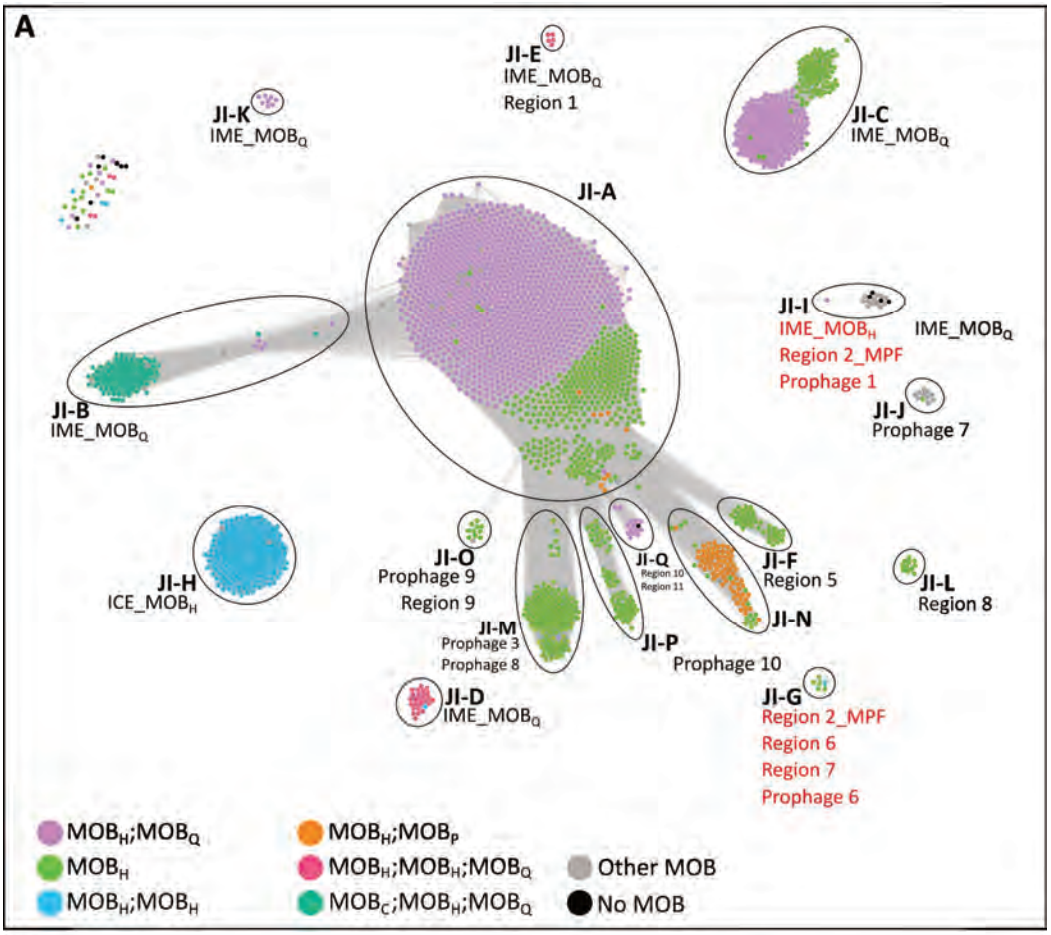
JI-A genomes. **(B)** Subclustering analysis of 114 of JI-B genomes. **(C)** Subclustering analysis of 265 JI-C genomes.

4.3 Pangenome population structure of U.S. Typhi

Exploring the U.S. Typhi dataset, we found that autonomous and integrated MGEs are ubiquitous in Typhi. A MOB relaxase gene, serving as a proxy for plasmids, ICEs and IMEs, was detected in 99.5% (n=2,380/2,392) of the isolates (**Figure 4.9A**).

JI-groups often correlated with the presence/absence of known MGEs. For example, several large (>80 kb) autonomous plasmids were found to underpin JI-group definitions. Members of JI-groups B and J all contained plasmids belonging to PTU-E50 (average size 90 kb), JI-group C contained PTU-E18 (average size 107 kb), JI-group D contained PTU-HI1A (average size 217 kb), and JI-group K contained PTU-Y plasmids (average size 100 kb) (**Table 4.2, Figure 4.9B**). Plasmids <40 kb, such as PTU-N1 and PTU-X1 in JI-groups A, H, and N, did not define JI-groups (at thresholds JI=0.983 and GLD=0.05) due to their relatively small size.

Many unknown MGEs and accessory regions were also responsible for the genetic difference between JI-groups: JI-E was defined by the presence of a 49 kb region of unknown function, while JI-P members all carried a unique 44 kb phage element (prophage 10) (**Figure 4.9A, Table 4.3**). Each JI-group was found to contain a unique complement of accessory genome elements, many of which were undetectable by current routine methods.


A
JI-E
IME_MOB_Q
Region 1
JI-K
IME_MOB_Q
JI-C
IME_MOB_Q
JI-A
JI-I
IME_MOB_H
IME_MOB_Q
Region 2_MPF
Prophage 1
JI-J
Prophage 7
JI-B
IME_MOB_Q
JI-H
ICE_MOB_H
JI-O
Prophage 9
Region 9
JI-Q
Region 10
Region 11
JI-F
Region 5
JI-L
Region 8
JI-N
JI-M
Prophage 3
Prophage 8
JI-P
Prophage 10
JI-D
IME_MOB_Q
JI-G
Region 2_MPF
Region 6
Region 7
Prophage 6
MOB_H;MOB_Q
MOB_H;MOB_P
Other MOB
MOB_H
MOB_H;MOB_H;MOB_Q
MOB_H;MOB_H
MOB_C;MOB_H;MOB_Q
No MOB

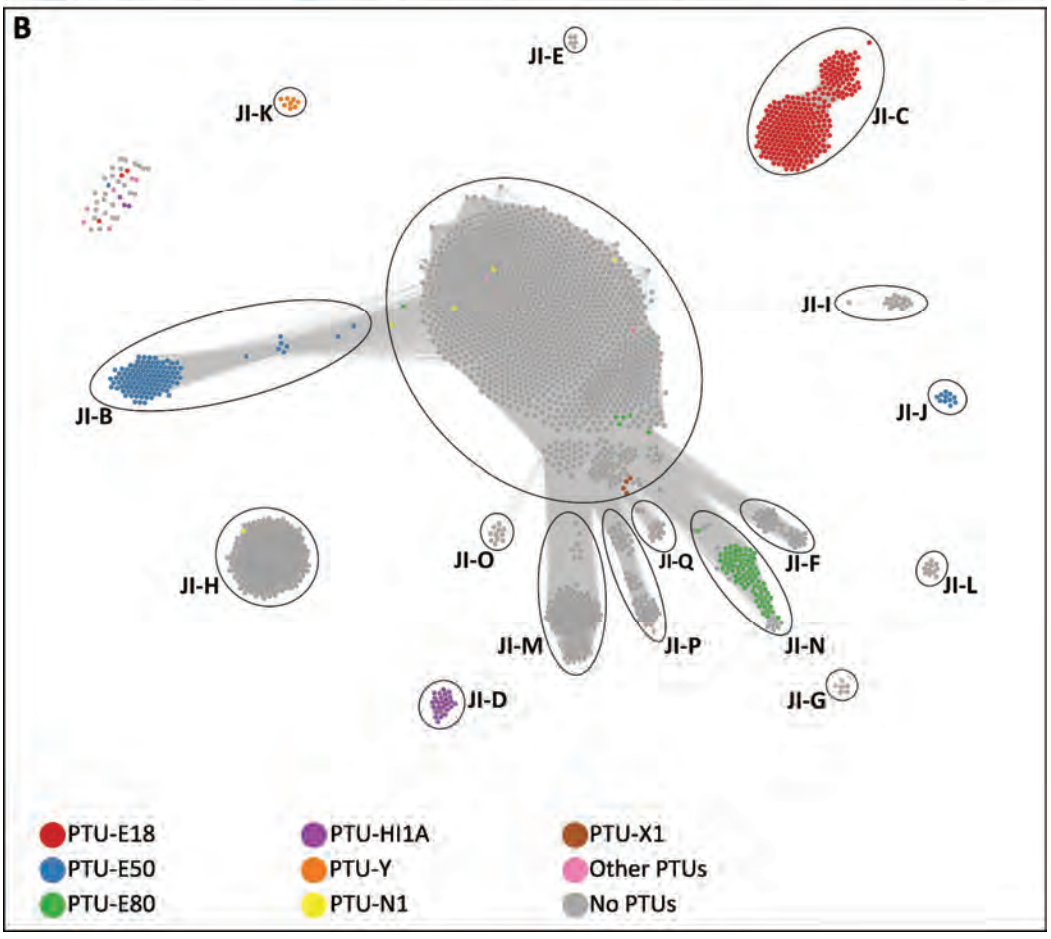

B
JI-E
JI-K
JI-C
JI-I
JI-J
JI-B
JI-H
JI-O
JI-Q
JI-F
JI-L
JI-M
JI-P
JI-N
JI-D
JI-G
PTU-E18
PTU-HI1A
PTU-X1
PTU-E50
PTU-Y
Other PTUs
PTU-E80
PTU-N1
No PTUs

Figure 4.9 | Distribution of accessory genome elements in the Typhi JI-groups. **(A)** Distribution of MOB relaxases in the JI-groups. Nodes are colored according to the MOB relaxase class present in each genome. Information on accessory elements (excluding plasmids) present in >=90% of the members of a given JI-group is included in black letters or in red letters when absent. **(B)** Distribution of PTUs in the JI-groups. Nodes are colored according to the PTUs present in each genome. Both (A) and (B) networks contain 2,392 nodes, connected when JI>=0.983 and GLD<=0.05. Seventeen clusters (named JI-A to JI-Q) are indicated by circles.

Table 4.2: Characteristics of plasmids identified in Typhi JI-groups.

JI-Groups (number of genomes)	**Number of genomes with plasmid**	**PTU** (grade host range) [a]	**Plasmid replicons** [b]	**Plasmid MOB type / MPF** (transmissibility) [b]	**AMR determinants** [b]	**Average plasmid size (kb)**
JI-A (1,320)	6	PTU-E80 (IV)	IncX1	MOB_P / - (mobilizable)	-	18
	4	PTU-N1 (III)	IncN	MOB_F / MPF_T (conjugative)	-	40
	4	PTU-X1 (III)	IncX1	MOB_P / MPF_T (conjugative)	-	30
	1	PTU-X3 (III)	IncX3	MOB_P / MPF_T (conjugative)	-	44
	1	PTU-E73 (IV)	IncFII(pCRY)	MOB_C / MPF_T (conjugative)	-	21
JI-B (114)	114	PTU-E50 (III)	IncY, IncFIB(K)	MOB_C / MPF_T (conjugative)	*bla* $_{TEM-1B}$, *qnrS1, sul2, tet(A), aph(3")-Ib, aph(6)-Id, dfrA14, bla* $_{CTX-M-15}$, *bla* $_{CTX-M-88}$	90
JI-C (265)	265	PTU-E18 (IV)	IncFIB(pHCM2)	- / - (non-tranmissible by conjugation)	-	107
JI-D (26)	26	PTU-HI1A (IV)	IncHI1A, IncHI1B(R27), IncFIA(HI1)	MOB_H / MPF_F (conjugative)	*aph(3")-Ib, aph(6)-Id, bla* $_{TEM-1B}$, *catA1, dfrA7, qacE, sul1, sul2, tet(B)*	217
JI-H (225)	1	PTU-N1 (III)	IncN	MOB_F / MPF_T (conjugative)	*aph(3")-Ib, aph(6)-Id, bla* $_{TEM-1B}$, *dfrA14, sul2, tet(A)*	50
JI-J (11)	11	PTU-E50 (III)	IncY	MOB_C / MPF_T (conjugative)	*aph(3")-Ib, aph(6)-Id, bla* $_{TEM-1B}$, *dfrA14, sul2, tet(A)*	115
JI-K (8)	8	PTU-Y (III)	IncY, p0111	- / - (non-tranmissible by conjugation)	*bla* $_{CTX-M-15}$	100
JI-N (90)	78	PTU-E80 (IV)	IncX1	MOB_P / - (mobilizable)	-	25

[a] | PTU and the grade host range were assigned using COPLA.

[b] | Plasmid replicons, MOB class, MPF type, and AMR determinants were calculated, respectively, using PlasmidFinder, MOBscan, CONJScan, and ResFinder. –, the absence of the specific trait indicated in the column.

Table 4.3: Summary of all MGEs (other than plasmids) detected in Typhi JI-groups.

JI-Groups	Number of genomes[a]	MGE detected [b]	MGE absent[c]	Average MGE size (kb)
JI-A	1320	IME_MOB$_Q$	-	21
JI-B	114	IME_MOB$_Q$	-	21
JI-C	265	IME_MOB$_Q$	-	21
JI-D	26	IME_MOB$_Q$	-	21
JI-E	5	IME_MOB$_Q$	-	21
		Region 1	-	49
JI-F	39	Region 5	-	34
		-	Prophage 5	11
JI-G	6	-	Region 2 (MPF typeG, MPF typeF)	61
		-	Region 6	62
		-	Region 7	21
		-	Prophage 6	16
JI-H	225	ICE_MOB$_H$	-	55
JI-I	17	IME_MOB$_Q$	IME_MOB$_H$	38
		-	Region 2 (MPF typeG, MPF typeF)	61
		-	Prophage 1	21
JI-J	11	Prophage 7	-	29
JI-K	8	IME_MOB$_Q$	-	21
JI-L	11	Region 8	-	67
JI-M	133	Prophage 3	-	38
		Prophage 8	-	26
JI-N	90	-	Prophage 1	21
JI-O	11	Prophage 9	-	40
		Region 9	-	40
JI-P	58	Prophage 10	-	44
JI-Q	15	Region 10	-	12
		Region 11	-	11
		-	Prophage 5	11

[a] | Number of genomes present in each JI-group.

[b] | MGEs identified in the majority of genomes within each JI-group, suggesting these elements may contribute to the genomic definition of these groups.

c | MGEs typically found across other JI-groups but notably missing in most genomes of the indicated JI-group, highlighting potential loss events specific to these groups.

To further explore the contribution of MGEs in the JI-group clustering, an *in silico* experiment was carried out by removing them from reference genomes. PTU-E50 plasmids present in the B subgroups, and *Salmonella* Genomic Island 11 (SGI11) encoded in B1 and A3 references were eliminated. The "cured" genomes segregated from their original JI-

groups and associated with the JI-A1 genomes in the network (**Figure 4.10A**). The progressive reintroduction of SGI11 (**Figure 4.10B**) and PTU-E50 sequences (**Figure 4.10C**) led to the partition of JI-A3, JI-B1, JI-B2, and JI-B3 genomes from the JI-A1 group, rendering new clusters.

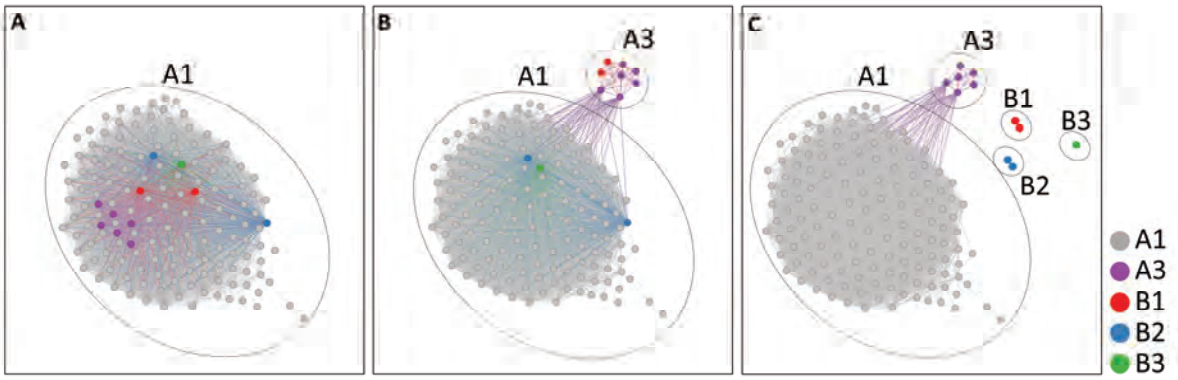


Figure 4.10 | Effect of MGEs on the JI-based genome clustering. The networks contain 154 nodes, connected when JI>=0.995. Nodes are colored according to the original JI subgroup of each genome (see Figure 4.8). **(A)** Clustering of genomes deprived of PTU-E50 and SGI11. PTU-E50 plasmids originally present in genomes of the B1, B2, and B3 subgroups, as well as the chromosomally-inserted element SGI11, encoded also in genomes of the B1 and A3 subgroups, were removed from the genome sequences. The resulting "pruned" genomes were used to calculate pairwise genome similarities. Genomes from all subgroups reassociate in a single cluster. **(B)** Clustering of genomes deprived of PTU-E50. The SGI11 elements were restituted to the A3, and B1 genomes and the network was recalculated. A3 and B1 genomes broke away from the previous cluster and grouped together. **(C)** Clustering of genomes with SGI11 and PTU-E50. The PTU-E50 plasmids were restituted to the B1, B2, and B3 genomes. The rebuilt network shows the emergence of distinctive clusters.

In a similar experiment, the plasmid sequences were removed from 13 RefSeq200 reference genomes present in JI-groups B, C, D, and K and 4 reference genomes of JI-groups C and J reconstructed by PLACNETw [203]. The pairwise JI values were recalculated, and a new network generated (**Figure 4.11**). The "cured" genomes segregated from their original JI-groups (B, C, D, J, and K) and associated with the JI-A and JI-M genomes (in this latter case they come originally from JI-C). This shift in JI-group associations demonstrated the role of these plasmids in shaping JI-groups, emphasizing their contribution to Typhi pangenome structure.

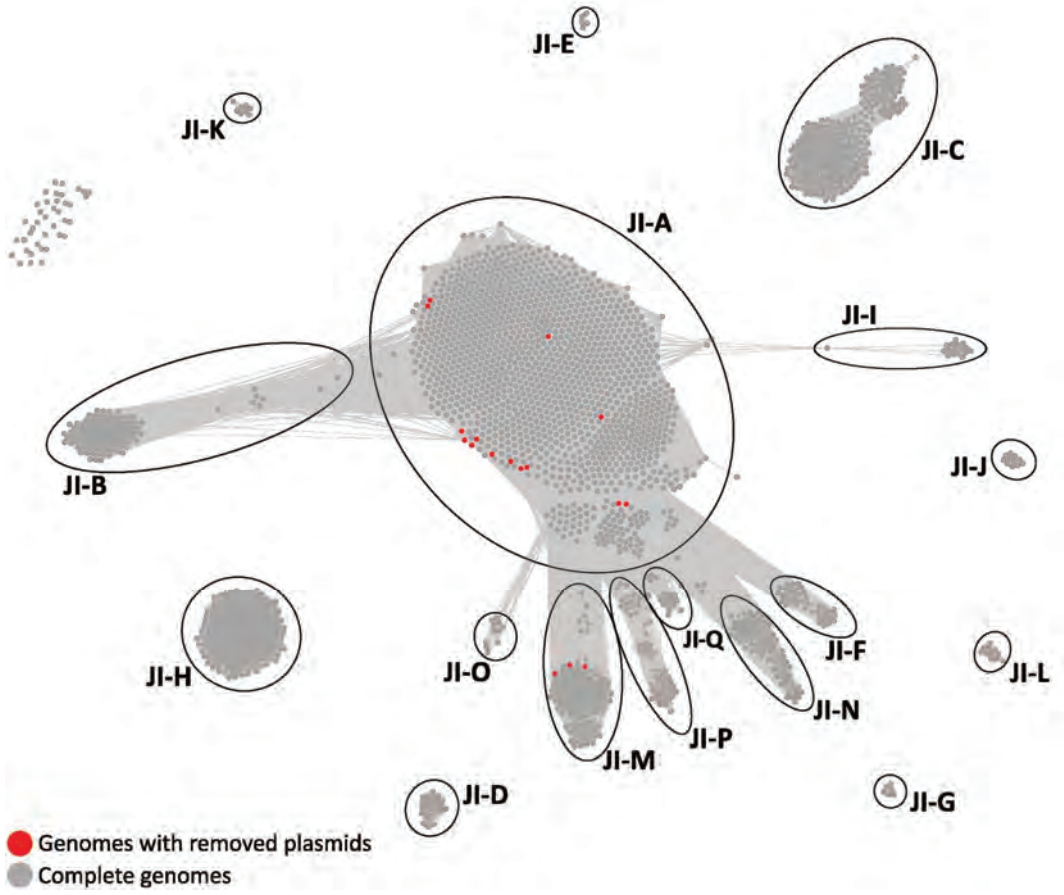


Figure 4.11 | Effect of plasmid removal on the JI network. The network contains 2,392 nodes, which are connected whenever JI >= 0.983 and GLD <= 0.05. Nodes colored in red (17) indicate reference genomes that contain plasmids larger than 80 kb in size, which were removed prior to JI calculation. Seventeen distinct clusters (named JI-A to JI-Q) identified by the Louvain method are indicated by circles.

GenoTyphi genotypes [193,229] were visualized against JI-groups to compare phylogenetic context to pangenome groupings. Primary clade 4 dominated the dataset, followed by primary clade 2, while primary clade 1 was barely represented, indicating that this lineage might be less prevalent in the dataset or has undergone population bottlenecks over time. The presence of primary clade 0 in JI-F suggests that this group may represent one of the most ancestral lineages of Typhi.

Most JI-groups (n=12/17) associated with a single GenoTyphi primary clade (**Table 4.1, Figure 4.12A**), whereas JI-A, JI-C, JI-D, JI-I, and JI-N contained isolates that fell into two or more GenoTyphi primary clades. JI-A contained genomes from all Typhi primary

clades (0, 1, 2, 3, and 4), with each primary clade mostly confined to distinct areas in the network map of JI-A (**Figure 4.12A**).

At the higher threshold used to determine JI-A subgroups (JI>=0.995), GenoTyphi lineages were largely resolved into their own cluster, with 15 of 17 JI-A subgroups containing genomes of a single GenoTyphi primary clade (**Figure 4.13A**). Similarly, group JI-C had members of primary clades 2, 3 and 4, but at the JI-subgroup resolution, all members within each of the six JI-C subgroups contained a single GenoTyphi primary clade (**Figura 4.13B**).

The distribution of GenoTyphi 4.3.1 subclade, the most common GenoTyphi in MDR Typhi, [159,235] and its derivatives was visualized in the JI-groups (**Figure 4.12B**), providing a higher resolution view of this subclade. Multiple lineages of 4.3.1 often coexisted within the same JI-group, indicating homogeneity in accessory genome content despite some phylogenetic differences. On the contrary, in some cases, the same lineages or sublineages of 4.3.1 (i.e. 4.3.1.1.P1 and 4.3.1.1.EA1) appeared in different JI-groups, indicating accessory genome differences despite being phylogenetically identical.

Membership to a JI-group does not necessarily imply vertical descent (as defined by GenoTyphi); since JI-grouping aggregates genomes of distinct vertical lineages if they share substantial accessory genome material, and partitions genomes of the same vertical lineage into separate groups according to their accessory genome content. However, pangenome groupings did tend to align with phylogenetic lineage, especially at the level of subgroups (n=32/40 JI-group or subgroup contained a single GenoTyphi primary clade) (**Figure 4.13, Table S1**).

Thus, coupling of pangenomic and phylogenetic methods can simultaneously offer information on horizontal and vertical evolutionary dimensions. In fact, coupling information from GenoTyphi and MOB typing methods already accounted for a substantial proportion of the genetic variance of JI-groups (combined variance partitioning R^2=0.725), suggesting much of the Typhi pangenome can be effectively identified with existing methods. Although MOB typing methods can detect the presence of accessory elements, they do not determine its exact nature, whether it is a plasmid, an IME, or an ICE, without further analysis.

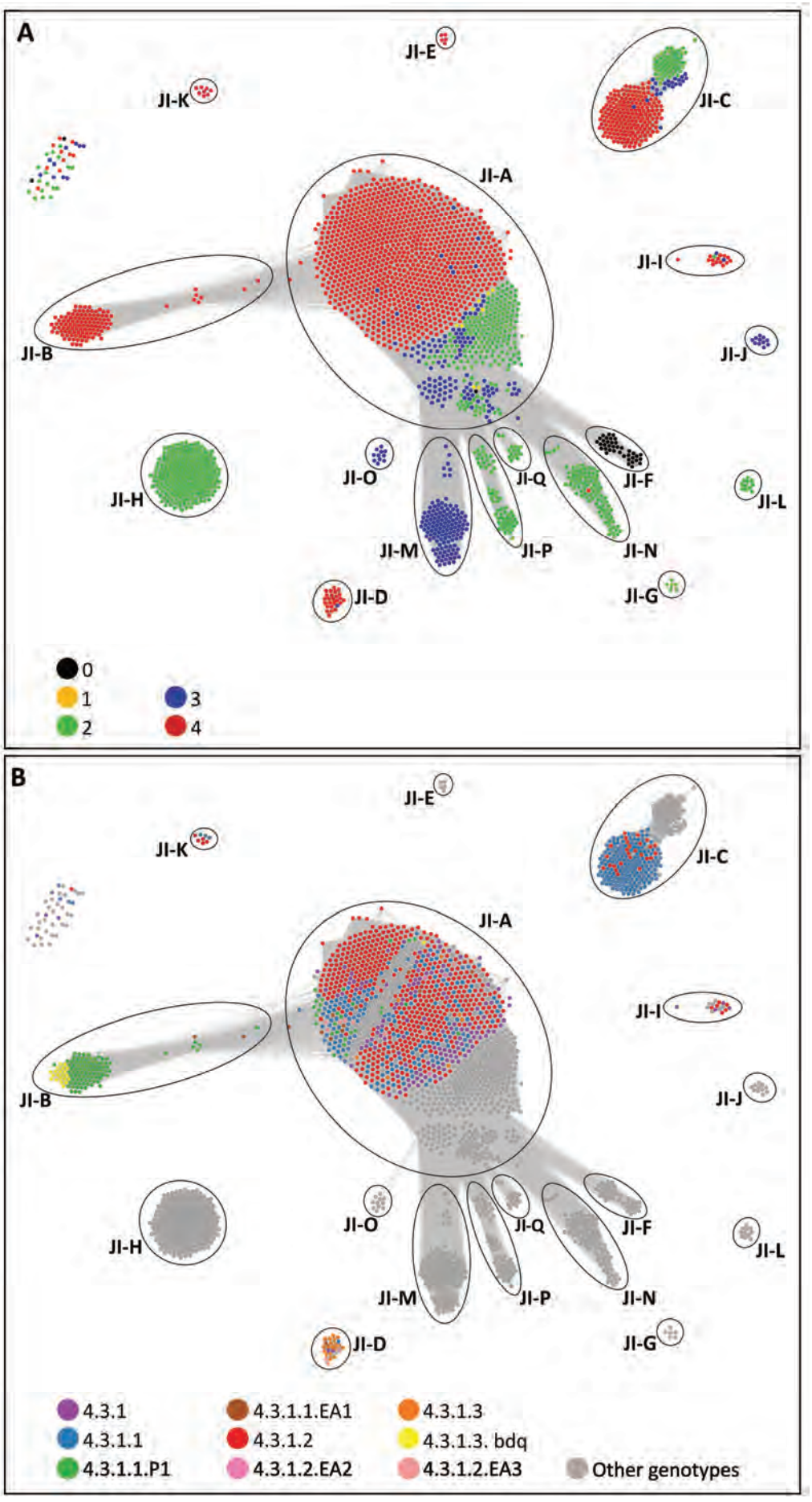

A
JI-E
JI-K
JI-C
JI-A
JI-I
JI-B
JI-J
JI-H
JI-O
JI-Q
JI-F
JI-L
JI-M
JI-P
JI-N
JI-D
JI-G
0
1
2
3
4
B
JI-E
JI-K
JI-C
JI-A
JI-I
JI-B
JI-J
JI-H
JI-O
JI-Q
JI-F
JI-L
JI-M
JI-P
JI-N
JI-D
JI-G
4.3.1
4.3.1.1
4.3.1.1.P1
4.3.1.1.EA1
4.3.1.2
4.3.1.2.EA2
4.3.1.3
4.3.1.3. bdq
4.3.1.2.EA3
Other genotypes

Figure 4.12 | Distribution of GenoTyphi in the JI-groups. (A) Distribution of GenoTyphi primary clades in the JI-groups. Nodes are colored according to the GenoTyphi primary clades. **(B)** Distribution of the 4.3.1 GenoTyphi genotype in the JI-groups. Nodes are colored according to the lineages and sublineages of the 4.3.1 genotype. Both (A) and (B) networks contain 2,392 nodes, connected when JI>=0.983 and GLD<=0.05. Seventeen clusters (named JI-A to JI-Q) are indicated by circles.

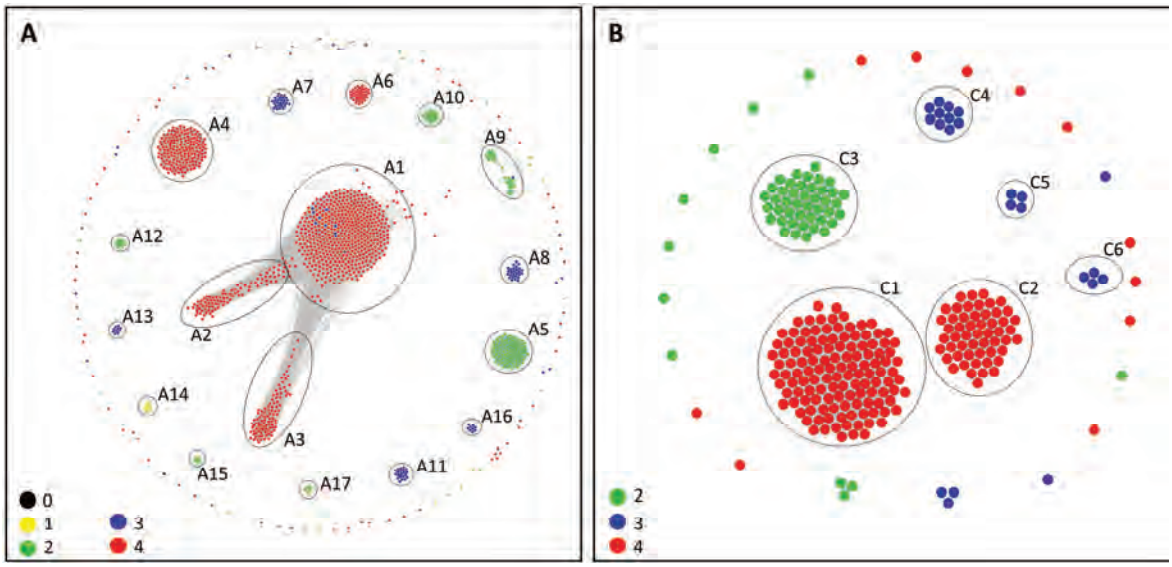


Figure 4.13 | Distribution of GenoTyphi primary clades in JI-subgroups A and C. (A) Subclustering analysis of JI-group A. A set of 1,320 genomes of the JI-group A were used to build the JI network, using JI>=0.995 as a threshold. Subgroups A1 to A17 identified by the Louvain method are defined by circles. Nodes are colored by the GenoTyphi primary clade they belong to. **(B)** Subclustering analysis of JI-group C. A set of 265 genomes of the JI-group C were used to build the JI network, using JI>=0.997 as a threshold. Subgroups C1 to C6 identified by the Louvain method are defined by circles. Nodes are colored by the GenoTyphi primary clade they belong to.

4.4 U.S. Typhi pangenome structure aligns with epidemiological patterns

To gain insights into the epidemiological patterns, epidemiological metadata was mapped onto the network to visualize temporal and geographical patterns. JI-A represented the largest and most persistent group, with a high concentration of genomes from 2015 onwards. On the other hand, other JI-groups emerged or became more prevalent in later

years, such as JI-B and JI-H. Smaller groups, such as JI-G, JI-E and JI-O, exhibited short-lived spikes, suggesting these groups could represent localized events (**Figure 4.14**).

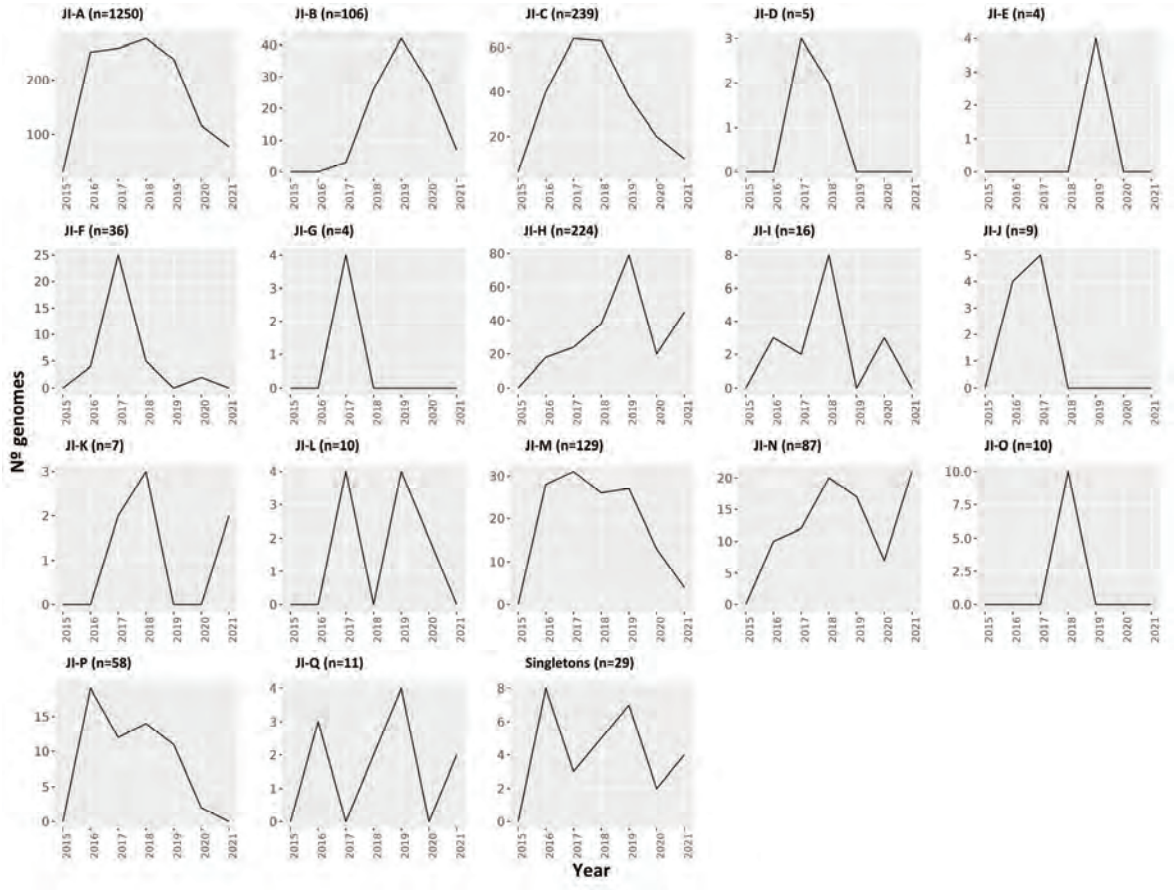


Figure 4.14 | Abundance of Typhi genomes of each JI-group over time. Number of genomes of each JI-group during the period 2015-2021. Only the genomes that contain the year of isolation were used to produce this figure.

All genomes analyzed in this study were isolated in the U.S. (a region where Typhi is not endemic) and travel history data was available for 866 genomes, providing insights into their likely geographic origins. A substantial portion of these genomes (634/866) was associated with travel to Asia, and they were predominantly clustered in groups JI-A, JI-B, JI-C, JI-M, and JI-P. This association suggests that these JI-groups are frequently introduced from regions such as Pakistan, India, and Bangladesh, aligning with previous findings on the prevalence of Typhi in these areas [236]. Although a smaller proportion of genomes were linked to non-Asian regions, certain patterns emerge: JI-H and JI-N genomes were associated with travel to America while JI-J genomes were mainly linked to Africa (**Figure 4.15**).

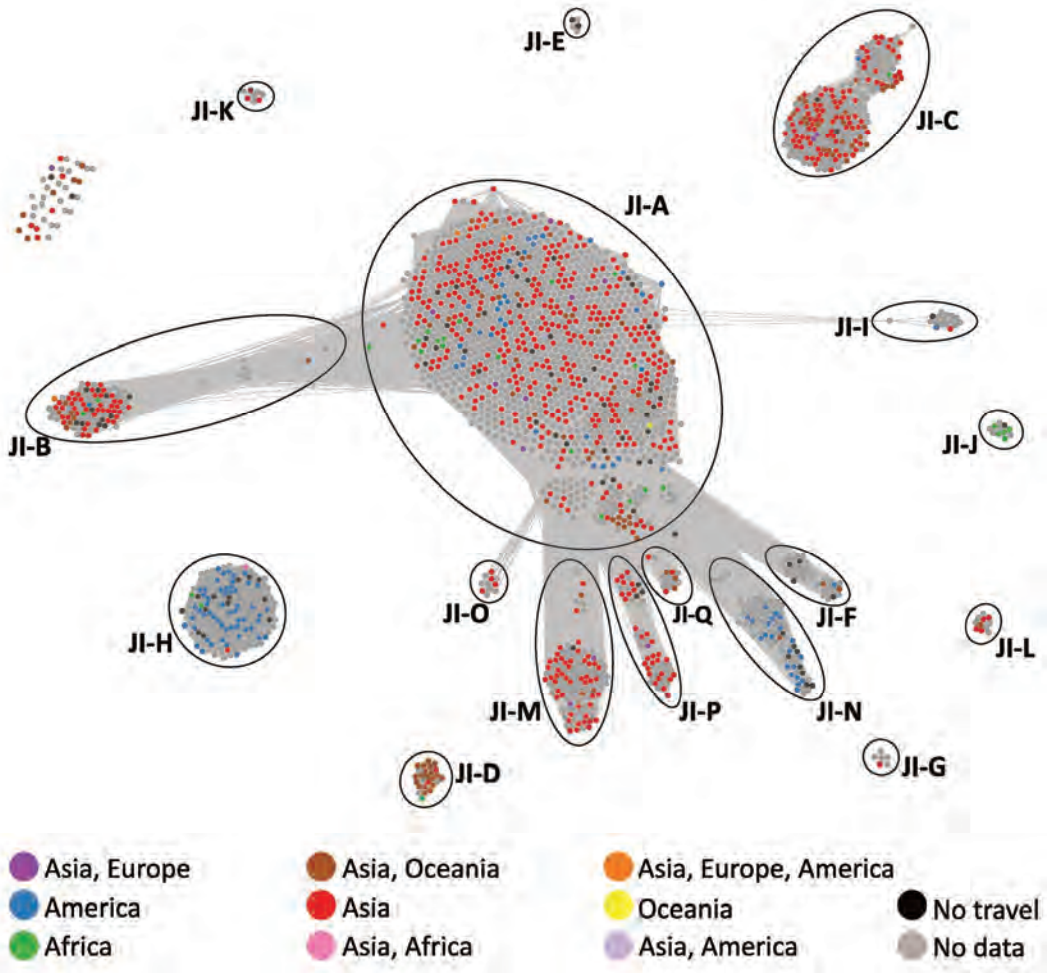


Figure 4.15 | Geographical data mapped onto the JI network of Typhi. The network contains 2,392 nodes, which are connected whenever JI >= 0.983 and GLD <= 0.05. Nodes are colored by United Nations Region of travel within 30 days of illness onset. Seventeen distinct clusters (named JI-A to JI-Q) identified by the Louvain method are indicated by circles.

4.5 U.S. Typhi pangenome structure aligns with and expands on known AMR

MDR in Typhi genomes were genetically defined by carriage of genes conferring resistance to ampicillin, chloramphenicol and co-trimoxazole, genes that are typically found in the genomic island SGI11. SGI11 contains AMR genes (*bla*$_{\text{TEM-1}}$, *catA1*, *aph(3')-Ib* [*strA*], *aph(6)-Id* [*strB*], *sul1*, *sul2* and *dfrA7*), a mercury resistance operon, and the *qacEΔ1* gene that encodes an ethidium-bromide resistance protein [237]. MDR in Typhi emerged several decades ago, driven by the expansion of a 4.3.1 (previously H58) strain carrying SGI11 on an IncHI1 (PTU-HI1A) plasmid [200]. Subsequent degradation of SGI11 [237], as well as

integration into the Typhi chromosome and loss of the IncHI1 (PTU-HI1A) plasmid [233] has occurred. The next major evolutionary step in Typhi AMR was the emergence of XDR Typhi, first reported in Pakistan in 2016. XDR Typhi evolved from the MDR 4.3.1 lineage through the acquisition of an IncY (PTU-E50) plasmid carrying *bla*$_{\text{CTX-M-15}}$, which confers resistance to cephalosporins and *qnrS,* which confers resistance to fluoroquinolone. More recent studies have reported XDR genomes isolated from 2018 onward that have lost the PTU-E50 plasmid [238], with the *bla*$_{\text{CTX-M-15}}$ gene integrating directly into the chromosome (**Figure 4.16**).

Genetic metadata was mapped onto the JI network to determine if JI-grouping could easily detect known AMR patterns. For example, XDR Typhi (genotype 4.3.1.1.P1), first reported [239] in the U.S. in 2018 among patients with travel history to Pakistan, corresponded to subgroup JI-B1. Genomes in JI-B1 were isolated from 2018 onward (**Figure 4.14**), were all genotype 4.3.1.1.P1 (**Figure 4.12B**), carried an IncY (PTU-E50) plasmid with *bla*$_{\text{CTX-M-15}}$, and were significantly associated with travel to Pakistan ($P < 0.01$, chi-squared test of independence), despite limited travel data for this group (n=47/88 have any travel information available) (**Figure 4.15**). Additionally, other XDR genomes from the same 4.3.1.1.P1 sublineage fell into JI-A (specifically JI-A3). Unlike JI-B1 genomes, these JI-A3 strains have recently lost the IncY (PTU-E50) plasmid and integrated *bla*$_{\text{CTX-M-15}}$, the gene that confers ceftriaxone resistance and defines XDR, into their chromosome [238]. Thus, both the original XDR 4.3.1.1.P1 Pakistan outbreak strain with an IncY (PTU-E50) plasmid [126] and its recent XDR variants (without the plasmid) were quickly identifiable in the JI-network as JI-B (JI-B1) and JI-A (JI-A3), respectively, supporting what is known about this sublineage (**Figures 4.16** and **4.17**).

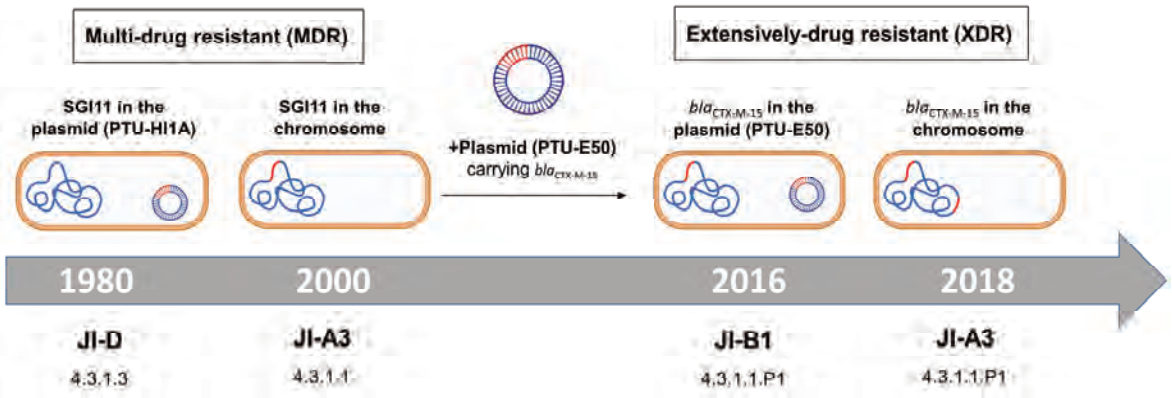


Figure 4.16 | Timeline depicting the major evolutionary steps leading to MDR and XDR Typhi. MDR emerged in the 1980s with strains carrying SGI11 on a PTU-HI1A plasmid (corresponding to

JI-group D, genotype 4.3.1.3). Around the year 2000, SGI11 integrated into the chromosome, exemplified by JI-group A3 (genotype 4.3.1.1). Subsequently, the XDR phenotype arose in 2016 via the acquisition of an IncY plasmid (PTU-E50) harboring the *bla*$_{CTX-M-15}$ gene, defining JI-group B1 (genotype 4.3.1.1.P1). More recently, since 2018, XDR strains without the PTU-E50 plasmid have been reported, integrating *bla*$_{CTX-M-15}$ directly into their chromosome, represented by JI-group A3 (genotype 4.3.1.1.P1).

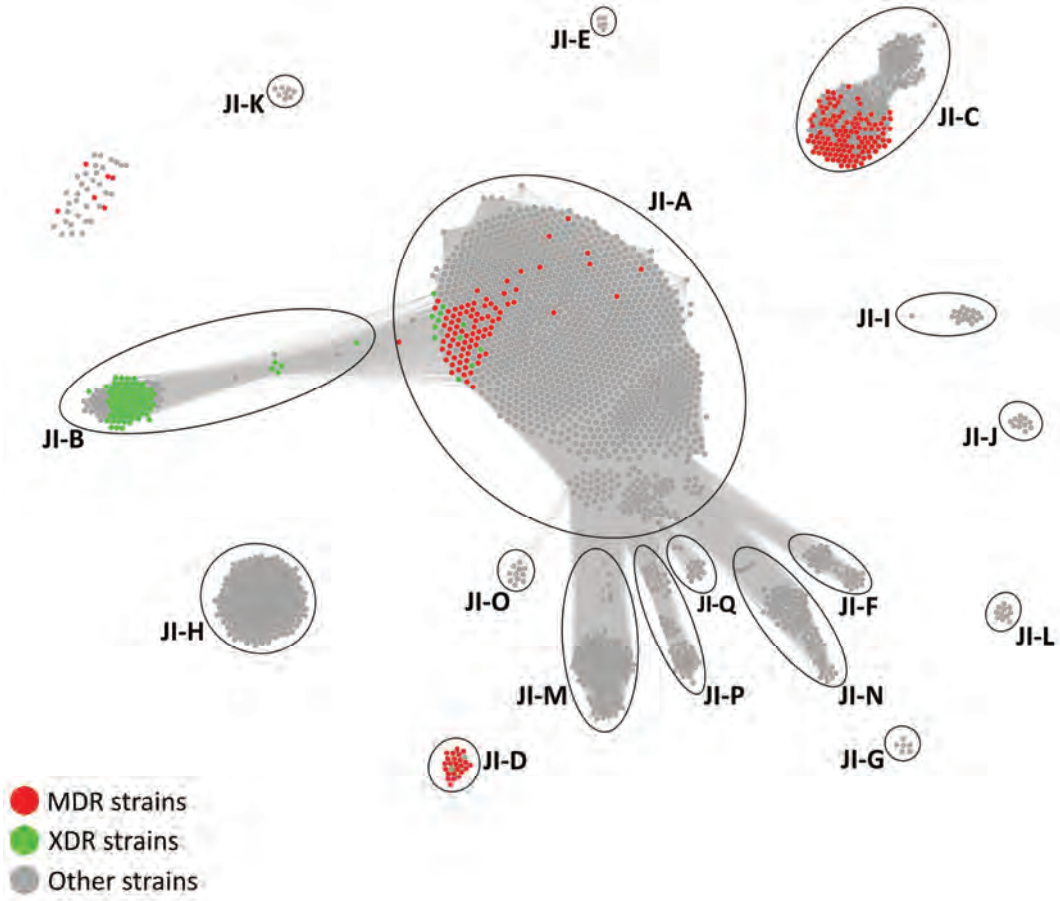


Figure 4.17 | Distribution of MDR and XDR Typhi genomes. The network contains 2,392 nodes, connected when JI ≥0.983 and GLD ≤0.05. Seventeen clusters (named JI-A to JI-Q) are indicated by circles. Nodes are colored according to either they belong to MDR or XDR strains or none of them.

Carriage of SGI11 (denoted as MDR or XDR in **Figure 4.17**) was identified in JI-A, JI-B, JI-C and JI-D, and the genetic location was consistent within each group, chromosomal in JI-A, JI-B and JI-C, or plasmid-mediated in JI-D (PTU-HI1A). At higher JI-thresholds, the presence of SGI11 was even confined to specific JI-subgroups JI-A1, JI-A3, JI-B1 and JI-C1 (**Figure 4.18**). JI-B1 and JI-C1 represented known epidemiological lineages, the

"XDR Pakistan" strain and MDR 4.3.1.1 Typhi strains with chromosomal SGI11, respectively.

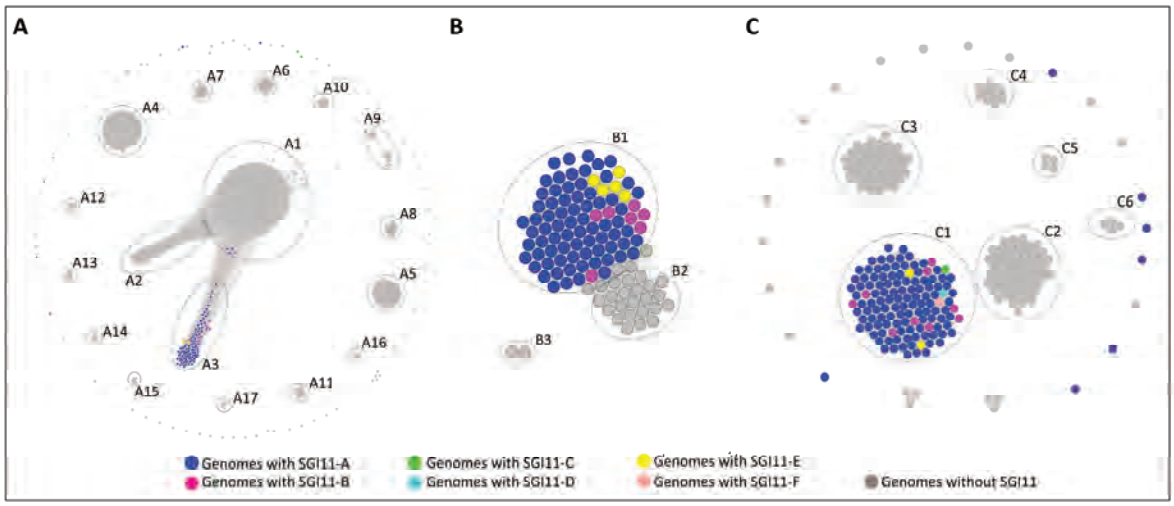


Figure 4.18 | Subclustering analysis of JI-groups A, B, and C colored by the SGI11 variant. (A) Subclustering analysis of JI-Group A. The network contains 1,320 genomes of JI-group A, using JI>=0.995 and GLD<=0.05 as a threshold. Subgroups A1 to A17 defined by the Louvain method are surrounded by circles. **(B)** Subclustering analysis of JI-group B. JI network of 114 JI-B using JI>=0.986 and GLD<=0.05 as a threshold. Subgroups determined by the Louvain method are indicated by circles. **(C)** Subclustering analysis of JI-group C. JI network of 265 JI-C genomes using JI>=0.997 and GLD<=0.05 as a threshold. Subgroups determined by the Louvain method are indicated by circles. For more details on SG11 variants, see Figure 4.19A.

Genetic context analysis of genomes with chromosomal SGI11 (28 reference and 300 U.S. genomes) detected six variants of SGI11 (previously described variants A-E [237] and a novel variant F described here). SGI11 was found inserted in two distinct genomic regions: either interrupting the *yidA* gene or in the intergenic region between genes *cyaA* and *cyaY* (**Figure 4.19**). To further investigate the relationship between SGI11 variants, insertion sites, and JI-groups, a core-genome phylogenetic tree was constructed, mapping the JI-groups and SGI11 insertions sites onto the rings (**Figure 4.19B**). However, neither SGI11 variant nor chromosomal insertion site were found to align with JI grouping, likely due to the minimal size differences among SGI11 variants. Interestingly, within this phylogenetic framework, we detected a likely event of SGI11 excision from the *yidA* gene in six JI-B isolates otherwise practically identical to other JI-B members encoding SGI11 interrupting *yidA*. These six isolates were represented in Ring 1 of **Figure 4.19B** as "no SGI11 and *yidA*

disrupted". In these cases, long-read sequencing of two of these genomes (PNUSAS224101 and PNUSAS195139) confirmed that the *yidA* gene was disrupted by IS*1*, suggesting that it could be either a precursor to the SGI11 acquisition, or most likely a derivative of SGI11 excision, both probably through IS*1*-mediated recombination. Nonetheless, mapping of SGI11 presence onto the JI-network (MDR and XDR strains contain SGI11, **Figure 4.17**) quickly revealed that some JI-groups more frequently host this MGE than others, which is likely driven but not entirely explained by the overrepresentation of 4.3.1 in these groups.

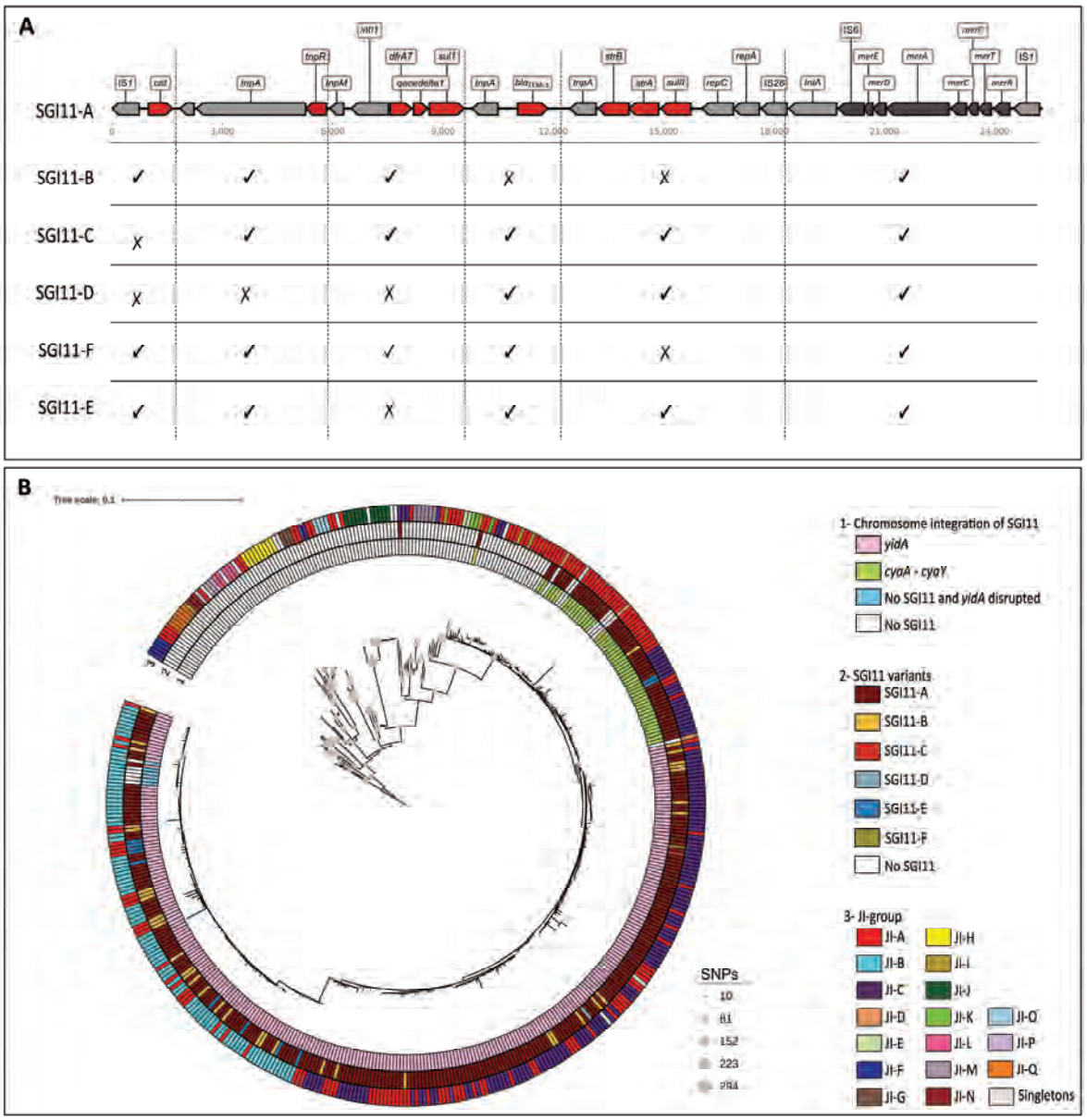


Figure 4.19 | SGI11 variants and their distribution in the Typhi genomes. (A) SGI11 variants found in the study genomes. Genes encoded by the SGI11 variant A, present in *S.* Typhi BD1380 (GenBank Acc. No. KM023773), are represented by arrows and named in the upper part. They are colored in red (antimicrobial resistance genes), dark grey (*mer* operon), or light grey (other genes).

In the lower part, a table records the presence or absence of different SGI11 segments in the different variants found in the study genomes. SGI11 variants A-E were previously described [237], while SGI11-F is a new variant described here. **(B)** Phylogenetic distribution of SGI11-containing genomes. Parsimony reconstruction based on the core SNPs using kSNP3.0 [217]. The tree includes 328 genomes that contain a chromosomal SGI11 insertion, 130 genomes that do not contain SGI11 (Table S1), and a non-Typhi genome (NZ_CP015724.1) that was used as an outgroup. Circles at the internal nodes indicate the SNPs shared exclusively by their descendants, according to the legend. Branch length scale represents changes per number of SNPs. Colored rings indicate the chromosomal SGI11 integration site (1), the SGI11 variant (2), and the JI-group of each genome (3). The tree was visualized with iTol v6 [215].

A total of 109 isolates harbored *bla*$_{\text{CTX-M-15}}$, predominantly in the XDR group JI-B1 (87/109), with smaller numbers in JI-A (16/109; specifically in subgroups JI-A1 and JI-A3) and JI-K (5/109), and in one singleton. In all cases, the gene was likely mobilized by IS*Ecp.* The *bla*$_{\text{CTX-M-15}}$ gene was plasmid-mediated in JI-B (PTU-E50) and JI-K (PTU-Y) but was integrated into the chromosome of all 16 JI-A genomes.

The chromosomal genetic context of *bla*$_{\text{CTX-M-15}}$ was further explored in JI-A1 (n=4) and JI-A3 (n=12) genomes. Three different sized regions (a-c) of the original IncY (PTU-E50) plasmid were detected, reflecting the incorporation of the plasmid´s drug resistance region into the chromosome, likely captured and mobilized by IS*Ecp1* (**Figure 4.20A**). ISMapper identified four possible IS*Ecp1*-*bla*$_{\text{CTX-M-15}}$ insertion sites (I-IV) (**Figure 4.20B**). Insertion sites were confirmed either by direct analysis of the *bla*$_{\text{CTX-M-15}}$-containing contigs (insertion sites I-III), or with additional long-read sequencing (insertion site IV). Nine genomes contained region a (~4 kb), which lacked the *qnrS1* gene, inserted between *gutQ* and *norA* (insertion I) and four genomes contained region c (~17 kb) inserted within SGI11 (insertion IV). These two insertion sites (I and IV) were previously reported [238]. Additional IS*Ecp1* insertion sites were found between *phsA* and *sopA* (region b, insertion II, two genomes), and interrupting *stgC* (region c, insertion III, one genome).

To further explore the distribution of *bla*$_{\text{CTX-M-15}}$, a core-genome phylogenetic tree was constructed, including all *bla*$_{\text{CTX-M-15}}$-containing genomes, along with representative genomes from all JI-groups that do not carry this gene (**Figure 4.21**). The distal position of the phylogenetic clades including the JI-A genomes harboring *bla*$_{\text{CTX-M-15}}$ in the tree suggests its chromosomal acquisition occurred recently. Isolates with core genomes differing by

fewer than 10 SNPs contain the *bla*$_{\text{CTX-M-15}}$ gene either in a PTU-E50 plasmid or in different chromosomal locations, or entirely lack it (**Figure 4.21**). This suggests that, indeed, the PTU-E50-borne *bla*$_{\text{CTX-M-15}}$ inserted in at least four independent events.

Genomes with a given, unique integration site, clustered together in the JI network when the JI threshold was increased (i.e., enhanced discrimination between genomes) (**Figure 4.21C**); however, these clusters were not distinct enough to be used for prediction of the genetic context from the JI-network alone.

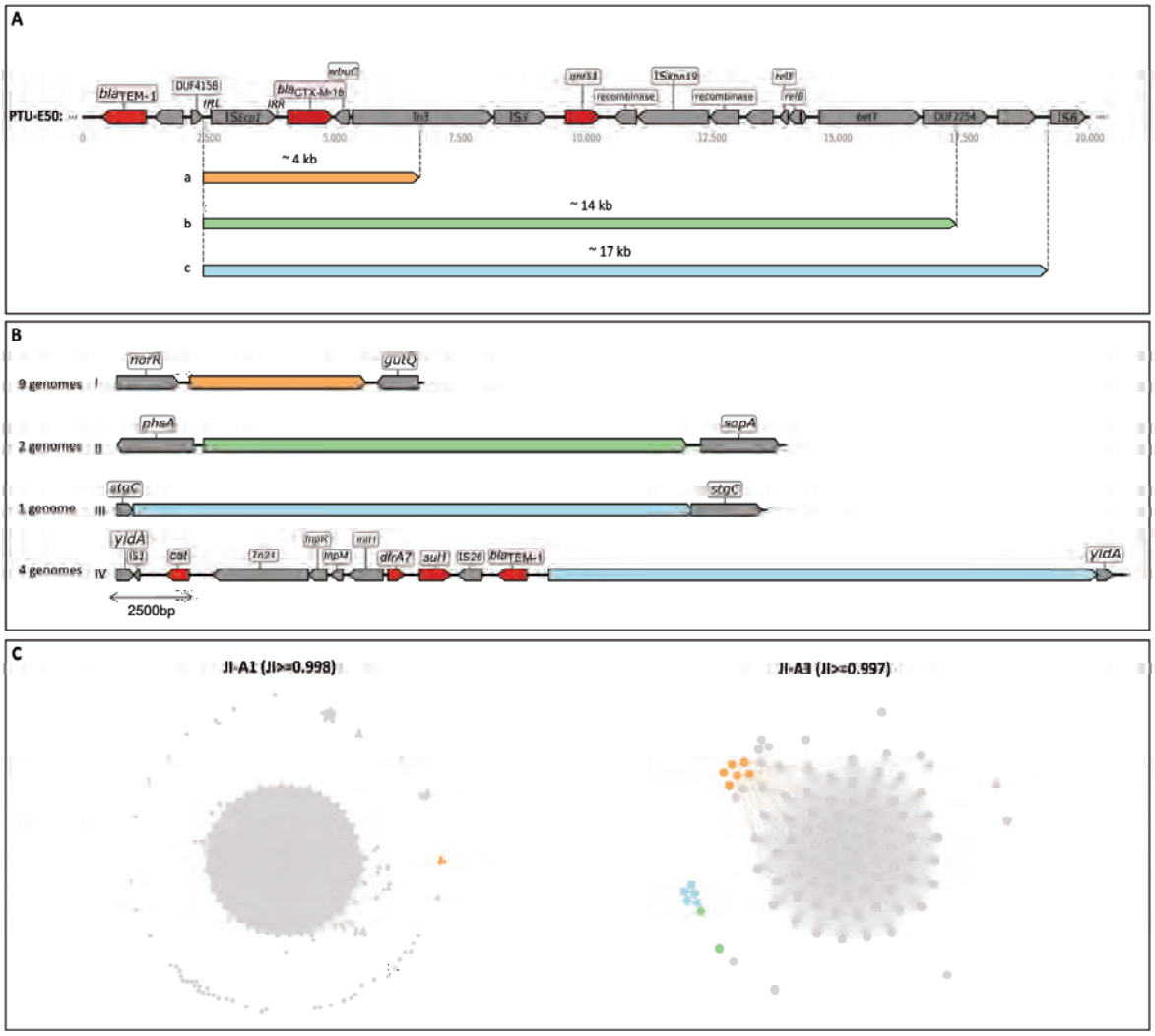


Figure 4.20 | Genomic context of *bla*$_{\text{CTX-M-15}}$. **(A)** The genetic vicinity of *bla*$_{\text{CTX-M-15}}$ in PTU-E50 plasmids. The region containing the *bla*$_{\text{CTX-M-15}}$ gene of plasmid NZ_CP046430 is depicted. Genes are represented by arrows and those encoding AMR are colored in red. Below, three arrows of different sizes, indicated by different colors represent the PTU-E50 regions (a-c) that were found integrated into the chromosomes. **(B)** Chromosomal integration sites of the *bla*$_{\text{CTX-M-15}}$-containing regions (I-IV). Insertion site I locates between genes *norR* and *gutQ*; site II between genes *phsA* and

sopA; site III interrupts gene *stgC*; site IV resides within SGI11. **(C)** JI networks of subgroups JI-A1 and JI-A3. Nodes colored in orange, green, and blue indicate genomes containing the different $bla_{CTX\text{-}M\text{-}15}$-encoding regions a, b and c, respectively.

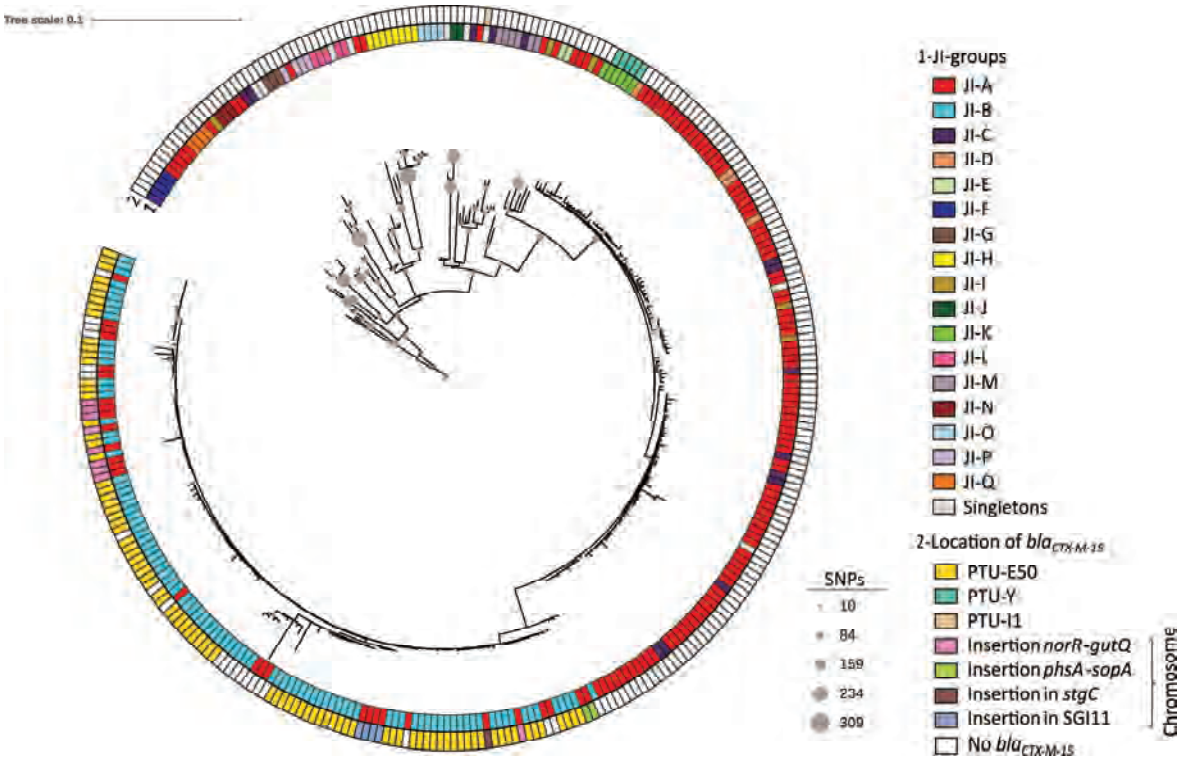


Figure 4.21 | Core genome phylogeny of Typhi genomes. The phylogenetic tree, parsimony reconstruction based on the core SNPs using kSNP3.0 [217], includes all Typhi genomes that contain the gene $bla_{CTX\text{-}M\text{-}15}$ (n=109), all genomes from JI-A3 that lack $bla_{CTX\text{-}M\text{-}15}$ (n=88), all genomes from JI-B1 that lack $bla_{CTX\text{-}M\text{-}15}$ (n=4), 122 representative genomes from the 17 JI-groups (Table S1), and one genome from serovar Indiana as an outgroup. Branch length scale represents changes per number of SNPs. Circles at the internal nodes indicate the number of SNPs distinctive of the corresponding clade. The colored rings indicate the JI-group of the corresponding genome (1), and the $bla_{CTX\text{-}M\text{-}15}$ gene location (2).

Chromosomal mutations in the quinolone-resistance determining region (QRDR), and presence of *acrB* mutations (azithromycin resistance) were mapped onto the JI network (**Figure 4.22**). Genomes with triple QRDR mutations tended to cluster within JI-subgroups JI-A1 and JI-A4 (GenoTyphi 4.3.1.2) but were also found in different JI-groups (JI-C, JI-I, JI-M), consistent with the observation that QRDR mutants have emerged spontaneously in different lineages [124,232]. Specific *acrB* mutations aligned with JI-group (n=1/6

acrB(R717L) in JI-B, n=5/6 *acrB*(R717Q) in JI-C), but with relatively low prevalence of these mutations, this observation may be anecdotal.

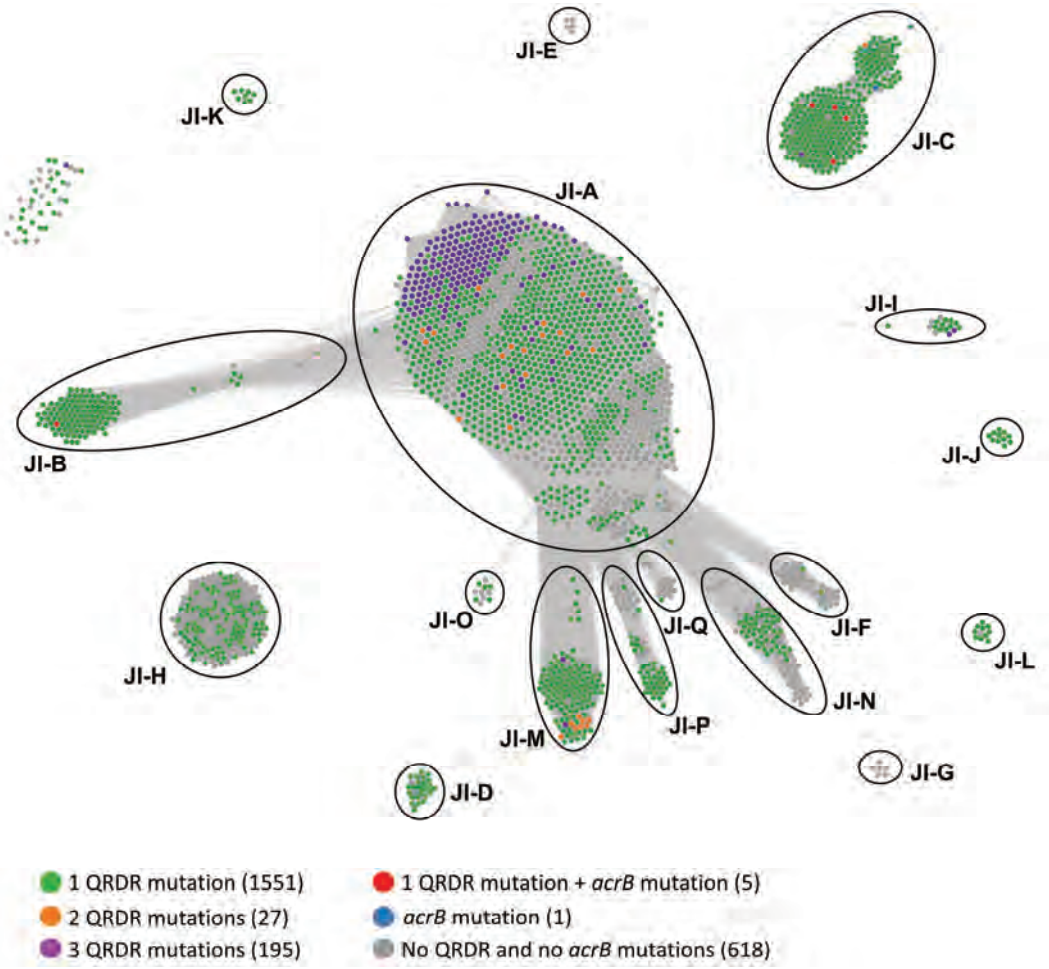


Figure 4.22 | Distribution of QRDR and *acrB* mutations in the JI network of Typhi. The network contains 2,392 nodes that are connected whenever JI>=0.983 and GLD>=0.05. Seventeen distinct clusters (named JI-A to JI-Q) are indicated by circles. Nodes are colored according to the pattern of quinolone resistance determining region (QRDR; mutations in genes *gyrA, gyrB,* and *parC*), and the *acrB* gene mutations. The number of genomes with mutations are indicated at the left.

4.6 U.S. Typhi pangenome structure reveals novel plasmid patterns

Nine different PTUs were detected in JI-groups, predominantly from MOB_P and MOB_H classes (**Table 4.2, Figure 4.9**). While some were well known (e.g. PTU-HI1A (IncHI1A) in JI-D), others were not well characterized (e.g. PTU-Y (IncY) in JI-K). All of them were graded as host range III or higher, including those that lacked a MOB relaxase

(PTU-E18 and PTU-Y are phage-plasmids). This is an indication of their broad ability to colonize bacteria from different genera of the same taxonomic family. Of particular interest were PTU-E50 (IncY) and PTU-Y (IncY) plasmids because of their carriage of *bla*$_{\text{CTX-M-15}}$ and association of the former PTU with XDR Typhi (**Table 4.2**).

Four different PTU-E50 plasmid variants were identified, each associated with a different group or subgroup: JI-B1 (IncY), JI-B2 (IncFIB(K)), JI-B3 (IncFIB(K)), and JI-J (IncY) (**Figure 4.23**). To investigate the phylogenetic relationships and protein content profiles of PTU-E50 plasmids from Typhi, a comparative analysis was conducted, incorporating PTU-E50 plasmids present in other *Enterobacteriaceae* species available in the RefSeq200 database.

The phylogenetic analysis (**Figure 4.23A**) showed that PTU-E50 plasmids of JI-B were very similar among them and similar to other plasmids of other hosts. However, JI-J plasmids appeared in a separate clade, indicating core differences. AcCNET proteome analysis (**Figure 4.23C**) revealed that plasmids from different JI-groups or JI-subgroups contain unique set of proteins. Interestingly, two plasmids from *E. coli* showed similar protein repertoires to the JI-B1 plasmids, as previously reported [126]. The genetic divergence between JI-groups by comparing genome structures (**Figure 4.23B**), showed that only partial homology exists between JI-B and JI-J plasmids. This indicated that even within the PTU-E50 classification, significant genomic variability exists across plasmids.

Further, these JI-groups have differing and significant ($P < 0.01$, chi-squared test of independence) geographic signals, despite extremely limited travel data; JI-B1 was linked to travel to Pakistan, JI-B2 to Bangladesh, and JI-J to Nigeria (**Figure 4.23D**). Interestingly, while most PTU-E50 (IncY) plasmids from JI-B1 ("XDR Pakistan" plasmid) harbored the *bla*$_{\text{CTX-M-15}}$ gene (n=84/88), four isolates did not. These genomes were likely variants of the original XDR Pakistan plasmid [126] that have subsequently lost the *bla*$_{\text{CTX-M-15}}$ gene, representing a novel lineage of the 4.3.1.1.P1 PTU-E50 (IncY)-containing strain. JI-grouping could be leveraged to link unique plasmids to geographic regions, in the same way that core genome SNPs are used to reflect geographical signals.

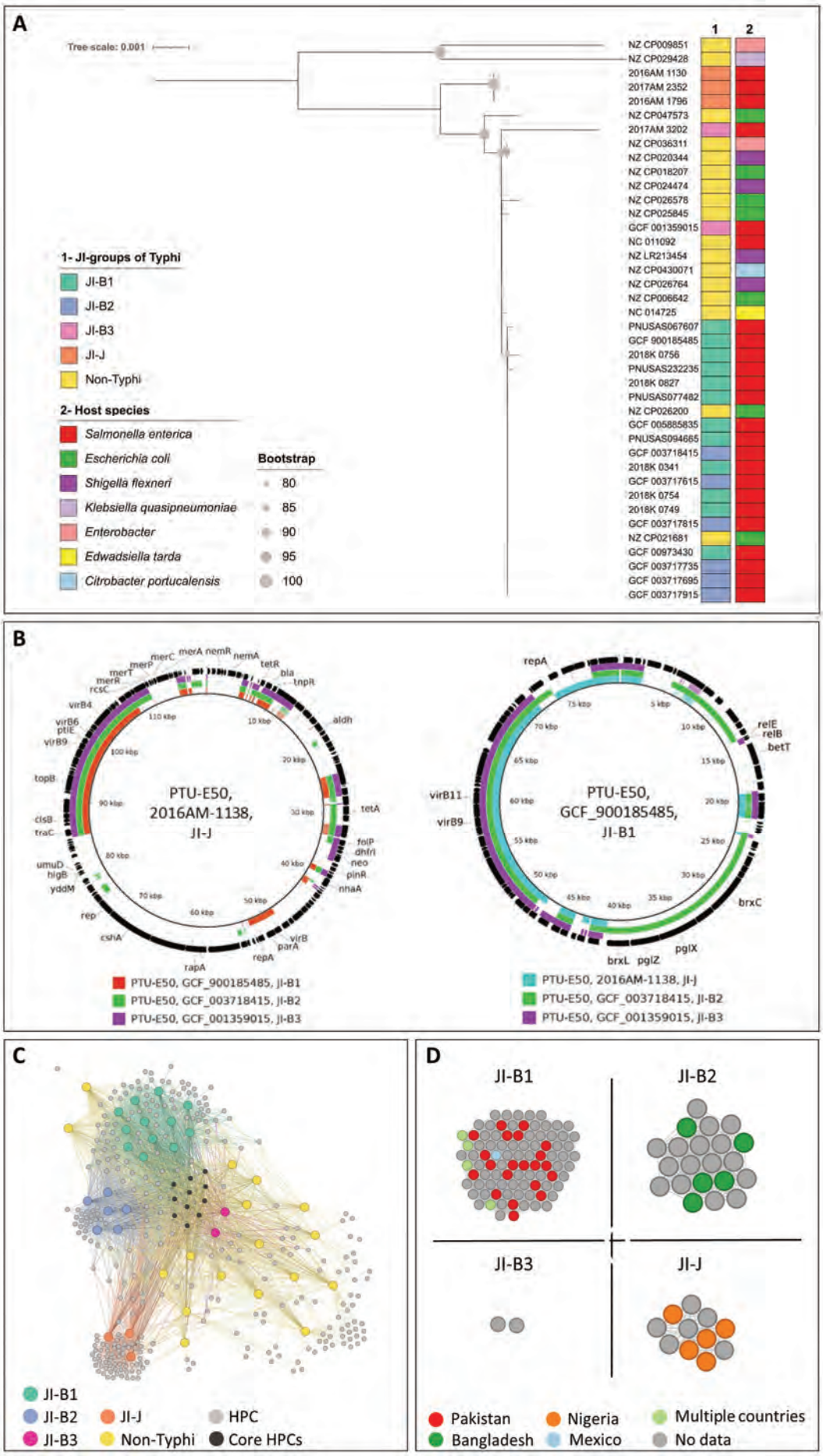

A
Tree scale: 0.001
1- JI-groups of Typhi
JI-B1
JI-B2
JI-B3
JI-J
Non-Typhi
2- Host species
Salmonella enterica
Escherichia coli
Shigella flexneri
Klebsiella quasipneumoniae
Enterobacter
Edwadsiella tarda
Citrobacter portucalensis
Bootstrap
80
85
90
95
100
1
2
NZ CP009851
NZ CP029428
2016AM 1130
2017AM 2352
2016AM 1796
NZ CP047573
2017AM 3202
NZ CP036311
NZ CP020344
NZ CP018207
NZ CP024474
NZ CP026578
NZ CP025845
GCF 001359015
NC 011092
NZ LR213454
NZ CP0430071
NZ CP026764
NZ CP006642
NC 014725
PNUSAS067607
GCF 900185485
2018K 0756
PNUSAS232235
2018K 0827
PNUSAS077482
NZ CP026200
GCF 005885835
PNUSAS094665
GCF 003718415
2018K 0341
GCF 003717615
2018K 0754
2018K 0749
GCF 003717815
NZ CP021681
GCF 00973430
GCF 003717735
GCF 003717695
GCF 003717915
B
PTU-E50, 2016AM-1138, JI-J
PTU-E50, GCF_900185485, JI-B1
PTU-E50, GCF_003718415, JI-B2
PTU-E50, GCF_001359015, JI-B3
C
JI-B1
JI-B2
JI-J
HPC
JI-B3
Non-Typhi
Core HPCs
D
JI-B1
JI-B2
JI-B3
JI-J
Pakistan
Nigeria
Multiple countries
Bangladesh
Mexico
No data

Figure 4.23 | Analysis of PTU-E50. (A) Core genome phylogenetic tree of PTU-E50 plasmids. 23 plasmids present in Typhi and 17 plasmids from RefSeq200 non-Typhi *Enterobacteriaceae* hosts were used to build a maximum likelihood tree based on the core genome by using IQ-TREE V2 [219] (with HKY+F model). The tree was midpoint rooted and visualized with iTol v6 [215]. UFBootstrap values > 80% are indicated by circles on the corresponding nodes. Branch length scale represents substitutions per site. Close to the tips, colored strips indicate the JI-group (1) and the taxonomic species (2) of the plasmid host. **(B)** Genome comparison of PTU-E50 plasmids. In the left panel, a PTU-E50 plasmid of the JI-J group was used as the reference and the colored rings represent regions of PTU-E50 plasmids from different JI-B subgroups shared with the reference. In the right panel, a PTU-E50 plasmid of the JI-B1 subgroup was used as the reference and the colored rings represent regions of PTU-E50 plasmids from JI-J, JI-B2, and JI-B3 groups shared with the reference. In both cases, the outermost ring depicts the genes of the reference plasmid. The genomic comparisons were carried out with BRIG v0.95 [210]. **(C)** Proteome network of PTU-E50 plasmids. The proteins of the PTU-E50 plasmids were clustered at 80% identity and 80% coverage using AcCNET [205]. The larger nodes correspond to plasmids and were colored according to the JI-group or subgroup of their bacterial host. The smaller nodes represent HPCs and were colored in black if containing members from all plasmids, or grey otherwise. Both kinds of nodes were connected if a plasmid contained a member in the corresponding protein cluster. **(D)** Probable origin of JI-groups containing PTU-E50 plasmids. JI-group or subgroups containing PTU-E50 plasmids are shown. Nodes were colored according to the country of travel.

PTU-Y was exclusively identified in JI-K. Two genotypes were present in this JI-group, 4.3.1.1 and 4.3.1.2 (**Figure 4.12B**), and only PTU-Y plasmids hosted in the latter genotype carried $bla_{\text{CTX-M-15}}$. This plasmid carried an IncY replicon, but rather than being a conjugative plasmid (as is PTU-E50 (IncY) in JI-B1), it was a large non-conjugative phage-plasmid whose transmission is governed by an entirely different mechanism [240,241]. In this case, relying on replicon typing alone (as commonly practiced) would have generated confusion, as two very distinct plasmid types (PTU-E50 and PTU-Y) carried the same replicon (IncY) (**Table 4.2**), and interestingly, in this case, both harbor $bla_{\text{CTX-M-15}}$. Of interest, carriage of PTU-Y was significantly associated with travel to Iraq ($P < 0.01$, chi-squared test of independence).

JI grouping has the advantage of accounting for all genetic material within the plasmid rather than a single replicon target, and therefore can simultaneously differentiate highly-related plasmids (as seen for PTU-E50 plasmids), and disintegrate seemingly similar plasmids (PTU-Y (IncY) versus PTU-E50 (IncY)). These plasmid subgroups can be rapidly

detected in a network (and overlaid with epidemiological data), preventing the continual need for separate plasmid core genome analysis.

In contrast to large plasmids (>80 kb), smaller plasmids (<80 kb) did not often contain enough genetic content to define individual JI-groups (**Table 4.2).** For example, JI-H had only one member with a 50 kb PTU-N1 (IncN) plasmid, and instead, was genetically distinct from other groups by the presence of a ~55 kb ICE (**Figure 4.9, Table 4.3**). Another small mobilizable plasmid, PTU-E80 (IncX1, ~25 kb, highly related with PTU-X1) (**Table 4.2**), was among the most common plasmids detected, predominantly in JI-N, and while it is likely important to this group (>85% of members carry PTU-E80), it did not exclusively underpin the genetic definition of JI-N. Instead, JI-N was genetically distinct from other JI-groups also due to the absence of a ~21 kb phage (prophage 1) and the absence of a 21Kb IME (MOB_Q) (**Figure 4.9, Table 4.3**). Of interest, JI-N was almost exclusively lineage 2.0.2 (one genome is 4.1), a genotype that was also detected in JI-A and JI-I. In this case, JI-grouping enables stratification of an epidemiologically important genotype [193] using "unknown" accessory genetic content.

4.7 U.S. Typhi pangenome structure offers avenues for further investigation

Co-visualization of phylogenetic lineages across the JI network enabled rapid detection of groups that are likely characterized by clonal expansion (homology-by-descent), versus groups that contain disparate genomes that have converged on their MGEs (homology-by-admixture). For example, genomes of lineage 3.1.1 fell into either JI-A or JI-J. JI-J genomes were exclusively of genotype 3.1.1 and differentiated from JI-A partially due to carriage of a unique PTU-E50 (IncY) plasmid (**Figure 4.9**). Thus, it is plausible that JI-A/3.1.1 genomes represent a precursor strain that subsequently acquired a PTU-E50 plasmid and clonally expanded to become group JI-J. JI-J was significantly ($P < 0.01$, chi-squared test of independence) associated with travel to Nigeria (despite limited travel data for U.S. isolates), an epidemiological signal that could prove useful as lineage 3.1.1 is the most common genotype in Western Africa [124].

Lineage 2.3.2 is also common in Western Africa and the Americas and was recently shown to separate into discrete geographic clades by distance-based phylogeny [124]. Sixty-

two genomes from this phylogeny were also present in the U.S. dataset; 61 genomes fell into the "Central American" clade and belonged to JI-H, while the remaining genome was in the "Western Africa" clade and belonged to JI-A. JI-H differed from JI-A by the presence of a ~55 kb ICE (**Figure 4.9, Table 4.3**), and was significantly associated with travel to the Americas ($P < 0.01$, chi-squared test of independence). With further confirmation, the presence of this ICE could potentially be used to stratify 2.3.2 lineages into geographically and epidemiologically meaningful groups without the need for phylogenetic analysis.

In contrast to differentiation, aggregation of disparate genomes by their pangenome is of interest. For example, JI-C genomes all carried a unique ~107 kb phage-like PTU-E18 (IncFIB (pHCM2)) plasmid (**Figure 4.9**), but belonged to a variety of GenoTyphi primary caldes and lineages (**Figure 4.12, Table S1**) with diverse geographic signals, including 4.3.1.1 dominant in Pakistan [124] and 3.5.4 exclusively associated with Samoa [242]. Convergence of these diverse lineages on a large non-mobilizable phage-plasmid that does not carry AMR genes is curious, since acquisition cannot simply be explained by conjugation under antibiotic selection pressure. Rather, acquisition of plasmid-phages relies on viral-like mechanisms (transduction or lysogenic conversion) [240] and is likely induced by different ecological factors than conjugation [243,244]. Grouping and investigating Typhi strains through the lens of shared MGEs provides an opportunity to uncover common environmental exposures between genomes that might otherwise appear disparate using phylogenetic methods, adding an exciting new dimension to Typhi epidemiology.

4.8 Pangenome structure of U.S. Typhi is generalizable

To assess whether the network obtained with U.S. genomes is generalizable to the global population structure of Typhi, a new network was generated with a large dataset from a distinct geographic region. It included 1,606 genomes isolated in the Indian subcontinent and 136 genomes (**Table S2**) from the U.S. dataset, representative of the 17 JI-groups previously identified (**Figure 4.24**). The new network organized into 17 JI-groups already delimited in the U.S. dataset (5 JI-groups (E, G, J, N, and P) were represented only by reference genomes in this network and thus absent in the Indian dataset), and two new JI-groups (JI-R and JI-S, containing 10 and 11 genomes, respectively). JI-R genomes contained a PTU-X1 plasmid and three chromosomal regions enriched in phage-related genes, while JI-S members contained two plasmids (PTU-E18, PTU-HI1A) and an IME.

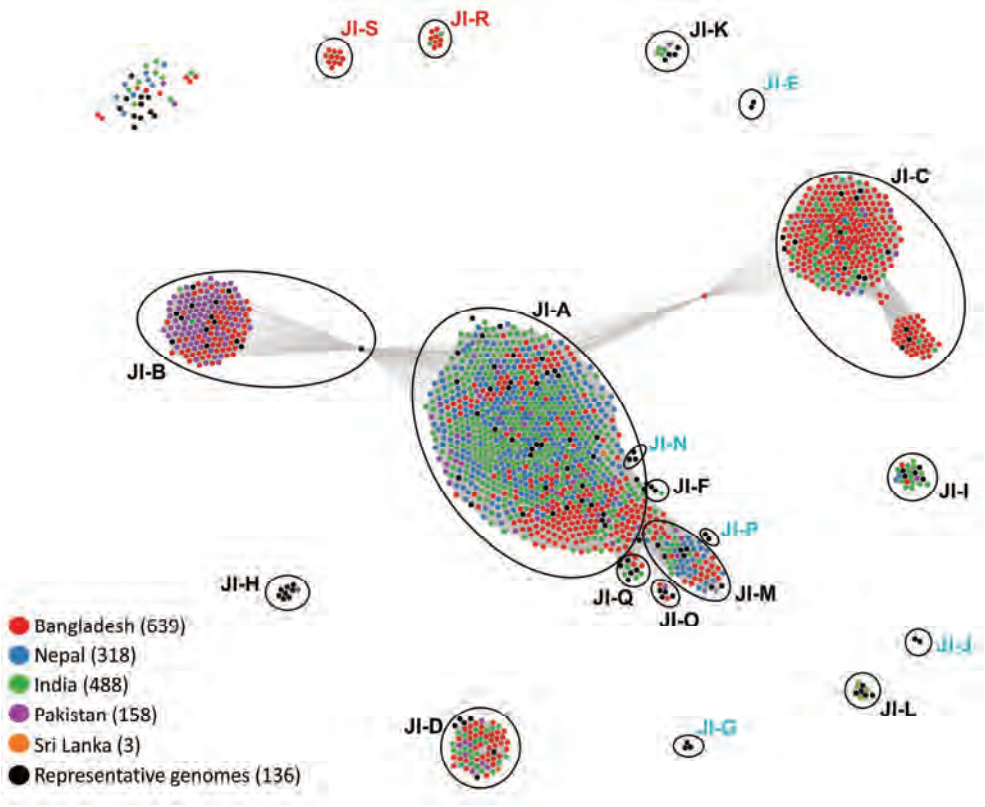


Figure 4.24 | Genomic diversity among Typhi genomes from the Indian subcontinent. The network comprises 1,606 Typhi genomes originating from the Indian subcontinent, along with 136 reference genomes selected from both U.S. dataset and RefSeq200. These reference genomes were chosen based on their high connectivity within each cluster of the original network and are representatives of the different JI-groups. Each genome is represented by a node, which is colored by its country of isolation as indicated in the legend. The number of isolates from each country is indicated between parentheses. Nodes are connected with thresholds JI=0.983 and GLD=0.05. Nineteen distinct clusters (JI-A to JI-S) detected by the Louvain method are indicated by circles. They are named in blue when only found in the representative genomes, in red when only present in the Indian subcontinent dataset, and in black when containing genomes from both datasets.

Since a quarter of the U.S. dataset was associated to travel to the Indian subcontinent, which could bias the comparison, we analyzed a different dataset, representative of the global Typhi diversity. We generated a JI network using 1,804 globally representative Typhi genomes spanning more than 60 countries, which were previously used to define the GenoTyphi typing nomenclature [229], and 136 reference genomes from the U.S. dataset (**Figure 4.25**). Emergence of novel clusters would be an expectable outcome, especially considering that they may emerge by the acquisition or loss of MGEs. Nevertheless, the vast majority (1,662/1,804, 92%) of the genomes in the GenoTyphi dataset clustered in 12 of the

originally defined JI-groups. The remaining genomes fell into one of eight small new JI-groups (98 genomes) or were singletons (44 genomes). The application of this method to these datasets demonstrates the robustness of the JI-groups identified here, despite being established using only genomes collected in the U.S. It suggests that the U.S. dataset effectively represents the global diversity of the Typhi pangenome and serves as a proxy for global sentinel surveillance.

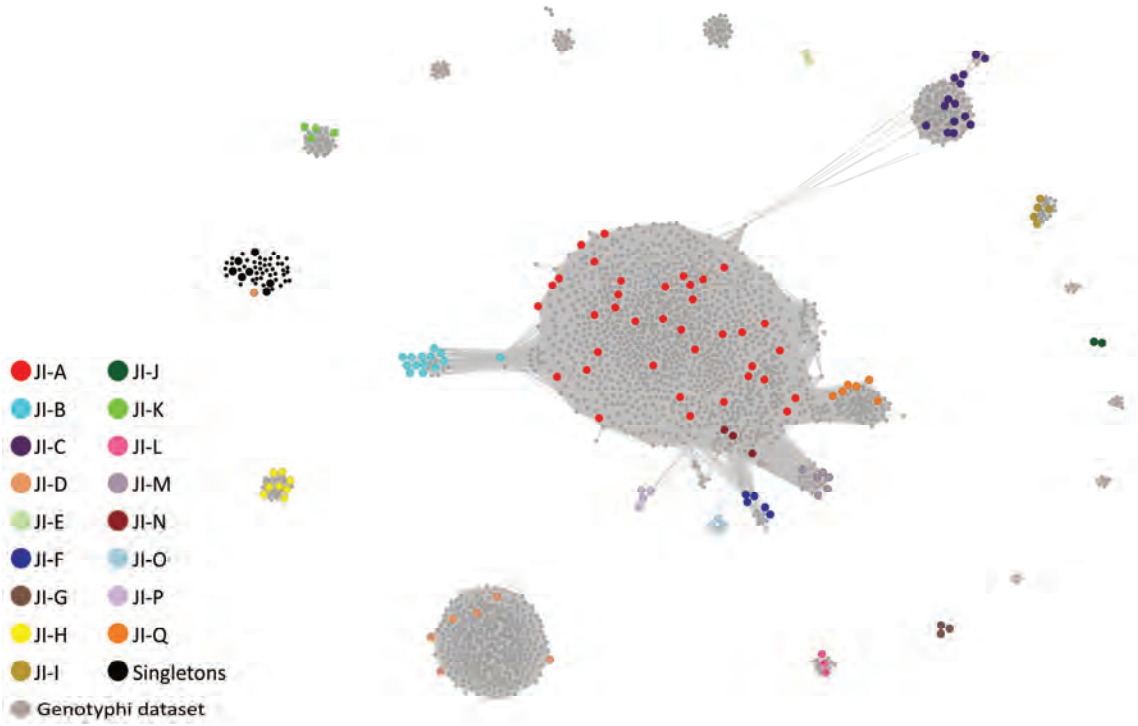


Figure 4.25 | Genomic diversity of globally representative Typhi genomes. The network contains 1,804 Typhi genomes from the GenoTyphi dataset and 136 genomes (from both U.S. dataset and RefSeq200 as representatives of the different JI-groups). Nodes in the network represent genomes, with larger nodes corresponding to representative genomes, which are color-coded based on the JI-group they are assigned to. Nodes are connected whenever JI>=0.983 and GLD<=0.05.

4.9 Typhi diversity in the pre-antibiotic era

Thirty eight Typhi genomes of the pre-antibiotic era obtained from the Murray collection [185] (**Table S4**) were incorporated to the U.S. genome network to explore the genetic relationship between strains from the pre-antibiotic era and those circulating today. The 38 Murray genomes were distributed across multiple genetic clusters, JI-A (20 genomes), JI-F (7 genomes), JI-M (5 genomes), JI-I (1 genome), JI-Q (1 genome) and four

isolates were singletons (**Figure 4.26A**). Whithin the JI-A group, the Murray genomes exhibited additional substructuring (**Figure 4.26B**). This distribution indicates that specific lineages of Typhi were already stablished prior to the introduction of antibiotics and have continued to circulate with minimal genetic changes.

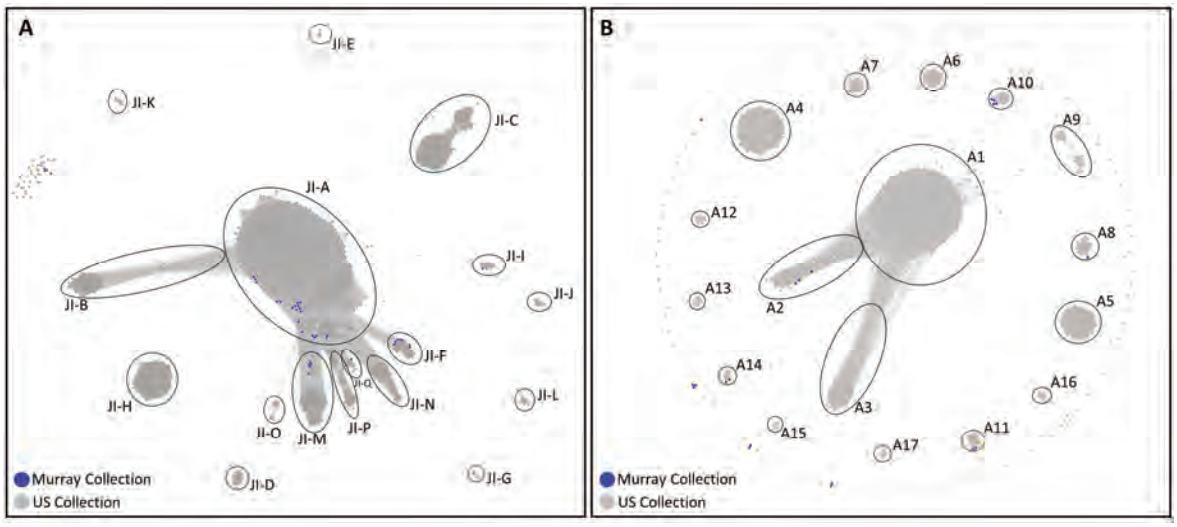


Figure 4.26 | Genomic diversity among Typhi from the Murray collection. (A) Distribution of Typhi genomes of the Murray collection in the overall JI-network of Typhi. The network contains all the Typhi genomes of the current study (n=2,392, colored in grey) and 38 Typhi genomes from the Murray collection, colored in blue (Table S4). Nodes are connected whenever JI>=0.983 and GLD>=0.05. **(B)** JI network of JI-A genomes. The network contains 1,320 JI-A genomes, and 20 JI-A genomes from the Murray collection. This network was generated at JI=0.995.

The core phylogenetic analysis (**Figure 4.27**) further illustrated this stability, showing that Murray genomes occupay distinct branches, with JI-A members appearing in more recent clades alongside contemporary isolates. In addition, the Murray genomes contained representative genomes from all GenoTyphi primary clades (**Table S4**).

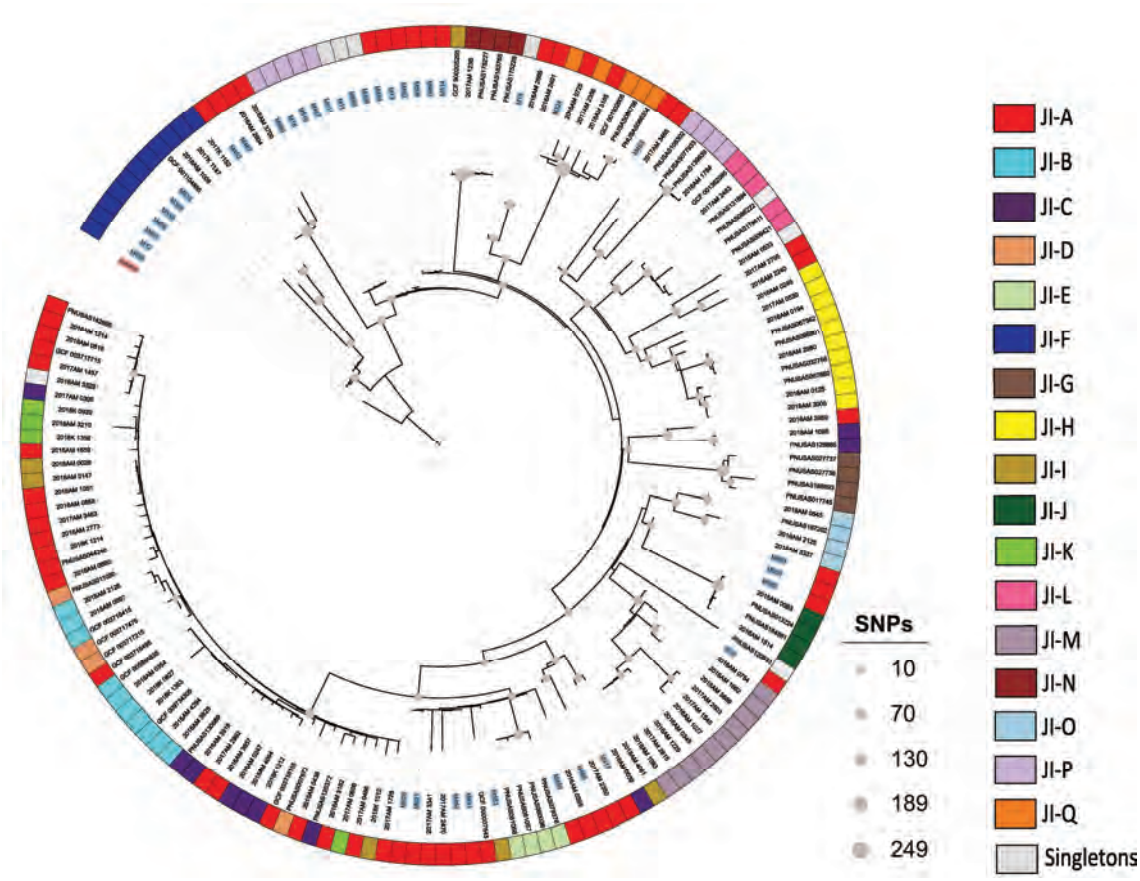
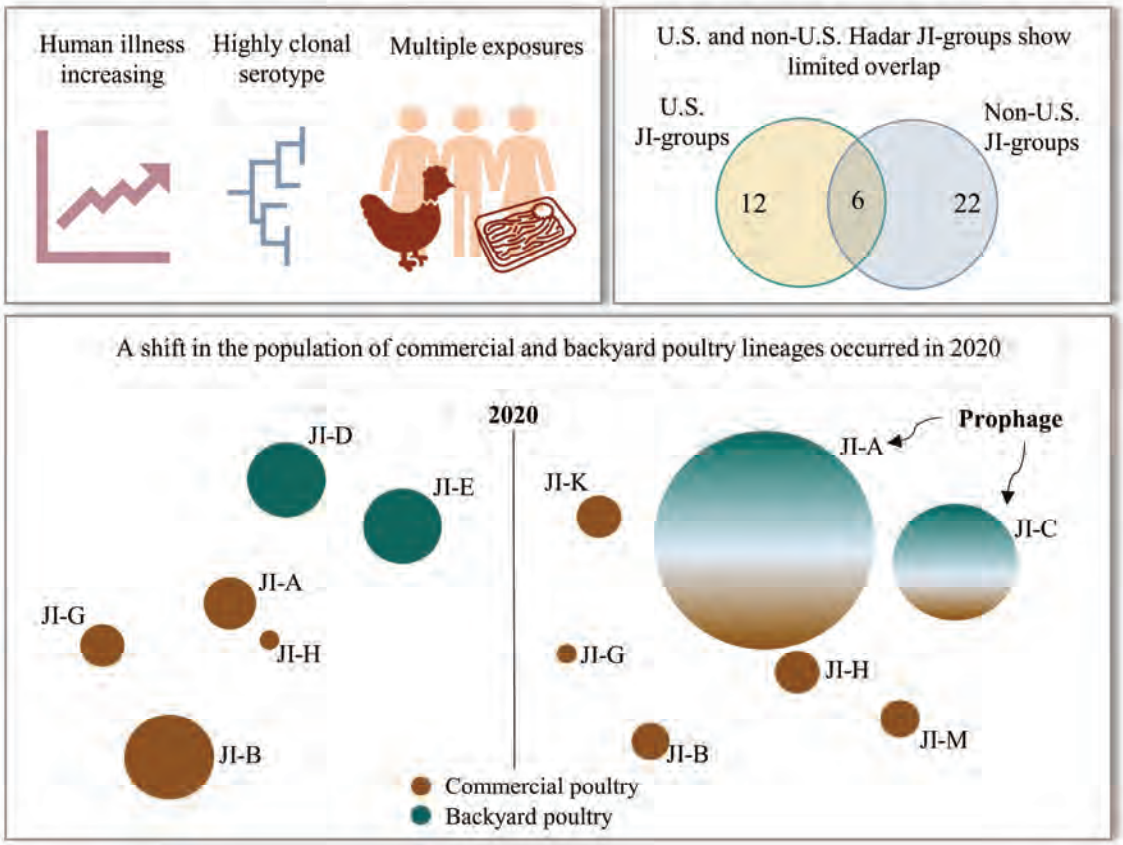


Figure 4.27 | Phylogenetic distribution of Typhi genomes from the Murray collection. Parsimony reconstruction based on the core SNPs using kSNP3.0 [217]. The tree includes 130 representative genomes of the U.S. dataset and 38 Typhi genomes of the Murray collection (names labeled in blue). Genome NZ_CP015724.1 (serovar Indiana) was used as an outgroup and is labeled in red. The JI-group for each genome is indicated in the outer ring. Circles at the internal nodes indicate the number of SNPs that are shared exclusively by the descendands of each node. Branch length scale represents changes per number of SNPs. The tree was visualized with iTol v6 [215].

CHAPTER 5: RESULTS II

SALMONELLA ENTERICA SEROVAR HADAR

Graphical abstract



Salmonella enterica subsp. *enterica* serovar Hadar (Hadar) is an emerging zoonotic pathogen in the United States (U.S.) associated with both commercial and backyard poultry. It is characterized by high clonality and is transmitted to humans through contaminated food or contact with animals. In this study, we explored the population structure and epidemiology of Hadar in the U.S. through a pangenome approach. Historically, genetically distinct lineages circulated in commercial versus backyard poultry populations. Around 2020, the U.S. Hadar population experienced a notable shift, driven by the rapid expansion of two clonal groups (JI-A and JI-C) harboring a previously uncommon prophage-like element. These groups were implicated in multiple outbreaks linked to both backyard and commercial poultry sources. Global genomic comparisons revealed that U.S. Hadar isolates form genetically distinct groups with minimal overlap with international strains, highlighting both the localized evolution and the unique dynamics shaping the U.S. Hadar population.

5.1 Background and specific objectives

In recent years, backyard poultry-associated salmonellosis (BYPAS) outbreaks have increased significantly (**Figure 5.1**), causing more illnesses in the U.S. than outbreaks linked to any other type of animal [186]. Compared to multistate BYPAS outbreaks from 1990 to 2014, the number of outbreak-associated illnesses has nearly tripled [245]. Among these outbreaks, *Salmonella* Enteritidis, Hadar, and Infantis have been responsible for the highest annual number of BYPAS outbreak-associated illnesses during 2015-2022. While extensive research has been conducted on Enteritidis [246,247] and Infantis [248,249] due to their high prevalence in various settings, Hadar remains comparatively under-studied. Notably, *Salmonella* Hadar has resulted in the highest proportion of BYPAS hospitalizations compared with other serotypes. Indeed, two of the five largest BYPAS outbreaks in the U.S. were caused by *Salmonella* Hadar (**Table 5.1**: Outbreak A and Outbreak C) [186].

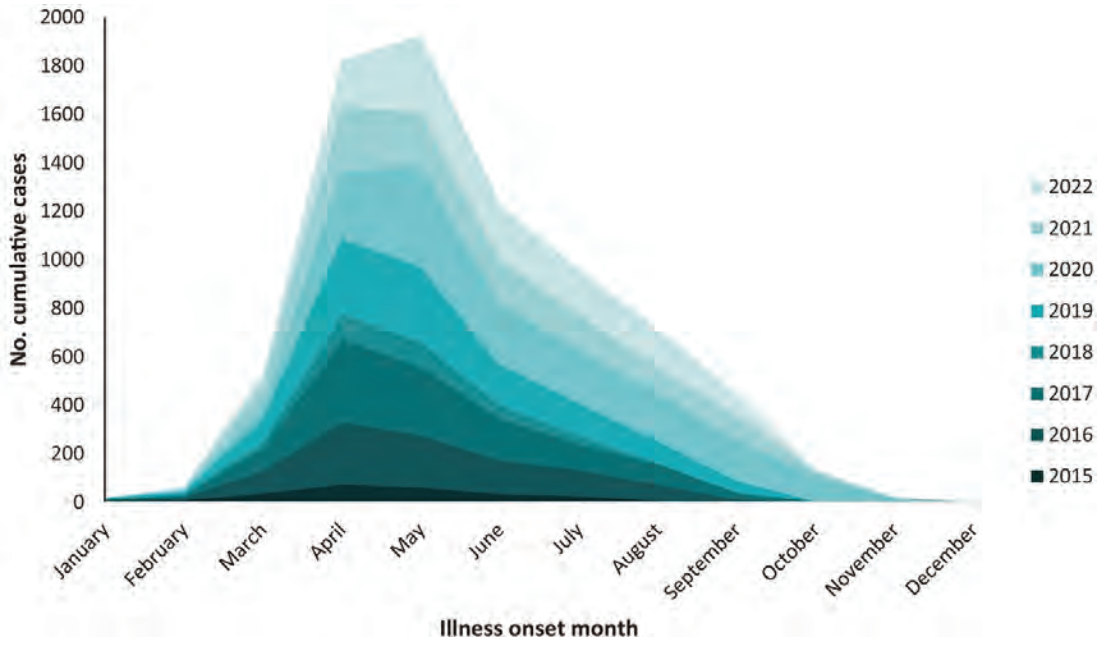


Figure 5.1 | Multistate outbreaks of *Salmonella* linked to contact with backyard poultry, United States, 2015-2022. Original legend "Cumulative number of backyard poultry-associated *Salmonella* illnesses by month of illness onset, United States, 2015-2022. The number of cases identified each month in a given year is indicated by the area shaded by each different colour, with the overall curve demonstrating the cumulative number of cases during 2015-2022 for each month. When patient illness onset date was not reported, an estimated onset date was determined as 3 days before the reported isolation date." Figure taken from [186].

Salmonella enterica subsp. *enterica* serovar Hadar (Hadar) is a non-typhoidal *Salmonella*. Hadar is transmitted to humans via contaminated food and contact with animals and has caused several outbreaks in U.S. during the last decade, linked to either ground turkey consumption or contact with backyard poultry (i.e., privately-owned, non-commercial poultry such as chickens, ducks, or turkeys) [183,186]. Although Hadar is considered a highly clonal serotype, exhibiting limited variability based on cgMLST [183], strains transmitted by these two different sources were historically differentiable with allele differences ranging from 25 to 50.

However, in 2020, despite decreased reporting of enteric illness during the early years of the COVID-19 pandemic, an emergent Hadar strain was linked to both ground turkey consumption and backyard poultry contact. These outbreaks have caused over 900 human illnesses in several states during the period 2020-2023 (**Table 5.1**), more than doubling the fewer than 500 total reported cases of Hadar in all years prior to 2020 [183,250]. Traceback investigations were not able to identify an epidemiological connection between the indistinguishable strains (as determined by cgMLST) from two seemingly distinct sources: commercial poultry and backyard poultry [183,186]. This emergent strain, now responsible for > 2000 human illnesses, continued to cause outbreaks into 2024. It has been designated by the U.S. CDC as a Reoccurring, Emerging, or Persisting (REP) strain REPTDK01, with a cgMLST range of 0–26 allele differences [251].

Table 5.1 | Summary of some multistate outbreaks caused by the REPTDK01 strain from April 2020 to May 2023. Taken from https://www.cdc.gov/salmonella/php/data-research/reptdk01.html.

Outbreak	Dates people got ill	Outbreak source	Reported ilnesses	Reported hospitalizations	Reported deaths	Number of states with illnesses
Outbreak A	April 2020-November 2020	Backyard poultry (confirmed)	848	186	0	49
Outbreak B	December 2020-April 2021	Ground turkey (confirmed)	33	4	0	14
Outbreak C	April 2021-October 2021	Backyard poultry (confirmed)	364	111	0	45
Outbreak D	April 2022-October 2022	Backyard poultry (confirmed)	273	74	1	40
Outbreak E	February 2023-May 2023	Multiple vehicles (confirmed backyard poultry and suspected ground turkey)	55	9	0	28

Therefore, the rapid increase in Hadar cases observed in recent years, along with the detection of a single strain across multiple, separate poultry industries in several U.S. states highlights a public health threat that warrants greater attention and further investigation.

Given the limitations in the discriminatory power of cgMLST for this strain, the Jaccard Index was employed to assess pangenome relatedness of Hadar along the U.S. farm-to-fork continuum. Building on the thesis-wide objectives (see *Objectives* chapter), the specific aims of the Hadar study are to:

- Construct a foundational landscape of Hadar diversity. Characterize both core and accessory genomic processes that contribute to Hadar's genetic heterogeneity within a U.S. Hadar dataset, and compare these findings with Hadar datasets from other regions.
- Evaluate whether pangenome analysis can differentiate REPTDK01 strains isolated from turkey consumption from those associated with backyard poultry contact.
- Explore the potential of pangenome analysis to provide actionable insights for outbreak tracking and improved source attribution.
- Assess how accessory genomic elements influence Hadar's short-term evolutionary dynamics, potentially driving the emergence of new sublineages.

5.2 Pangenome analysis of U.S. Hadar population

JI was used as a similarity measure between all genome pairs within the Hadar dataset, calculated with BinDash [62]. This analysis involved pairwise comparisons of 3,384 Hadar genomes. The distribution of JI values showed that the majority of genome pairs had JI values greater than 0.9 (**Figure 5.2**).

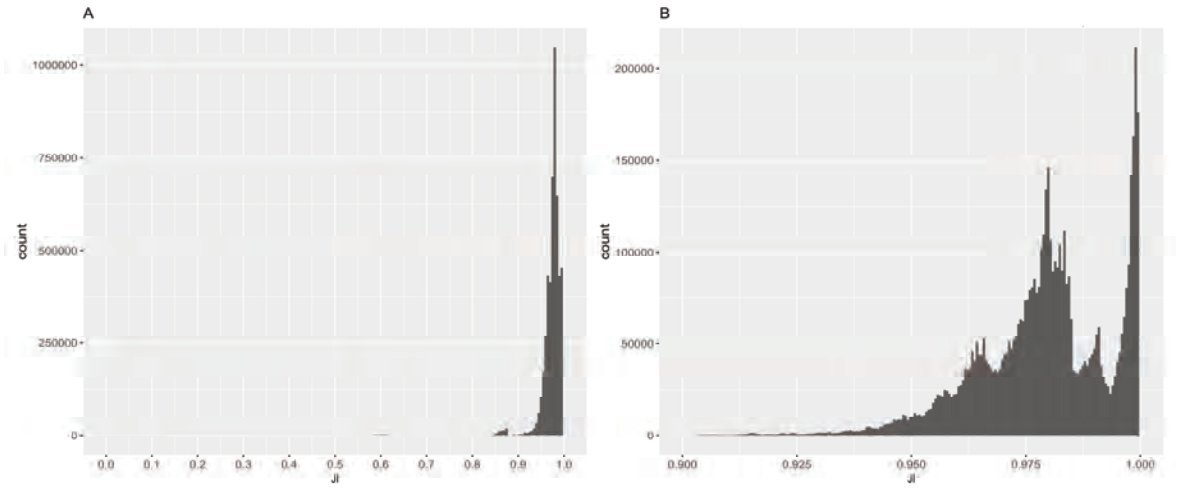


Figure 5.2 | JI distribution obtained from the pairwise comparison of Hadar genomes. (A) Histogram displaying the distribution of JI values ranging from 0 to 1. **(B)** Zoom in on JI values between 0.9 and 1.

Following the approach used for Typhi, pairwise JI comparisons of Hadar genomes were visualized as a network, where each node represents a genome and edges denote similarity based on JI values. The threshold used for this analysis was carefully determined through the examination of multiple network properties across different cutoff values (**Figure 5.3**). This analysis revealed a threshold range between 0.985 and 0.990, where the network's transitivity remained stable with values above 0.95. This stability indicates consistent internal cluster connectivity, making any threshold within this interval a reasonable choice. However, other network properties within this range showed subtle variations in community structure. The optimal threshold for analyzing Hadar genomes was determined to be JI=0.988, at which point the network resolved into 18 well-defined communities that encompassed more than 95% the genomes. This threshold balances avoiding excessive fragmentation at higher thresholds and preventing over-clustering at lower values. This empirically derived threshold provides a robust foundation for downstream comparative genomic analyses.

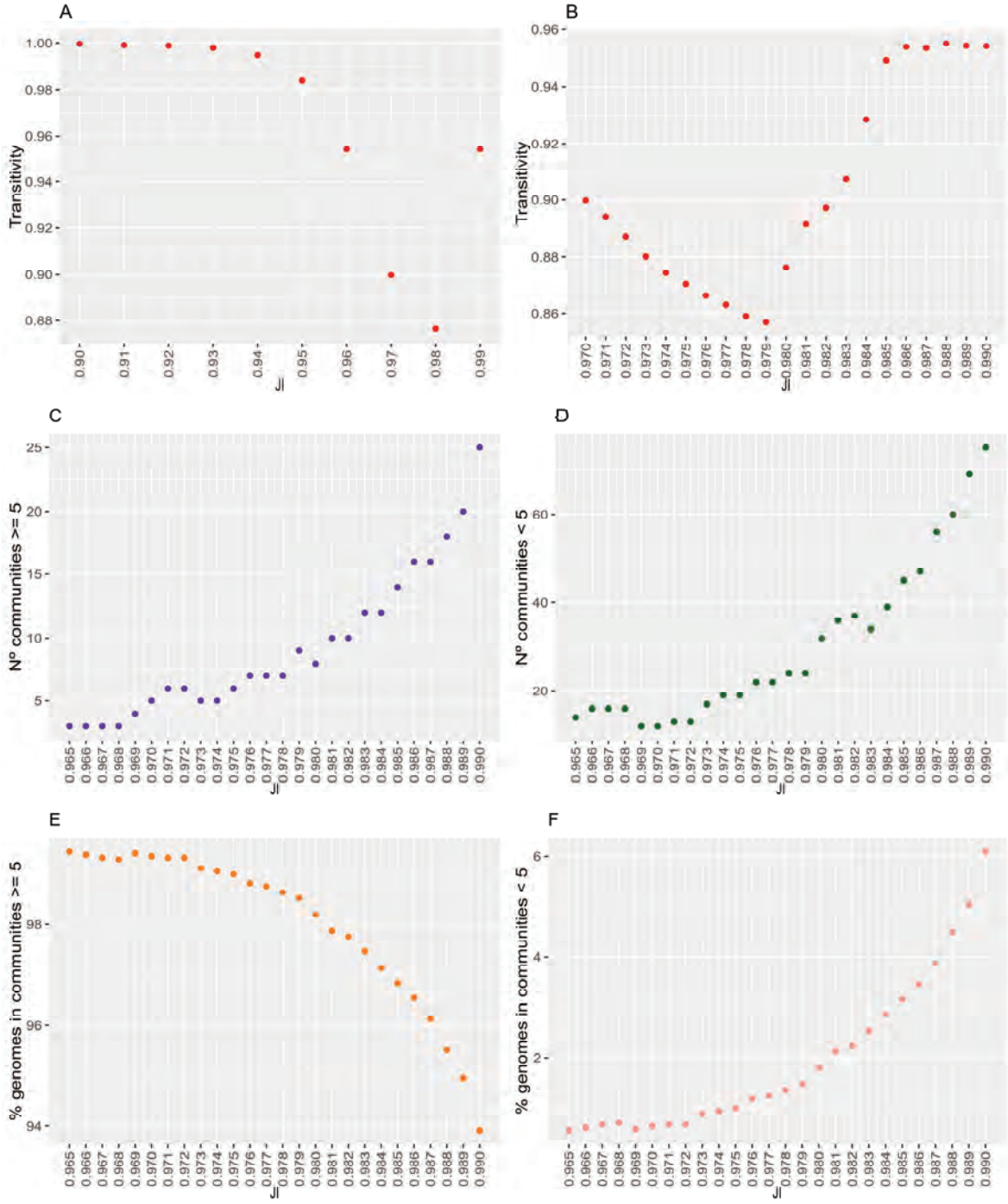


Figure 5.3 | Analysis of different networks parameters in the Hadar dataset. A range of JI thresholds was applied to the original network and several criteria were analyzed. **(A)** Transitivity for JI values in the range from 0.9 to 1. **(B)** Transitivity for JI values between 0.97 and 0.99. **(C)** Number of communities containing at least five members. **(D)** Number of communities containing less than five members. **(E)** Percentage of genomes contained in communities with at least five members. **(F)** Percentage of genomes contained in communities with less than five members.

To independently analyze differences in the core and accessory genome of Hadar, PopPUNK was employed (**Figure 5.4**). Hadar exhibited several distinct populations with varying degrees of core genome divergence. The majority of Hadar genomes had very low core genome distances, as indicated by the dense concentration of points near the y-axis. Within this range, there was little variability, as reflected by the tightly clustered contour lines. However, some minority populations displayed higher core genome distances, while the extent of accessory genome variation remained comparable to the majority of genomes. This suggests that certain groups of genomes share similar core genome distances but exhibit varying degrees of accessory genome divergence. Another subset of genomes displayed both high core genome and high accessory genome distances, representing highly divergent lineages that may have undergone significant evolutionary changes through both vertical divergence and accessory gene acquisition.

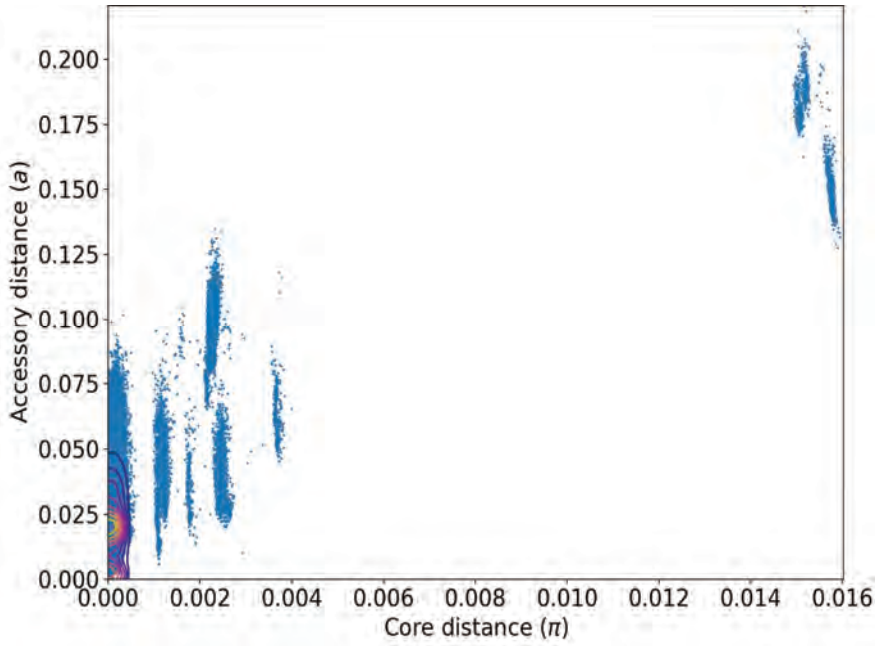


Figure 5.4 | PopPUNK analysis of Hadar isolates showing core versus accessory genome distances. Each dot represents a pairwise comparison between two genomes, with the x-axis indicating core distance (primarily driven by SNP differences) and the y-axis indicating accessory distance (reflecting variation in gene content). Contour lines highlight density regions, illustrating that most isolates cluster at very low core distances while displaying a broader range of accessory distances.

Hadar genomes self-organized into 18 clusters using a JI threshold of 0.988, labeled JI-A through R; less than 5% of genomes (n=158/3,384) did not cluster with a JI-group and were considered singletons (**Figure 5.5**).

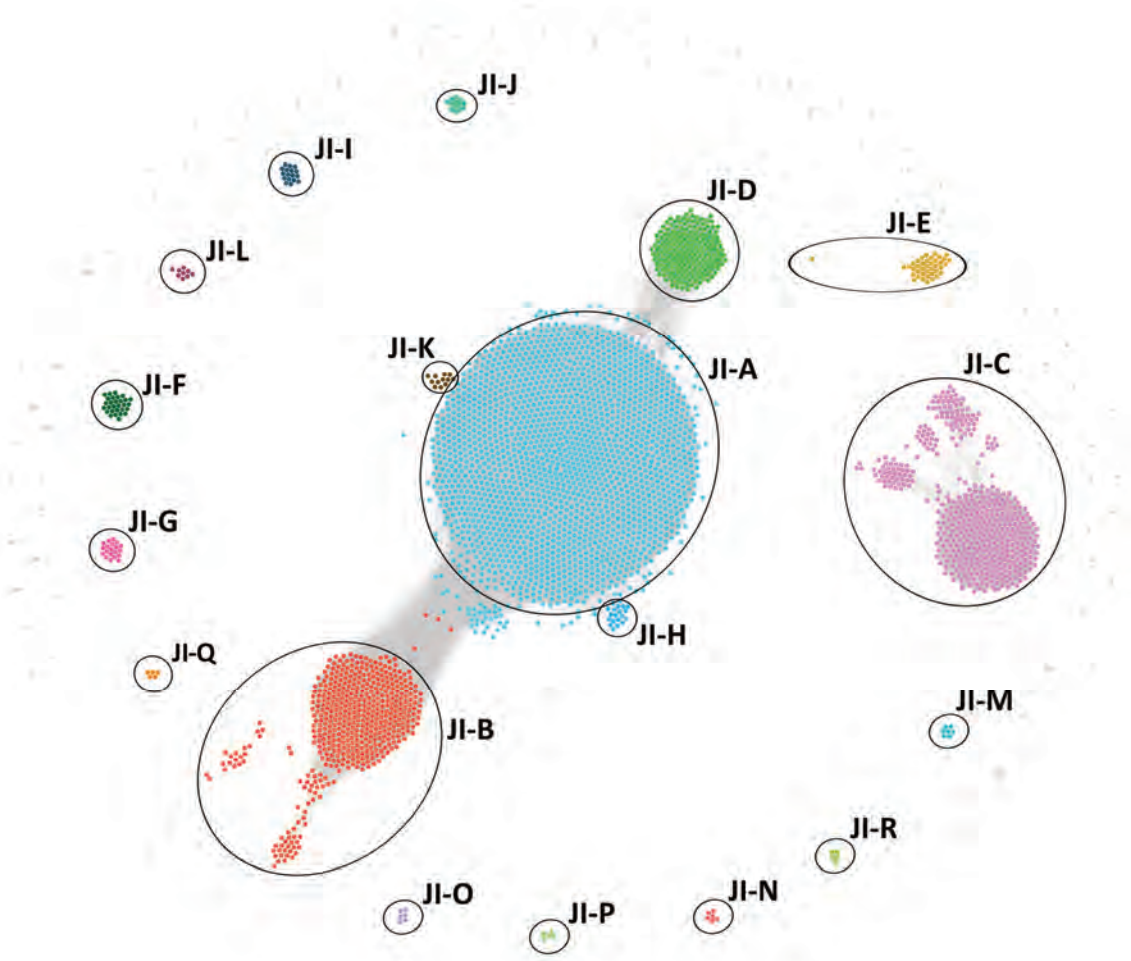


Figure 5.5 | Distribution of Hadar genomes by JI. The network contains 3,384 nodes, connected when JI≥=0.988. Eighteen groups (named JI-A to JI-R) are indicated by circles. The nodes of each JI-group are represented by a distinct color.

JI-group A was the largest group (n=1,899/3,384), with all other JI-groups represented by at least five genomes (**Table 5.2**).

Table 5.2: Summary of Hadar JI-group information for 3,384 genomes.

JI group	Count[a]	%[b]
A	1899	56,1
B	489	14,5
C	453	13,4
D	191	5,6
E	40	1,2
F	29	0,9
G	20	0,6
H	20	0,6
I	17	0,5
J	13	0,4
K	12	0,4
L	9	0,3
M	7	0,2
N	6	0,2
O	6	0,2
P	5	0,1
Q	5	0,1
R	5	0,1
Singleton	158	4,7
Total	**3384**	**100**

[a] | Number of genomes present in each JI group.
[b] | Percentage of genomes from the total data set that belong to each JI group.

Most of the groups showed median JI values between their genomes well above 0.99, indicating high genomic similarity among group members (**Figure 5.6A**). Additionally, genomes within each JI-group exhibited an ANI greater than 99.8% (**Figure 5.6B**). For JI-groups A, B, C and D, while the central tendency of JI and ANI values appeared significantly high, as indicated by the upward displacement of the box plots, the presence of scattered data points at lower JI values within these distributions was notable (**Figure 5.6**). This observation suggests potential genomic divergence or substructuring within these groups, where certain genomes exhibit lower similarity to the majority of the group.

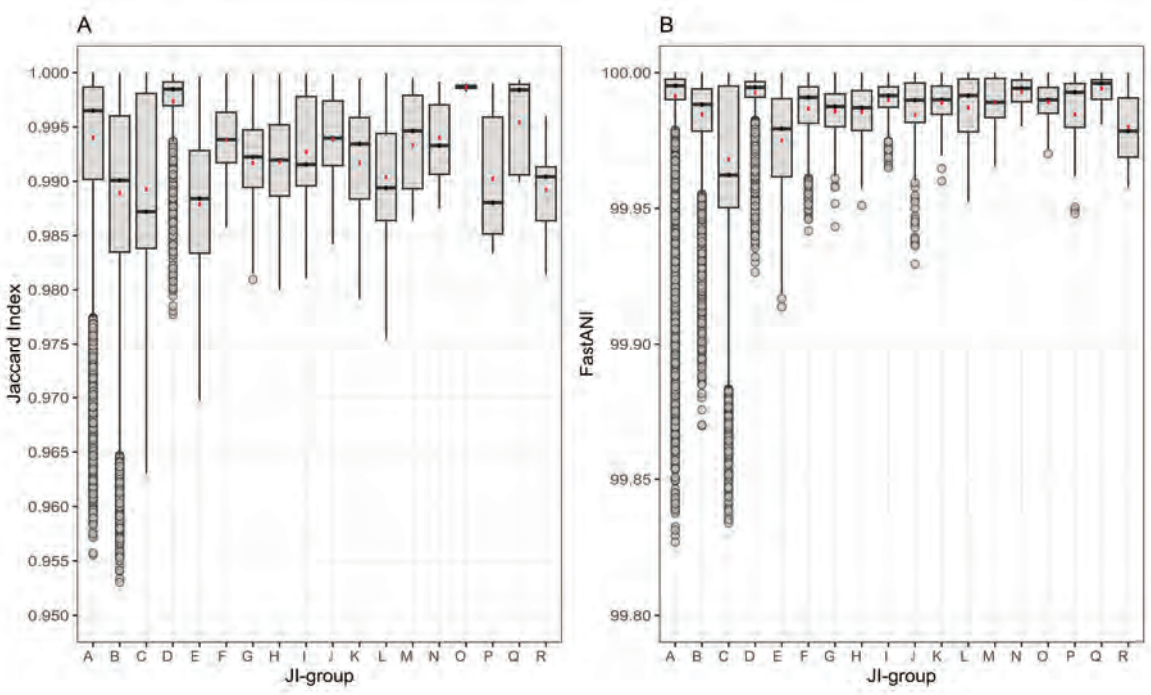


Figura 5.6 | Relatedness of Hadar genomes within each JI-group. Boxplot illustrating the distribution of **(A)** Jaccard Index and **(B)** FastANI values across different JI-groups. The boxplot displays the IQR of JI or ANI values within each JI-group, with the lower and upper edges of the box indicating the first quartile and third quartile, respectively. Within each boxplot, horizontal lines represent the median (black) and the average (red) values. The 'whiskers' of the boxplot extend to the most extreme values within 1.5 times the IQR from the edges of the box, while outliers are depicted as individual points beyond the whiskers.

The three largest groups, JI-A, JI-B, and JI-C, were further divided into subgroups using an increased JI threshold (**Figure 5.7**). Subgroups were identified using the Louvain method, with a minimum size criterion of five members. JI-A subgroups A1-15 were defined at JI=0.995; JI-B subgroups B1-6 and JI-C subgroups C1-9 were defined at JI=0.992. By applying a more stringent threshold, we were able to detect subtle genetic variations and structural patterns that the broader JI threshold did not reveal, thereby enabling a deeper analysis of the genomic relationships among these primary groups. The chosen thresholds were determined through visual inspection of the network, striking a balance between the number of groups and network transitivity.

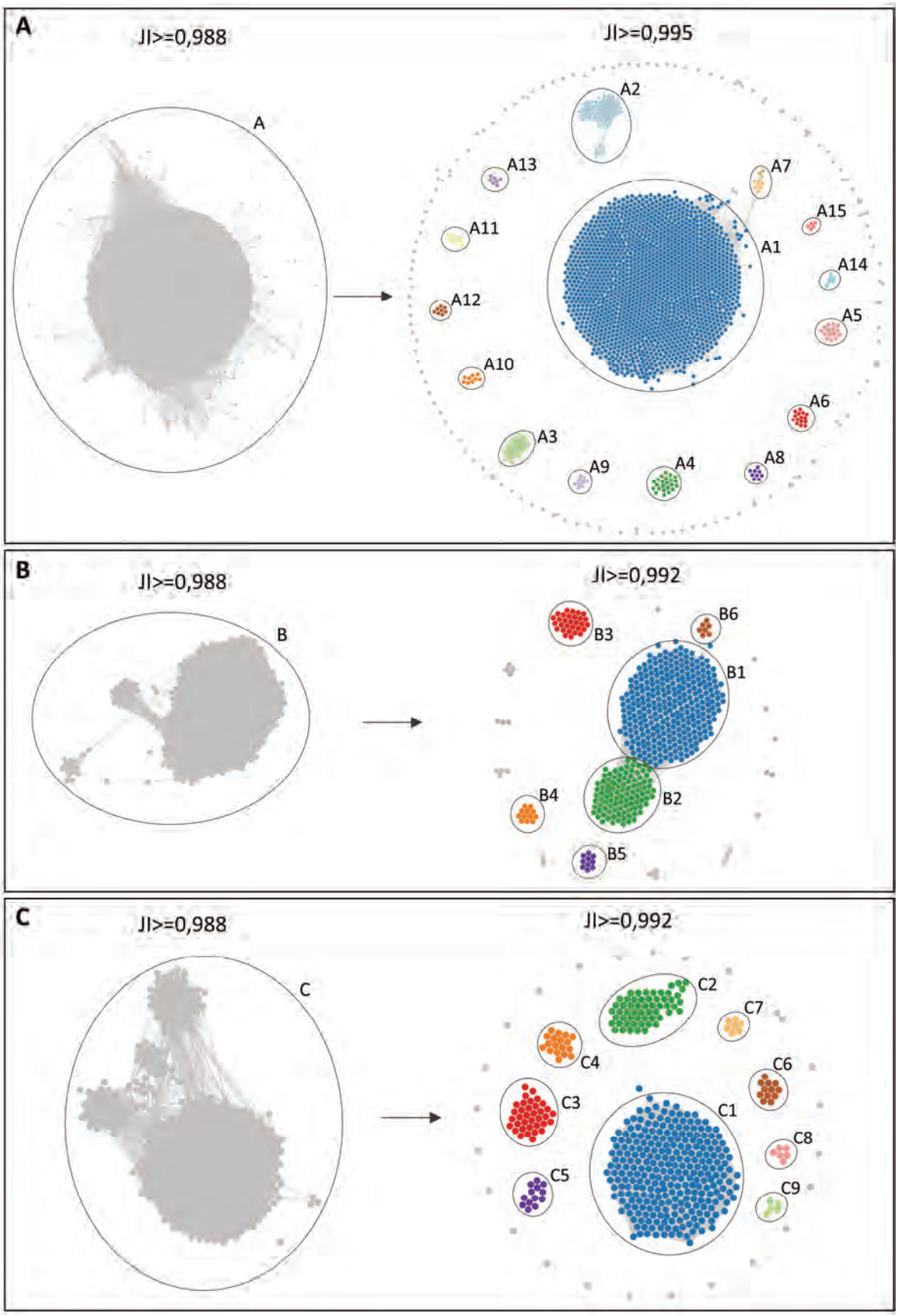


Figure 5.7 | Subclustering analysis of Hadar JI-groups A, B, and C. Each panel contains two networks filtered at different JI thresholds. In the left network, a lower JI threshold displays the entire

group. In the right network, a higher JI threshold reveals subgroups, which are highlighted with circles and assigned distinct colors. **(A)** Subclustering of 1,899 JI-A genomes. **(B)** Subclustering of 489 of JI-B genomes. **(C)** Subclustering of 453 JI-C genomes.

5.3 Pangenome structure of U.S. Hadar population

Analysis of the U.S. Hadar dataset revealed that the variations delineating each JI-group involve a mix of genomic features, including plasmids larger than 30 kb, prophages, AMR regions, or regions of unknown function (**Figure 5.8**). In some cases, two JI-groups differed only by the presence of a large plasmid (e.g., JI-A and JI-C; JI-B and JI-G; JI-D and JI-E), while others displayed more differences in their pangenome content (e.g., JI-I) (**Figure 5.8**).

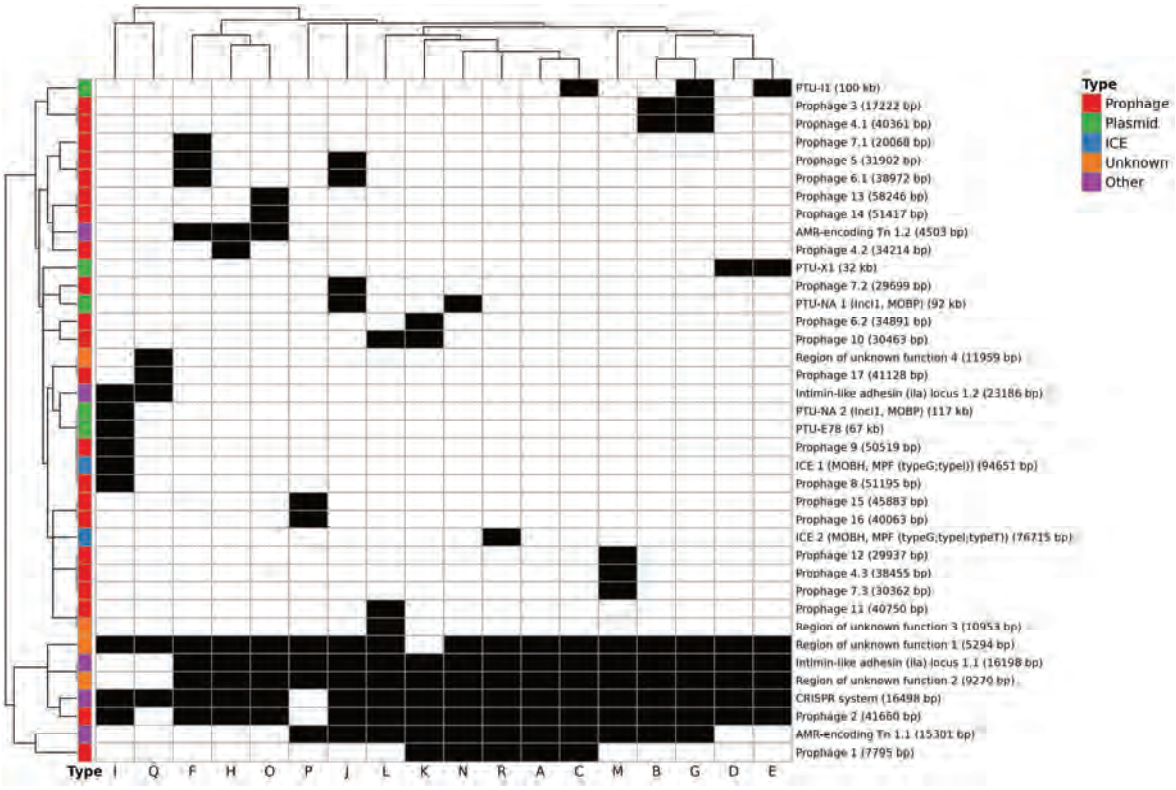


Figure 5.8 | Differential distribution of accessory genome elements in the Hadar JI-groups. The heatmap was built with the presence or absence of accessory elements in the JI-groups detected by PanGraph [77]. Accessory elements larger than 5kb and their derivatives were included in the analysis. Each column represents a JI-group. Each row corresponds to an element (with its size in bp indicated in parentheses after its name), whose presence in the corresponding JI-group is indicated in black while the absence in white. The left bar categorizes the accessory elements as "plasmid",

"prophages", "ICE", "other", or "unknown", as represented in the legend. A digit differentiates derivative accessory elements (elements highly similar) sharing otherwise the same name.

ST (sequence type, based on 7 core loci) and cgMLST allele code (based on n=3002 core loci) [39] were separately visualized on the network to contextualize the pangenome with core lineage information. Over 98% of Hadar genomes in this analysis were ST33 (n=3,326/3,384); only JI-I (ST473), JI-L (ST5130 and ST9222), and JI-Q (ST473) contained genomes of a different ST (**Figure 5.9A**). cgMLST allele codes aligned well with JI-groups, with the majority of groups (n=12/18) containing a single condensed allele code (**Figure 5.9B**). Phylogenetic analyses suggest membership within certain JI-groups is due to convergence in accessory genome content rather than core genome similarity.

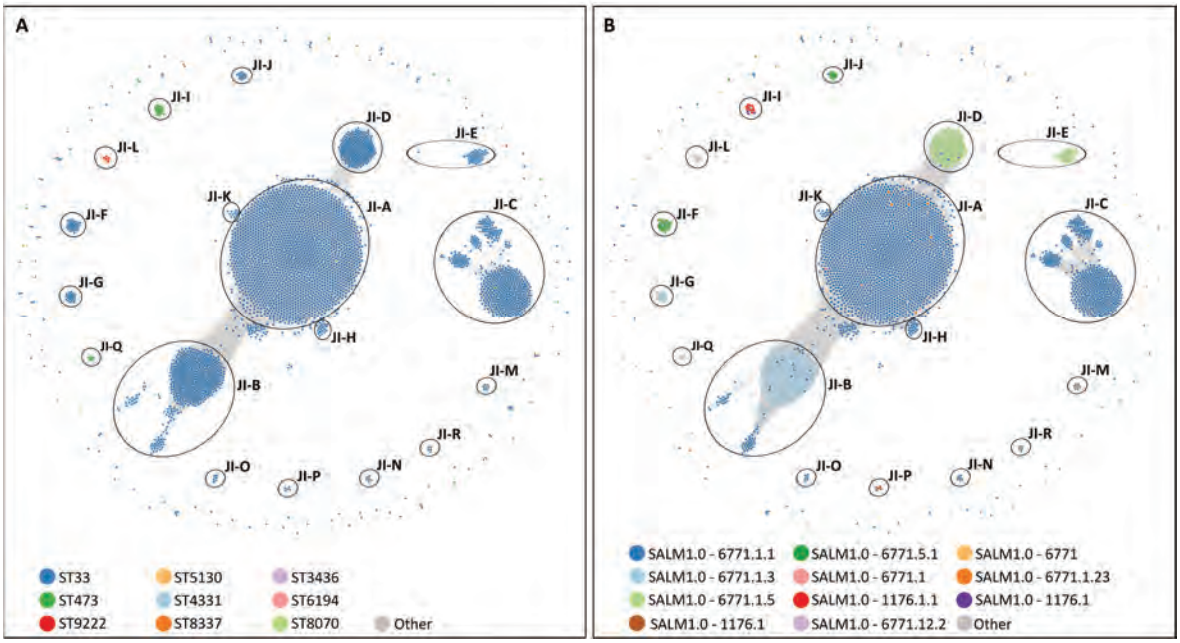


Figure 5.9 | Distribution of core lineage information in the Hadar JI-groups. The two networks contain 3,384 nodes, which are connected whenever JI >= 0.988. Nodes are colored according to **(A)** Sequence type; **(B)** Condensed allele code. Eighteen groups (named JI-A to JI-R) are indicated by circles.

Plasmids were common in U.S. Hadar genomes, with 60% (n=2,047/3,384) containing one or more Col-like plasmids and 25% (n=740/3,384) carrying at least one conjugative plasmid larger than 30 kb (**Figure 5.10**). IncI1 was the most common replicon found in large plasmids, detected in three different PTUs: PTU-I1, present in JI-C and JI-E;

a newly identified PTU-NA (IncI1, MOB_P) in JI-J and JI-N; and another newly identified PTU-NA (2) (IncI1, MOB_P) in JI-I (**Figure 5.10**). Although these three PTUs shared the same replicon and relaxase, they exhibited differences in their sequence that indicate they belong to different taxonomic units. JI-I also contained PTU-E78, a non-mobilizable PTU.

However, nearly 30% of genomes (n=1,011/3,384) contained neither plasmid replicons nor MOB relaxase genes (**Figures 5.10 and 5.11**); these genomes predominantly fell into JI-A. PanGraph analysis revealed that integrated MGEs were also common in several JI-groups, including prophages and ICEs (**Figure 5.8**).

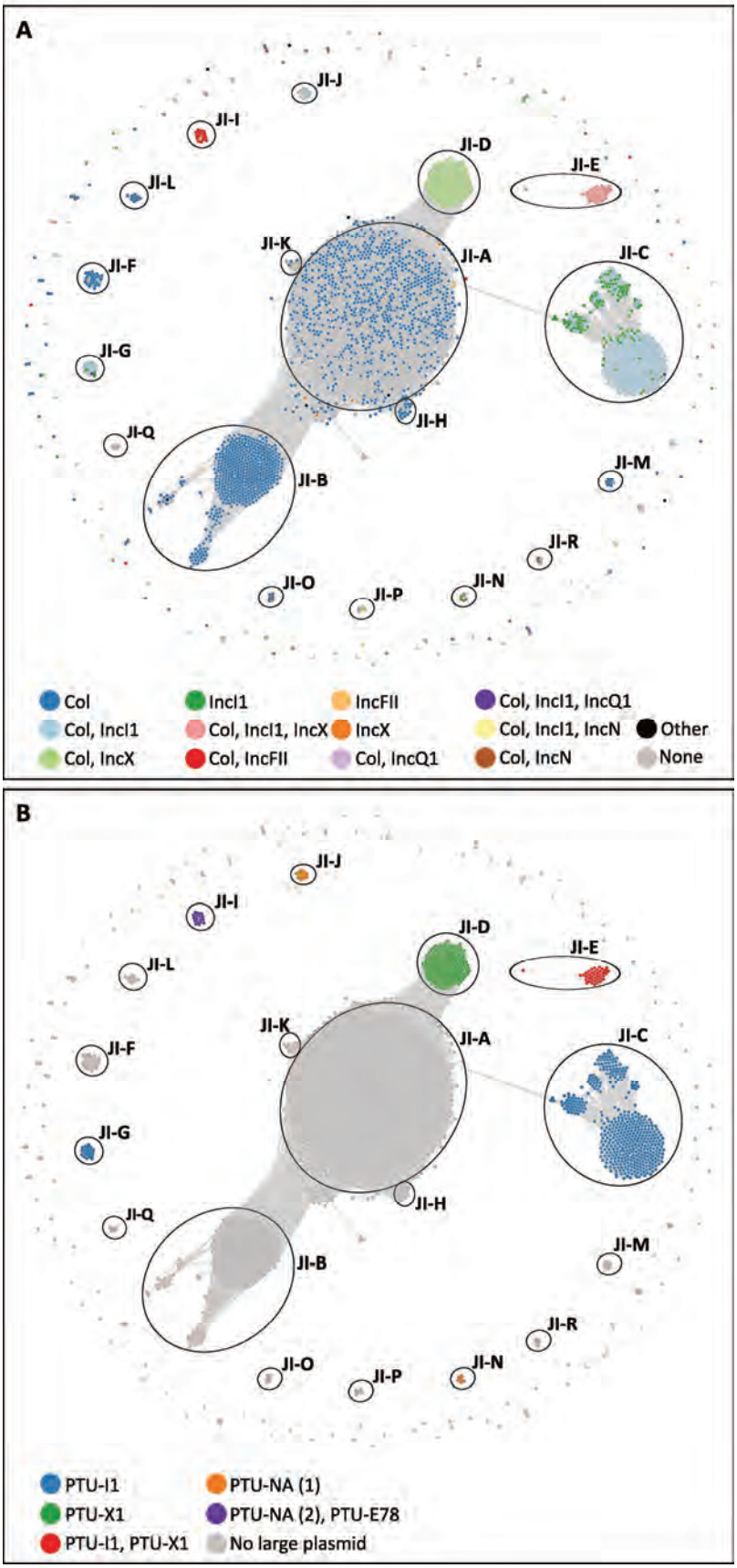


Figure 5.10 | Distribution of plasmids in the Hadar JI-groups. The networks contain 3,384 nodes, connected when JI≥0.988. Eighteen groups (named JI-A to JI-R) are indicated by circles. **(A)**

Distribution of the 12 most common plasmid replicons patterns across JI-groups. Nodes are colored according to the plasmid replicon pattern present in each genome. **(B)** Distribution of large (>30kb) plasmids (classified by PTUs) across JI-groups. Nodes are colored according to the PTUs present in each genome.

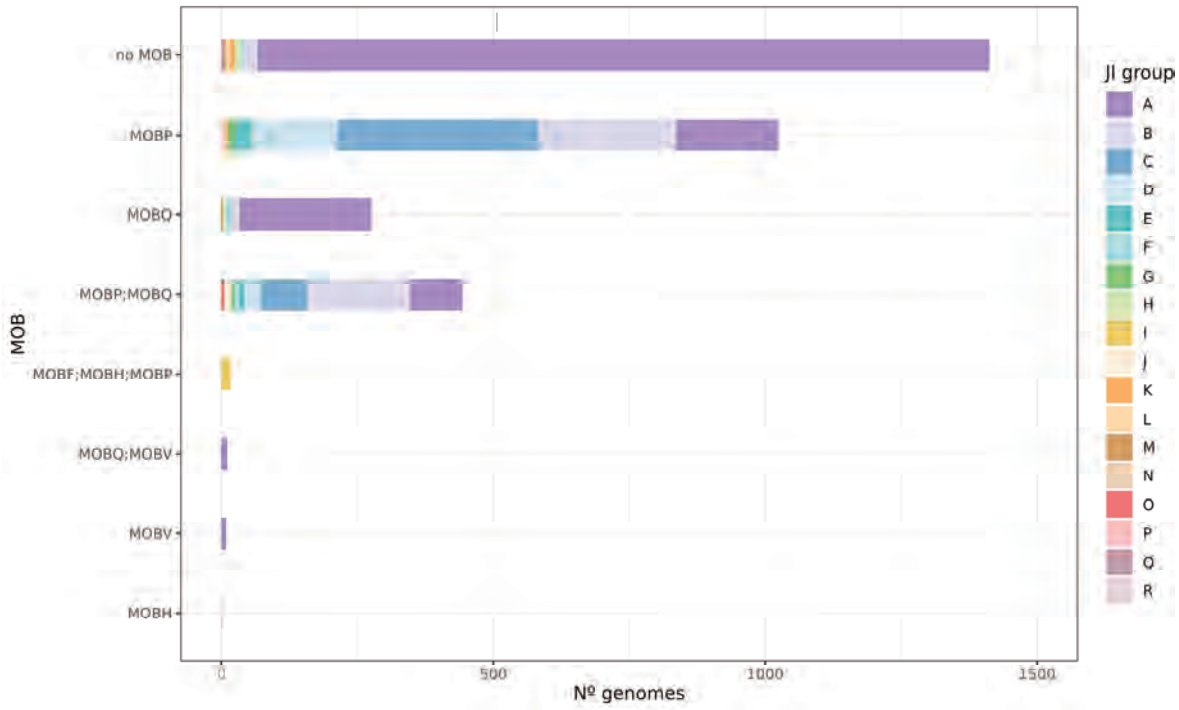


Figure 5.11 | Distribution of MOB classes across different Hadar JI-groups. The bar chart illustrates the number of genomes for each MOB class. Only the MOB patterns that are present in more than three genomes within a JI-group are represented. For simplicity, repeated occurrences of the same MOB within a genome (e.g., MOB_P;MOB_P) are counted only once (MOB_P).The color of each bar corresponds to a specific JI-group, with the color mapping provided in the legend.

Over 90% (n=3,055/3,384) of genomes contained at least one AMR determinant. Predicted resistance to aminoglycosides (specifically, streptomycin) and tetracyclines was the most common profile, mediated by *aph(3'')-Ib*, *aph(6)-Id*, and *tet(A)*), all integrated in the chromosome (**Figures 5.12A and 5.13**). Predicted resistance to penicillins was less common (4%, n=128/3,384) (**Figure 5.12B**) and was predominantly mediated by bla_{TEM-1} (**Figure 5.13**). While rare, cephalosporin resistance mediated by bla_{CMY-2} was detected in plasmids of groups JI-C and JI-E (0.4%, n=12/3,384; **Figure 5.12B**). Members of JI-D, JI-I, and JI-Q were predicted to be pansusceptible, with no known AMR determinants detected. **Figure 5.13** further detailed all the AMR genes present in all the JI-groups.

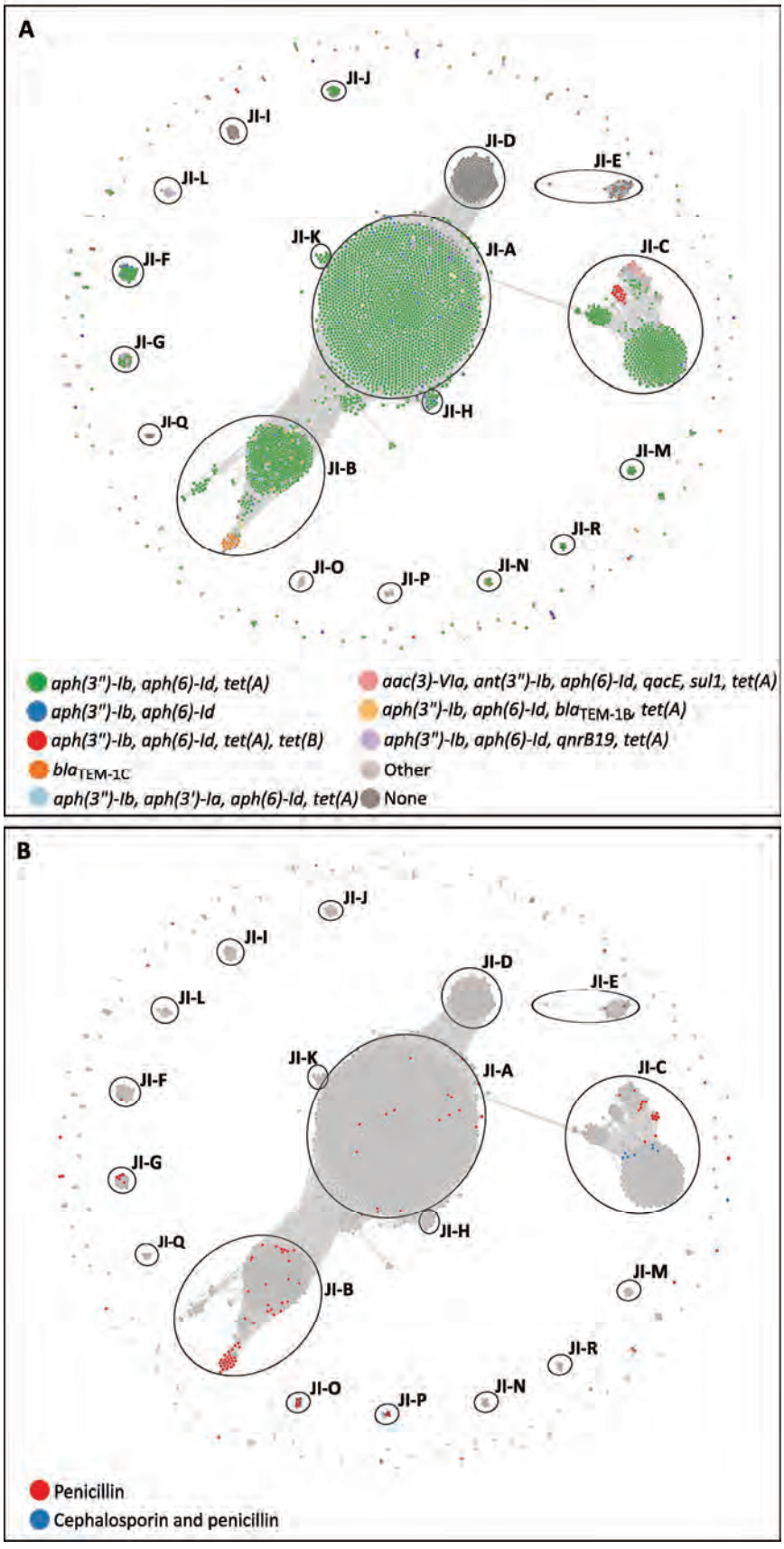


Figure 5.12 | Distribution of resistance determinants in the Hadar JI-groups. The networks contain 3,384 nodes, connected when JI≥0.988. Eighteen groups (named JI-A to JI-R) are indicated

by circles. **(A)** Distribution of the most common AMR genes patterns across JI-groups. Nodes are colored based on the AMR determinants. **(B)** Distribution of genomes with predicted resistance to penicillin and cephalosporin across JI-groups. Nodes are colored according to their predicted resistance to these antibiotics.

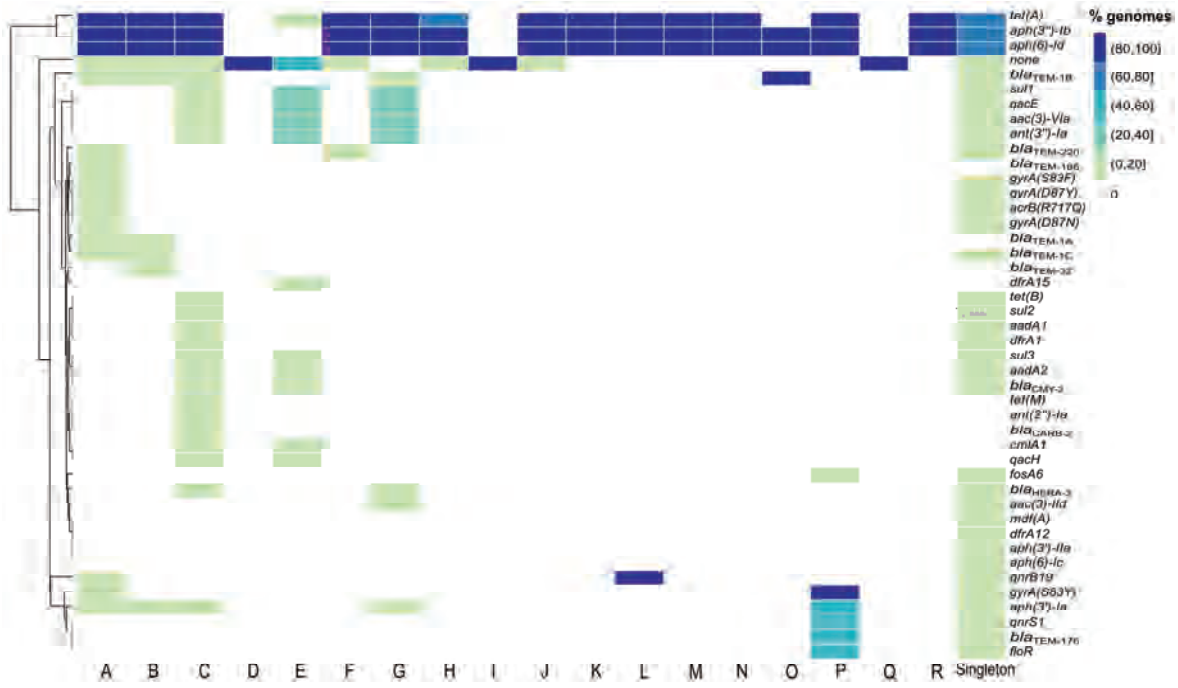


Figure 5.13 | Heatmap showing the distribution of antimicrobial resistance genes across JI-groups. The heatmap was constructed with the presence or absence of AMR genes in the JI-groups. Each column represents a JI-group. Each row corresponds to an AMR determinant. The color scale indicates the percentage of genomes within each group that harbor the specific AMR gene.

5.4 Genetic and epidemiological differences between most abundant pangenome groups

The dominant pangenome groups changed substantially between 2016 and 2023, most notably between 2019 and 2020 (**Figures 5.14** and **5.15**). This shift was particularly pronounced for human and retail meat samples, where JI-A and JI-C were rare prior to 2020 yet comprise between 56% and 100% of samples collected in years 2020-2023. JI-B was the most common group detected in retail meat and animal (cecal) sampling prior to 2020 but decreased in detection substantially in 2020-2023; JI-B was not detected at all in 2023 retail meat sampling (**Figure 5.14**). Groups JI-D and JI-E made up more than half of human Hadar

samples in 2016 and 2017 but have not been detected since 2019; these groups were not found in retail meat or animal sampling throughout the study years. JI-A and JI-C are the most common JI-groups in all three sampling systems from 2020-2023 (**Figure 5.14**).

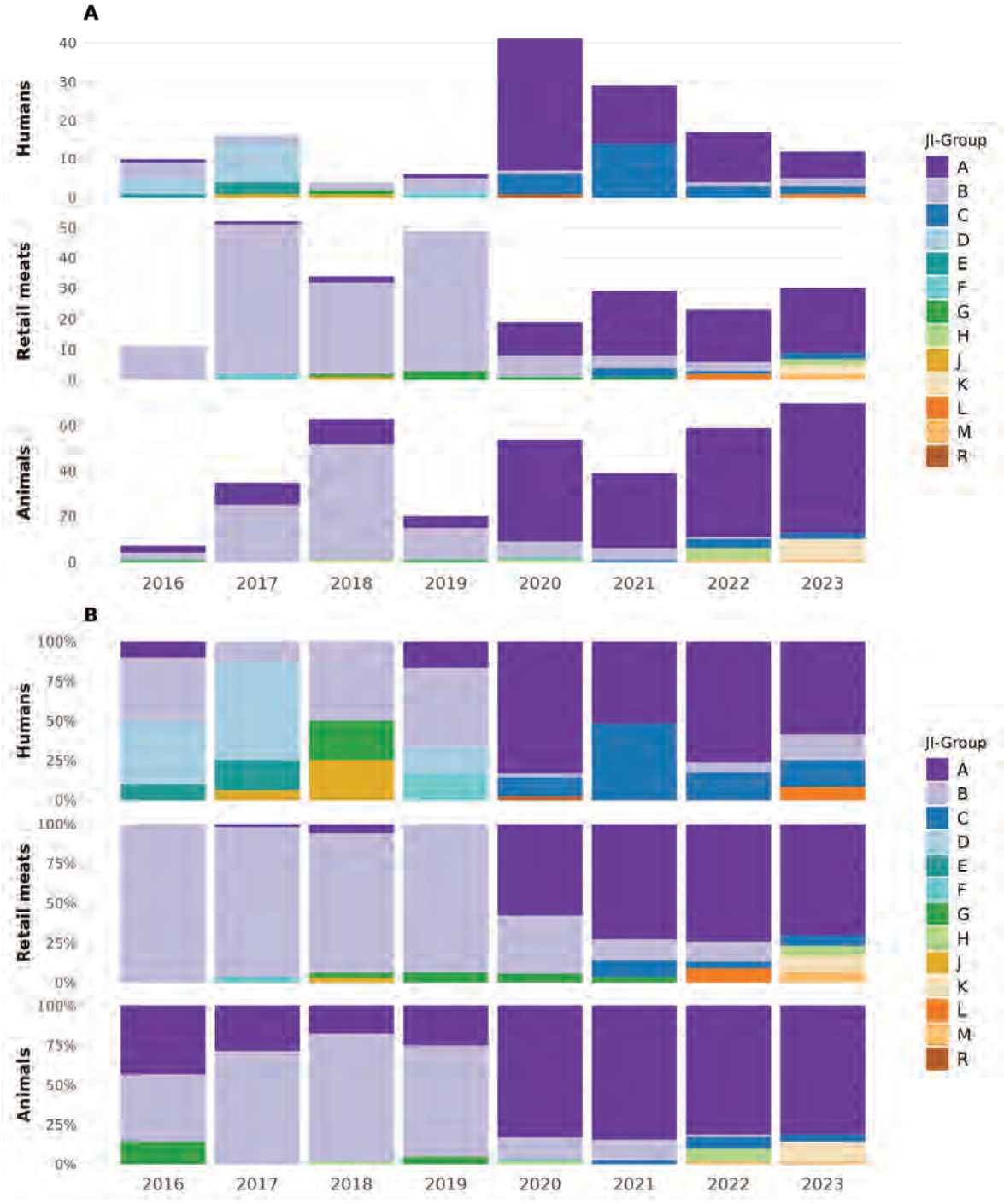


Figure 5.14 | Abundance of Hadar JI-groups over time. The bar charts display either the number of genomes **(A)** or the percentages of genomes relative to each year and source **(B)** detected in humans, retail meats, and animals from 2016 to 2023, categorized by JI-groups. Data are sourced

from NARMS, including CDC (humans), FDA (retail meats), and FSIS (animals). Counts and percentages are plotted on the y-axis, while years on the x-axis.

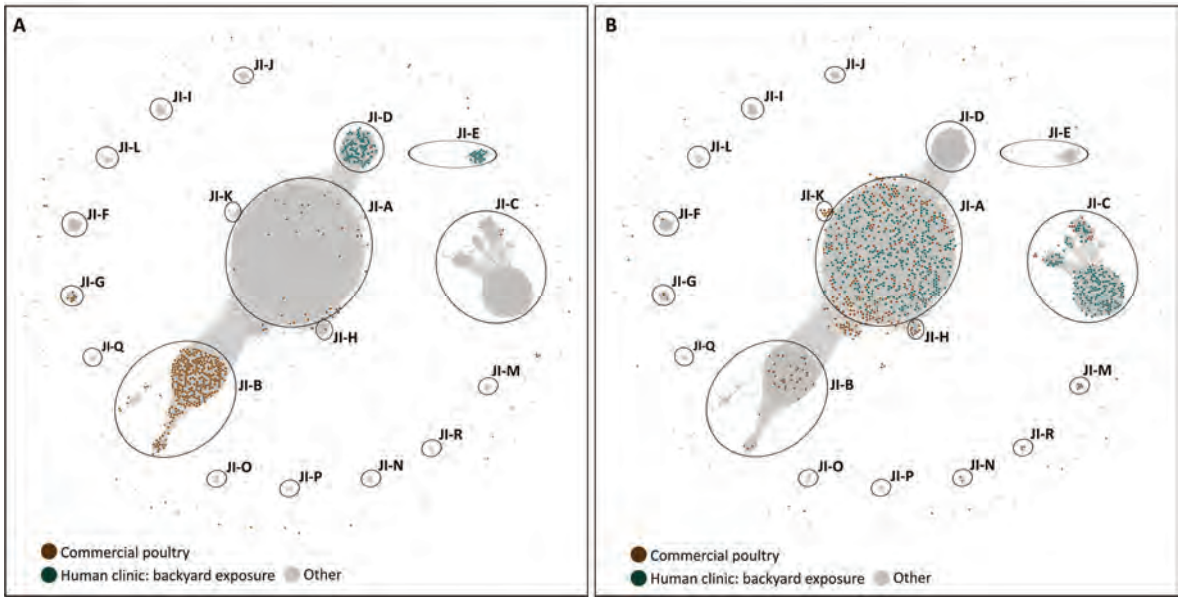


Figure 5.15 | Distribution of Hadar genomes by variables of interest. The networks contain 3,384 nodes, connected when JI≥0.988. Eighteen groups (named JI-A to JI-R) are indicated by circles. Only the genomes that were isolated either from commercial poultry or in the human clinic with reported backyard exposure are colored. **(A)** Distribution of the genomes isolated before 2020 in the JI-groups. **(B)** Distribution of the genomes isolated between 2020 and 2023 in the JI-groups.

JI-A and JI-C were indistinguishable by cgMLST-based phylogeny (**Figure 5.16**: Ring 1) but differed in their pangenome due to the presence of a ~100 kb PTU-I1 (IncI1) plasmid, which underpins the separation of these two JI-groups (**Figures 5.8** and **5.10**). Most JI-A and JI-C genomes fell within a comparatively tight "emergent" clade that forms the CDC-defined REPTDK01 strain (**Figures 5.16** and **5.17**), associated with ground turkey consumption and backyard poultry contact based on previous multistate outbreak investigations [251]. This emergent clade contained a ~8 kb prophage, labeled here prophage 1 (**Figure 5.16**), that forms part of the core pangenome of JI-A and JI-C (**Figure 5.8**). Prophage 1 was detected as early as 2004 in singleton Hadar genomes (imported "sweet good without custard or cream filling" from Pakistan) (**Table S5**), and it was seen in genomes from swine and commercial poultry samples from 2015 (**Table S5**) yet remained uncommon until the 2020 emergence of REPTDK01 (**Figures 5.16** and **5.18**).

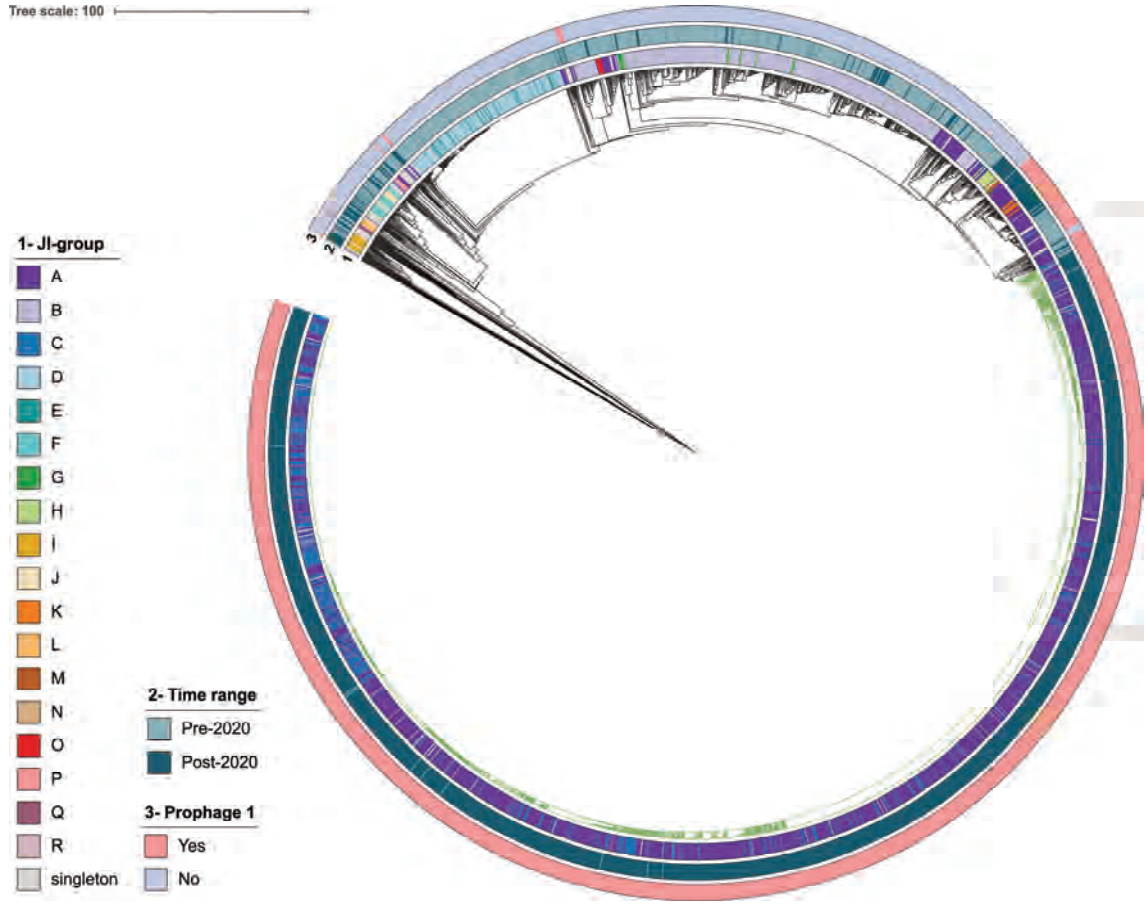


Figure 5.16 | cgMLST-based phylogenetic tree of Hadar genomes. Tree generated using Bionumerics v7.6.3 and visualized in iTol v6 [215]. Ring 1 displays JI-group, Ring 2 displays time range (1990-2019 versus 2020-2023), and Ring 3 displays presence of prophage 1 detected in this study. The large clade colored in green represents REPTDK01 strains.

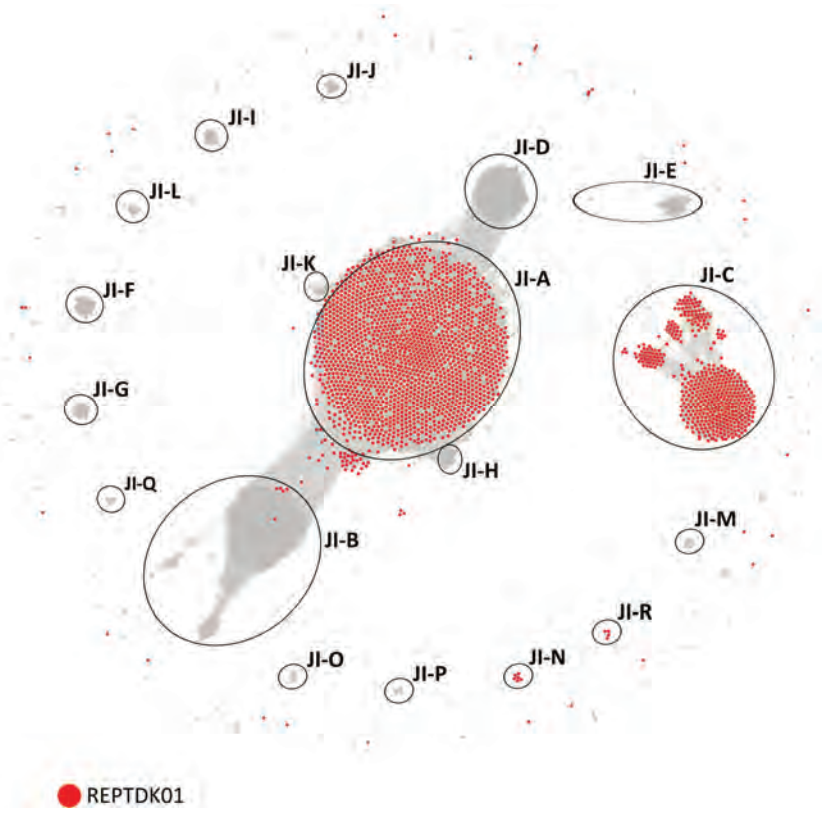


Figure 5.17 | Distribution of the REPTDK01 strains in the JI network. The network contains 3,384 nodes, connected when JI≥0.988. Eighteen groups (named JI-A to JI-R) are indicated by circles. Genomes that belong to REPTDK01 strain are colored in red.

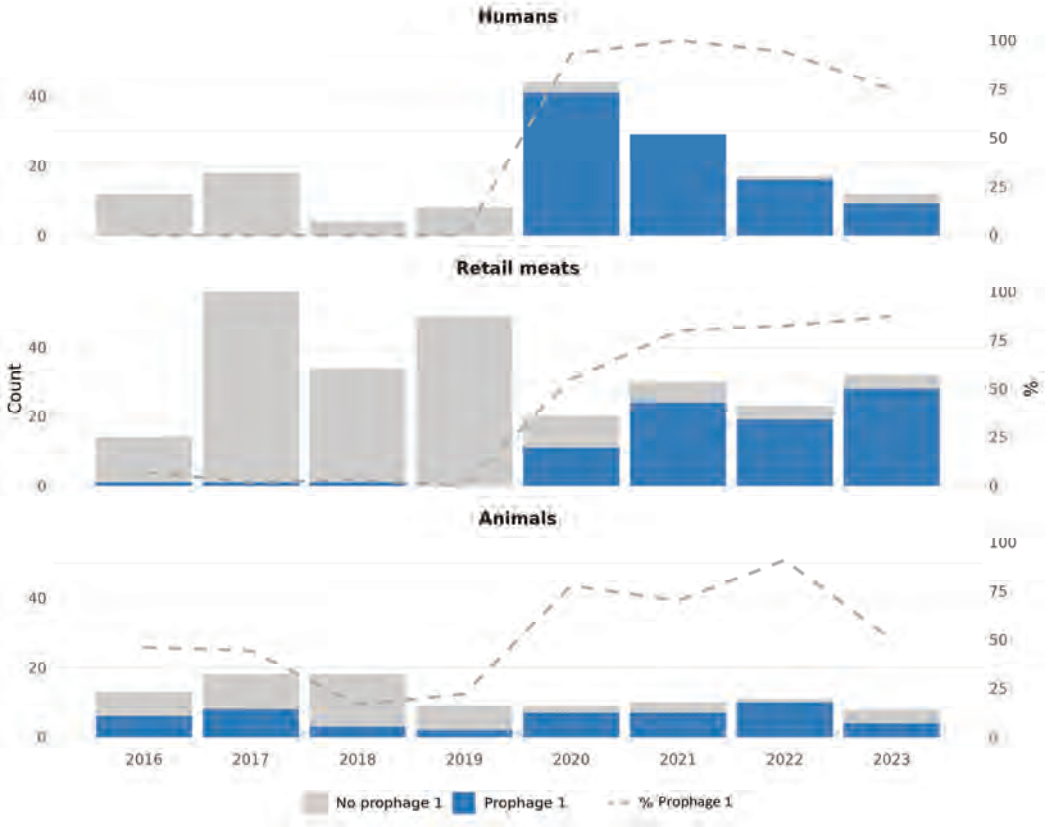


Figure 5.18 | Occurrence of prophage 1 in NARMS surveillance sequencing over time. The bar charts represent the number of genomes detected in humans, retail meats, and animals from 2016 to 2023, categorized by the presence (blue) or absence (grey) of prophage 1. Counts are displayed on the left y-axis, year is displayed on the x-axis. The dashed line indicates the percentage of genomes (right y-axis) carrying prophage 1 within each category over time.

According to PHASTEST [197], Prophage 1 is related to filamentous phages I2-2 and Ike, and contains a protein with N-terminal homology to the zonular occludens toxin protein (Zot) (**Figure 5.19**). The phage-encoded Zot proteins in *Vibrio cholerae* [252] and *Campylobacter* spp. [253,254] have demonstrated a pathogenic role, attributable to a C-terminal enterotoxic domain [255]. While homology with Zot proteins does not imply toxigenic function, the Hadar Zot-like protein identified here was bioinformatically predicted as an exotoxin using ToxinPred3.0 [256], hinting at a putative role in pathogenesis. Thus, prophage 1 presence is notable both from an epidemiological and biological perspective, and its pathogenic and adaptive capacity is being assessed with functional analysis.

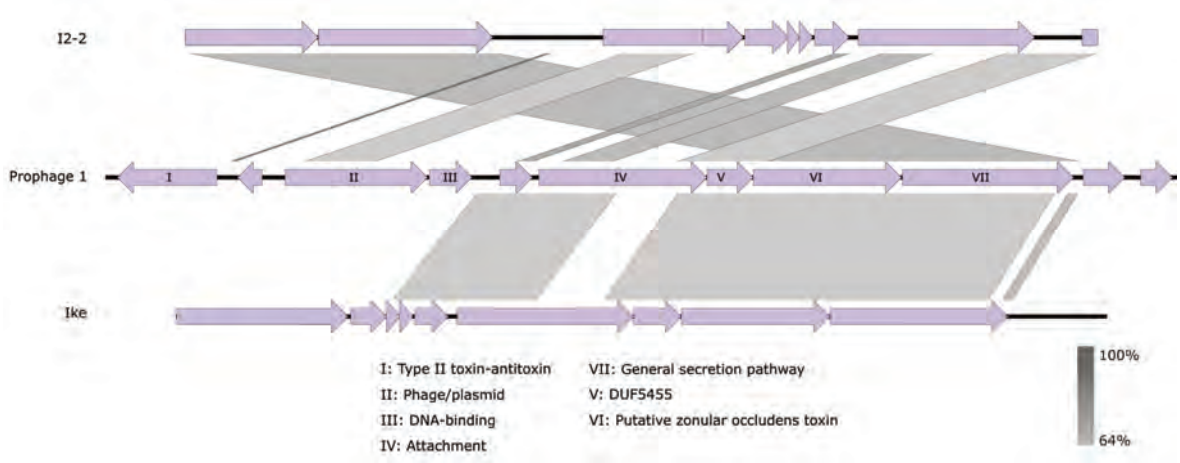


Figure 5.19 | Genomic alignment of prophages I2-2 (NC_001332.1), Prophage 1, and Ike (NC_002014.1). This figure was obtained using Easyfig. Purple arrows represent annotated genes, with arrowheads indicating the direction of transcription. Shaded gray regions between the prophages denote areas of sequence homology, with shading intensity reflecting the percentage of sequence identity (64%–100%, as indicated by the scale bar on the right). Gene annotations of prophage 1 was obtained with Bakta.

JI-B was the second most abundant pangenome group (**Table 5.2**), predominantly encompassing genomes from commercial poultry isolated before 2020 (**Figures 5.14** and **5.15**). A smaller group, JI-G, that also contained genomes from commercial poultry (12/20) isolated before 2020 (**Figures 5.14** and **5.15**), was indistinguishable from JI-B phylogenetically (**Figure 5.16**: Ring 1) but can be differentiated by the presence of PTU-I1 plasmids (**Figure 5.10**). JI-B genomes appeared more diverse in their core genome relative to those from other dominant pangenome groups (e.g., JI-A, JI-C, JI-D and JI-E) (**Figures 5.9B** and **5.16**), which may be a reflection of time and environmental factors (genomes in JI-B were isolated as early as 2011 from poultry sources across the country). Analysis of JI-B subgroups did not reveal any geographic association or link to specific processing facilities (data not shown). Of note, genomes from human samples that were part of a 2019 multistate Hadar outbreak linked to ground turkey consumption (internal CDC investigation) all fell into JI-B or JI-G, suggesting Hadar strains from these groups are transmitted via food.

In contrast, groups JI-D and JI-E were almost always from ill humans (rather than animal or meat samples), often with reported contact with backyard poultry and isolated before 2020 (**Figure 5.14**). JI-D and JI-E genomes displayed relatively little core diversity (**Figure 5.16**) and differed from each other only by the carriage of PTU-I1 (IncI1) plasmids (**Figure 5.10**). They differed from other JI-groups phylogenetically and pangenomically; phylogenetically, they were encompassed in a single clade by core SNP analysis (**Figure 5.16**) and they belonged to a different allele code from cgMLST (**Figure 5.9B**); pangenomically, they lacked a common AMR region ("AMR-encoding Tn 1.1", **Figures 5.8** and **5.12A**) and were the only groups to carry PTU-X1 (IncX1) plasmids (**Figure 5.10**). Genomes in these groups were part of 2016 and 2017 multistate outbreaks linked to contact with backyard poultry (https://archive.cdc.gov/#/details?url=https://www.cdc.gov/salmonella/live-poultry-05-16/index.html).

Two small pangenome groups, JI-H and JI-K, were of interest because of their connectivity to JI-A in the network, indicating pangenomic relatedness (**Figure 5.5**). JI-H genomes were all from commercial chicken sampling, or from ill humans (no exposure information available), representing a statistically significant "chicken-source cluster" ($p < 0.00001$, chi-squared) that is unique among the more common commercial turkey source (**Figure 5.15, Table S5**). JI-K genomes were all isolated throughout 2023, were almost exclusively from turkey product samples (n=11/12) (**Figure 5.15, Table S5**) and were

predominantly from a single state (n=8/12 were isolated in California). JI-K genomes carried prophage 1, along with two other larger prophages unique to this group (prophage 6.2 and prophage 10; **Figure 5.8**), potentially representing recent divergence from REPTDK01.

Several pangenome groups harbored PTU-I1 (IncI1) plasmids, including JI-C, JI-E and JI-G (**Figure 5.10**). PTU-I1 (IncI1) plasmids are common in avian environments, often carry AMR genes, and may play a role in virulence and growth inhibition of competing bacteria [257,258]; thus, their presence and diversity in this dataset were of interest. Core plasmid SNP analysis coupled with AcCNET [205] plasmid proteome analysis were used to assess the relatedness of PTU-I1 (IncI1) plasmids between and within JI-groups (**Figure 5.20**). In addition, this comparative analysis was also conducted with PTU-I1 plasmids present in other *Enterobacteriaceae* genus available in the RefSeq200 database (**Table S8**). PTU-I1 plasmids from all three JI-groups were diverse in their core and proteome and intermingled phylogenetically with PTU-I1 plasmids from other *Enterobacteriaceae* genus (**Figure 5.20**).

Proteomic analyses further emphasized the diversity of PTU-I1 plasmids. While certain HPCs were conserved across all PTU-I1 plasmids, subgroup-specific clusters suggest functional specialization (**Figure 5.20B**). For instance, JI-C1 plasmids form a subgroup characterized by their high similarity, sharing the same set of proteins while also carrying proteins exclusive to this subgroup (**Figure 5.20C**). Upon investigating these proteins, none stood out as particularly noteworthy. Furthermore, within this subgroup, a plasmid from *E. coli* was identified, indicating a high degree of similarity to the JI-C1 plasmids.

Plasmids from the same JI-C subgroups clustered together phylogenetically (**Figure 5.20A**) and proteomically (**Figure 5.20C**), indicating that plasmid content was responsible for JI-C subgrouping. A phylogenetic comparison of JI-C plasmids and their host chromosomes (**Figure 5.21**) revealed that JI-C chromosomes were highly conserved, with greater similarity to one another than the plasmid did to each other. Notably, the largest subgroup, JI-C1, formed a distinct clonal lineage with tightly related plasmids and chromosomal genomes, suggesting a multiyear clonal expansion. In contrast, other JI-C subgroups chromosomes were intermingled in the tree, suggesting that the subgroup distinctions are not chromosome-driven. JI-C1 plasmids were primarily associated with human clinical isolates of unknown exposure and commercial poultry. Other subgroups, such as JI-C2 and JI-C9, were more related to backyard poultry (**Figure 5.21**).

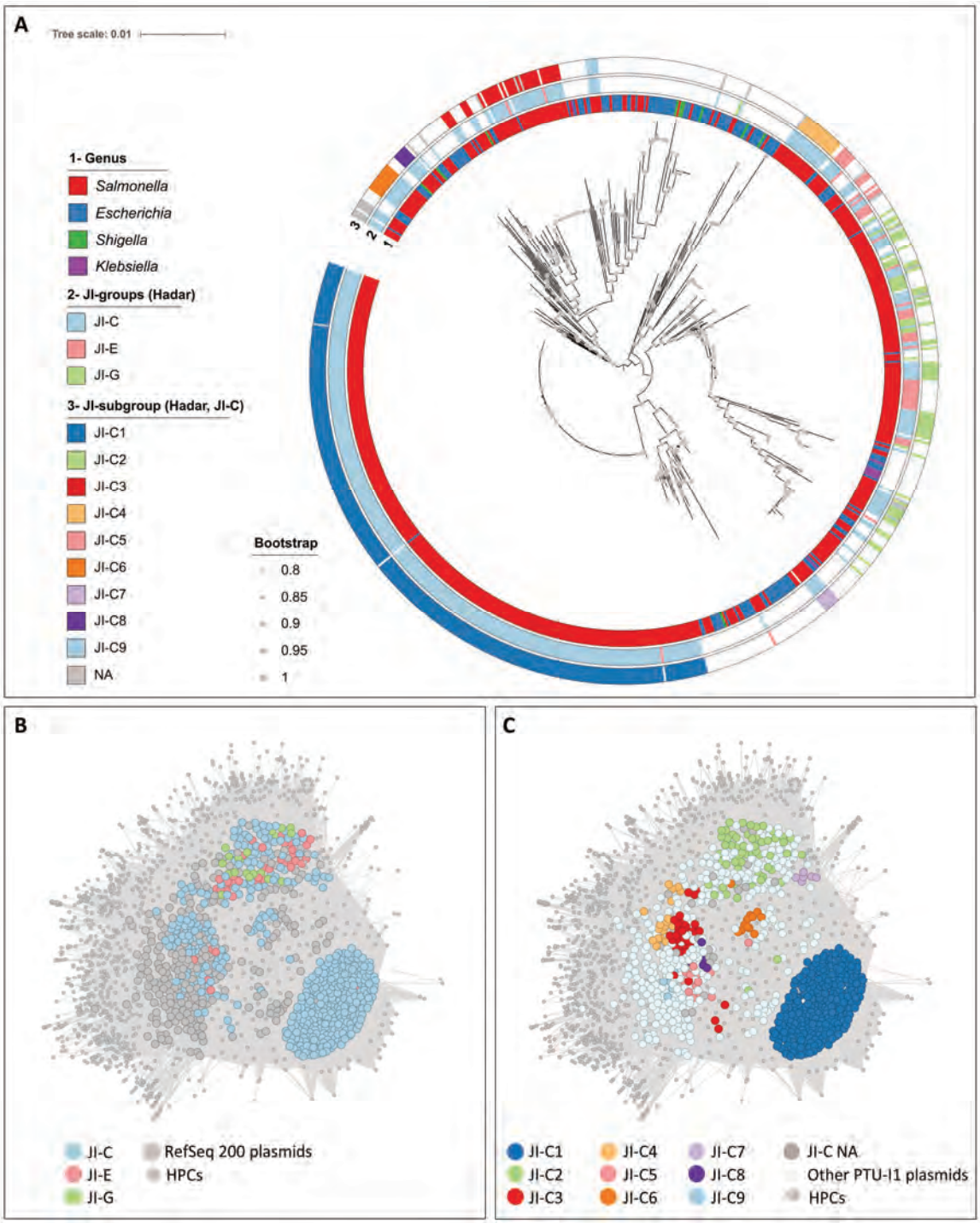


Figure 5.20 | Analysis of PTU-I1. **(A)** Core genome phylogenetic tree of PTU-I1 plasmids. A ML tree was constructed based on core SNPs obtained using Snippy from the core genome of 512 PTU-I1 plasmids from the Hadar dataset and 259 PTU-I1 plasmids from RefSeq200. The tree was built with IQ-TREE [219] using the suggested model, midpoint-rooted, and visualized in iTol v6 [215]. UFBootstrap values > 80% are indicated by circles on the corresponding nodes. Branch length scale represents substitutions per site. Colored rings indicate the plasmid host (1), the JI-group of Hadar plasmid hosts (2), and the JI-subgroup of the JI-C plasmids (3). **(B)** Proteome network of PTU-I1 plasmids colored by JI-group. The proteins of the PTU-I1 plasmids were clustered at 80% identity

and 80% coverage using AcCNET [205]. The larger nodes correspond to plasmids and are colored according to the JI-group of the Hadar plasmids (JI-C, JI-E, JI-G), or in grey if they belong to other plasmids from RefSeq200. The smaller nodes represent HPCs and are colored in grey. Both kinds of nodes are connected if a plasmid contains a member in the corresponding protein cluster. HPCs present a single plasmid were removed. **(C)** Proteome network of PTU-I1 plasmids colored by JI-C-subgroup. The network was constructed as indicated in (B). The JI-C Hadar plasmids are colored based on their JI-subgroup, and the other plasmids have the same color (light blue).

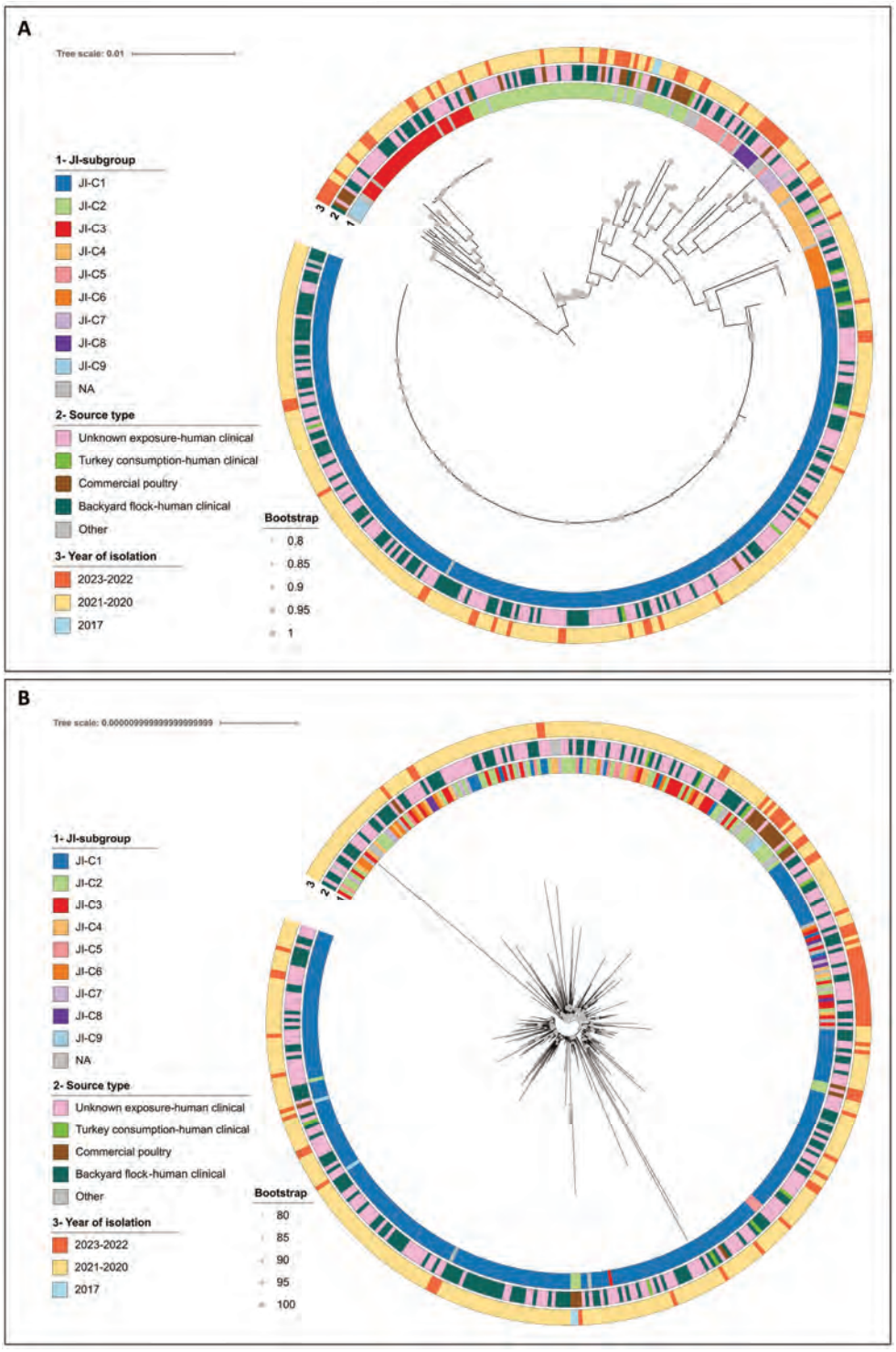


Figure 5.21 | Phylogenetic analysis of Hadar JI-group C isolates. The ML trees were generated from core-genome SNP alignments obtained using Snippy for either 453 JI-C plasmids (PTU-I1) **(A)** or JI-C chromosomes **(B).** The trees were built using IQ-TREE [219] with the substitution models TVM+F+I+R4 (A) or TIMe+ASC (B) and visualized in iTol v6 [215]. UFBootstrap values > 80% are indicated by circles on the corresponding nodes. Colored rings indicate the JI-subgroup (1), the source type (2), and the time range (3).

5.5 Hadar pangenome offers increased discriminatory power for retrospective and prospective public health investigations

REPTDK01 was clearly detectable in the pangenome network, 98% (n=2,148/2,194) of these genomes fell into JI-A, JI-C, JI-N and JI-R (**Figure 5.17**), genetically corroborating and adding confidence to the REPTDK01 definition using pangenomic data. Additionally, REPTDK01 was further stratified by JI-grouping and JI-subgrouping, revealing clear epidemiological patterns. For example, while JI-A itself was not statistically associated with either commercial or backyard poultry (**Figure 5.15, Table 5.3**), JI-A2 was predominantly commercial poultry related genomes from the U.S. and Canada (n=42/68) and none of the human clinical cases in this group (n=24/68) reported backyard poultry contact (**Figure 5.22**). In contrast, JI-A3 was almost exclusively human clinical samples (n=27/28), a third of which reported backyard poultry contact, and zero commercial poultry related genomes fell into this group (**Figure 5.22**). JI-N genomes were all human clinical, mostly isolated from the northeast (n=4/6) (**Tables 5.4 and S5**), may represent a closely related subcluster of illnesses that differ from JI-A REPTDK01 strains only by the carriage of a large plasmid (PTU-NA, IncI1) (**Figure 5.8**).

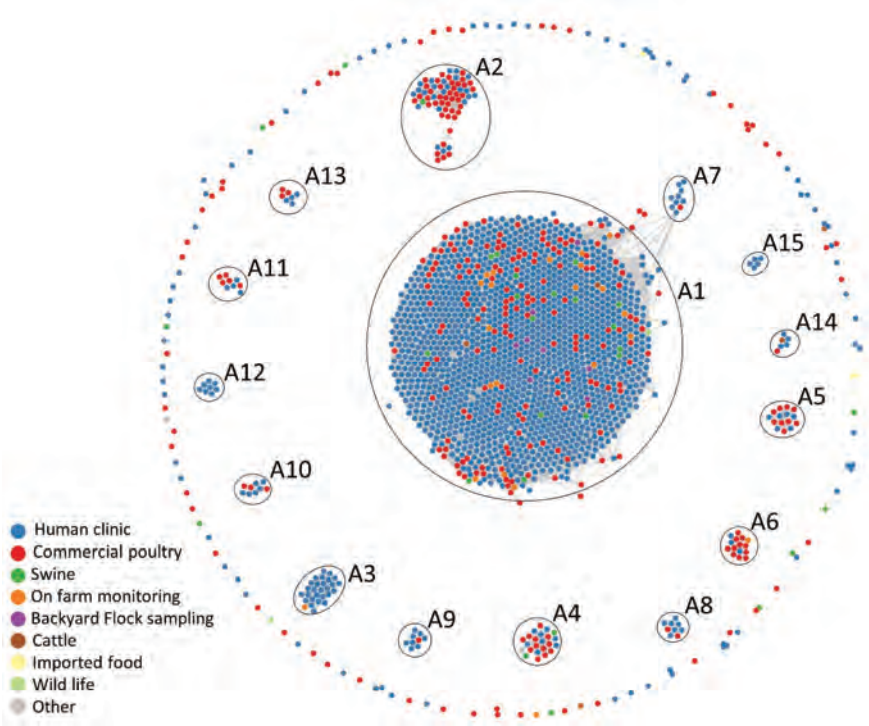


Figure 5.22 | Subclustering analysis of Hadar JI-group A. The network contains 1,899 nodes of JI-group A connected when JI≥=0.995. Subgroups A1 to A15 defined by the Louvain method are surrounded by circles, singleton genomes that do not associate with a JI-subgroup are displayed around the outside of the network. Nodes are colored by source.

JI-C was significantly associated with backyard poultry ($p < 0.00001$, chi-squared), representing a subgroup of REPTDK01 (defined by the carriage of PTU-I1 plasmids) that was likely transmitted to humans via animal contact rather than food. More specifically, epidemiological traceback data available for clonal subgroup JI-C1 disproved the involvement of a single backyard poultry supply store chain or hatchery, instead suggesting a common reservoir of Hadar upstream of hatcheries. Coupling pangenome data and epidemiological data, REPTDK01 strains can be further differentiated for both retrospective and prospective investigations.

Several other non-REPTDK01 pangenome groups were statistically associated with a specific source or exposure. JI-B and JI-G were each significantly associated with commercial turkey ($p < 0.00001$, chi-squared); JI-B genomes had 17.5 times (95% CI: 13.7-22.3) and JI-G genomes had 5.9 times (95% CI: 2.1-17.1) higher odds of being from commercial turkey compared with all other JI-groups (**Table 5.3**). Coupled with the absence of human cases reporting backyard poultry contact in these groups, it is likely that Hadar

strains from JI-B and JI-G, isolated from human cases with unknown exposure, were acquired through foodborne transmission.

In contrast, JI-D and JI-E were each significantly associated with backyard poultry contact ($p < 0.00001$, chi-squared); JI-D genomes had 2.6 times (95% CI: 1.9-3.6) and JI-E genomes had 5.2 times (95% CI: 2.7-10.6) greater odds of backyard poultry contact, relative to all other JI-groups (**Table 5.3**). The stark lack of genomes from commercial poultry sources (only JI-D had a single commercial chicken source genome), and the predominance of backyard poultry-associated outbreak genomes in these groups (n=140/191 in JI-D, n=35/40 in JI-E), strongly suggests JI-D and JI-E strains of Hadar were transmitted through animal contact. It is important to note that cgMLST differentiates JI-B and JI-G genomes from JI-D and JI-E (**Figure 5.16**), thus, the pangenome analysis performed here provides additional genomic confidence in these attributions.

Table 5.3: Statistical analysis between JI-groups and backyard poultry and/or turkey genomes in Hadar. *This table was prepared by Kaitlin Tagg, whose contribution is gratefully acknowledged.*

JI-group[a]	Source							Total[d]
	Backyard poultry related			Commercial turkey			Other sources	
	Nº genomes	OR[b]	CI (lower-upper)[c]	Nº genomes	OR[b]	CI (lower-upper)[c]		
JI-A	431	1,1	0,9-1,3	219	0,4	0,3-0,5	1188	1838
JI-B	-	-	-	285	17,5	13,7-22,3	142	427
JI-C	184	2,8	2,3-3,5	-	-	-	265	449
JI-D	74	2,6	1,9-3,6	-	-	-	104	178
JI-E	24	5,2	2,7-10,6	-	-	-	16	40
JI-G	-	-	-	10	5,9	2,1-17,1	8	18

[a] | Statistical analysis was conducted only for JI-groups with at least 20 genomes and those related to backyard poultry or turkey. Consequently, JI groups JI-I to JI-R were excluded for having fewer than 20 genomes, while JI-G and JI-H were excluded because they lack genomes associated with these sources, as explained in *Materials and Methods*.

[b] | OR indicates Odds ratios for associations between the JI-group and the corresponding source.

[c] | CI indicates 95% confidence intervals for associations between the JI-group and the corresponding source.

[d] | Total genomes in each group that meet the requirement for statistical analysis as explained in Materials and Methods.

A handful of small JI-groups contained genomes from humans with limited epidemiological information, but with one or two genomes from a known source (**Tables 5.4 and S5**). Specifically, both JI-F and JI-J contained a genome from raw dog food

(containing duck), obtained from *ad hoc* pet food sampling. JI-L contained two genomes from imported shrimp (Ecuador) isolated in 2022 (**Table 5.4, Table S5**). Given the close relatedness of genomes within JI-groups (ANI ≥ 99.9%, **Figure 5.6**), the presence of the pet food and imported food genomes alongside genomes from human samples is suggestive of an epidemiological connection, though without exposure information reported by these ill people this link cannot be confirmed. Prospectively, the relatedness of additional human cases found within the JI-F and JI-J groups could inform which food items to assess during supplementary interviews of ill people included in an outbreak investigation.

Although this study primarily focused on the dominant groups, the surveillance of minor groups remains important, as they may carry advantageous traits or signal emerging trends (**Table 5.4**). Capturing the full spectrum of genomic diversity contributes to a more complete picture of evolutionary dynamics and potential sources of infection.

Table 5.4: Distribution of smaller JI groups by isolation source and year.

JI-group[a]	Source[b]	Year[c]
F (29)	Commercial poultry (chicken)	2020 (1)
	Human clinical (unkwnown exposure)	2015 (1), 2017 (3), 2018 (6), 2019 (5), 2020 (3), 2021 (2), 2022 (4), 2023 (2)
	Animal feed	2017 (2)
G (20)	Commercial poultry (turkey)	2014 (2), 2016 (2), 2018 (1), 2019 (4), 2020 (1), 2021 (1), 2023 (1)
	Human clinical (unkwnown exposure)	2018 (3), 2019 (2), 2020 (1), 2021 (2)
H (20)	Commercial poultry (chicken)	2018 (1), 2020 (1), 2021 (1), 2022 (5), 2023 (3)
	Human clinical (unkwnown exposure)	2017 (1), 2018 (1), 2023 (5)
	On farm monitoring (chicken)	2018 (1), 2020 (1)
I (17)	Human clinical (unkwnown exposure)	2016 (3), 2017 (1), 2018 (1), 2019 (2), 2020 (2), 2021 (1), 2022 (2), 2023 (5)
J (13)	Animal feed	2018 (1)
	Human clinical (unkwnown exposure)	2016 (1), 2017 (2), 2018 (3), 2020 (1), 2022 (3), 2023 (2)
K (12)	Commercial poultry (chicken)	2023 (1)
	Commercial poultry (turkey)	2023 (11)
L (9)	Human clinical (unkwnown exposure)	2019 (1), 2020 (3), 2023 (3)
	Imported food	2022 (2)
M (7)	Commercial poultry (chicken)	2022 (1)
	Commercial poultry (turkey)	2023 (4)
	Human clinical (unkwnown exposure)	2022 (2)
N (6)	Human clinical (backyard poultry contact)	2020 (2)
	Human clinical (unkwnown exposure)	2020 (4)
O (6)	Swine	2011 (6)
P (5)	Human clinical (unkwnown exposure)	2018 (1), 2019 (2), 2021 (1), 2022 (1)
Q (5)	Human clinical (unkwnown exposure)	2017 (1), 2019 (2), 2020 (1), 2022 (1)
R (5)	Commercial poultry (chicken)	2021 (1)
	Human clinical (backyard poultry contact)	2020 (1), 2022 (1)
	Human clinical (unkwnown exposure)	2020 (1), 2022 (1)

[a] | JI-groups containing fewer than 30 genomes. The number in parentheses indicates the genome count within each JI group.

[b] | Source of isolation.

[c] |Year of isolation. The number in parentheses indicates the genome count for each year.

As mentioned above, several pairs of JI-groups differed only by the presence of PTU-I1 (IncI1) plasmids: JI-A and JI-C, JI-D and JI-E, JI-B and JI-G. We further assessed these pairs for epidemiological patterns associated with plasmid presence, including source of isolation, geographic region, and patient demographics (age, sex, site of infection, hospitalization), but no variables were significantly different between paired groups ($V < 0.3$, corrected Cramer's V; $p > 0.005$, chi-squared). However, PTU-I1 (IncI1) plasmids were independently associated with backyard poultry-related sources (PTU-I1 n=208, no PTU-I1 n=526) when compared with commercial poultry sources (PTU-I1 n=32, no PTU-I1 n=699), and when compared with all other sources (PTU-I1 n=305, no PTU-I1 n=2,345) ($p < 0.00001$, chi-squared). Thus, PTU-I1 plasmids have statistical support to serve as a genetic marker to distinguish strains transmitted via backyard poultry contact versus those more likely attributed to another source, which is of particular value for differentiating REPTDK01 strains that can be transmitted via several pathways.

5.6 U.S. Hadar pangenome structure reflects a subset of global diversity

A dataset of Hadar genomes (n=1,145) from 33 countries other than the U.S., isolated from 1950 through 2023, was used to assess differences in pangenome structure between separate geographical locations (**Figure 5.23**). The non-U.S. dataset partially overlapped with U.S. genomes: 33% of non-U.S. genomes clusterd within JI-groups identified in the U.S. pangenome data, while 47% formed distinct JI-groups not present in the U.S. dataset and the remaining genomes were singletons (**Figure 5.24, Table 5.5**).

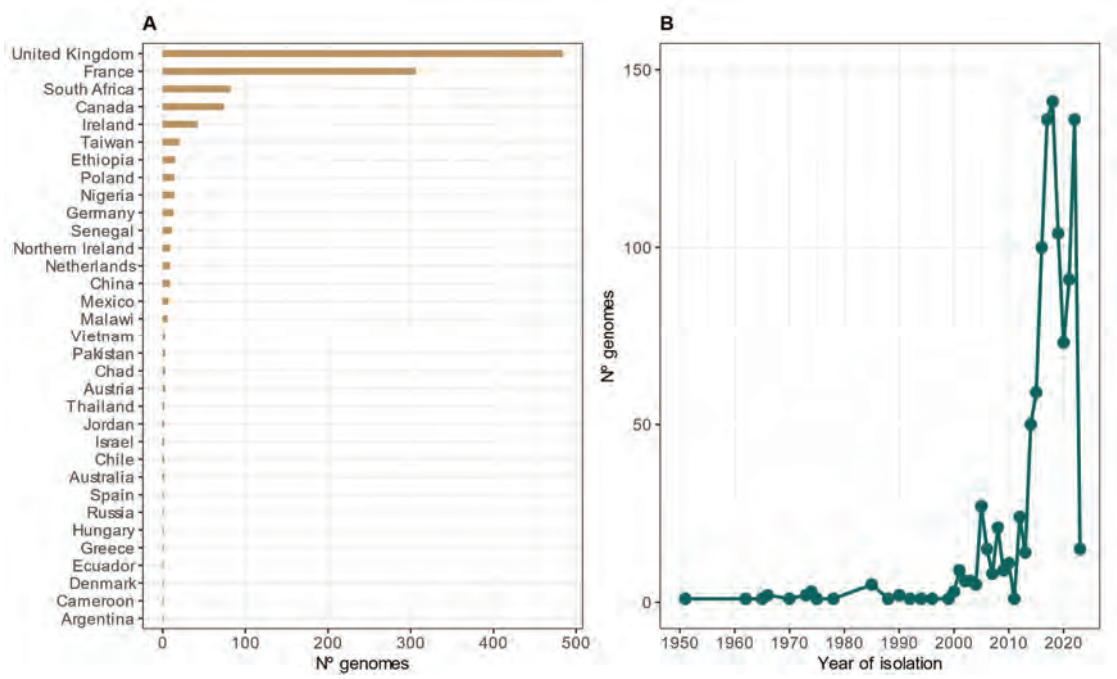


Figure 5.23 | Metadata overview of the non-U.S. Hadar dataset. 1,145 Hadar genomes present in EnteroBase comprise the non-U.S. dataset. **(A)** Number of genomes per country. **(B)** Number of genomes per year of isolation.

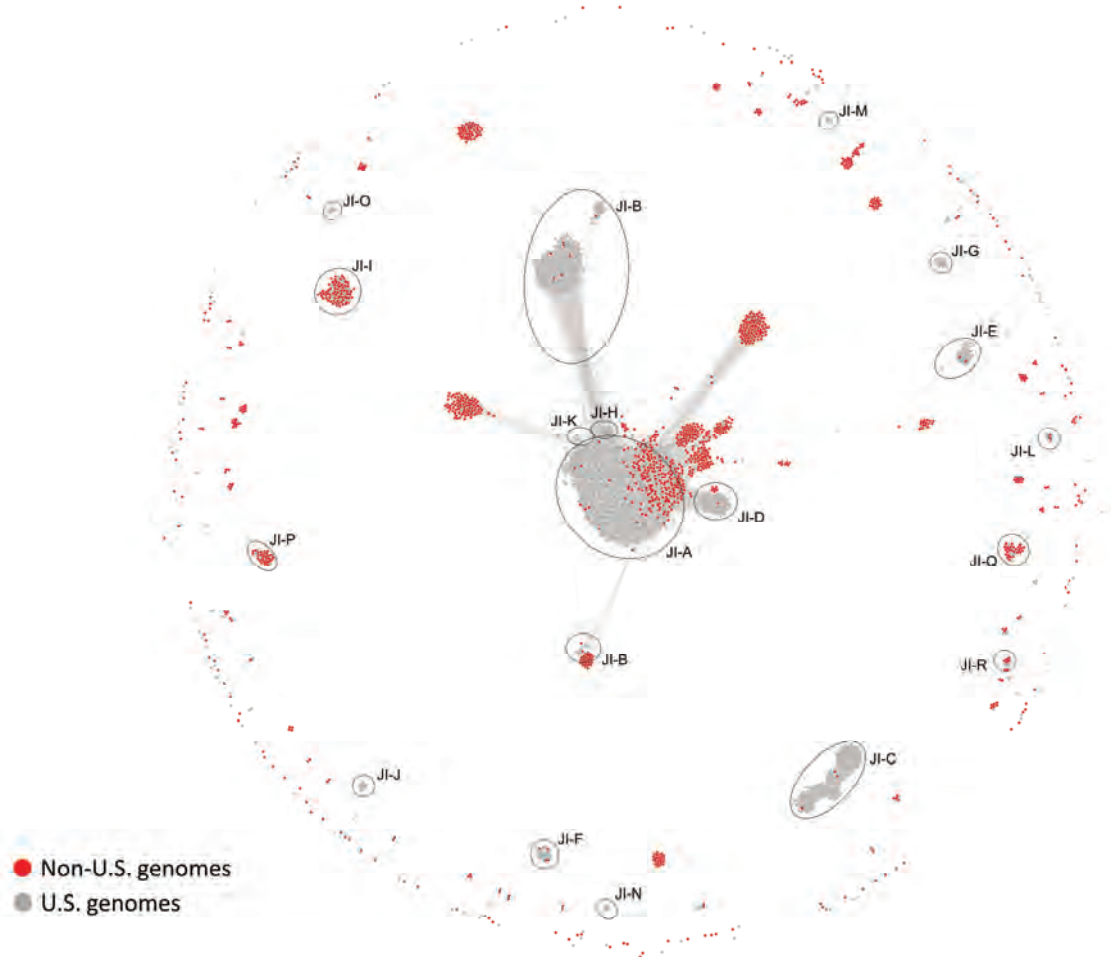


Figure 5.24 | Distribution of U.S. and non-U.S. Hadar genomes by JI. The JI network contains 1,145 non-U.S. Hadar genomes from EnteroBase and a reduced U.S. genomes dataset (n=1,516), using JI>=0.988 as a threshold. U.S. genomes are represented by grey nodes and circled when belonging to a JI-group, while non-U.S. genomes are represented by red nodes. To select the representatives U.S. genomes, the complete U.S. Hadar dataset was first clustered at JI>=0.99916, a threshold in which only practically identical genomes were connected. A greedy set algorithm was implemented to reduce the dataset. The connectivity degree (number of connections) of each node was calculated to select the most connected genome as a representative. All nodes connected to the representative were removed. This process was repeated until the network exclusively contains unconnected representative nodes.

Table 5.5: Distribution of Hadar genomes from U.S. and non-U.S. dataset across JI-groups.

JI-groups[a]	Nº U.S. genomes[b]	% U.S. genomes[c]	Nº non-U.S. genomes[d]	% non-U.S. genomes[e]
A + K	1911	56,42	208	18,17
B	489	14,44	36	3,14
C	453	13,37	2	0,17
D	191	5,64	1	0,09
E	40	1,18	1	0,09
F	29	0,86	4	0,35
G	20	0,59	0	0
H	20	0,59	0	0
I	17	0,5	70	6,11
J	13	0,38	0	0
L	9	0,27	3	0,26
M	7	0,21	0	0
N	6	0,18	0	0
O	6	0,18	0	0
P	5	0,15	28	2,45
Q	5	0,15	26	2,27
R	5	0,15	5	0,44
S	0	0	90	7,86
T	0	0	46	4,02
U	0	0	27	2,36
V	0	0	22	1,92
W	0	0	19	1,66
X	0	0	17	1,48
Y	0	0	10	0,87
Z	0	0	8	0,7
AA	0	0	8	0,7
AB	0	0	8	0,7
AC	0	0	7	0,61
AD	0	0	7	0,61
AE	0	0	7	0,61
AF	0	0	6	0,52
AG	0	0	5	0,44
AH	6	0,18	96	8,38
AI	4	0,12	81	7,07
AJ	1	0,03	36	3,14
AK	2	0,06	16	1,4
AL	1	0,03	8	0,7
AM	2	0,06	9	0,79
AN	1	0,03	8	0,7
Singletons	141	4,16	220	19,21

[a] | JI-groups defined in the U.S. dataset (JI-A to JI-R) and JI-groups defined in the non-U.S. dataset (JI-S to JI-AN). JI-A and JI-K are fused because when adding the non-U.S. dataset, Louvain grouped both groups in a single cluster. JI-groups JI-AH to JI-AN include genomes from the U.S. dataset; however, they were not previously defined as JI-groups because they did not meet the minimum threshold of at least five genomes. JI-AH contains six genomes from the U.S. dataset, but this group was not defined when analyzing only the U.S. genomes. This is because these six genomes did not

initially cluster together. However, when additional non-U.S. genomes were included, they formed a connected cluster with other genomes, leading to the definition of JI-AH in the combined dataset.

[b] | Number of genomes from the U.S. dataset present in each JI-group. JI-groups with 0 genomes are highlighted in blue.

[c] | Percentage of genomes from the U.S. dataset present in each JI-group. JI-groups with 0 genomes are highlighted in blue.

[d] | Number of genomes from the non-U.S. dataset present in each JI-group. JI-groups with 0 genomes are highlighted in blue.

[e] | Percentage of genomes from the non-U.S. dataset present in each JI-group. JI-groups with 0 genomes are highlighted in blue.

Both datasets were further compared using Roary [67]. This analysis showed differences and overlaps between the U.S. and non-U.S. Hadar pangenomes (**Figure 5.25**). Both datasets shared a robust core genome comprising 4,187 genes (**Figure 5.25A**). Notably, separate analysis of each dataset revealed similar core gene counts, further highlighting the robustness of the core genome across different geographic populations. However, the breakdown of core, shell, and cloud genes highlighted significant geographic variation. Genes present in over 80% of genomes were classified as core, those present in 15-79% as shell, and those in less than 15% as cloud. The non-U.S. dataset contained a larger number of cloud genes (8,391) compared to the U.S. dataset (6,791), suggesting higher accessory genome diversity among global isolates (**Figure 5.25A**). This is particularly notable given that the non-U.S. dataset includes only about one-third as many genomes as the U.S. dataset.

To compare the diversity of each dataset at equivalent sample sizes, pangenome accumulation curves (**Figure 5.25B**) were generated to show number of genes across increasing numbers of genomes. When comparing 1,000 genomes, the pangenome of non-U.S. dataset surpassed 9,000 genes, while the pangenome of the U.S. dataset remained below this number. Both datasets exhibited an open pangenomes (Heaps' law γ value ~ 0.2) and due to this fact, it is expected that if we add more genomes to the non-U.S. dataset, the pangenome will be much larger than the U.S. dataset. Additionally, the Venn diagram (**Figure 5.25C**) illustrated the unique genetic contributions of each dataset. The non-U.S. dataset contained 3,095 genes entirely absent from the U.S. pangenome, whereas the U.S. dataset had 1,628 unique genes not found in the non-U.S. collection. This distribution suggests that the U.S. Hadar population represents a subset of the global genetic diversity, with certain accessory genes potentially linked to region-specific adaptations.

The gene frequency distribution provided an alternative cumulative overview of core, shell, and cloud genes (**Figure 5.25D**). This panel reaffirms that the majority of accessory genes fell into the cloud category, particularly in the non-U.S. dataset, indicating a large pool of genes that were sporadically present and likely linked to environmental or host-specific adaptations. The gradual slope of the curve for non-core genes reinforces the impression of a highly variable accessory genome in non-U.S. isolates.

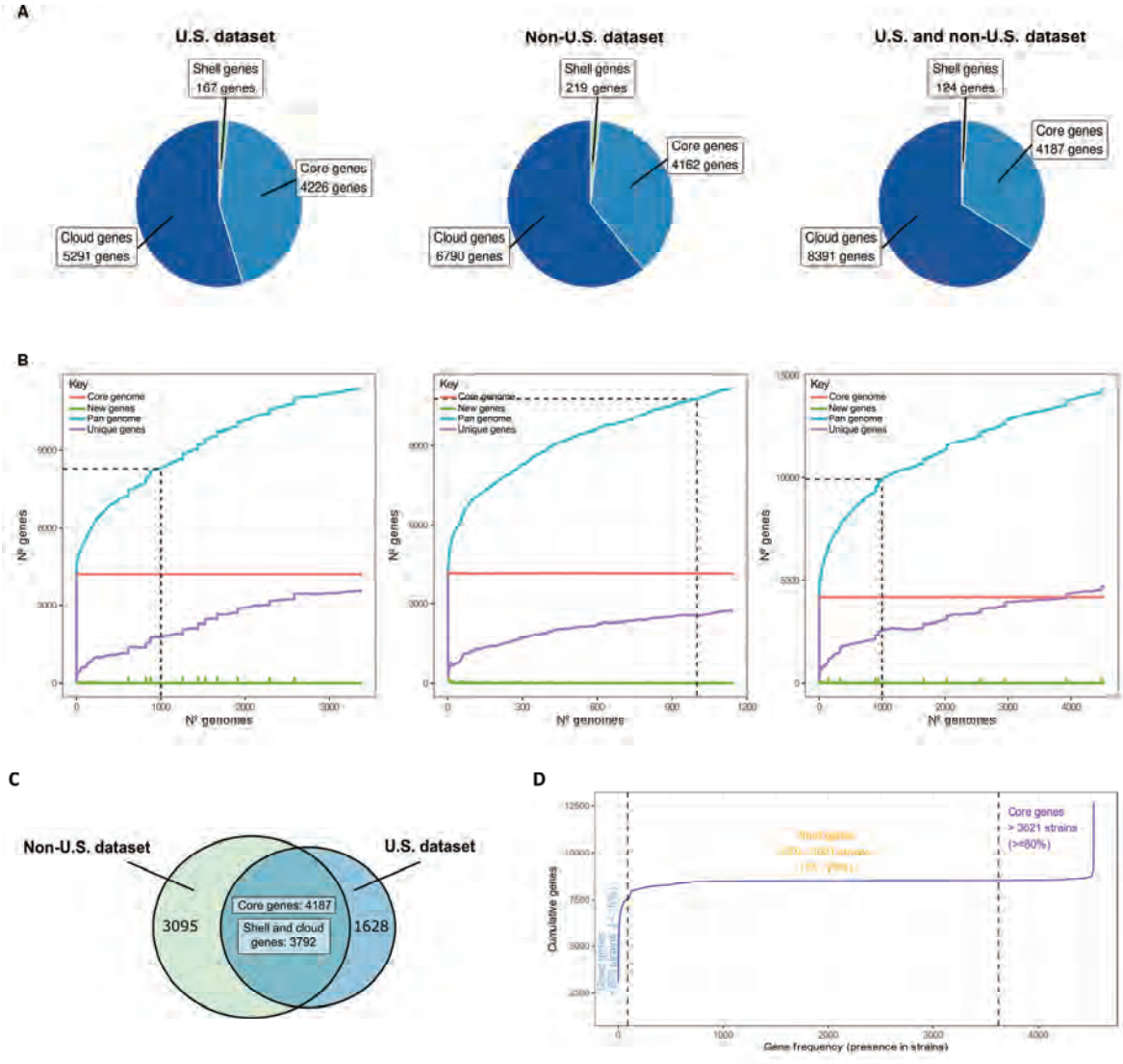


Figure 5.25 | Comparative analysis of pangenome distribution across U.S. and non-U.S. Hadar datasets using Roary. (A) Pie charts show the distribution of core, shell, and cloud genes within each dataset. Core genes, defined as those present in at least 80% of genomes, are shown in dark blue; shell genes, present in 15-79% of genomes, are shown in medium blue, while cloud genes, found in a maximum of 15% of genomes, are shown in light blue. The number of genes in each category is indicated on each chart. **(B)** Accumulation curves for core, new (previously unseen), pan, and unique (observed only once) genes across the increasing number of genomes analyzed. The x-

axis represents the number of genomes, while the y-axis represents the cumulative number of genes in each category. **(C)** Venn diagram comparing the U.S. and non-U.S. dataset. **(D)** Cumulative gene frequency distribution function. The curve represents the accumulated abundance of genomes in which a gene is present (gene frequency). Discontinuous lines divide the cumulative gene frequency curve into the three gene categories: cloud, shell, and core.

Separate analyses of genomes from France (n=306) and the United Kingdom (U.K.) (n=484) were performed since they represented more than half of the non-U.S. genomes. Of 18 JI-groups defined in the U.S. dataset, the U.K. and France datasets shared only seven (170 genomes, 35%) and six (74 genomes, 24%) JI-groups, respectively. On the other hand, nine France JI-groups (162 genomes, 53%) (**Figure 5.26**) and seventeen U.K. JI-groups (228 genomes, 47%) were distinct from those isolated in the U.S (**Figure 5.27**).

While no temporal shift was observed for pangenome groups from U.K. data, a notable increase in genomes belonging to one of the novel groups in France, JI-S, beginning in 2019 was observed (**Figure 5.26**). JI-S genomes contained a prophage closely related to prophage 1, highlighting an intriguing parallel dynamic to the recent proliferation of prophage 1-containing groups JI-A and JI-C in the U.S. Thus, these analyses suggest Hadar pangenomic diversity is largely geographically defined, with potentially important genetic overlaps that will be further investigated.

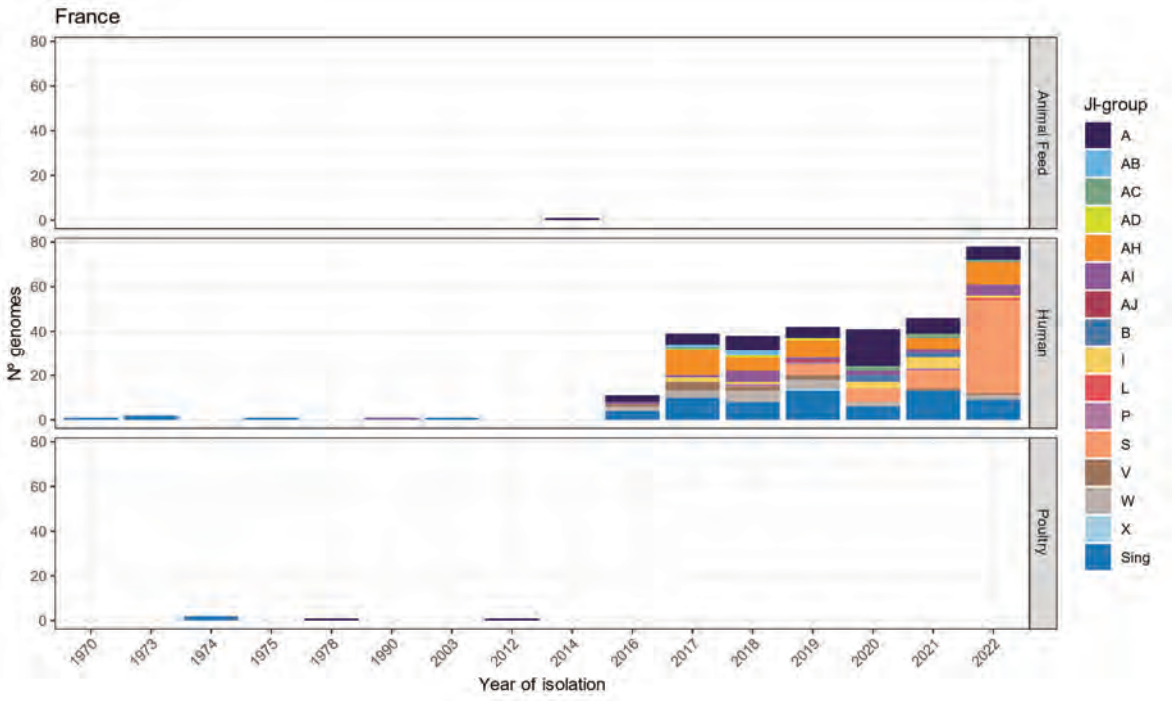


Figure 5.26 | Distribution of France Hadar dataset by source and year of isolation. The bar plot shows the number of genomes (y-axis) isolated each year (x-axis) from different sources: environment, food, human, poultry, and wild animal. Each bar is color-coded according to the JI-group classification, as indicated in the legend on the right. JI-groups A-R were defined in the U.S. dataset, the remaining groups are exclusively found in the non-U.S. dataset.

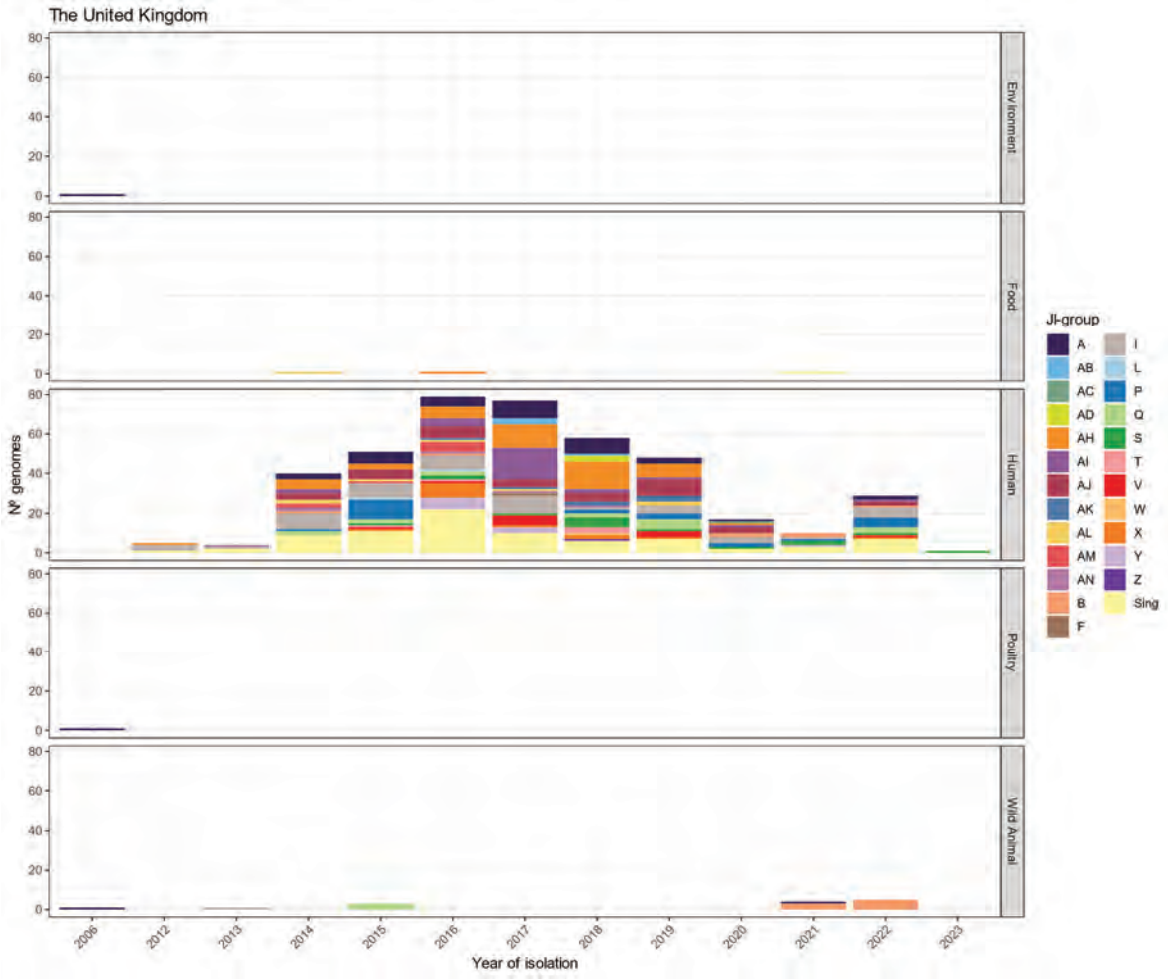


Figure 5.27 | Distribution of U.K. Hadar dataset by source and year of isolation. The bar plot shows the number of genomes (y-axis) isolated each year (x-axis) from different sources: environment, food, human, poultry, and wild animal. Each bar is color-coded according to the JI-group classification, as indicated in the legend on the right. JI-groups A-R were defined in the U.S. dataset, the remaining groups are exclusively found in the non-U.S. dataset.

CHAPTER 6: DISCUSSION

This thesis investigated two serovars of *Salmonella* at the population genomic level using a pangenome approach. Recognizing that bacterial evolution occurs in both vertical and horizontal dimensions, inclusion of both core and accessory genetic material is a logical step toward understanding pathogen dynamics, not to mention a more holistic use of increasingly available molecular datasets.

Current genomic surveillance methods predominantly rely on core genome or gene-by-gene strategies, which often overlook the extensive genetic diversity inherent in bacterial populations. In contrast, the primary objective of this work was to apply a comparative genomic framework that fully leverages all information contained within whole-genome assemblies. This strategy aimed to uncover subtle population structures and evolutionary dynamics in serovars of *Salmonella,* insights that would remain hidden when relying solely on traditional gene-by-gene or core genome strategies.

This study specifically examined short-term population shifts and the role of the accessory genome in generating genetic diversity within *Salmonella*. Given that outbreak dynamics and the epidemic success of pathogens are typically driven by rapid genetic changes, often mediated by MGE acquisition, the accessory genome becomes a critical focus. Consequently, incorporating the accessory genome into public health analyses can provide unparalleled epidemiological resolution.

Before this project began, a new computational tool, PopPUNK [65], was published whose functionality initially matched our needs for analyzing *Salmonella* serovars. This tool calculates two distinct *k*-mer-based distances, one for the core genome and another for the accessory genome, and it is designed to identify natural clusters within bacterial populations by applying spatial clustering algorithms. For some species, predefined bacterial models exist, providing established optimal clustering thresholds based on previous population studies. In cases where no predefined bacterial model exists, as is the case for the *Salmonella* serovars analyzed here, PopPUNK requires manual optimization of clustering thresholds following the steps and recommendations described in the paper.

When evaluating the quality of the clustering, the user must interpret network parameters such as density, transitivity score, and average entropy and make several decisions to obtain the most suitable output. The PopPUNK developers generally recommend evaluating clustering performance by visually inspecting the output and comparing the assigned clusters to a core phylogenetic tree. This approach uses

differences in the core genome as a guide to determine whether the clusters accurately reflect biological relationships and achieve the desired resolution, thereby informing the selection of an appropriate threshold for a given project.

Although PopPUNK offered an attractive framework, and, in theory, should allow clustering at any resolution, it presented several limitations for the objectives of this project, particularly in defining robust clusters within our dataset, which exhibited low genetic variability. Our datasets did not display a clear bimodal separation between within-strain and between-strain comparisons. As a result, the distance distribution formed a diffuse "cloud," which made it difficult for the clustering algorithm to converge on a definitive solution.

This issue is illustrated in **Figure 6.1**, which shows the distance distributions for *Salmonella enterica* serovars Typhi and Hadar (analyzed in this study) alongside examples from the original PopPUNK publication, including *Salmonella enterica* and *Mycobacterium tuberculosis*. The former of the original examples shows clearly separated clusters, representing a typical case where PopPunk performs well (**Figure 6.1A**), while the latter exemplifies a low diversity species in which the distance distributions do not exhibit a clear within-strain versus between-strain separation and instead form a cloud of points (**Figure 6.1B**). The PopPUNK authors acknowledged this limitation, noting that in cases like *M. tuberculosis*, network-based model refinement is needed but may result in over-splitting into many substrains. A similar pattern was observed in Typhi and Hadar (**Figure 6.1C** and **6.1D**), where a dense cluster of points near the origin of the graph (**Figures 4.5**, **5.4,** and **6.1**) and the lack of a clear boundary between within- and between-strain distances compromised convergence, leading to unstable clustering outputs.

Additionally, the distinct patterns observed between core and accessory distances in our data suggest that these two genomic components may be evolving somewhat independently. This situation, also described in the original PopPUNK framework for the case of *M. tuberculosis*, can reduce the effectiveness of using a combined distance metric for clustering. In such cases, alternative strategies that rely on a single dimension may offer better resolution than a combined metric.

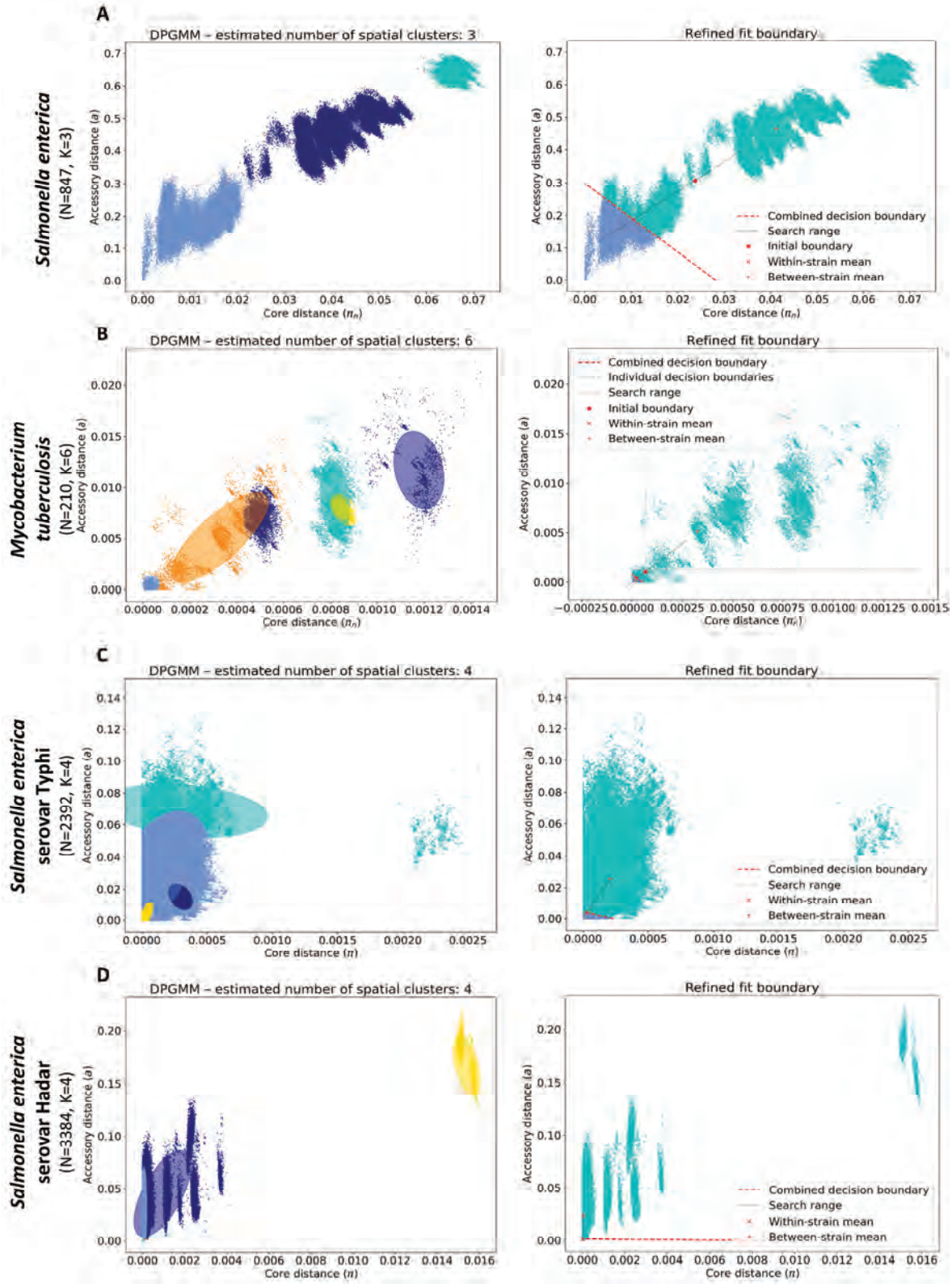


Figure 6.1 | PopPUNK model fitting output. Each row is a species (A and B) or a serovar (C and D), with each plot showing the distribution of core and accessory distances. The left column illustrates the 2D GMM fits including ellipses representing the mean and covariance of the fitted mixture components and points are colored by their predicted cluster. The cluster closest to the origin

represents the within-strain cluster. The right column shows the fits after refining the 2D GMM to maximize the network score. Initially, a linear boundary was estimated from the GMM clustering to separate within-strain from between-strain comparisons. This boundary was subsequentially adjusted to maximize the network score. The red dashed line indicates the final optimized decision boundary, while the black line represents the search range used during this optimization. In panel B, the grey lines depict the optimization with a vertical boundary (using core distances only) or a horizontal boundary (using accessory distances only). Red crosses indicate the mean of within-strain distances, and blue crosses denote the mean of between-strain distances.

Another limitation of the first PopPUNK version is that it provides only a single threshold and, consequently, a single network. A recent version, Iterative-PopPUNK [66], addresses this limitation by implementing a multi-level clustering strategy (as described in section *1.2.5.2*). It leverages a partially resolved core genome phylogeny to guide the selection of multiple thresholds, allowing clustering to reflect different levels of sequence similarity. This provides greater flexibility compared to the original approach, as it allows adjusting the clustering resolution depending on the specific aims of the analysis. Consequently, the thresholds selected by PopPUNK are based on differences observed in the core genome. However, this project aimed to take a different approach. Instead, we interpreted Jaccard Index values in terms of differences in both kilobases and SNPs, which can be more informative for choosing the threshold than relying solely on core genome differences. In this study, PopPUNK provided valuable insights into assessing the independent diversities of the core and accessory genomes.

Consequently, this study adopted an alternative approach by using the JI directly. The JI provides a single global distance for each genome pair by computing overall similarity based on the full set of *k*-mers, without separating core and accessory components. The most convenient threshold for analyzing each dataset was determined through direct inspection of the JI data combined with network properties such as transitivity, the number of clusters formed, and the percentage of genomes grouped within each cluster. This approach was designed to provide a primary, informative landscape that was not overly fragmented. By manually adjusting the JI threshold, it becomes evident how strains or groups differentiate, thereby enabling analysis at the desired level of resolution and uncovering both broad and fine-scale population structures.

The results of this study align with several observations made by PopPUNK authors, confirming that there is no universally optimal clustering solution. Instead, the most appropriate clustering resolution depends on the specific dataset and the study's objectives. Moreover, both approaches demonstrate that selecting an appropriate threshold necessitates careful consideration of network properties to achieve meaningful stratification of the bacterial population. Additionally, incorporating relevant metadata into the network analysis can further ensure that clustering reflects the context and characteristics of the data.

Although JI captures both core and accessory genome diversity, including indels and SNPs, it has a limitation in that it relies exclusively on unique *k*-mers. Consequently, JI cannot distinguish between two genomes that are identical except for the presence of repeated sequences.

Given the study's particular interest in indels, the new metric GLD was used to highlight differences between genomes attributed to indels. GLD serves as an optional layer, applied as needed based on the dataset and the desired resolution. In Hadar, most genomes differ by only a few SNPs; therefore, GLD had minimal impact on clustering. In contrast, in Typhi, some groups of genomes differ not only in indels but also in a number of SNPs (i.e., belonging to different primary clades of GenoTyphi), resulting in GLD having a more substantial impact on the clustering process. By applying GLD with a lower JI threshold in the Typhi network, the analysis emphasizes indel-driven differences. At the same time, it preserves the grouping of genomes that differ by a combination of indels and significant SNPs, thereby maintaining overall network connectivity. In summary, selectively integrating GLD with a JI threshold captures significant accessory genome changes without unnecessarily fragmenting clusters defined by SNP differences.

Importantly, in highly clonal bacteria, short-term changes driven by point mutations are typically minimal and may not significantly affect the JI. In contrast, the acquisition of MGEs can introduce a large number of new *k*-mers in a single event. For example, a 5 kb insertion could add roughly 5,000 new *k*-mers (with k=21), which is comparable in effect to about 238 SNPs ($238 \times 21 = 4{,}998$ *k*-mers). Thus, while the gradual accumulation of SNPs in the core genome over long periods is detectable and contributes more to the JI than a small indel, small numbers of mutations occurring over short timeframes may be overlooked, with indel events disproportionately influencing the JI. Therefore, for detecting minimal changes in the core genome, phylogenetic trees are the best option.

Finally, the JI method used here is not immediately implementable within the U.S. enteric surveillance system, PulseNet, due to current computational infrastructure. However, recent efforts to modernize PulseNet's genomic surveillance (https://www.aphl.org/aboutAPHL/publications/Documents/PulseNet-2.0-White-Paper.pdf) may provide an opportunity to incorporate JI-based methods, offering pangenomic analysis closer to "real-time", and simplifying the detection of unknown MGEs that can be explored with targeted genetic analysis. The ultimate public health goal is to provide a practical approach for enhanced genetic discrimination that improves surveillance and outbreak detection of otherwise indistinguishable enteric pathogens.

6.1 The *Salmonella enterica* serovar Typhi pangenome

The emergence of AMR is overwhelmingly driven by the acquisition of MGEs, particularly in Typhi. Thus, a comprehensive view of Typhi epidemiology necessitates a focus on the accessory genome. JINA allowed us to represent U.S. Typhi epidemiology as a reticulate network, revealing non-random structure in the pangenome and offering additional insights into Typhi epidemiology, ecology and evolutionary dynamics.

MGEs (both known and unknown) in Typhi were universally present, and each JI-group displayed a distinct profile corresponding to the presence or absence of particular plasmids or integrated MGEs (**Figure 4.9**), highlighting HGT as a significant mechanism of short-term diversification in Typhi. While large detectable plasmids were often the unique distinguishing feature of a JI-group, many unknown ICEs, phage-like elements, or hypothetical regions, which are generally overlooked in genomic analysis, were also responsible for JI-group differentiation (**Figure 4.9, Table 4.3**). These regions would not have been detected by routine screening methods (PulseNet USA screens for AMR determinants and plasmid replicons only), nor would they be of interest in investigations focused on AMR. Yet, these known and unknown MGEs are key features defining the structure of U.S. and global Typhi populations. Further knowledge of the transmission dynamics, functional capacity, and environmental reservoirs of these "cryptic" MGEs could offer valuable insight into the differing ecological predictors of Typhi occurrence and persistence.

Stratification of Typhi populations by accessory genome material, alongside existing core genome methods corroborated historical and recent epidemiological patterns. JI-grouping of Typhi genomes detected the previously globally dominant 4.3.1 MDR lineage carrying SGI11 on an IncHI1 (PTU-HI1A) plasmid (JI-D) [200], the 4.3.1.1 MDR lineage with chromosomal SGI11 (JI-A1, JI-C1), clonal and de novo emergence of triple QRDR mutants in different lineages (JI-A, JI-C, JI-I, JI-M) [124], clonal expansion of XDR 4.3.1.1.P1 strains (JI-B1) associated with travel to Pakistan circa 2018 [126], recent chromosomal integration of *bla*$_{\text{CTX-M-15}}$ into the chromosome of 4.3.1.1.P1 strains (JI-A3) [238], and a Nigeria-associated lineage of the 3.1.1 West African genotype (JI-J). These epidemiologically relevant JI-groups support the use of the Typhi pangenome for public health purposes. Specifically, in cases where travel data is unavailable, high genetic homology (>99.9% ANI) within a JI-subgroup can be leveraged to make travel-related inferences, potentially ameliorating the frequent lack of travel information on U.S. cases.

Pangenomic analysis expanded our understanding of Typhi plasmids and MGEs and suggests that AMR emergence and epidemiology in this pathogen are shaped by complex gene exchange networks and dynamics. First, two AMR-associated PTUs have emerged relatively recently in Typhi populations (PTU-E50 in JI-B and PTU-Y in JI-K), seemingly in distinct geographic regions. Given the host range of these PTUs [25] and the similar plasmids found in other genera, it is plausible that these acquisitions are the result of active genetic exchange between diverse genera within the *Enterobacteriaceae* family. Secondly, the high prevalence of chromosomally integrated MGEs (**Figure 4.9**) indicates the presence of "hotspots" for AMR region integration in the Typhi chromosome. This is supported by the detection of several unique integration events documented in this report (**Figures 4.19** and **4.20**), and in previous studies [238]. Thus, we should expect to see continual "stabilization" of AMR phenotypes in the chromosome, which may in turn create opportunities for new AMR plasmids to enter. With this in mind, it is tempting to speculate that long-established Typhi lineages (e.g. JI-A, represented in the Murray collection) may "sample" the mobile gene pool for plasmids and other MGEs of benefit (e.g. PTU-HI1A with SGI11, or PTU-E50 with *bla*$_{\text{CTX-M-15}}$) before eventually incorporating their advantageous cargo into the chromosome for reliable expression and long-term stability.

Although substantial effort is focused on understanding AMR, much of the Typhi population is neither MDR nor XDR (**Figure 4.17**), and most genomes do not carry known plasmids or any AMR genes at all (**Figures 4.9** and **4.16, Table S5**). Stratifying Typhi

populations based on “invisible” or “cryptic” MGEs can offer an additional layer of molecular resolution for exploration alongside epidemiological variables (e.g. geographic origin). This approach enables further differentiation of highly related genomes, as seen with JI-H and JI-N which belong to the same GenoTyphi but contain different MGEs. Alternatively, it can also enable the grouping of genomes that have converged on a single MGE. For instance, JI-C includes genomes from different GenoTyphi, yet all share the same plasmid. Just as individual SNPs serve as unique molecular signatures for identifying subpopulations [193], we can exploit the unknown accessory genome for enhanced discriminatory power or source attribution [180]. With the flexibility to “toggle” the JI threshold for increasing differentiation, JI-grouping proves to be a valuable analytical method for this purpose.

The potential biases introduced by utilizing Typhi genomes from a single country were addressed by analyzing multiple datasets from different geographic locations and time ranges. These datasets allowed us to evaluate whether the groups identified in the U.S. dataset remained consistent across geographic locations and time periods. Importantly, most of the U.S. JI-groups were also observed in these datasets. This consistency demonstrates that the genomic stratification and clustering observed in the U.S. dataset are not unique to that population but are reflective of broader global patterns in Typhi evolution and epidemiology. One remaining limitation of this analysis is the lack of very recent genomes (2022-2024). Given the rapid evolution of Typhi populations, new JI-groups may emerge in a relatively short time frame. Additionally, this analysis detected many previously unknown MGEs that may prove epidemiologically relevant; however, in-depth genetic characterization of each MGE was beyond the scope of this analysis.

Genomic analysis of pre-antibiotic era isolates revealed that certain lineages were already established prior to the introduction of antibiotics and have since circulating with minimal genetic changes (**Figures 4.26** and **4.27**). These findings underscore the long-term persistence and evolutionary stability of some Typhi lineages.

In summary, this analysis revealed a non-random structure in the Typhi pangenome, driven both by differences in the core genome and by the gain and loss of mobile genetic elements. These findings confirm and expand upon known epidemiological patterns, reveal novel plasmid dynamics, and identify new avenues for further genomic epidemiological exploration.

6.2 The *Salmonella enterica* serovar Hadar pangenome

Identifying the molecular mechanisms underlying shifts in bacterial populations is key to understanding the adaptive forces that drive evolution of human bacterial pathogens. Analysis of the Hadar pangenome confirmed known epidemiological and microevolutionary dynamics while also revealing previously unknown ones.

Before 2020, two distinct lineages dominated in commercial poultry (JI-B and JI-G) and backyard poultry environments (JI-D and JI-E) separately. In 2020, an emergent lineage closely related to previously circulating strains became dominant, displacing the historical commercial poultry lineage. Around the same time, coinciding with a surge in backyard poultry ownership during the COVID-19 pandemic [259], this same emergent lineage became dominant among backyard poultry-associated human cases, confirming through high-resolution pangenomic analysis a link between two presumably separate industries. Epidemiological and biological evidence suggests that the presence of a novel prophage in the emergent lineage may have contributed to its recent expansion. Interestingly, a similar genetic shift underpinned by an emergent prophage-containing lineage was seen in the French genomes analyzed here, suggesting this phenomenon is not restricted to the U.S. The adaptive capacity of this prophage in Hadar, and specifically, the putative pathogenic role of the phage-encoded Zot-like protein, is still under evaluatation in U.S. Hadar genomes.

These new findings can be leveraged to mitigate further spread of this emergent strain in several ways. First, comparative plasmid analysis identified a clonal subcluster within this lineage (JI-C1), suggesting a reservoir upstream of backyard poultry suppliers and hatcheries, one that likely interfaces with commercial poultry (**Figure 5.20**). Backyard poultry hatchery practices, such as drop-shipping and outsourcing to larger commercial hatcheries to meet demand [82,250], could explain this connection. This data can inform conversations between industry and government stakeholders, as it promotes collective action with the goal of eliminating shared reservoirs affecting multiple industries. Second, functional analyses to determine the contribution of prophage 1 to avian gut colonization could inform intervention strategies in both commercial and backyard poultry settings; for example, by minimizing bacterial burden in birds, which is considered a control strategy to reduce risk of transmission to humans [260]. Third, this analysis highlighted the importance of known MGE (e.g., PTU-I1 plasmids) and identified previously uncharacterized MGE (e.g., prophage 1) that can potentially be incorporated into source attribution models and

molecular case definitions. For example, PTU-I1 plasmids could serve as a genetic marker that distinguishes backyard poultry-related strains from those transmitted via other sources. More accurate prediction of foodborne versus animal contact transmission pathways and refinement of outbreak and REP strain case definitions both contribute to timelier epidemiological traceback, and ultimately, a reduction in human illness [183].

More generally, this analysis enabled high-resolution genomic linking of human cases to potential sources (e.g., pet food, imported shrimp), helping to identify specific vehicles for further investigation and refining supplementary interviews or traceback efforts when exposure information is limited and no transmission vehicles are otherwise suspected. Additionally, this study identified avenues for investigating the ecological dynamics that underpin persistence of Hadar in different environments.

For example, PTU-I1 and other large plasmids are associated with backyard poultry rather than commercial poultry environments, and some JI-groups (with distinct MGE profiles) display a unique chicken association rather than the more common turkey signal. Along with the previously unreported role of prophages in Hadar diversification and microevolution, this comprehensive description of MGE in the U.S. Hadar population is foundational information for pathogen risk modeling, especially as it pertains to carriage of AMR. The presence of "risky" MGE related to AMR, virulence, or colonization capacity, can be proactively monitored through existing surveillance programs, allowing emergent threats to be addressed before they become systematically disseminated, as has previously occurred with *Salmonella* serotypes Infantis [251] and Reading [261].

Due to the nature of the pangenomic approach employed here, the exact timing and location of this persisting REP strain were not determined. Furthermore, while this approach cannot definitively determine the source of human illnesses with unknown exposures, or multiple exposures (e.g., both commercial and backyard poultry), the findings from this study will be assessed within ongoing source attribution modeling to estimate the added value of inclusion of accessory genome content. Additionally, while efforts were made to obtain genomes representing diverse environments (wildlife, imported foods, commercial poultry production, backyard poultry environments, ill humans), several sources remain underrepresented (e.g., live animals on farm) or absent (e.g., hatcheries), potentially missing dominant pangenomic groups in these settings. Expanded analyses that include genomes

from underrepresented sources, along with deeper investigations into the global diversity of Hadar, will fill important gaps in the pangenome landscape described here.

Unraveling pathogen epidemiology and microevolutionary dynamics is a complex challenge, and the plethora of available data is both an opportunity and a challenge. Leveraging existing genomic data, we demonstrate the value of JI-based pangenomic analysis for delineating a highly clonal serotype and uncover actionable data to mitigate the spread of an emergent and potentially more pathogenic lineage of Hadar. We paint a pangenome landscape of this previously understudied serotype, highlighting the importance of both known and unknown MGE, and revealing surprising geographic patterns and dynamics. These findings will inform future risk and source attribution modeling, reducing public health burdens and mitigating impacts on implicated food and animal industries.

6.3 Final discussion

Salmonella Typhi and *Salmonella* Hadar are clonal pathogens, yet notable differences in their pangenomes underscore their evolutionary and epidemiological divergence. While most genomes in both serovars differ by only a few SNPs, the accessory genome substantially contributes to overall genetic variability, with prophages and plasmids playing important roles. A key difference is that plasmids in Hadar usually do not carry antibiotic resistance genes. Even though over 90% of Hadar genomes harbor resistance genes, they are generally limited to aminoglycosides and tetracyclines, making resistance a less pressing concern. Nevertheless, the detection of certain PTU-I1 plasmids, known for their broad host range, that harbor some AMR genes suggests the potential for horizontal gene transfer of other resistance traits. Conversely, resistance genes in Typhi represent a serious global issue, as treatment options in some regions are very limited. Moreover, AMR genes not only reside on plasmids but also frequently integrate into the chromosome.

These genomic distinctions are further reflected in the ecological niches of the two serovars. Typhi is a human-adapted pathogen with a globally distributed pangenome driven by the extensive mobility of its human host. In contrast, Hadar, primarily associated with poultry, exhibits more geographical genomic differentiation due to limited animal movement and the regional nature of commercial supply chains. This genomic footprint not only

confirms the known specialization of these serovars but also provides tangible genomic evidence of their distinct evolutionary paths.

Methodologically, this work expands on traditional phylogenetic analyses by integrating both vertical inheritance and horizontal gene transfer. By leveraging Jaccard Index Network Analysis to integrate both core and accessory genetic material, the study presents a multidimensional framework that complements existing phylogenetic methods. This approach redefines our understanding of "homology" by accounting for both vertical and horizontal genetic relationships, with implications for analyzing short-term bacterial evolution and public health applications. Additionally, it offers a powerful tool for surveillance, outbreak management, and tracking antimicrobial resistance by mapping complex reticulate genomic networks.

Furthermore, these networks facilitate the incorporation of newly isolated genomes of these serovars. As new genomes are added to the network, they are integrated into one of the predefined groups, allowing for the rapid identification of certain characteristics, such as the presence of specific MGEs, without requiring detailed individual analysis. Conversely, if a new genome does not closely match any existing groups, it indicates the emergence of a new lineage. This was exemplified when we added two new Typhi isolates to the network and grouped independently (data not shown). Detailed analysis revealed that these genomes contained a PTU-FE plasmid conferring resistance to azithromycin [189] and exhibited chromosomal similarity to strain JI-subgroup A1, confirming them as a distinct lineage.

In summary, this thesis not only elucidates the genomic nuances of pathogen adaptation and specialization of Typhi and Hadar but also establishes a framework for analyzing bacterial population structure. By demonstrating the significant role of the accessory genome, the study paves the way for improved public health strategies and a more comprehensive approach to bacterial typing and epidemiological surveillance.

CONCLUSIONS

1. Applying the JI as an agnostic measure of genomic distance allowed for the capture of both vertical (core genome) and horizontal (accessory genome) evolutionary relationships. This enabled high-resolution stratification of closely related genomes and revealed epidemiologically relevant subgroups.

2. Our pangenome analysis revealed fine-scale population structures, emerging lineages, and the impact of mobile genetic elements driving epidemiologically relevant shifts in Typhi and Hadar in short periods of time. This high-resolution approach enhances outbreak detection, facilitates source attribution, and strengthens surveillance of antimicrobial resistance.

3. In Typhi, JI-based clustering linked specific groups to globally significant lineages, including MDR and XDR strains. It also distinguished between different genetic contexts (e.g., chromosomal versus plasmid integrations of resistance genes) and identified both known and previously unknown MGEs. Therefore, JI-based networks offer a valuable complement to traditional GenoTyphi typing by providing an additional layer of genomic information that enhances strain discrimination and deepens our understanding of Typhi's evolutionary dynamics, with clear implications for public health.

4. Pangenome analysis revealed a shift in Hadar populations. Before 2020, distinct groups existed in commercial and backyard poultry, but in 2020, a new lineage (designated as REPTDK01) emerged and rapidly expanded, displacing the previously established lineages. This analysis refined the definition of REPTDK01 into two groups based on the presence or absence of a plasmid. Additionally, we identified a novel prophage in all these isolates that may confer a selective advantage, potentially driving this expansion. Notably, no clear pangenomic differences were identified between strains from backyard and commercial settings, suggesting shared reservoirs and transmission routes between these traditionally separate environments.

5. The U.S. Typhi pangenome structure closely resembled global populations, suggesting that extensive human mobility drives the widespread circulation of similar lineages worldwide. In contrast, non-U.S. Hadar populations showed geographical genomic differentiation that may be influenced by limited animal movement and

localized supply chains. These findings underscore the distinct evolutionary paths and ecological niches of Typhi and Hadar, enhancing our understanding of their diversity and public health implications.

6. Continuous pangenomic surveillance is essential to monitor emerging MGEs and evolving lineages. JI-based networks provide a robust tool for this purpose, as the appearance of a new group signals the emergence of a new lineage, guiding timely interventions and enriching our understanding of *Salmonella* evolution.

REFERENCES

1. Cheung MK, Kwan HS. Fighting Outbreaks with Bacterial Genomics: Case Review and Workflow Proposal. Public Health Genomics. 2012;15(6):341–51.

2. Taylor LH, Latham SM, Woolhouse MEJ. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci. 2001 Jul 29;356(1411):983–9.

3. Lombard JE, Patton EA, Gibbons-Burgener SN, Klos RF, Tans-Kersten JL, Carlson BW, et al. Human-to-Cattle Mycobacterium tuberculosis Complex Transmission in the United States. Front Vet Sci. 2021 Jul 12;8:691192.

4. Messenger AM, Barnes AN, Gray GC. Reverse Zoonotic Disease Transmission (Zooanthroponosis): A Systematic Review of Seldom-Documented Human Biological Threats to Animals. PLoS ONE. 2014 Feb 28;9(2):e89055.

5. Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. 2008 Jun;6(6):431–40.

6. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. Nat Rev Genet. 2018 Sep;19(9):549–65.

7. Almeida RPP, Nunney L. How Do Plant Diseases Caused by *Xylella fastidiosa* Emerge? Plant Dis. 2015 Nov;99(11):1457–67.

8. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. Nat Rev Microbiol. 2016 Mar;14(3):150–62.

9. Balloux F, Van Dorp L. Q&A: What are pathogens, and what have they done to and for us? BMC Biol. 2017 Dec;15(1):91.

10. Lawrence JG. Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol. 1999 Oct;2(5):519–23.

11. Pallen MJ, Wren BW. Bacterial pathogenomics. Nature. 2007 Oct;449(7164):835–42.

12. Andersson JO, Andersson SG. Insights into the evolutionary process of genome degradation. Curr Opin Genet Dev. 1999 Dec;9(6):664–71.

13. Brown EW, Bell R, Zhang G, Timme R, Zheng J, Hammack TS, et al. Salmonella Genomics in Public Health and Food Safety. EcoSal Plus. 2021 Dec 15;9(2):eESP-0008-2020.

14. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2009 Feb 1;3(2):199–208.

15. Haudiquet M, De Sousa JM, Touchon M, Rocha EPC. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. Philos Trans R Soc B Biol Sci. 2022 Oct 10;377(1861):20210234.

16. De La Cruz F, Frost LS, Meyer RJ, Zechner EL. Conjugative DNA metabolism in Gram-negative bacteria. FEMS Microbiol Rev. 2010 Jan;34(1):18–40.

17. Chiang YN, Penadés JR, Chen J. Genetic transduction by phages and chromosomal islands: The new and noncanonical. PLOS Pathog. 2019 Aug 8;15(8):e1007878.

18. De La Cruz F, Davies J. Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol. 2000 Mar;8(3):128–33.

19. Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: Convergence or divergence by homologous recombination? Genome Res. 2007 Jan;17(1):61–8.

20. Dixit PD, Pang TY, Maslov S. Recombination-Driven Genome Evolution and Stability of Bacterial Species. Genetics. 2017 Sep 1;207(1):281–95.

21. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci. 2010 Jan 5;107(1):127–32.

22. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, De La Cruz F. Mobility of Plasmids. Microbiol Mol Biol Rev. 2010 Sep;74(3):434–52.

23. San Millan A. Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context. Trends Microbiol. 2018 Dec;26(12):978–85.

24. Alekshun MN, Levy SB. Molecular Mechanisms of Antibacterial Multidrug Resistance. Cell. 2007 Mar;128(6):1037–50.

25. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. Nat Commun. 2020 Jul 17;11(1):3602.

26. Ares-Arroyo M, Coluzzi C, Rocha EPC. Origins of transfer establish networks of functional dependencies for plasmid transfer by conjugation. Nucleic Acids Res. 2023 Apr 24;51(7):3001–16.

27. Garcillán-Barcia MP, Francia MV, De La Cruz F. The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol Rev. 2009 May;33(3):657–87.

28. Garcillán-Barcia MP, Redondo-Salvo S, Vielva L, de la Cruz F. MOBscan: Automated Annotation of MOB Relaxases. In: Horizontal Gene Transfer [Internet]. New York, NY: Springer US; 2020 [cited 2022 Dec 16]. p. 295–308. (Methods in Molecular Biology; vol. 2075). Available from: http://link.springer.com/10.1007/978-1-4939-9877-7_21

29. Del Solar G, Giraldo R, Ruiz-Echevarría MJ, Espinosa M, Díaz-Orejas R. Replication and Control of Circular Bacterial Plasmids. Microbiol Mol Biol Rev. 1998 Jun;62(2):434–64.

30. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, et al. *In Silico* Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrob Agents Chemother. 2014 Jul;58(7):3895–903.

31. Garcillán-Barcia MP, Redondo-Salvo S, De La Cruz F. Plasmid classifications. Plasmid. 2023 May;126:102684.

32. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genomics. 2018 Aug 1;4(8):e000206.

33. Shapiro BJ. How clonal are bacteria over time? Curr Opin Microbiol. 2016 Jun;31:116–23.

34. Spratt B. The relative contributions of recombination and point mutation to the diversification of bacterial clones. Curr Opin Microbiol. 2001 Oct 1;4(5):602–6.

35. Feil EJ, Spratt BG. Recombination and the Population Structures of Bacterial Pathogens. Annu Rev Microbiol. 2001 Oct;55(1):561–90.

36. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017 Apr;2(4):17040.

37. Costa SS, Guimarães LC, Silva A, Soares SC, Baraúna RA. First Steps in the Analysis of Prokaryotic Pan-Genomes. Bioinforma Biol Insights. 2020 Jan;14:117793222093806.

38. Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, et al. Typing methods based on whole genome sequencing data. One Health Outlook. 2020 Dec;2(1):3.

39. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. PLOS Genet. 2018 Apr 5;14(4):e1007261.

40. Hunter SB, Vauterin P, Lambert-Fair MA, Van Duyne MS, Kubota K, Graves L, et al. Establishment of a Universal Size Standard Strain for Use with the PulseNet Standardized Pulsed-Field Gel Electrophoresis Protocols: Converting the National Databases to the New Size Standard. J Clin Microbiol. 2005 Mar;43(3):1045–50.

41. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci. 1998 Mar 17;95(6):3140–5.

42. Nadon CA, Trees E, Ng LK, Møller Nielsen E, Reimer A, Maxwell N, et al. Development and application of MLVA methods as a tool for inter-laboratory surveillance. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull. 2013 Aug 29;18(35):20565.

43. Maiden MCJ, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013 Oct;11(10):728–36.

44. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data. J Bacteriol. 2004 Mar;186(5):1518–30.

45. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Agama Study Group, Achtman M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res. 2020 Jan;30(1):138–52.

46. Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MC. A gene-by-gene population genomics platform: de novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes. BMC Genomics. 2014 Dec;15(1):1138.

47. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, et al. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes. J Clin Microbiol. 2015 Sep;53(9):2869–76.

48. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Eurosurveillance. 2017 Jun 8;22(23).

49. Den Bakker HC, Moreno Switt AI, Cummings CA, Hoelzer K, Degoricija L, Rodriguez-Rivera LD, et al. A Whole-Genome Single Nucleotide Polymorphism-Based Approach To Trace and Identify Outbreaks Linked to a Common Salmonella enterica subsp. enterica Serovar Montevideo Pulsed-Field Gel Electrophoresis Type. Appl Environ Microbiol. 2011 Dec 15;77(24):8648–55.

50. Taylor AJ, Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ, Medus C, et al. Characterization of Foodborne Outbreaks of Salmonella enterica Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection. J Clin Microbiol. 2015 Oct;53(10):3334–40.

51. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987–93.

52. Chan CX, Ragan MA. Next-generation phylogenomics. Biol Direct. 2013 Dec;8(1):3.

53. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016 Dec;17(1):132.

54. Gardner SN, Slezak T. Scalable SNP Analyses of 100+ Bacterial or Viral Genomes. J Forensic Res. 2010;01(03).

55. Harris SR. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology [Internet]. 2018 [cited 2025 Jan 2]. Available from: http://biorxiv.org/lookup/doi/10.1101/453142

56. Hall BG, Nisbet J. Building Phylogenetic Trees From Genome Sequences With kSNP4. Mol Biol Evol. 2023 Nov 3;40(11):msad235.

57. Derelle R, Von Wachsmann J, Mäklin T, Hellewell J, Russell T, Lalvani A, et al. Seamless, rapid, and accurate analyses of outbreak genomic data using split *k* -mer analysis. Genome Res. 2024 Oct;34(10):1661–73.

58. Eliades SJ, Brown JC, Colston TJ, Fisher RN, Niukula JB, Gray K, et al. Gut microbial ecology of the Critically Endangered Fijian crested iguana (*Brachylophus vitiensis*): Effects of captivity status and host reintroduction on endogenous microbiomes. Ecol Evol. 2021 May;11(9):4731–43.

59. Leclaire S, Nielsen JF, Drea CM. Bacterial communities in meerkat anal scent secretions vary with host sex, age, and group membership. Behav Ecol. 2014 Jul 1;25(4):996–1004.

60. Puspaningrum EY, Nugroho B, Setiawan A, Hariyanti N. Detection of Text Similarity for Indication Plagiarism Using Winnowing Algorithm Based K-gram and Jaccard Coefficient. J Phys Conf Ser. 2020 Jul 1;1569(2):022044.

61. Temma S, Sugii M, Matsuno H. The Document Similarity Index based on the Jaccard Distance for Mail Filtering. In: 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC) [Internet]. JeJu, Korea (South): IEEE; 2019 [cited 2022 Dec 19]. p. 1–4. Available from: https://ieeexplore.ieee.org/document/8793419/

62. Zhao X. BinDash, software for fast genome distance estimation on a typical personal laptop. Bioinformatics. 2019 Feb 15;35(4):671–3.

63. Acman M, van Dorp L, Santini JM, Balloux F. Large-scale network analysis captures biological features of bacterial plasmids. Nat Commun. 2020 May 15;11(1):2452.

64. Prokopenko D, Hecker J, Silverman EK, Pagano M, Nöthen MM, Dina C, et al. Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. Bioinformatics. 2016 May 1;32(9):1366–72.

65. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019 Feb;29(2):304–16.

66. Zhao B, Lees JA, Wu H, Yang C, Falush D. Genealogical inference and more flexible sequence clustering using iterative-PopPUNK. Genome Res. 2023 Jun;33(6):988–98.

67. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3.

68. Perrin A, Rocha EPC. PanACoTA: a modular tool for massive microbial comparative genomics. NAR Genomics Bioinforma. 2021 Jan 12;3(1):lqaa106.

69. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020 Dec;21(1):180.

70. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. Nucleic Acids Res. 2018 Jan 9;46(1):e5–e5.

71. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59–60.

72. Bradley P, Den Bakker HC, Rocha EPC, McVean G, Iqbal Z. Ultrafast search of all deposited bacterial and viral genomic data. Nat Biotechnol. 2019 Feb;37(2):152–9.

73. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150–2.

74. Fernández-de-Bobadilla MD, Talavera-Rodríguez A, Chacón L, Baquero F, Coque TM, Lanza VF. PATO: Pangenome Analysis Toolkit. Bioinformatics. 2021 Dec 7;37(23):4564–6.

75. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017 Nov;35(11):1026–8.

76. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

77. Noll N, Molari M, Shaw LP, Neher RA. PanGraph: scalable bacterial pan-genome graph construction. Microb Genomics. 2023 Jun 6;9(6).

78. Loharikar A, Briere E, Schwensohn C, Weninger S, Wagendorf J, Scheftel J, et al. Four Multistate Outbreaks of Human *Salmonella* Infections Associated with Live Poultry Contact, United States, 2009. Zoonoses Public Health. 2012 Aug;59(5):347–54.

79. Gaffga NH, Behravesh CB, Ettestad PJ, Smelser CB, Rhorer AR, Cronquist AB, et al. Outbreak of Salmonellosis Linked to Live Poultry from a Mail-Order Hatchery. N Engl J Med. 2012 May 31;366(22):2065–73.

80. Harris JR, Neil KP, Behravesh CB, Sotir MJ, Angulo FJ. Recent Multistate Outbreaks of Human *Salmonella* Infections Acquired from Turtles: A Continuing Public Health Challenge. Clin Infect Dis. 2010 Feb 15;50(4):554–9.

81. Hale CR, Scallan E, Cronquist AB, Dunn J, Smith K, Robinson T, et al. Estimates of Enteric Illness Attributable to Contact With Animals and Their Environments in the United States. Clin Infect Dis. 2012 Jun 1;54(suppl_5):S472–9.

82. Behravesh CB, Brinson D, Hopkins BA, Gomez TM. Backyard Poultry Flocks and Salmonellosis: A Recurring, Yet Preventable Public Health Challenge. Clin Infect Dis. 2014 May 15;58(10):1432–8.

83. Hoelzer K, Moreno Switt A, Wiedmann M. Animal contact as a source of human non-typhoidal salmonellosis. Vet Res. 2011;42(1):34.

84. Silva C, Calva E, Maloy S. One Health and Food-Borne Disease: *Salmonella* Transmission between Humans, Animals, and Plants. Microbiol Spectr. 2014 Jan 17;2(1):2.1.08.

85. Brenner FW, Villar RG, Angulo FJ, Tauxe R, Swaminathan B. *Salmonella* Nomenclature. J Clin Microbiol. 2000 Jul;38(7):2465–7.

86. Grimont PAD, Weill FX. Antigenic formulae of the Salmonella serovars. 9th ed, WHO Collaborating Centre for Reference and Research on Salmonella, Institut Pasteur, Paris, France; 2007.

87. Crosa JH, Brenner DJ, Ewing WH, Falkow S. Molecular Relationships Among the Salmonelleae. J Bacteriol. 1973 Jul;115(1):307–15.

88. Reeves MW, Evins GM, Heiba AA, Plikaytis BD, Farmer JJ. Clonal nature of Salmonella typhi and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of Salmonella bongori comb. nov. J Clin Microbiol. 1989 Feb;27(2):313–20.

89. Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M. Mismatch induced speciation in *Salmonella* : model and data. Philos Trans R Soc B Biol Sci. 2006 Nov 29;361(1475):2045–53.

90. Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, et al. Multilocus Sequence Typing as a Replacement for Serotyping in Salmonella enterica. PLoS Pathog. 2012 Jun 21;8(6):e1002776.

91. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res. 2018 Sep;28(9):1395–404.

92. Achtman M, Zhou Z, Alikhan NF, Tyne W, Parkhill J, Cormican M, et al. Genomic diversity of Salmonella enterica -The UoWUCC 10K genomes project. Wellcome Open Res. 2021 Feb 1;5:223.

93. Jones TF, Ingram LA, Cieslak PR, Vugia DJ, Tobin-D'Angelo M, Hurd S, et al. Salmonellosis Outcomes Differ Substantially by Serotype. J Infect Dis. 2008 Jul;198(1):109–14.

94. Blaser MJ, Newman LS. A Review of Human Salmonellosis: I. Infective Dose. Clin Infect Dis. 1982 Nov 1;4(6):1096–106.

95. El Sayed F, Sapriel G, Fawal N, Gruber A, Bauer T, Heym B, et al. In-Host Adaptation of Salmonella enterica Serotype Dublin during Prosthetic Hip Joint Infection. Emerg Infect Dis. 2018 Dec;24(12):2360–3.

96. Lichtensteiger CA, Vimr ER. Systemic and enteric colonization of pigs by a hilA signature-tagged mutant of Salmonella choleraesuis. Microb Pathog. 2003 Mar;34(3):149–54.

97. Edsall G, Gaines S, Landy M, Tigertt WD, Sprinz H, Trapani RJ, et al. Studies on infection and immunity in experimental typhoid fever. J Exp Med. 1960 Jul 1;112(1):143–66.

98. Uzzau S, Brown DJ, Wallis T, Rubino S, Leori G, Bernard S, et al. Host adapted serotypes of *Salmonella enterica*. Epidemiol Infect. 2000 Oct;125(2):229–55.

99. Liu SL, Sanderson KE. Rearrangements in the genome of the bacterium Salmonella typhi. Proc Natl Acad Sci. 1995 Feb 14;92(4):1018–22.

100. Gal-Mor O, Boyle EC, Grassl GA. Same species, different diseases: how and why typhoidal and non-typhoidal Salmonella enterica serovars differ. Front Microbiol. 2014 Aug 4;5.

101. Haraga A, Ohlson MB, Miller SI. Salmonellae interplay with host cells. Nat Rev Microbiol. 2008 Jan;6(1):53–66.

102. Feasey NA, Dougan G, Kingsley RA, Heyderman RS, Gordon MA. Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa. The Lancet. 2012 Jun;379(9835):2489–99.

103. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018 Nov;392(10159):1736–88.

104. European Food Safety Authority (EFSA), European Centre for Disease Prevention and Control (ECDC). The European Union One Health 2023 Zoonoses report. EFSA J. 2024 Dec;22(12).

105. Centers for Disease Control and Prevention (U.S.). Antibiotic resistance threats in the United States, 2019 [Internet]. Centers for Disease Control and Prevention (U.S.); 2019 Nov [cited 2022 Dec 16]. Available from: https://stacks.cdc.gov/view/cdc/82532

106. Stanaway JD, Reiner RC, Blacker BF, Goldberg EM, Khalil IA, Troeger CE, et al. The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Infect Dis. 2019 Apr;19(4):369–81.

107. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ. Typhoid Fever. N Engl J Med. 2002 Nov 28;347(22):1770–82.

108. Mogasale V, Maskery B, Ochiai RL, Lee JS, Mogasale VV, Ramani E, et al. Burden of typhoid fever in low-income and middle-income countries: a systematic, literature-based update with risk-factor adjustment. Lancet Glob Health. 2014 Oct;2(10):e570–80.

109. European Centre for Disease Prevention and Control. Typhoid and paratyphoid infection. In: ECDC, editor. Annual epidemiological report for 2020. Stockholm: ECDC; 2024.

110. Han J, Aljahdali N, Zhao S, Tang H, Harbottle H, Hoffmann M, et al. Infection biology of *Salmonella enterica*. EcoSal Plus. 2024 Dec 12;12(1):eesp-0001-2023.

111. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, et al. Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature. 2001 Oct 25;413(6858):852–6.

112. Raffatellu M, Chessa D, Wilson RP, Dusold R, Rubino S, Bäumler AJ. The Vi Capsular Antigen of *Salmonella enterica* Serotype Typhi Reduces Toll-Like Receptor-

Dependent Interleukin-8 Expression in the Intestinal Mucosa. Infect Immun. 2005 Jun;73(6):3367–74.

113. Hu X, Chen Z, Xiong K, Wang J, Rao X, Cong Y. Vi capsular polysaccharide: Synthesis, virulence, and application. Crit Rev Microbiol. 2017 Jul 4;43(4):440–52.

114. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, et al. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature. 2001 Oct;413(6858):848–52.

115. Winter SE, Thiennimitr P, Winter MG, Butler BP, Huseby DL, Crawford RW, et al. Gut inflammation provides a respiratory electron acceptor for Salmonella. Nature. 2010 Sep;467(7314):426–9.

116. D'Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, et al. Antibiotic resistance is ancient. Nature. 2011 Sep;477(7365):457–61.

117. Aslam B, Wang W, Arshad MI, Khurshid M, Muzammil S, Rasool MH, et al. Antibiotic resistance: a rundown of a global crisis. Infect Drug Resist. 2018 Oct;Volume 11:1645–58.

118. Woodward TE. Management of Typhoid Fever and Its Complications. Ann Intern Med. 1964 Jan 1;60(1):144.

119. Anderson ES. The problem and implications of chloramphenicol resistance in the typhoid bacillus. J Hyg (Lond). 1975 Apr;74(2):289–99.

120. Olarte J, Galindo E. *Salmonella typhi* resistant to Chloramphenicol, Ampicillin, and Other Antimicrobial Agents: Strains Isolated During an Extensive Typhoid Fever Epidemic in Mexico. Antimicrob Agents Chemother. 1973 Dec;4(6):597–601.

121. Wain J, Hendriksen RS, Mikoleit ML, Keddy KH, Ochiai RL. Typhoid fever. The Lancet. 2015 Mar;385(9973):1136–45.

122. Mirza SH, Beechmg NJ, Hart CA. Multi-drug resistant typhoid: a global problem. J Med Microbiol. 1996 May 1;44(5):317–9.

123. Andrews JR, Qamar FN, Charles RC, Ryan ET. Extensively Drug-Resistant Typhoid — Are Conjugate Vaccines Arriving Just in Time? N Engl J Med. 2018 Oct 18;379(16):1493–5.

124. Carey ME, Dyson ZA, Ingle DJ, Amir A, Aworh MK, Chattaway MA, et al. Global diversity and antimicrobial resistance of typhoid fever pathogens: Insights from a meta-analysis of 13,000 Salmonella Typhi genomes. eLife. 2023 Sep 12;12:e85867.

125. Day MR, Doumith M, Do Nascimento V, Nair S, Ashton PM, Jenkins C, et al. Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of Salmonella enterica serovars Typhi and Paratyphi. J Antimicrob Chemother. 2018 Feb 1;73(2):365–72.

126. Klemm EJ, Shakoor S, Page AJ, Qamar FN, Judge K, Saeed DK, et al. Emergence of an Extensively Drug-Resistant *Salmonella enterica* Serovar Typhi Clone Harboring a

Promiscuous Plasmid Encoding Resistance to Fluoroquinolones and Third-Generation Cephalosporins. mBio. 2018 Mar 7;9(1):e00105-18.

127. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, et al. Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. Nat Genet. 2012 Nov;44(11):1215–21.

128. Boyd D, Peters GA, Cloeckaert A, Boumedine KS, Chaslus-Dancla E, Imberechts H, et al. Complete Nucleotide Sequence of a 43-Kilobase Genomic Island Associated with the Multidrug Resistance Region of *Salmonella enterica* Serovar Typhimurium DT104 and Its Identification in Phage Type DT120 and Serovar Agona. J Bacteriol. 2001 Oct;183(19):5725–32.

129. Mulvey MR, Boyd DA, Olson AB, Doublet B, Cloeckaert A. The genetics of Salmonella genomic island 1. Microbes Infect. 2006 Jun;8(7):1915–22.

130. Hawkey J, Le Hello S, Doublet B, Granier SA, Hendriksen RS, Fricke WF, et al. Global phylogenomics of multidrug-resistant Salmonella enterica serotype Kentucky ST198. Microb Genomics. 2019 Jul 1;5(7).

131. Ghoddusi A, Nayeri Fasaei B, Zahraei Salehi T, Akbarein H. Prevalence and characterization of multidrug resistance and variant *Salmonella* genomic island 1 in *Salmonella* isolates from cattle, poultry and humans in Iran. Zoonoses Public Health. 2019 Sep;66(6):587–96.

132. Ahmed AM, Hussein AIA, Shimamoto T. Proteus mirabilis clinical isolate harbouring a new variant of Salmonella genomic island 1 containing the multiple antibiotic resistance region. J Antimicrob Chemother. 2007 Feb;59(2):184–90.

133. Hamidian M, Holt KE, Hall RM. Genomic resistance island AGI1 carrying a complex class 1 integron in a multiply antibiotic-resistant ST25 Acinetobacter baumannii isolate. J Antimicrob Chemother. 2015 Sep;70(9):2519–23.

134. Cummins ML, Roy Chowdhury P, Marenda MS, Browning GF, Djordjevic SP. Salmonella Genomic Island 1B Variant Found in a Sequence Type 117 Avian Pathogenic Escherichia coli Isolate. mSphere. 2019 May 22;4(3):e00169-19.

135. Doublet B, Boyd D, Mulvey MR, Cloeckaert A. The *Salmonella* genomic island 1 is an integrative mobilizable element. Mol Microbiol. 2005 Mar;55(6):1911–24.

136. Douard G, Praud K, Cloeckaert A, Doublet B. The Salmonella Genomic Island 1 Is Specifically Mobilized In Trans by the IncA/C Multidrug Resistance Plasmid Family. PLoS ONE. 2010 Dec 20;5(12):e15302.

137. Kariuki S, Gordon MA, Feasey N, Parry CM. Antimicrobial resistance and management of invasive Salmonella disease. Vaccine. 2015 Jun;33:C21–9.

138. Tack B, Vanaenrode J, Verbakel JY, Toelen J, Jacobs J. Invasive non-typhoidal Salmonella infections in sub-Saharan Africa: a systematic review on antimicrobial resistance and treatment. BMC Med. 2020 Dec;18(1):212.

139. Date KA, Bentsi-Enchill A, Marks F, Fox K. Typhoid fever vaccination strategies. Vaccine. 2015 Jun 19;33 Suppl 3(Suppl 3):C55-61.

140. World Health Organization (2018) Typhoid vaccines: WHO position paper Week Epidemiol Rec 13:153–172.

141. Batool R, Tahir Yousafzai M, Qureshi S, Ali M, Sadaf T, Mehmood J, et al. Effectiveness of typhoid conjugate vaccine against culture-confirmed typhoid in a peri-urban setting in Karachi: A case-control study. Vaccine. 2021 Sep;39(40):5858–65.

142. Siddique A, Wang Z, Zhou H, Huang L, Jia C, Wang B, et al. The Evolution of Vaccines Development across Salmonella Serovars among Animal Hosts: A Systematic Review. Vaccines. 2024 Sep 18;12(9):1067.

143. Crouch CF, Nell T, Reijnders M, Donkers T, Pugh C, Patel A, et al. Safety and efficacy of a novel inactivated trivalent Salmonella enterica vaccine in chickens. Vaccine. 2020 Oct;38(43):6741–50.

144. Farzan A, Friendship RM. A clinical field trial to evaluate the efficacy of vaccination in controlling Salmonella infection and the association of Salmonella-shedding and weight gain in pigs. Can J Vet Res Rev Can Rech Veterinaire. 2010 Oct;74(4):258–63.

145. Van Der Pol L, Stork M, Van Der Ley P. Outer membrane vesicles as platform vaccine technology. Biotechnol J. 2015 Sep;10(11):1689–706.

146. Cui H, Sun Y, Lin H, Zhao Y, Zhao X. The Outer Membrane Vesicles of Salmonella enterica Serovar Typhimurium Activate Chicken Immune Cells through Lipopolysaccharides and Membrane Proteins. Pathogens. 2022 Mar 11;11(3):339.

147. Jiang X, Chu C, Wang Z, Gu J, Hong Y, Li Q, et al. Preclinical evaluation of OMVs as potential vaccine candidates against Salmonella enterica serovar Enteritidis infection. Front Cell Infect Microbiol. 2022 Oct 27;12:1037607.

148. Ochman H, Wilson AC. Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. J Mol Evol. 1987 Nov;26(1–2):74–86.

149. McClelland M. Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three Salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. Nucleic Acids Res. 2000 Dec 15;28(24):4974–86.

150. Blount ZD. The unexhausted potential of E. coli. eLife. 2015 Mar 25;4:e05826.

151. Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, Davis BR, et al. Hemorrhagic Colitis Associated with a Rare *Escherichia coli* Serotype. N Engl J Med. 1983 Mar 24;308(12):681–5.

152. Bäumler AJ. The record of horizontal gene transfer in Salmonella. Trends Microbiol. 1997 Aug;5(8):318–22.

153. Bäumler AJ, Tsolis RM, Ficht TA, Adams LG. Evolution of Host Adaptation in *Salmonella enterica*. Infect Immun. 1998 Oct;66(10):4579–87.

154. Winfield MD, Groisman EA. Evolution and Ecology of *Salmonella*. EcoSal Plus. 2004 Dec 31;1(1):10.1128/ecosalplus.6.4.6.

155. Dyer NP, Päuker B, Baxter L, Gupta A, Bunk B, Overmann J, et al. EnteroBase in 2025: exploring the genomic epidemiology of bacterial pathogens. Nucleic Acids Res. 2025 Jan 6;53(D1):D757–62.

156. Miller EA, Elnekave E, Flores-Figueroa C, Johnson A, Kearney A, Munoz-Aguayo J, et al. Emergence of a Novel Salmonella enterica Serotype Reading Clonal Group Is Linked to Its Expansion in Commercial Turkey Production, Resulting in Unanticipated Human Illness in North America. mSphere. 2020 Apr 29;5(2):e00056-20.

157. Carroll LM, Pierneef R, Mathole M, Matle I. Genomic Characterization of Endemic and Ecdemic Non-typhoidal Salmonella enterica Lineages Circulating Among Animals and Animal Products in South Africa. Front Microbiol. 2021 Oct 4;12:748611.

158. Octavia S, Wang Q, Tanaka MM, Sintchenko V, Lan R. Genomic Variability of Serial Human Isolates of Salmonella enterica Serovar Typhimurium Associated with Prolonged Carriage. J Clin Microbiol. 2015 Nov;53(11):3507–14.

159. Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, et al. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter- and intracontinental transmission events. Nat Genet. 2015 Jun;47(6):632–9.

160. Laing CR, Whiteside MD, Gannon VPJ. Pan-genome Analyses of the Species Salmonella enterica, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar. Front Microbiol. 2017 Jul 31;8:1345.

161. Aguirre-Sánchez JR, Chaidez C, Castro-del Campo N. The pangenome analysis of the environmental source *Salmonella enterica* highlights a diverse accessory genome and a distinct serotype clustering. FEMS Microbiol Lett. 2024 Jan 9;371:fnae090.

162. Cuypers WL, Meysman P, Weill FX, Hendriksen RS, Beyene G, Wain J, et al. A global genomic analysis of Salmonella Concord reveals lineages with high antimicrobial resistance in Ethiopia. Nat Commun. 2023 Jun 14;14(1):3517.

163. Commichaux S, Rand H, Javkar K, Molloy EK, Pettengill JB, Pightling A, et al. Assessment of plasmids for relating the 2020 Salmonella enterica serovar Newport onion outbreak to farms implicated by the outbreak investigation. BMC Genomics. 2023 Apr 4;24(1):165.

164. Liu CC, Hsiao WWL. Machine learning reveals the dynamic importance of accessory sequences for *Salmonella* outbreak clustering. mBio. 2025 Jan 28;e02650-24.

165. Robertson J, Schonfeld J, Bessonov K, Bastedo P, Nash JHE. A global survey of Salmonella plasmids and their associations with antimicrobial resistance. Microb Genomics. 2023 May 18;9(5).

166. McMillan EA, Jackson CR, Frye JG. Transferable Plasmids of Salmonella enterica Associated With Antibiotic Resistance Genes. Front Microbiol. 2020 Oct 8;11:562181.

167. Wahl A, Battesti A, Ansaldi M. Prophages in *Salmonella enterica* : a driving force in reshaping the genome and physiology of their bacterial host? Mol Microbiol. 2019 Feb;111(2):303–16.

168. Bobay LM, Rocha EPC, Touchon M. The Adaptation of Temperate Bacteriophages to Their Host Genomes. Mol Biol Evol. 2013 Apr;30(4):737–51.

169. Yates CR, Nguyen A, Liao J, Cheng RA. What's on a prophage: analysis of *Salmonella* spp. prophages identifies a diverse range of cargo with multiple virulence- and metabolism-associated functions. mSphere. 2024 Jun 25;9(6):e00031-24.

170. Sly LM, Guiney DG, Reiner NE. *Salmonella enterica* Serovar Typhimurium Periplasmic Superoxide Dismutases SodCI and SodCII Are Required for Protection against the Phagocyte Oxidative Burst. Infect Immun. 2002 Sep;70(9):5312–5.

171. Cardim Falcao R, Edwards MR, Hurst M, Fraser E, Otterstatter M. A Review on Microbiological Source Attribution Methods of Human Salmonellosis: From Subtyping to Whole-Genome Sequencing. Foodborne Pathog Dis. 2024 Mar 1;21(3):137–46.

172. Lupolova N, Lycett SJ, Gally DL. A guide to machine learning for bacterial host attribution using genome sequence data. Microb Genomics. 2019 Dec 1;5(12).

173. Zhang S, Li S, Gu W, Den Bakker H, Boxrud D, Taylor A, et al. Zoonotic Source Attribution of *Salmonella enterica* Serotype Typhimurium Using Genomic Surveillance Data, United States. Emerg Infect Dis. 2019 Jan;25(1).

174. Guillier L, Gourmelon M, Lozach S, Cadel-Six S, Vignaud ML, Munck N, et al. AB_SA: Accessory genes-Based Source Attribution – tracing the source of Salmonella enterica Typhimurium environmental strains. Microb Genomics. 2020 Jul 1;6(7).

175. Thomas CM, Nielsen KM. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nat Rev Microbiol. 2005 Sep 1;3(9):711–21.

176. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, et al. The use of next generation sequencing for improving food safety: Translation into practice. Food Microbiol. 2019 Jun;79:96–115.

177. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 2022 Apr;20(4):206–18.

178. Whelan FJ, Hall RJ, McInerney JO. Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome. Mol Biol Evol. 2021 Aug 23;38(9):3697–708.

179. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. PLOS Genet. 2016 Sep 12;12(9):e1006280.

180. Liu CM, Aziz M, Park DE, Wu Z, Stegger M, Li M, et al. Using source-associated mobile genetic elements to identify zoonotic extraintestinal E. coli infections. One Health. 2023 Jun;16:100518.

181. Ribot EM, Freeman M, Hise KB, Gerner-Smidt P. PulseNet: Entering the Age of Next-Generation Sequencing. Foodborne Pathog Dis. 2019 Jul;16(7):451–6.

182. Tolar B, Joseph LA, Schroeder MN, Stroika S, Ribot EM, Hise KB, et al. An Overview of PulseNet USA Databases. Foodborne Pathog Dis. 2019 Jul;16(7):457–62.

183. Brandenburg JM, Stapleton GS, Kline KE, Khoury J, Mallory K, Machesky KD, et al. *Salmonella* Hadar linked to two distinct transmission vehicles highlights challenges to enteric disease outbreak investigations. Epidemiol Infect. 2024;152:e86.

184. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, et al. An extended genotyping framework for Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun. 2016 Oct 5;7(1):12827.

185. Baker KS, Burnett E, McGregor H, Deheer-Graham A, Boinett C, Langridge GC, et al. The Murray collection of pre-antibiotic era Enterobacteriacae: a unique research resource. Genome Med. 2015 Dec;7(1):97.

186. Stapleton GS, Habrun C, Nemechek K, Gollarza L, Ellison Z, Tolar B, et al. Multistate outbreaks of salmonellosis linked to contact with backyard poultry—United States, 2015–2022. Zoonoses Public Health. 2024 Sep;71(6):708–22.

187. McMillan EA, Gupta SK, Williams LE, Jové T, Hiott LM, Woodley TA, et al. Antimicrobial Resistance Genes, Cassettes, and Plasmids Present in Salmonella enterica Associated With United States Food Animals. Front Microbiol. 2019 Apr 17;10:832.

188. Webb HE, Kim JY, Tagg KA, de la Cruz F, Peñil-Celis A, Tolar B, et al. Genome Sequences of 18 Salmonella enterica Serotype Hadar Strains Collected from Patients in the United States. Microbiol Resour Announc. 2022 Oct 20;11(10):e00522-22.

189. Tagg KA, Kim JY, Henderson B, Birhane MG, Snyder C, Boutwell C, et al. Azithromycin-resistant mph(A)-positive Salmonella enterica serovar Typhi in the United States. J Glob Antimicrob Resist. 2024 Dec;39:69–72.

190. Li C, Tate H, Huang X, Hsu CH, Harrison LB, Zhao S, et al. The spread of pESI-mediated extended-spectrum cephalosporin resistance in Salmonella serovars—Infantis, Senftenberg, and Alachua isolated from food animal sources in the United States. PLOS ONE. 2024 Mar 14;19(3):e0299354.

191. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, et al. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. PLOS ONE. 2016 Jan 22;11(1):e0147101.

192. Zhang S, Den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, et al. SeqSero2: Rapid and Improved *Salmonella* Serotype Determination Using Whole-Genome Sequencing Data. Appl Environ Microbiol. 2019 Dec;85(23):e01746-19.

193. Dyson ZA, Holt KE. Five Years of GenoTyphi: Updates to the Global *Salmonella* Typhi Genotyping Framework. J Infect Dis. 2021 Dec 20;224(Supplement_7):S775–80.

194. Cury J, Abby SS, Doppelt-Azeroual O, Néron B, Rocha EPC. Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan. In: Horizontal Gene Transfer [Internet]. New York, NY: Springer US; 2020 [cited 2022 Dec 16]. p. 265–83. (Methods in Molecular Biology; vol. 2075). Available from: http://link.springer.com/10.1007/978-1-4939-9877-7_19

195. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genomics. 2021 Nov 30;7(11).

196. Wang RH, Yang S, Liu Z, Zhang Y, Wang X, Xu Z, et al. PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations. Nucleic Acids Res. 2024 Jan 5;52(D1):D756–61.

197. Wishart DS, Han S, Saha S, Oler E, Peters H, Grant JR, et al. PHASTEST: faster than PHASTER, better than PHAST. Nucleic Acids Res. 2023 Jul 5;51(W1):W443–50.

198. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011 Apr 1;27(7):1009–10.

199. Bharat A, Petkau A, Avery BP, Chen JC, Folster JP, Carson CA, et al. Correlation between Phenotypic and In Silico Detection of Antimicrobial Resistance in Salmonella enterica in Canada Using Staramr. Microorganisms. 2022 Jan 26;10(2):292.

200. Holt KE, Phan MD, Baker S, Duy PT, Nga TVT, Nair S, et al. Emergence of a Globally Dominant IncHI1 Plasmid Type Associated with Multiple Drug Resistant Typhoid. PLoS Negl Trop Dis. 2011 Jul 19;5(7):e1245.

201. Hughes MJ, Birhane MG, Dorough L, Reynolds JL, Caidi H, Tagg KA, et al. Extensively Drug-Resistant Typhoid Fever in the United States. Open Forum Infect Dis. 2021 Dec 1;8(12):ofab572.

202. Hawkey J, Hamidian M, Wick RR, Edwards DJ, Billman-Jacobe H, Hall RM, et al. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. BMC Genomics. 2015 Dec;16(1):667.

203. Vielva L, de Toro M, Lanza VF, de la Cruz F. PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. Bioinformatics. 2017 Dec 1;33(23):3796–8.

204. Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-López R, et al. COPLA, a taxonomic classifier of plasmids. BMC Bioinformatics. 2021 Dec;22(1):390.

205. Lanza VF, Baquero F, de la Cruz F, Coque TM. AcCNET (Accessory Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks. Bioinformatics. 2017 Jan 15;33(2):283–5.

206. Hauser M, Mayer CE, Söding J. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics. 2013 Dec;14(1):248.

207. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. Proc Int AAAI Conf Web Soc Media. 2009 Mar 19;3(1):361–2.

208. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE. 2014 Jun 10;9(6):e98679.

209. Coluzzi C, Garcillán-Barcia MP, De La Cruz F, Rocha EPC. Evolution of Plasmid Mobility: Origin and Fate of Conjugative and Nonconjugative Plasmids. Mol Biol Evol. 2022 Jun 2;39(6):msac115.

210. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011 Dec;12(1):402.

211. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403–10.

212. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011 Mar 15;27(6):764–70.

213. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018 Nov 30;9(1):5114.

214. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9.

215. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021 Jul 2;49(W1):W293–6.

216. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. 2024 Jul 5;52(W1):W78–82.

217. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome: Table 1. Bioinformatics. 2015 Sep 1;31(17):2877–8.

218. Katoh K. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005 Jan 19;33(2):511–8.

219. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020 May 1;37(5):1530–4.

220. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018 Feb 1;35(2):518–22.

221. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun;14(6):587–9.

222. Leeper MM, Tolar BM, Griswold T, Vidyaprakash E, Hise KB, Williams GM, et al. Evaluation of whole and core genome multilocus sequence typing allele schemes for Salmonella enterica outbreak detection in a national surveillance network, PulseNet USA. Front Microbiol. 2023 Sep 21;14:1254777.

223. Bergsma W. A bias-correction for Cramér's and Tschuprow's. J Korean Stat Soc. 2013 Sep;42(3):323–8.

224. Hooda Y, Sajib MSI, Rahman H, Luby SP, Bondy-Denomy J, Santosham M, et al. Molecular mechanism of azithromycin resistance among typhoidal Salmonella strains in Bangladesh identified through passive pediatric surveillance. PLoS Negl Trop Dis. 2019 Nov 15;13(11):e0007868.

225. Nizamuddin S, Khan EA, Chattaway MA, Godbole G. Case of Carbapenem-Resistant *Salmonella* Typhi Infection, Pakistan, 2022. Emerg Infect Dis. 2023 Nov;29(11).

226. Duy PT, Dongol S, Giri A, Nguyen To NT, Dan Thanh HN, Nhu Quynh NP, et al. The emergence of azithromycin-resistant *Salmonella* Typhi in Nepal. JAC-Antimicrob Resist. 2020 Oct 16;2(4):dlaa109.

227. Carey ME, Jain R, Yousuf M, Maes M, Dyson ZA, Thu TNH, et al. Spontaneous Emergence of Azithromycin Resistance in Independent Lineages of *Salmonella* Typhi in Northern India. Clin Infect Dis. 2021 Mar 1;72(5):e120–7.

228. Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, et al. Evolutionary History of *Salmonella* Typhi. Science. 2006 Nov 24;314(5803):1301–4.

229. International Typhoid Consortium, Wong VK, Baker S, Connor TR, Pickard D, Page AJ, et al. An extended genotyping framework for Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun. 2016 Nov;7(1):12827.

230. Carey ME, Thi Nguyen TN, Tran DHN, Dyson ZA, Keane JA, Pham Thanh D, et al. The origins of haplotype 58 (H58) Salmonella enterica serovar Typhi. Commun Biol. 2024 Jun 28;7(1):775.

231. Achtman M. Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. Annu Rev Microbiol. 2008 Oct 1;62(1):53–70.

232. Da Silva KE, Tanmoy AM, Pragasam AK, Iqbal J, Sajib MSI, Mutreja A, et al. The international and intercontinental spread and expansion of antimicrobial-resistant Salmonella Typhi: a genomic epidemiology study. Lancet Microbe. 2022 Aug;3(8):e567–77.

233. Chiou CS, Alam M, Kuo JC, Liu YY, Wang PJ. Chromosome-Mediated Multidrug Resistance in Salmonella enterica Serovar Typhi. Antimicrob Agents Chemother. 2015 Jan;59(1):721–3.

234. Nair S, Chattaway M, Langridge GC, Gentle A, Day M, Ainsworth EV, et al. ESBL-producing strains isolated from imported cases of enteric fever in England and Wales reveal multiple chromosomal integrations of *bla* CTX-M-15 in XDR *Salmonella* Typhi. J Antimicrob Chemother. 2021 May 12;76(6):1459–66.

235. Onken A, Moyo S, Miraji MK, Bohlin J, Marijani M, Manyahi J, et al. Predominance of multidrug-resistant Salmonella Typhi genotype 4.3.1 with low-level ciprofloxacin resistance in Zanzibar. PLoS Negl Trop Dis. 2024 Apr 17;18(4):e0012132.

236. Walker J, Chaguza C, Grubaugh ND, Carey M, Baker S, Khan K, et al. Assessing the global risk of typhoid outbreaks caused by extensively drug resistant Salmonella Typhi. Nat Commun. 2023 Oct 16;14(1):6502.

237. Lima NCB, Tanmoy AM, Westeel E, de Almeida LGP, Rajoharison A, Islam M, et al. Analysis of isolates from Bangladesh highlights multiple ways to carry resistance genes in Salmonella Typhi. BMC Genomics. 2019 Dec;20(1):530.

238. Nair S, Chattaway M, Langridge GC, Gentle A, Day M, Ainsworth EV, et al. ESBL-producing strains isolated from imported cases of enteric fever in England and Wales reveal multiple chromosomal integrations of *bla* CTX-M-15 in XDR *Salmonella* Typhi. J Antimicrob Chemother. 2021 May 12;76(6):1459–66.

239. François Watkins LK, Winstead A, Appiah GD, Friedman CR, Medalla F, Hughes MJ, et al. Update on Extensively Drug-Resistant *Salmonella* Serotype Typhi Infections Among Travelers to or from Pakistan and Report of Ceftriaxone-Resistant *Salmonella* Serotype Typhi Infections Among Travelers to Iraq — United States, 2018–2019. MMWR Morb Mortal Wkly Rep. 2020 May 22;69(20):618–22.

240. Pfeifer E, Bonnin RA, Rocha EPC. Phage-Plasmids Spread Antibiotic Resistance Genes through Infection and Lysogenic Conversion. mBio. 2022 Oct 26;13(5):e01851-22.

241. Greig DR, Bird MT, Chattaway MA, Langridge GC, Waters EV, Ribeca P, et al. Characterization of a P1-bacteriophage-like plasmid (phage-plasmid) harbouring bla CTX-M-15 in Salmonella enterica serovar Typhi. Microb Genomics. 2022 Dec 20;8(12).

242. Sikorski MJ, Hazen TH, Desai SN, Nimarota-Brown S, Tupua S, Sialeipata M, et al. Persistence of Rare Salmonella Typhi Genotypes Susceptible to First-Line Antibiotics in the Remote Islands of Samoa. mBio. 2022 Oct 26;13(5):e01920-22.

243. Pattenden T, Eagles C, Wahl LM. Host life-history traits influence the distribution of prophages and the genes they carry. Philos Trans R Soc B Biol Sci. 2022 Jan 17;377(1842):20200465.

244. Nair S, Barker CR, Bird M, Greig DR, Collins C, Painset A, et al. Presence of phage-plasmids in multiple serovars of Salmonella enterica. Microb Genomics. 2024 May 8;10(5).

245. Basler C, Nguyen TA, Anderson TC, Hancock T, Behravesh CB. Outbreaks of Human *Salmonella* Infections Associated with Live Poultry, United States, 1990–2014. Emerg Infect Dis. 2016 Oct;22(10):1705–11.

246. Rounds JM, Taylor AJ, Eikmeier D, Nichols MM, Lappi V, Wirth SE, et al. Prospective Salmonella Enteritidis surveillance and outbreak detection using whole genome sequencing, Minnesota 2015-2017. Epidemiol Infect. 2020 Jun 16;148:e254.

247. Zheng J, Pettengill J, Strain E, Allard MW, Ahmed R, Zhao S, et al. Genetic diversity and evolution of Salmonella enterica serovar Enteritidis strains with different phage types. J Clin Microbiol. 2014 May;52(5):1490–500.

248. Guzinski J, Potter J, Tang Y, Davies R, Teale C, Petrovska L. Geographical and temporal distribution of multidrug-resistant Salmonella Infantis in Europe and the Americas. Front Microbiol. 2024 Feb 13;14:1244533.

249. Piña-Iturbe A, Díaz-Gavidia C, Álvarez FP, Barron-Montenegro R, Álvarez-Espejo DM, García P, et al. Genomic characterisation of the population structure and antibiotic resistance of Salmonella enterica serovar Infantis in Chile, 2009–2022. Lancet Reg Health - Am. 2024 Apr;32:100711.

250. Nichols M, Stevenson L, Whitlock L, Pabilonia K, Robyn M, Basler C, et al. Preventing Human *Salmonella* Infections Resulting from Live Poultry Contact through Interventions at Retail Stores. J Agric Saf Health. 2018;24(3):155–66.

251. Centers for Disease Control and Prevention. Data Summary: Persistent Strain of Salmonella Infantis (REPJFX01) Linked to Chicken. [Internet]. 2024 [cited 2024 Dec 2]. Available from: https://www.cdc.gov/salmonella/php/data-research/repjfx01.html

252. Fasano A, Baudry B, Pumplin DW, Wasserman SS, Tall BD, Ketley JM, et al. Vibrio cholerae produces a second enterotoxin, which affects intestinal tight junctions. Proc Natl Acad Sci. 1991 Jun 15;88(12):5242–6.

253. Liu F, Lee H, Lan R, Zhang L. Zonula occludens toxins and their prophages in Campylobacter species. Gut Pathog. 2016 Dec;8(1):43.

254. Mahendran V, Liu F, Riordan SM, Grimm MC, Tanaka MM, Zhang L. Examination of the effects of Campylobacter concisus zonula occludens toxin on intestinal epithelial cells and macrophages. Gut Pathog. 2016 Dec;8(1):18.

255. Di Pierro M, Lu R, Uzzau S, Wang W, Margaretten K, Pazzani C, et al. Zonula Occludens Toxin Structure-Function Analysis. J Biol Chem. 2001 Jun;276(22):19160–5.

256. Rathore AS, Choudhury S, Arora A, Tijare P, Raghava GPS. ToxinPred 3.0: An improved method for predicting the toxicity of peptides. Comput Biol Med. 2024 Sep;179:108926.

257. Foley SL, Kaldhone PR, Ricke SC, Han J. Incompatibility Group I1 (IncI1) Plasmids: Their Genetics, Biology, and Public Health Relevance. Microbiol Mol Biol Rev. 2021 May 19;85(2):e00031-20.

258. Kaldhone PR, Carlton A, Aljahdali N, Khajanchi BK, Sanad YM, Han J, et al. Evaluation of Incompatibility Group I1 (IncI1) Plasmid-Containing Salmonella enterica and Assessment of the Plasmids in Bacteriocin Production and Biofilm Development. Front Vet Sci. 2019 Sep 6;6:298.

259. Nichols M, Gollarza L, Palacios A, Stapleton GS, Basler C, Hoff C, et al. *Salmonella* illness outbreaks linked to backyard poultry purchasing during the COVID-19 pandemic: United States, 2020. Epidemiol Infect. 2021;149:e234.

260. Galán-Relaño Á, Valero Díaz A, Huerta Lorenzo B, Gómez-Gascón L, Mena Rodríguez M[a] Á, Carrasco Jiménez E, et al. Salmonella and Salmonellosis: An Update on Public Health Implications and Control Strategies. Animals. 2023 Nov 27;13(23):3666.

261. Hassan R, Buuck S, Noveroske D, Medus C, Sorenson A, Laurent J, et al. Multistate Outbreak of *Salmonella* Infections Linked to Raw Turkey Products — United States, 2017–2019. MMWR Morb Mortal Wkly Rep. 2019 Nov 22;68(46):1045–9.

APPENDIX: PUBLICATIONS

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Genome Sequences of 18 *Salmonella enterica* Serotype Hadar Strains Collected from Patients in the United States

Hattie E. Webb,[a,b] Justin Y. Kim,[a,b] Kaitlin A. Tagg,[a,b] Fernando de la Cruz,[c] Arancha Peñil-Celis,[c] Beth Tolar,[b] Zachary Ellison,[b,d] Colin Schwensohn,[b] Joshua Brandenburg,[b,d] Megin Nichols,[b] Jason P. Folster[b]

[a]ASRT, Inc., Suwanee, Georgia, USA
[b]Division of Foodborne, Waterborne, and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
[c]Instituto de Biomedicina y Biotecnología de Cantabria, Universidad de Cantabria, Santander, Spain
[d]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA

ABSTRACT Despite being linked to a number of recent poultry-associated outbreaks in the United States, few reference genomes are available for *Salmonella enterica* serotype Hadar. Here, we address this need by reporting 18 *Salmonella* Hadar genomes from samples collected from patients in the United States between 2014 and 2020.

Salmonella enterica serotype Hadar infections in humans in the United States increased in 2020 and 2021, compared with previous years, despite an overall decline in reported salmonellosis cases (1). Many infections occurred as part of recent outbreaks linked to either backyard poultry flocks (e.g., chickens and ducks) or consumption of ground turkey, but isolates linked to these different sources demonstrated a high degree of core genome relatedness (1, 2). Exploring the accessory genome may improve strain differentiation, as well as our understanding of the recent increase and evolution of this serotype. Here, we generated assemblies for 18 *S.* Hadar isolates collected from U.S. patients to serve as references for future investigations.

Briefly, isolates originated from clinical diagnostic laboratories or public health laboratories (PHLs) as part of the Centers for Disease Control and Prevention (CDC) national passive *Salmonella* surveillance (https://www.cdc.gov/nationalsurveillance/salmonella-surveillance.html); therefore, isolation methods varied by site (3). Isolates underwent short-read sequencing (https://www.cdc.gov/pulsenet/pathogens/wgs.html), and serotypes were confirmed using SeqSero2 v0.1 (4). Genomes were screened for resistance determinants and plasmids using the ResFinder database (downloaded 30 July 2020) (90% identity and a 50% cutoff value), the PointFinder scheme for *Salmonella* spp. (downloaded 30 August 2019) implemented in staramr v0.4.0 (5), a modified PlasmidFinder database (90% identity and 60% coverage) (https://cge.cbs.dtu.dk/services), and COPLA (6). Sequence types (STs) were determined using staramr v0.4.0 (with MLST software [https://github.com/tseemann/mlst] and PubMLST [7]). This report is a product of activities approved by the CDC internal review board (approval number 7172).

Isolates were selected for long-read sequencing based on diverse accessory genome content. Genomic DNA was extracted (Wizard genomic DNA purification kit [Promega, Madison, WI, USA], with a modification of the manufacturer's protocol) from cultures that had been incubated on tryptic soy agar-sheep blood overnight at 37°C. Libraries were prepared using the rapid barcoding kit (SQK-RBK004; Oxford Nanopore Technologies [ONT], Oxford, UK) according to the manufacturer's protocol and sequenced for 72 h on a GridION sequencing platform (R9.4.1 flow cells; ONT). Reads were base called using Guppy v4.2.2 and filtered for quality using MinKNOW (ONT). Hybrid assemblies were generated, polished, circularized, and rotated using Unicycler v0.4.8 (conservative option) (8); corresponding Illumina short reads that had been previously generated at the PHL (BioNumerics v7.6 [Applied Maths NV,

Editor David Rasko, University of Maryland School of Medicine

Address correspondence to Jason P. Folster, jpfolster@cdc.gov.

The authors declare no conflict of interest.

Received 22 June 2022
Accepted 16 August 2022
Published 29 August 2022

TABLE 1 Summary information for 18 *Salmonella enterica* serotype Hadar (ST33) genomes from samples collected from patients in the United States

Strain	Collection yr	Accession no.				PTU (plasmid replicon)[a]	Antimicrobial resistance determinants	Short-read findings		Long-read findings				GC content (%)	Total size (bp)	Mean coverage (×)
		BioSample	Short-read SRA	Long-read SRA	GenBank			Mean read length (bp)	No. of reads	Contig N_{50} (bp)	Mean read length (bp)	No. of reads	No. of contigs			
2014AM-1331	2014	SAMN05596322	SRR4044556	SRR19768540	CP093126	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	295.9	1,617,299	4,702,738	4,302	391,542	3	52.26	4,741,847	170
					CP093127	PTU-N3	—									
					CP093128	PTU-E10	—									
2014AM-2067	2014	SAMN05596277	SRR4044454	SRR19768539	CP093122	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	286.7	1,838,942	4,763,043	5,225.5	165,516	4	52.22	4,777,204	192
					CP093123	PTU-E22	—									
					CP093124	PTU-E1 (ColE1)	—									
					CP093125	PTU-E19 (Col156)[b]	—									
2015AM-0414	2015	SAMN07268462	SRR5740069	SRR19768530	CP093120	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	278.3	1,430,061	4,801,674	5,580.5	174,520	2	52.25	4,805,578	148
					CP093121	PTU-E19 (Col156)	—									
2015AM-0511	2015	SAMN07415498	SRR5868150	SRR19768529	CP093140	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	274.7	1,566,516	4,778,352	5,993	223,197	7	52.21	4,805,332	162
					CP093141	PTU-E3 [Col440II,Col (pHAD28)]	bla_{TEM-1C}									
					CP093142	PTU-E7 (Col156)	—									
					CP093143	PTU-E1 (ColE1)	—									
					CP093144	PTU-E58 (ColE1)	—									
					CP093145	PTU-E58	—									
					CP093146	PTU-E11 (ColpVC)	—									
2016AM-0673	2016	SAMN13512702	SRR10607636	SRR19768528	CP093116	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	277.9	1,363,160	4,703,663	5,196.8	324,412	4	52.29	4,712,319	143
					CP093117	PTU-E10	—									
					CP093118	PTU-NA	—									
					CP093119	PTU-E11	—									
2016K-0377	2016	SAMN05250424	SRR3667804	SRR19768533	CP093076	—	—	274.0	1,081,951	4,685,556	5,555.7	234,342	5	52.19	4,730,499	102
					CP093077	PTU-X1 (IncX1)	—									
					CP093078	PTU-E7 (Col156)	—									
					CP093079	PTU-E3 [Col(pHAD28)]	—									
					CP093080	PTU-E11 (ColpVC)	—									
2017AM-0493	2017	SAMN17129770	SRR13277812	SRR19768526	CP093109	—	—	298.6	866,278	4,819,027	5,104.1	104,476	3	52.18	4,829,291	90
					CP093110	PTU-E22	—									
					CP093111	PTU-E1 (ColE1)	—									
2021K-0017	2020	SAMN17478013	SRR13496954	SRR19768531	CP093072	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	272.7	662,153	4,711,128	4,785.5	459,695	1	52.27	4,711,128	67
PNUSAS002131	2016	SAMN04961841	SRR3499746	SRR19768527	CP093112	—	—	289.4	2,242,755	4,683,655	5,252.9	345,694	4	52.17	4,807,169	173
					CP093113	PTU-I1 (IncI1-Iγ)	—									
					CP093114	PTU-X1 (IncX1)	—									
					CP093115	PTU-E10 (Col8282)	—									
					—	(ColpVC)[c]	—									
					—	[Col(pHAD28)][c]	—									
PNUSAS018090	2017	SAMN07427456	SRR6014222	SRR19768524	CP093096	—	—	294.5	977,909	4,685,429	5,517.2	241,397	7	52.17	4,764,800	88
					CP093097	PTU-X1 (IncX1)	—									
					CP093098	PTU-N3	—									
					CP093099	PTU-NA [Col(pHAD28)]	—									
					CP093100	PTU-E19 (Col156)	—									
					CP093101	PTU-E3	—									
					CP093102	PTU-E11 (ColpVC)	—									
PNUSAS021403	2017	SAMN07521433	SRR5951569	SRR19768525	CP093103	—	—	284.6	828,218	4,685,480	4,125.7	280,865	6	52.17	4,837,812	80
					CP093104	PTU-I1 (IncI1-Iγ)	*aadA2*, *ant(3″)-I*, *cmlA1*, *sul3*									
					CP093105	PTU-X1 (IncX1)	—									
					CP093106	PTU-E3	—									
					CP093107	PTU-E3 [Col(pHAD28)]	—									
					CP093108	PTU-E11 (ColpVC)	—									
PNUSAS037609	2018	SAMN08815166	SRR6916443	SRR19768523	CP093093	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	243.3	960,837	4,771,555	5,272.1	147,247	3	52.24	4,775,081	90
					CP093094	PTU-E58	—									
					CP093095	PTU-NA	—									
PNUSAS039582	2018	SAMN09011259	SRR7093175	SRR19768538	CP093088	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	276.7	2,708,767	4,801,678	6,900.7	314,036	5	52.27	4,908,883	252

(Continued on next page)

TABLE 1 (Continued)

Strain	Collection yr	Accession no.				PTU (plasmid replicon)[a]	Antimicrobial resistance determinants	Short-read findings		Long-read findings				GC content (%)	Total size (bp)	Mean coverage (×)
		BioSample	Short-read SRA	Long-read SRA	GenBank			Mean read length (bp)	No. of reads	Contig N_{50} (bp)	Mean read length (bp)	No. of reads	No. of contigs			
					CP093089	PTU-NA (IncI1-Iγ)	—									
					CP093090	PTU-E1 [Col(pHAD28)][b]	—									
					CP093091	PTU-E58 (ColE1)	—									
					CP093092	PTU-E10 (Col8282)	—									
					—	[Col(pHAD28)][c]	—									
					—	[Col(pHAD28)][c]	—									
PNUSAS067730	2019	SAMN11029788	SRR8643922	SRR19768537	CP093086	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	272.5	859,691	4,763,126	4,853.2	357,105	2	52.23	4,767,759	88
					CP093087	PTU-E11 (ColE1)	—									
PNUSAS074905	2019	SAMN11618493	SRR9040334	SRR19768532	CP093073	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	230.7	1,282,945	4,766,350	7,143.7	112,075	3	52.21	4,887,846	102
					CP093074	PTU-I1 (IncI1-Iγ)	*aac(3)-IId*, *ant(3″)-Ia*, bla_{TEM-1B}, *tet*(A)									
					CP093075	PTU-E1 (ColE1)	—									
PNUSAS127695	2019	SAMN13795856	SRR10856733	SRR19768536	CP093135	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	263.0	612,925	4,801,677	5,894.7	376,137	5	52.24	4,819,091	62
					CP093136	PTU-E1 [Col(pHAD28)]	—									
					CP093137	PTU-E3 [Col(pHAD28)]	—									
					CP093138	PTU-E58 (ColE1)	—									
					CP093139	PTU-E1 (ColE1)	—									
PNUSAS147811	2020	SAMN15239347	SRR12017862	SRR19768535	CP093084	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	271.5	755,934	4,719,084	5,320.8	289,486	2	52.2	4,813,237	75
					CP093085	PTU-I1 (IncI1-Iγ)	*tet*(B)									
PNUSAS148096	2020	SAMN15249649	SRR12024307	SRR19768534	CP093081	—	*aph(3″)-Ib*, *aph(6)-Id*, *tet*(A)	129.2	737,381	4,708,042	4,754.8	293,015	3	52.23	4,802,435	41
					CP093082	PTU-I1 (IncI1-Iγ)	—									
					CP093083	PTU-NA	—									

[a] PTU, plasmid taxonomic unit; PTU-NA, plasmid taxonomic unit not assigned; —, no information.
[b] Plasmid assigned using updated version of COPLA.
[c] Plasmid replicon missing from long-read assembly but present in short reads.

Sint-Martens-Latem, Belgium] quality control metrics: quality score, ≥30; coverage, ≥30×) were accessed through NCBI (Table 1). Assemblies were quality controlled using QUAST v5.0.2 (9) and BLASTn v2.9.0 (10) and were annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v6.1 (11). Default parameters were used for all software unless otherwise specified.

All 18 *S.* Hadar strains were found to be ST33. Resistance determinants and plasmid types are summarized in Table 1. The most common resistance genes were *aph(3″)-Ib*, *aph(6)-Id*, and *tet*(A), which were always located on the chromosome ($n = 13$). When present, other resistance genes were associated with IncI1-Iγ or Col(pHAD28) plasmids. High levels of small plasmids with no known resistance genes were observed, some of which had not been previously characterized, as indicated by small, circular genetic elements not containing a known plasmid replicon. More generally, the hybrid assembly method employed here recovered small plasmids at a higher rate than did long-read-only assembly methods (data not shown). For two genomes, however, some small plasmids were not recovered despite the use of a hybrid assembly method (Table 1), a known artifact of the long-read sequencing process (12).

Data availability. The sequences discussed here have been deposited in GenBank and the SRA under the accession numbers listed in Table 1.

ACKNOWLEDGMENTS

This work was supported through the CDC.

The findings and conclusions of this article are those of the authors and do not necessarily represent the views of the CDC.

We acknowledge the state and local PHLs that participated in the National Antimicrobial Resistance Monitoring System (NARMS) and PulseNet.

REFERENCES

1. Nichols M, Gollarza L, Palacios A, Stapleton GS, Basler C, Hoff C, Low M, McFadden K, Koski L, Leeper M, Brandenburg J, Tolar B. 2021. *Salmonella* illness outbreaks linked to backyard poultry purchasing during the COVID-19 pandemic: United States, 2020. Epidemiol Infect 149:e234. https://doi.org/10.1017/S0950268821002132.
2. Centers for Disease Control and Prevention. 2021. *Salmonella* outbreak linked to ground turkey. https://www.cdc.gov/salmonella/hadar-04-21/index.html. Accessed 31 January 2022.
3. Tolar B, Joseph LA, Schroeder MN, Stroika S, Ribot EM, Hise KB, Gerner-Smidt P. 2019. An overview of PulseNet USA databases. Foodborne Pathog Dis 16:457–462. https://doi.org/10.1089/fpd.2019.2637.
4. Zhang S, den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng X. 2019. SeqSero2: rapid and improved *Salmonella* serotype determination using whole-genome sequencing data. Appl Environ Microbiol 85:e01746-19. https://doi.org/10.1128/AEM.01746-19.
5. Bharat A, Petkau A, Avery BP, Chen JC, Folster JP, Carson CA, Kearney A, Nadon C, Mabon P, Thiessen J, Alexander DC, Allen V, El Bailey S, Bekal S, German GJ, Haldane D, Hoang L, Chui L, Minion J, Zahariadis G, Domselaar GV, Reid-Smith RJ, Mulvey MR. 2022. Correlation between phenotypic and in silico detection of antimicrobial resistance in *Salmonella enterica* in Canada using Staramr. Microorganisms 10:292. https://doi.org/10.3390/microorganisms10020292.
6. Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-López R, de la Cruz F. 2021. COPLA, a taxonomic classifier of plasmids. BMC Bioinformatics 22:390. https://doi.org/10.1186/s12859-021-04299-x.
7. Jolley KA, Maiden MC. 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. https://doi.org/10.1186/1471-2105-11-595.
8. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
9. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34:i142–i150. https://doi.org/10.1093/bioinformatics/bty266.
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
11. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI Prokaryotic Genome Annotation Pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569.
12. Wick RR, Judd LM, Wyres KL, Holt KE. 2021. Recovery of small plasmid sequences via Oxford Nanopore sequencing. Microb Genom 7:000631. https://doi.org/10.1099/mgen.0.000631.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Open Peer Review | Public Health | Research Article

Mobile genetic elements define the non-random structure of the *Salmonella enterica* serovar Typhi pangenome

Arancha Peñil-Celis,[1] Kaitlin A. Tagg,[2] Hattie E. Webb,[2] Santiago Redondo-Salvo,[1,3] Louise Francois Watkins,[2] Luis Vielva,[4] Chelsey Griffin,[2,5] Justin Y. Kim,[2,6] Jason P. Folster,[2] M. Pilar Garcillan-Barcia,[1] Fernando de la Cruz[1]

AUTHOR AFFILIATIONS See affiliation list on p. 16.

ABSTRACT Bacterial relatedness measured using select chromosomal loci forms the basis of public health genomic surveillance. While approximating vertical evolution through this approach has proven exceptionally valuable for understanding pathogen dynamics, it excludes a fundamental dimension of bacterial evolution—horizontal gene transfer. Incorporating the accessory genome is the logical remediation and has recently shown promise in expanding epidemiological resolution for enteric pathogens. Employing *k*-mer-based Jaccard index analysis, and a novel genome length distance metric, we computed pangenome (i.e., core and accessory) relatedness for the globally important pathogen *Salmonella enterica* serotype Typhi (Typhi), and graphically express both vertical (homology-by-descent) and horizontal (homology-by-admixture) evolutionary relationships in a reticulate network of over 2,200 U.S. Typhi genomes. This analysis revealed non-random structure in the Typhi pangenome that is driven predominantly by the gain and loss of mobile genetic elements, confirming and expanding upon known epidemiological patterns, revealing novel plasmid dynamics, and identifying avenues for further genomic epidemiological exploration. With an eye to public health application, this work adds important biological context to the rapidly improving ways of analyzing bacterial genetic data and demonstrates the value of the accessory genome to infer pathogen epidemiology and evolution.

IMPORTANCE Given bacterial evolution occurs in both vertical and horizontal dimensions, inclusion of both core and accessory genetic material (i.e., the pangenome) is a logical step toward a more thorough understanding of pathogen dynamics. With an eye to public, and indeed, global health relevance, we couple contemporary tools for genomic analysis with decades of research on mobile genetic elements to demonstrate the value of the pangenome, known and unknown, annotated, and hypothetical, for stratification of *Salmonella enterica* serovar Typhi (Typhi) populations. We confirm and expand upon what is known about Typhi epidemiology, plasmids, and antimicrobial resistance dynamics, and offer new avenues of exploration to further deduce Typhi ecology and evolution, and ultimately to reduce the incidence of human disease.

KEYWORDS *Salmonella* Typhi, pangenome, plasmids, antimicrobial resistance

Enteric foodborne surveillance has benefited from the integration of whole genome sequencing (WGS) data with traditional epidemiological methods, rapidly improving outbreak detection, source attribution, and our understanding of antimicrobial resistance (AMR) (1, 2). Many genomic surveillance and outbreak detection systems rely on measuring core-genome relatedness; for example, the United States national molecular subtyping network for foodborne disease surveillance, PulseNet USA, uses core-genome multi-locus sequence typing to detect single nucleotide polymorphisms (SNPs) or indels (insertions and deletions) within specific core loci (3).

Editor Sima Tokajian, Lebanese American University, Byblos, Lebanon

Ad Hoc Peer Reviewer Álvaro San Millán, Centro Nacional de Biotecnología-CSIC, Madrid, Spain

Address correspondence to Fernando de la Cruz, delacruz@unican.es, or M. Pilar Garcillan-Barcia, garcilmp@unican.es.

Arancha Peñil-Celis and Kaitlin A. Tagg contributed equally to this article. Author order was determined both alphabetically and in order of increasing seniority.

The authors declare no conflict of interest.

See the funding table on p. 17.

Received 15 March 2024
Accepted 30 June 2024
Published 26 July 2024

Core genome-based methods approximate vertical evolution—homology-by-descent—and offer a genetic framework for understanding epidemiological patterns (2). However, non-core genes—the accessory genome—reflect horizontal gene transfer (HGT), another fundamental dimension of bacterial evolution (4).

The accessory genome is composed of plasmids, phages, and a variety of mobile genetic elements (MGEs) that exist as autonomous or chromosomally integrated molecules (5). It is highly variable and flexible, enabling rapid adaptation of bacterial species to new niches and environmental selection pressures (4, 6). While inclusion of accessory content is often thought to confound genomic epidemiological analyses (2, 7), the accessory genome is not randomly structured, nor is it under neutral selection (4, 6, 8, 9). In fact, accessory genome content has been shown to improve our understanding of bacterial population structure (10) and offer deeper epidemiological resolution for foodborne pathogen investigations (7, 11–13). For *Salmonella enterica*, demonstrated concordance of core and accessory phylogenies suggests routine incorporation of accessory genetic material could add substantial value to genomic surveillance (14).

In this genomic epidemiology analysis, we characterize core and accessory genome (i.e., pangenome) structure and diversity within *S. enterica* serovar Typhi (herein referred to as Typhi), using *k*-mer-based Jaccard index (JI) analysis coupled with MGE characterization. JI is a common proximity measurement used to compute the similarity between two objects, with wide use in numerous domains, such as ecology (15, 16), text mining (17, 18), and genome comparison (19–22). It is particularly useful for comparing and discriminating between very similar genomes (e.g., within a clonal serotype such as Typhi) because it is optimized for values well over 99.9% average nucleotide identity (ANI). Typhi is a globally distributed human pathogen, causing an estimated 9.2 million cases of typhoid fever annually (23); and while cases in the United States are relatively infrequent (24), their tendency to be travel-related means detailed epidemiological information is often unavailable. This limits our ability to pinpoint—or even begin to resolve—outbreaks, and to identify "short-" and "long-cycle" transmission pathways of Typhi (25). Analyzing the largest collection of U.S. Typhi genomes to date, we leverage the Typhi pangenome to explore and deduce the epidemiology of U.S, Typhi cases, providing a generalized overview of the pangenome structure of this pathogen, and offering insights into its evolutionary and ecological dynamics.

RESULTS

Pangenome analysis of U.S. Typhi population

JI is a measure of similarity between genomes. It is determined by dividing the size of the intersection of two sets of nucleotide *k*-mers by the size of their union. This metric captures both SNPs, either due to point mutations or recombination, and gene content differences that arise as the result of gain and loss of genetic material (indels), although it does not account for differences due to duplicated sequences (Fig. S1 in S1 Appendix). The pairwise genome similarities can be represented in an undirected network in which the nodes (genomes) are connected if the pairwise JI equals or exceeds the set JI threshold. At the initial network stage, genomes sharing any JI value greater than 0 will be linked by an edge, resulting in most genomes connected in a single component. By increasing the stringency of the JI threshold, separate connected components emerge.

Using BinDash (19), we calculated the exact JI values from pairwise comparisons of a 2,392 Typhi genome data set, comprising 2,272 study genomes that were isolated in the United States from 2008 to 2021 and assembled in this work, along with 120 RefSeq reference genomes (Table S1). Their JI value distribution showed that most comparisons (99.84%) produced JI values greater than 0.90 (Fig. S2 in S1 Appendix). To ensure the strength of cluster definition, networks should exhibit a community structure characterized by highly internally connected subgraphs and sparser connections between them. Furthermore, to provide informative insights, most genomes should belong to non-singleton communities (Fig. S3 in S1 Appendix). Following this approach, the optimum threshold for analyzing Typhi genomes was set at JI = 0.983 (see Materials and Methods).

Since both SNPs and indels are reflected in JI, their individual contribution cannot be estimated from the JI value alone. As SNPs do not contribute to genome length, but indels do, a new unit of measurement, genome length distance (GLD), was defined. It uses the difference between the unique *k*-mer counts of two given genomes as a proxy for genome size variation due to differential gene content (supporting information in S1 Appendix). On top of a given JI threshold, pairwise GLD values can be used as a proxy of indels to emphasize (and further explore) differences in genome size. At thresholds JI = 0.983 and GLD = 0.05, between-group differences can be explained by indels larger than 50 kb in size, or by ≥2,050 SNPs across the entire genome, or a mix of both (supporting information text in S1 Appendix).

At the abovementioned thresholds, the 2,392 Typhi genomes self-organized into 17 distinct clusters according to the Louvain method, named JI groups A–Q, with only 38/2,392 (1.6%) nodes not assigned (singletons or JI clusters with less than five members) (Fig. 1A). The relatedness of genomes within each cluster is >99.8% ANI (Fig. S4 in S1 Appendix). JI group A was the largest (*n* = 1,320/2,392), with all other JI groups represented by at least five genomes (Table 1). Three of the largest JI groups (A, B, and C) were further divided into JI subgroups using an increased JI threshold. JI-A subgroups A1 to A17 were defined at JI = 0.995; JI-B subgroups B1 to B3 at JI = 0.986; and JI-C subgroups C1 to C6 at JI = 0.997 (Fig. S5 in S1 Appendix).

Pangenome population structure of U.S. Typhi is non-random

Exploring the U.S. Typhi data set, we found that autonomous and integrated MGEs (detected by indels) are ubiquitous in Typhi. A proxy for integrative and conjugative elements (ICEs) and integrative and mobilizable elements (IMEs), a MOB relaxase gene, was detected in 99.5% (*n* = 2,380/2,392) of the isolates (Fig. 1B). JI groups often correlated with the presence/absence of known MGEs. For example, several large (>80 kb) autonomous plasmids were found to underpin JI group definitions: members of JI groups B and J all contain plasmids belonging to PTU-E50 (average size 90 kb), JI group C contains PTU-E18 (average size 107 kb), JI group D contains PTU-HI1A (average size 217 kb), and JI group K contains PTU-Y plasmids (average size 100 kb) (Table 2; Fig. 1C). Plasmids <40 kb, such as PTU-N1 and PTU-X1 in JI groups A, H, and N, did not define JI groups (at thresholds JI = 0.983 and GLD = 0.05) due to their relatively small

TABLE 1 Summary of JI group information for 2,272 U.S. CDC and 120 RefSeq200 genomes

JI group	Count[a]	%[b]	GenoTyphi primary cluster
A	1,320	55.1	0, 1, 2, 3, 4
B	114	4.8	4
C	265	11.1	2, 3, 4
D	26	1.1	3, 4
E	5	0.2	4
F	39	1.6	0
G	6	0.3	2
H	225	9.4	2
I	17	0.7	2, 3, 4
J	11	0.5	3
K	8	0.3	4
L	11	0.5	2
M	133	5.6	3
N	90	3.8	2, 4
O	11	0.5	3
P	58	2.4	2
Q	15	0.6	2
Singletons	38	1.6	0, 2, 3, 4

[a]Number of genomes present in each JI group.
[b]Percentage of genomes from the total data set that belong to each JI group.

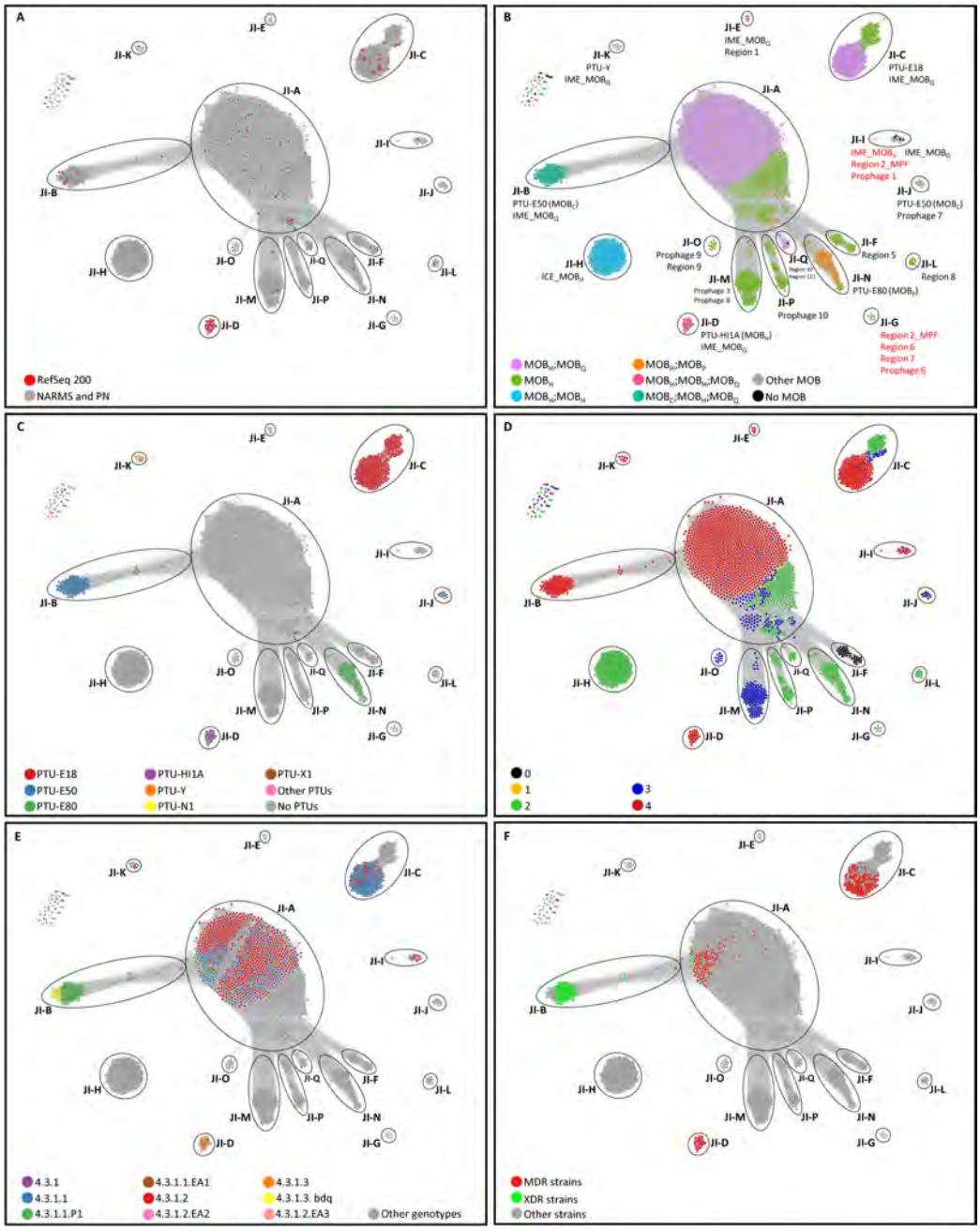


FIG 1 Distribution of Typhi genomes by JI. The networks contain 2,392 nodes, connected when JI ≥0.983 and GLD ≤0.05. Seventeen clusters (named JI-A to JI-Q) are indicated by circles. (A) Distribution of references and CDC study genomes in the JI groups. Nodes depicted in red represent RefSeq200 genomes (references) and those in gray represent study genomes. (B) Distribution of MOB relaxases in the JI groups. Nodes are colored according to the MOB relaxase class present in each genome. Information on the PTUs, as well as other accessory elements present in >90% of the members of a given JI group, are included in black letters when present in a given cluster or in red letters when absent (see also Fig. S6 in S1 Appendix). (C) Distribution of PTUs in the JI groups. Nodes are colored according to the PTUs present in each genome. (D) Distribution of GenoTyphi primary clades in the JI groups. Nodes are colored according to the GenoTyphi primary clades. (E) Distribution of the 4.3.1 GenoTyphi genotype in the JI groups. Nodes are colored according to the lineages and sublineages of the 4.3.1 genotype. (F) Distribution of multidrug-resistant (MDR) and extensive drug-resistant (XDR) genomes. Nodes are colored according to AMR categories. A Gephi file containing the JI network is available at https://github.com/PenilCelis/Salmonella_Typhi_JINA.

size. Many unknown MGEs and accessory regions, identified through BLAST, were also responsible for the genetic difference between JI groups (Fig. S6 in S1 Appendix and S2 Appendix); JI-E is defined by the presence of a 49 kb region of unknown function, while JI-P members all carry a unique 44 kb phage element (prophage 10) (Fig. 1B; Fig. S6 and S7A in S1 Appendix; S2 Appendix). Each JI group was found to contain a unique complement of accessory genome elements, many of which were undetectable by current routine methods.

To further explore the contribution of MGEs in the JI group clustering, an *in silico* experiment was carried out by removing them from reference genomes. PTU-E50 plasmids present in the B subgroups and SGI11 encoded in B1 and A3 references were eliminated. The "cured" genomes segregated from their original JI groups and associated with the JI-A1 genomes in the network (Fig. 2A). The progressive reintroduction of SGI11 (Fig. 2B) and PTU-E50 sequences (Fig. 2C) led to the partition of JI-A3, JI-B1, JI-B2, and JI-B3 genomes from the JI-A1 group, rendering new clusters. In a similar experiment, large plasmids in JI-B, JI-C, JI-D, JI-J, and JI-K reference genomes were shown to shape these JI groups (Fig. S8 in S1 Appendix), emphasizing their contribution to Typhi pangenome structure.

GenoTyphi genotypes (27, 28) were visualized against JI groups to compare phylogenetic context to pangenome groupings. Most JI groups (n = 12/17) associated with a single subclade, clade, or GenoTyphi primary clade (Table 1; Fig. 1D and E; Table S1), whereas JI-A, JI-C, JI-D, JI-I, and JI-N contained isolates that fell into two or more primary clades. JI-A contains genomes from the ancestral (0) and all four Typhi primary clades (1, 2, 3, and 4), with each primary clade mostly confined to distinct areas in the

TABLE 2 Characteristics of plasmids identified in JI groups

JI groups (number of genomes)	Number of plasmids	PTU (grade host range)[a]	Plasmid replicons[b]	Plasmid MOB type/MPF (transmissibility)[b]	AMR determinants[b]	Average plasmid size (kb)
JI-A (1,320)	6	PTU-E80 (IV)	IncX1	MOB_P/– (mobilizable)	–[c]	18
	4	PTU-N1 (III)	IncN	MOB_F/MPF_T (conjugative)	–	40
	4	PTU-X1 (III)	IncX1	MOB_P/MPF_T (conjugative)	–	30
	1	PTU-X3 (III)	IncX3	MOB_P/MPF_T (conjugative)	–	44
	1	PTU-E73 (IV	IncFII(pCRY)	MOB_C/MPF_T (conjugative)	–	21
JI-B (114)	114	PTU-E50 (III)	IncY, IncFIB(K)	MOB_C/MPF_T (conjugative)	blaTEM-1B, qnrS1, sul2, tet(A), aph(3")-Ib, aph(6)-Id, dfrA14, blaCTX-M-15, blaCTX-M-88	90
JI-C (265)	265	PTU-E18 (IV)	IncFIB(pHCM2)	–/– (non-tranmissible by conjugation)	–	107
JI-D (26)	26	PTU-HI1A (IV)	IncHI1A, IncHI1B(R27), IncFIA(HI1)	MOB_H/MPF_F (conjugative)	aph(3")-Ib, aph(6)-Id, blaTEM-1B, catA1, dfrA7, qacE, sul1, sul2, tet(B)	217
JI-K (8)	8	PTU-Y (III)	IncY, p0111	–/– (non-tranmissible by conjugation)	blaCTX-M-15	100
JI-J (11)	11	PTU-E50 (III)	IncY	MOB_C/MPF_T (conjugative)	aph(3")-Ib, aph(6)-Id, blaTEM-1B, dfrA14, sul2, tet(A)	115
JI-H (225)	1	PTU-N1 (III)	IncN	MOB_F/MPF_T (conjugative)	aph(3")-Ib, aph(6)-Id, blaTEM-1B, dfrA14, sul2, tet(A)	50
JI-N (90)	78	PTU-E80(IV)	IncX1	MOB_P/– (mobilizable)	–	25

[a]PTU and the grade host range were assigned using COPLA (Materials and Methods). The grade host range definition is specified in reference 26.
[b]Plasmid replicons, MOB class, MPF type, and AMR determinants were calculated, respectively, using PlasmidFinder, MOBscan, CONJScan, and ResFinder, according to the parameters specified in Materials and Methods.
[c]–, the absence of the specific trait indicated in the column.

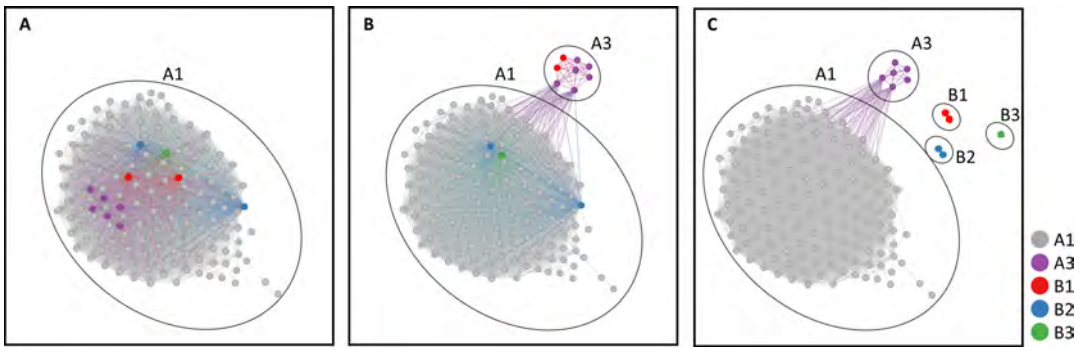


FIG 2 Effect of MGEs on the JI-based Typhi genome clustering. The networks contain 154 nodes, connected when JI ≥0.995. Nodes are colored according to the original JI subgroup of each genome (see also Fig. S5 in S1 Appendix). (A) Clustering of genomes deprived of PTU-E50 and SGI11. PTU-E50 plasmids originally present in genomes of the B1, B2, and B3 subgroups, as well as the chromosomally inserted element SGI11, encoded also in genomes of the B1 and A3 subgroups, were removed from the genome sequences. The resulting "pruned" genomes were used to calculate pairwise genome similarities. Genomes from all subgroups reassociate in a single cluster. (B) Clustering of genomes deprived of PTU-E50. The SGI11 elements were restituted to the A3, and B1 genomes and the network were recalculated. A3 and B1 genomes broke away from the previous cluster and grouped together. (C) Clustering of genomes with SGI11 and PTU-E50. The PTU-E50 plasmids were restituted to the B1, B2, and B3 genomes. The rebuilt network shows the emergence of distinctive clusters.

network map of JI-A (Fig. 1D). At the higher threshold used to determine JI-A subgroups (JI = 0.995), GenoTyphi lineages were largely resolved into their own cluster, with 15 of 17 JI-A subgroups containing genomes of a single GenoTyphi subclade, clade, or primary clade (Fig. S9 in S1 Appendix, Table S1). Similarly, group JI-C has members of primary clades 2, 3, and 4, but at the JI subgroup resolution, all members within each of the six JI-C subgroups contain a single GenoTyphi subclade or clade (Table S1). Membership to a JI group does not necessarily imply vertical descent (as defined by GenoTyphi), since JI grouping aggregates genomes of distinct vertical lineages if they share substantial accessory genome material and partitions genomes of the same vertical lineage into separate groups according to their accessory genome content. However, pangenome groupings did tend to align with phylogenetic lineage, especially at the level of subgroups (n = 32/40 JI group or subgroup contained a single GenoTyphi clade or subclade). Thus, coupling of pangenomic and phylogenetic methods can simultaneously offer information on horizontal and vertical evolutionary dimensions. In fact, coupling information from GenoTyphi and MOB typing methods already accounts for a substantial proportion of the genetic variance of JI groups (combined variance partitioning R^2 = 0.725), suggesting much of the Typhi pangenome can be effectively identified with existing methods.

U.S. Typhi pangenome structure aligns with and expands on known AMR and epidemiological patterns

Genetic and epidemiological metadata was mapped onto the JI network to determine if JI grouping could easily detect known AMR and epidemiological patterns. For example, extensively drug-resistant (XDR) Typhi (genotype 4.3.1.1.P1) among patients with a history of travel to Pakistan first appeared in the United States in 2018 (29). JI-B1 genomes first appeared in the United States in 2018 (Fig. S10A in S1 Appendix), are all genotype 4.3.1.1.P1, carry an IncY (PTU-E50) plasmid with $bla_{\text{CTX-M-15}}$, and are significantly associated with travel to Pakistan (P < 0.01, chi-squared test of independence), despite limited travel data for this group (n = 47/88 have any travel information available) (Table S1; Fig. S11A and B in S1 Appendix). Genomes of the 4.3.1.1.P1 lineage also fall into JI-A (specifically JI-A1 and JI-A3), representing XDR 4.3.1.1.P1 strains that have recently lost the IncY (PTU-E50) plasmid and integrated $bla_{\text{CTX-M-15}}$, the gene that confers additional ceftriaxone resistance and defines XDR, into their chromosome (30). Thus, both the original XDR 4.3.1.1.P1 Pakistan outbreak strain with an IncY (PTU-E50) plasmid (31) and its recent XDR variants (without the plasmid) are quickly identifiable

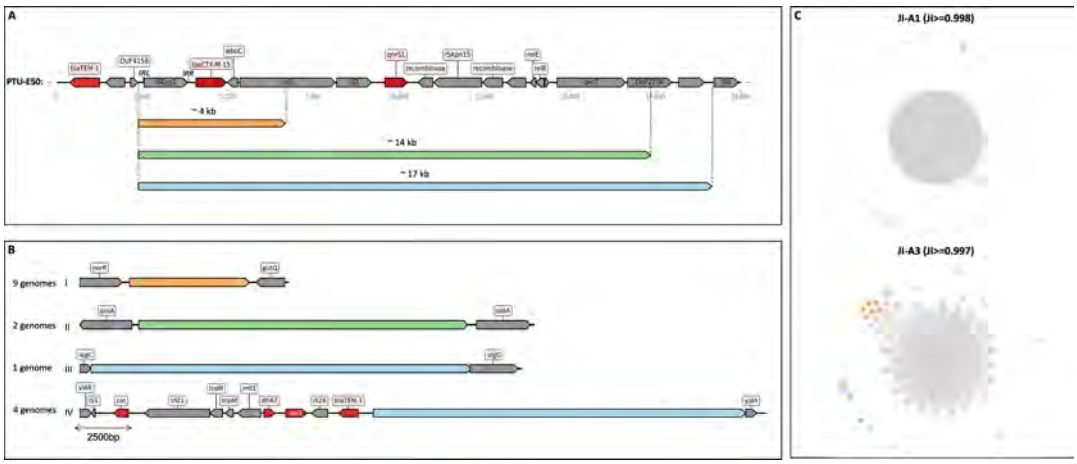


FIG 3 Genomic context of *bla*$_{CTX-M-15}$. (A) The genetic vicinity of *bla*$_{CTX-M-15}$ in PTU-E50 plasmids. The region containing the *bla*$_{CTX-M-15}$ gene of plasmid NZ_CP046430 is depicted. Genes are represented by arrows and those encoding AMR are colored in red. Below, three arrows of different sizes, indicated by different colors, represent the PTU-E50 regions that are found integrated into the chromosomes. (B) Chromosomal integration sites of the *bla*$_{CTX-M-15}$-containing regions (I–IV). Insertion site I locates between genes *norR* and *gutQ*; site II between genes *phsA* and *sopA*; site III interrupts gene *stgC*; site IV resides within SGI11. (C) JI networks of subgroups JI-A1 (upper panel) and JI-A3 (lower panel). Nodes colored in orange, green, and blue indicate genomes containing the different *bla*$_{CTX-M-15}$-encoding regions.

in the JI network as JI-B1 and within JI-A1/JI-A3, respectively, supporting what is known about this lineage.

The chromosomal genetic context of *bla*$_{CTX-M-15}$ was further explored in JI-A1 (n = 4) and JI-A3 (n = 12) genomes. Three different sized regions of the original IncY (PTU-E50) plasmid were detected, likely captured and mobilized by IS*Ecp1* (Fig. 3A). ISMapper identified four possible IS*Ecp1*-*bla*$_{CTX-M-15}$ insertion sites (I–IV) (Fig. 3B). Insertion sites were confirmed either by direct analysis of the *bla*$_{CTX-M-15}$-containing contigs (insertion sites I-III) or with additional long-read sequencing (insertion site IV) (PNUSAS198714, SAMN18813804). Despite being highly related by phylogenetic analysis (Fig. 4; Fig. S12), the difference in size and chromosomal location of these insertions confirms the PTU-E50-borne *bla*$_{CTX-M-15}$ inserted in at least four independent events. Genomes with a given, unique integration site clustered together in the JI network when the JI threshold was increased (i.e., enhanced discrimination between genomes) (Fig. 3C); however, these clusters were not distinct enough to be used for prediction of the genetic context from the JI network alone.

Multidrug resistant (MDR) in Typhi emerged several decades ago, driven by the expansion of a 4.3.1 (previously H58) strain carrying SGI11 [containing *bla*$_{TEM-1}$, *catA1*, *aph(3')-Ib* (*strA*), *aph(6)-Id* (*strB*), *sul1*, *sul2*, and *dfrA7*, a mercury resistance operon, and the antiseptic resistance gene *qacEΔ1*] (32) on an IncHI1 plasmid (33). Subsequent degradation of SGI11 (32), as well as integration into the Typhi chromosome and loss of the IncHI1 (PTU-HI1A) plasmid (34), has occurred. Carriage of SGI11 (denoted as MDR or XDR in Fig. 1F) was identified in JI-A, JI-B, JI-C, and JI-D, and the genetic location was consistent within each group—chromosomal in JI-A, JI-B, and JI-C, or plasmid-mediated in JI-D (PTU-HI1A) (Fig. 1F). At higher JI thresholds, presence of SGI11 was even confined to specific JI-subgroups JI-A1, JI-A3, JI-B1, and JI-C1 (Fig. S5 in S1 Appendix). JI-B1 and JI-C1 represent known epidemiological lineages, the "XDR Pakistan" strain (discussed above) and MDR 4.3.1.1 Typhi strains with chromosomal SGI11, respectively. Genetic context analysis of genomes with chromosomal SGI11 (28 reference and 300 U.S. genomes) detected six variants (previously described variants A–E and a novel variant F described here) in two different genetic locations, the *yidA* gene and the intergenic region between genes *cyaA* and *cyaY* (Fig. S13 in S1 Appendix). However, neither SGI11 variant nor chromosomal insertion site was found to align with JI subgrouping, likely due to the relatively small size of SGI11 (~25 kb or smaller). Indeed, we detected a

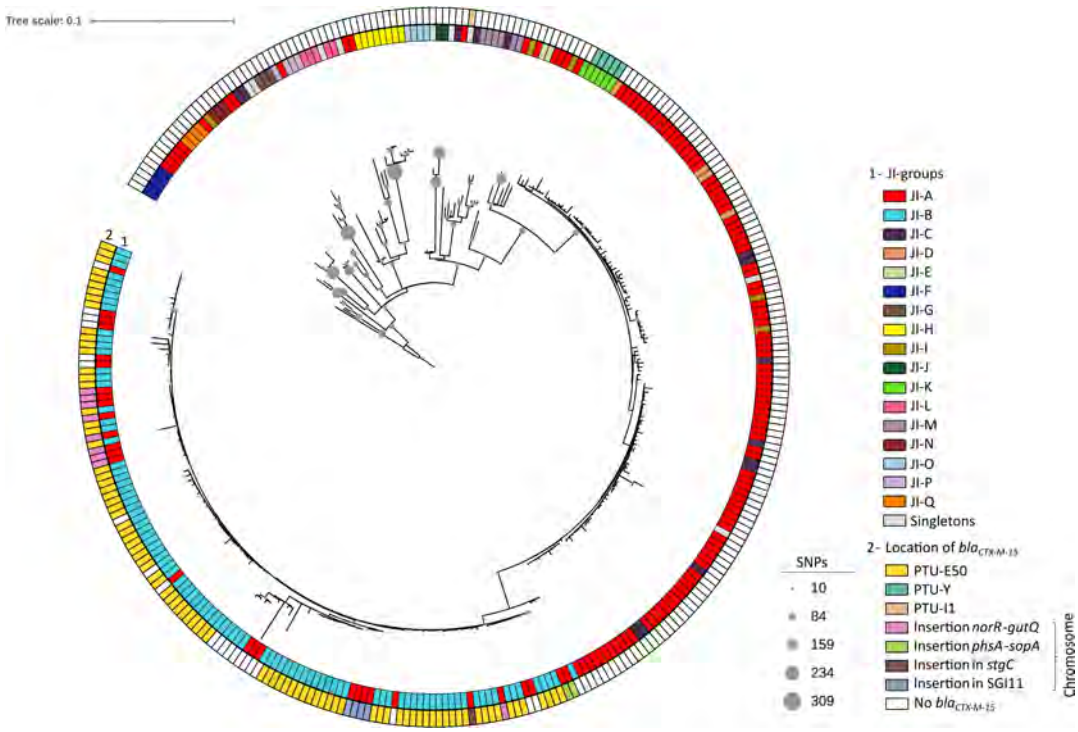


FIG 4 Core-genome phylogeny of Typhi genomes. The cladogram includes all Typhi genomes that contain the gene $bla_{\text{CTX-M-15}}$ (n = 109), all genomes from JI-A3 that lack $bla_{\text{CTX-M-15}}$ (n = 88), all genomes from JI-B1 that lack $bla_{\text{CTX-M-15}}$ (n = 4), 122 representative genomes from the 17 JI groups (Table S1, see column "Genomes in Fig. 4"), and one genome from serovar Indiana as an outgroup. Branch length scale represents changes per number of SNPs. Circles at the internal nodes indicate the number of SNPs distinctive of the corresponding clade. The colored rings indicate the JI group of the corresponding genome (i), and the $bla_{\text{CTX-M-15}}$ gene location (ii). A phylogenetic tree of the representative genomes from the 17 JI groups is shown in Fig. S12 in S1 Appendix).

likely event of SGI11 excision from the *yidA* gene in six JI-B isolates otherwise practically identical to other JI-B members encoding SGI11 interrupting *yidA* (Fig. S13b). In these cases, long-read sequencing of two of these genomes (PNUSAS224101, SAMN21040479; PNUSAS195139, SAMN18332688) confirmed that the *yidA* gene is disrupted by IS*1*, suggesting that it could be either a precursor to the SGI11 acquisition, or most likely a derivative of SGI11 excision, both probably through IS*1*-mediated recombination. Nonetheless, mapping of SGI11 presence onto the JI network (Fig. 1F) quickly reveals that some JI groups more frequently host this MGE than others, which is likely driven but not entirely explained by the overrepresentation of 4.3.1 in these groups.

Chromosomal mutations in the quinolone-resistance determining region (QRDR) and presence of *acrB* mutations (azithromycin resistance [35]) were mapped onto the JI network (Fig. S14 in S1 Appendix). Genomes with triple QRDR mutations tended to cluster within JI subgroups JI-A1 and JI-A4 (GenoTyphi 4.3.1.2) but were also found in different JI groups (JI-C, JI-I, JI-M), consistent with the observation that QRDR mutants have emerged spontaneously in different lineages (36, 37). Interestingly, specific *acrB* mutations aligned with JI group, rather than GenoTyphi lineage (Table S1) [n = 1/6 *acrB*(R717L) in JI-B, n = 5/6 *acrB*(R717Q) in JI-C], but with relatively low prevalence of these mutations, this observation may be anecdotal.

U.S. Typhi pangenome structure reveals novel plasmid patterns

Nine different PTUs were detected in JI groups, predominantly from MOB_P and MOB_H classes (Table 2; Fig. 1C). While some are well known (e.g., PTU-HI1A [IncHI1A] in JI-D), others are not well characterized (e.g., PTU-Y [IncY] in JI-K). The plasmid copy number of all PTUs did not exceed that of the chromosome. All of them were graded as host range III or higher, including those that lacked a MOB relaxase (PTU-E18 and PTU-Y are

phage-plasmids). This is an indication of their broad ability to colonize bacteria from different genera of the same taxonomic family. Of particular interest are PTU-E50 (IncY) and PTU-Y (IncY) plasmids because of their carriage of $bla_{\text{CTX-M-15}}$ and association of the former PTU with XDR Typhi.

Four different PTU-E50 plasmid variants were identified in JI-B1 (IncY), JI-B2 [IncFIB(K)], JI-B3 [IncFIB(K)], and JI-J (IncY) (Fig. S15A in S1 Appendix), each containing a unique set of proteins (Fig. S15B and C in S1 Appendix). Hence, PTU-E50 core gene analysis would not detect these plasmid clusters (Fig. S15A in S1 Appendix). Furthermore, these JI groups have differing and significant ($P < 0.01$, chi-squared test of independence) geographic signals, despite extremely limited travel data (Table S1); JI-B1 is linked to travel to Pakistan, JI-B2 to Bangladesh, and JI-J to Nigeria (Fig. S11B in S1 Appendix). Interestingly, while most PTU-E50 (IncY) plasmids from JI-B1 ("XDR Pakistan" plasmid) harbor the $bla_{\text{CTX-M-15}}$ gene ($n = 84/88$), four isolates do not. These genomes are likely variants of the original p60006 (31) plasmid that have subsequently lost the $bla_{\text{CTX-M-15}}$ gene, representing a novel lineage of the 4.3.1.1.P1 PTU-E50 (IncY)-containing strain. JI grouping could be leveraged to link unique plasmids to geographic regions, in the same way that core genome SNPs are used to reflect geographical signals.

PTU-Y was exclusively identified in JI-K. Two genotypes were present in this JI group, 4.3.1.1 and 4.3.1.2, and only PTU-Y plasmids hosted in the latter genotype carried $bla_{\text{CTX-M-15}}$ (Table S1; Fig. S15D through F in S1 Appendix). This plasmid carries an IncY replicon, but rather than being a conjugative plasmid (as is PTU-E50 [IncY] in JI-B1), it is a large non-conjugative phage-plasmid whose transmission is governed by an entirely different mechanism (38, 39). In this case, relying on replicon typing alone (as commonly practiced) would generate confusion, as two very distinct plasmid types (PTU-E50 and PTU-Y) carry the same replicon (IncY) (Table 2), and interestingly, in this case, both carry $bla_{\text{CTX-M-15}}$. Of interest, carriage of PTU-Y is significantly associated with travel to Iraq ($P < 0.01$, chi-squared test of independence). JI grouping has the advantage of accounting for all genetic material within the plasmid rather than a single replicon target, and therefore can simultaneously differentiate highly related plasmids (as seen for PTU-E50 plasmids), and disintegrate seemingly similar plasmids (PTU-Y [IncY] versus PTU-E50 [IncY]). These plasmid subgroups can be rapidly detected in a network (and overlaid with epidemiological data), preventing the continual need for separate plasmid core genome analysis.

In contrast to large (>90 kb) plasmids, smaller plasmids (<50 kb) did not often contain enough genetic content to define individual JI groups (Table 2). For example, JI-H has only one member with a 50 kb PTU-N1 (IncN) plasmid, and instead is genetically distinct from other groups by the presence of a ~55 kb ICE (Fig. S6 in S1 Appendix and S2 Appendix). Another small mobilizable plasmid, PTU-E80 (IncX1, ~25 kb, highly related with PTU-X1) (Table 2), was among the most common plasmids detected, predominantly in JI-N, and while it is likely important to this group (>85% of members carry PTU-E80), it did not exclusively underpin the genetic definition of JI-N. Instead, JI-N was genetically distinct from other JI groups also due to the absence of a ~21 kb phage (prophage 1) and the absence of a 21 Kb IME (MOB_{Q}) (Fig. S6 and S7B in S1 Appendix, and S2 Appendix). Of interest, JI-N is almost exclusively lineage 2.0.2 (one genome is 4.1), a genotype that was also detected in JI-A and JI-I. In this case, JI grouping enables stratification of an epidemiologically important genotype (28) using "unknown" accessory genetic content.

U.S. Typhi pangenome structure offers avenues for further investigation

Co-visualization of phylogenetic lineages across the JI network enabled rapid detection of groups that are likely characterized by clonal expansion (homology-by-descent) versus groups that contain disparate genomes that have converged on their MGEs (homology-by-admixture). For example, genomes of lineage 3.1.1 fall into either JI-A or JI-J. JI-J genomes are exclusively of genotype 3.1.1 and differentiate from JI-A partially due to carriage of a unique PTU-E50 (IncY) plasmid (Fig. S6 in S1 Appendix). Thus, it is plausible that JI-A/3.1.1 genomes represent a precursor strain that subsequently acquired a

PTU-E50 plasmid and clonally expanded to become group JI-J. JI-J is significantly (P < 0.01, chi-squared test of independence) associated with travel to Nigeria (despite limited travel data for U.S. isolates), an epidemiological signal that could prove useful as lineage 3.1.1 is the most common genotype in western Africa (36).

Lineage 2.3.2 is common in western Africa and the Americas and was recently shown to separate into discrete geographic clades by distance-based phylogeny (36). Sixty-two genomes from this phylogeny are also present in the U.S. data set; 61 genomes fall into the "Central American" clade and belong to JI-H, while the remaining genome is in the "Western Africa" clade and belongs to JI-A. JI-H differs from JI-A by the presence of an ~55 kb ICE (Fig. S6 in S1 Appendix, and S2 Appendix), and is significantly associated with travel to the Americas (P < 0.01, chi-squared test of independence). With further confirmation, the presence of this ICE could potentially be used to stratify 2.3.2 lineages into geographically and epidemiologically meaningful groups without the need for phylogenetic analysis.

In contrast to differentiation, aggregation of disparate genomes by their pangenome is of interest. For example, JI-C genomes all carry a unique ~107 kb phage-like PTU-E18 [IncFIB(pHCM2)] plasmid (38, 40) (Fig. 1C and D), but belong to a variety of GenoTyphi lineages with diverse geographic signals, including 4.3.1.1 dominant in Pakistan (36) and 3.5.4 exclusively associated with Samoa (41). Convergence of these diverse lineages on a large non-mobilizable phage-plasmid that does not carry AMR genes is curious, since acquisition cannot simply be explained by conjugation under antibiotic selection pressure. Rather, acquisition of plasmid-phages relies on viral-like mechanisms (transduction or lysogenic conversion) (38) and is likely induced by different ecological factors than conjugation (42, 43). Grouping and investigating Typhi strains through the lens of shared MGEs provides an opportunity to uncover common environmental exposures between genomes that might otherwise appear disparate using phylogenetic methods, adding an exciting new dimension to Typhi epidemiology.

Pangenome structure of U.S. Typhi is generalizable

To assess whether the network obtained with U.S. genomes is generalizable to the global population structure of Typhi, a new network was generated with a large data set from a distinct geographic region. It included 1,606 genomes isolated in the Indian subcontinent and 136 genomes (Table S2) from the U.S. data set, representative of the 17 JI groups previously identified (Fig. S16 in S1 Appendix). The new network organized into 17 JI groups already delimited in the U.S. data set (5 JI groups [E, G, J, N, and P] are represented only by reference genomes in this network and thus absent in the Indian data set) and two new JI groups (JI-R and JI-S, containing 10 and 11 genomes, respectively). JI-R genomes contain a PTU-X1 plasmid and three chromosomal regions enriched in phage-related genes, while JI-S members contain two plasmids (PTU-E18, PTU-HI1A) and an IME. In a similar experiment, 38 Typhi genomes of the pre-antibiotic era obtained from the Murray collection (44) (Table S3) were incorporated to the U.S. genome network (Table S1). They were distributed in groups JI-A (20 genomes), JI-F (seven genomes), JI-M (five genomes), JI-I (one genome), and JI-Q (one genome) and four isolates were singletons (Fig. S17A and B in S1 Appendix), with JI-A members belonging to different subgroups (Fig. S17C in S1 Appendix). Since a quarter of the U.S. data set is associated to travel to the Indian subcontinent (Table S1), which could bias the comparison, we analyzed a different data set, representative of the global Typhi diversity. We generated a JI network using 1,804 globally representative Typhi genomes, which were previously used to define the GenoTyphi typing nomenclature (27), and 136 reference genomes from the U.S. data set (Table S4; Fig. S18 in S1 Appendix). Emergence of novel clusters would be an expectable outcome, especially considering that they may emerge by the acquisition or loss of MGEs. Nevertheless, the vast majority (1,662/1,804, 92%) of the genomes in the GenoTyphi data set clustered in 12 of the originally defined JI groups. The remaining genomes fell into one of eight small new JI groups (98 genomes) or were singletons (44 genomes). The application of this method to these data sets demonstrates

the robustness of the JI groups identified here, despite being established using only genomes collected in the United States. It suggests that the U.S. data set effectively represents the global diversity of the Typhi pangenome and serves as a proxy for global sentinel surveillance.

DISCUSSION

Bacterial genomes frequently evolve by HGT, and the emergence of AMR is overwhelmingly driven by acquisition of MGEs, particularly in Typhi. Thus, a comprehensive view of Typhi epidemiology necessitates a focus on the accessory genome. Employing JI values, calculated from entire genome assembly pairwise comparisons with BinDash (19), we incorporated both vertical (SNPs) and horizontal (indels) evolutionary mechanisms to simultaneously analyze homology-by-descent and homology-by-admixture. This allowed us to represent U.S. Typhi epidemiology as a reticulate network, revealing non-random structure in the pangenome and offering additional information toward Typhi epidemiology, ecology, and evolutionary dynamics.

MGEs (both known and unknown) were universally present, and each JI group displayed a distinct profile corresponding to the presence or absence of particular plasmids or integrated MGEs (Fig. 1B and C), highlighting HGT as a significant mechanism of short-term diversification in Typhi. While large detectable plasmids were often the unique distinguishing feature of a JI group, many unknown ICEs, phage-like elements, or hypothetical regions, which are generally overlooked in genomic analysis, were also responsible for JI group differentiation (Fig. 1B; Fig. S6 and S7 in S1 Appendix and S2 Appendix). These regions would not have been detected by routine screening methods (PulseNet USA screens for AMR determinants and plasmid replicons only), nor would they be of interest in investigations focused on AMR. Yet these known and unknown MGEs are key features defining the structure of U.S. and global Typhi populations. Further knowledge of the transmission dynamics, functional capacity, and environmental reservoirs of these "cryptic" MGEs could offer valuable insight into the differing ecological predictors of Typhi occurrence and persistence.

Stratification of Typhi populations by accessory genome material alongside existing core-genome methods corroborated historical and recent epidemiological patterns. JI grouping of Typhi genomes detected the previously globally dominant 4.3.1 MDR lineage carrying SGI11 on an IncHI1 (PTU-HI1A) plasmid (JI-D) (33), the 4.3.1.1 MDR lineage with chromosomal SGI11 (JI-C1), clonal and *de novo* emergence of triple QRDR mutants in different lineages (JI-A, JI-C, JI-I, JI-M) (36), clonal expansion of XDR 4.3.1.1.P1 strains (JI-B1) associated with travel to Pakistan circa 2018 (31), recent chromosomal integration of $bla_{\text{CTX-M-15}}$ into the chromosome of 4.3.1.1.P1 strains (JI-A3) (30), and a Nigeria-associated lineage of the 3.1.1 West African genotype (JI-J). These epidemiologically relevant JI groups support the use of the Typhi pangenome for public health purposes. Namely in cases where travel data are unavailable, high genetic homology (>99.9% ANI) within a JI subgroup can be leveraged to make travel-related inferences, potentially ameliorating the frequent lack of travel information on U.S. cases.

Pangenomic analysis expanded our understanding of Typhi plasmids and MGEs, and suggests AMR emergence and epidemiology in this pathogen are subject to complex gene exchange networks and dynamics. First, two AMR-associated PTUs have emerged relatively recently in Typhi populations (PTU-E50 in JI-B and PTU-Y in JI-K), seemingly in distinct geographic regions. Given the host range of these PTUs (26), it is plausible these acquisitions are the result of active genetic exchange between diverse genera within the Enterobacteriaceae family. Secondly, the abundance of chromosomally integrated MGEs (Fig. 1B) suggests the existence of "hotspots" for integration of AMR regions in the Typhi chromosome, supported by the detection of several unique integration events described in this report (Fig. 3; Fig. S13 in S1 Appendix), and previously (30). Thus, we should expect to see continual "stabilization" of AMR phenotypes in the chromosome, which may in turn create opportunities for new AMR plasmids to enter. With this in mind, it is tempting to speculate that long-standing established lineages of Typhi (e.g., JI-A represented in

the Murray collection) may "sample" the mobile gene pool for plasmids and other MGEs of benefit (e.g., PTU-HI1A with SGI11, or PTU-E50 with $bla_{CTX\text{-}M\text{-}15}$) before eventually incorporating their advantageous cargo into the chromosome for reliable expression and long-term stability (45).

Although substantial effort is focused toward understanding AMR, much of the Typhi population is not MDR or XDR, nor do most genomes carry known plasmids or any AMR genes at all (Table S1). Stratification of Typhi populations using "invisible" or "cryptic" MGEs can offer an additional layer of molecular resolution for exploration alongside epidemiological variables (e.g., geographic origin), either by further partitioning of highly related genomes (see JI-H and JI-N) or aggregation of genomes that have converged on a single MGE. In the same way that we look to individual SNPs as unique molecular signatures for identifying subpopulations (28), we can exploit the unknown accessory genome for enhanced discriminatory power, or source attribution (12). With the flexibility to "toggle" the JI threshold for increasing differentiation, JI grouping is a useful method of analysis for this purpose.

JI analysis and network visualization as performed here is a powerful approach for pangenome exploration, enabling high-resolution stratification of thousands of genomes without the need for references, existing databases, or genomic annotation. While this analysis was performed on short-read data collected as part of routine surveillance, additional long-read sequencing is required for confirmation of genetic differences between JI groups. In fact, JI grouping facilitates prudent selection of representative genomes for long-read sequencing, minimizing sequencing resources, and maximizing coverage of population diversity. Similar clustering approaches harnessing *k*-mer-based genome distance estimation (20, 46, 47), as well as refined implementations (10, 48), are increasingly available and diminish considerably the computational challenges associated with incorporating both core and accessory genetic material into genomic analyses. However, as demonstrated here, it is essential to perform downstream analyses, leveraging existing tools and the rich body of knowledge on MGEs, for maximal interpretation of bacterial genomic clusters, especially for meaningful application in the public health space.

The potential biases introduced by utilizing Typhi genomes from a single country were addressed by analyzing multiple data sets from different geographic locations and time ranges. A remaining limitation of this analysis is the lack of very recent genomes (2022–2024). Since Typhi populations can rapidly evolve, new JI groups may emerge in a relatively short time period. Additionally, many previously unknown MGEs were detected in this analysis that may prove epidemiologically relevant, but in-depth genetic characterization of every MGE was outside the scope of this analysis. Finally, the JI method used here is not immediately implementable within the U.S. enteric surveillance system, PulseNet, due to existing computational infrastructure. However, recent efforts to modernize PulseNet's genomic surveillance (https://www.aphl.org/aboutAPHL/publications/Documents/PulseNet-2.0-White-Paper.pdf) may offer an opportunity for incorporation of JI-based methods, offering pangenomic analysis closer to "real-time" and simplifying the detection of unknown MGEs that can be explored with targeted genetic analysis. The ultimate public health goal is to provide a practical approach for enhanced genetic discrimination that improves surveillance and outbreak detection of otherwise indistinguishable enteric pathogens.

Given bacterial evolution occurs in both vertical and horizontal dimensions, inclusion of both core and accessory genetic material is a logical step toward understanding pathogen dynamics, not to mention a more holistic usage of increasingly available molecular data sets. With an eye to public, and indeed, global health relevance, we couple contemporary tools for genomic analysis with decades of research on MGEs to demonstrate the value of the pangenome, known and unknown, annotated, and hypothetical, for stratification of bacterial populations. We confirm and expand upon what is known about Typhi epidemiology, MGEs and AMR dynamics, and offer new

avenues for exploration to unravel Typhi reservoirs, "short" and "long-cycle" transmission pathways, and ultimately reduce incidence of human disease.

MATERIALS AND METHODS

Isolate collection and metadata

Typhi is a nationally notifiable disease in the United States (https://ndc.services.cdc.gov/case-definitions/salmonella-typhi-infection-2019/). The Centers for Disease Control and Prevention (CDC) requests state and participating local public health laboratories (PHL) (https://www.cdc.gov/narms/index.html) to submit all Typhi isolates that they receive from clinical laboratories to the National Antimicrobial Resistance Monitoring System (NARMS). Since 2016, NARMS and PulseNet USA, an enteric disease surveillance network of state and local PHL, have routinely performed WGS on Typhi isolates (3). CDC's National Typhoid and Paratyphoid Fever Surveillance system collects metadata on all Typhi cases reported to PHL, including history of international travel in the 30 days before illness onset (https://www.cdc.gov/typhoid-fever/surveillance.html).

Whole genome sequencing

WGS data were available for 2,272 Typhi isolates collected from 1 January 2008 through 30 September 2021 (Table S1). For years prior to routine WGS (2008–2015), all Typhi isolates in the PulseNet national database with WGS data available were included (n = 68); these isolates represent a small proportion of total isolates from this time period. For years 2016–2018, all Typhi isolates sent to NARMS for WGS were included (n = 1,343), which is representative of U.S. Typhi cases reported to CDC for these years. Due to logistics and delays in shipping for NARMS surveillance isolates in recent years, the 2019–2021 time period is represented by Typhi isolates in PulseNet with WGS data available (n = 861), with expected underreporting due to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic-related factors. WGS performed through NARMS and PulseNet followed standard operating procedures for the Illumina Miseq platform (https://www.aphl.org/programs/food_safety/Documents/PNL38_WGS%20on%20MiSeq%20SOP_v3.pdf). Reads with a base call quality score ≥28 and coverage ≥40× were assembled in this study using shovill v.1.0.9 (https://github.com/tseemann/shovill); and contigs with coverage below 10% average genome coverage were excluded from final assemblies. Typhi serotype for all study genomes was confirmed using SeqSero2 v.0.1 (49), and genomes were further genotyped using the updated GenoTyphi scheme.

Long-read sequencing was performed in this study on select isolates (PNUSAS224101, SAMN21040479; PNUSAS195139, SAMN18332688; PNUSAS198714, SAMN18813804; see Table S1) for indel verification (the first two isolates to confirm the absence of SGI11 and the *yidA* gene disruption, and the third one to detect the integration of $bla_{\text{CTX-M-15}}$ in SG11) as previously described (50). Corresponding Illumina short reads were generated from the same DNA extraction; libraries were prepared using the Illumina DNA Flex preparation kit per the PulseNet protocol (https://www.aphl.org/programs/food_safety/Documents/PNL35%20Illumina%20DNA%20Prep%20SOP_v4.pdf) and sequenced on the Illumina MiSeq platform as described above. Hybrid assemblies were generated as previously described (50) and uploaded to the National Center for Biotechnology Information (NCBI).

Molecular subtyping and characterization

Typhi study genomes were typed using the updated GenoTyphi scheme (27, 28) (https://github.com/katholt/genotyphi). AMR determinants were detected using staramr v.0.4.0 (51), which employs the ResFinder database (updated 30 July 2020; 90% identity, 50% gene coverage) and the *Salmonella* spp. PointFinder scheme (updated 30 August 2019). Accessory genome elements were detected using a database adapted from

PlasmidFinder (52) (90% identity, 60% gene coverage) for plasmid replicons, MOBscan (53) for conjugative relaxases, CONJScan (54) for conjugative systems, and COPLA (55) to assign plasmids to a given PTU (26). Reconstruction of plasmids from Illumina reads was performed using PLACNETw (56). Bakta was used for gene annotation of indel regions (57). Indel regions containing phage-related proteins were assigned as prophages and further re-annotated using PhageScope (58).

MDR was defined as the presence of genes conferring resistance to ampicillin, chloramphenicol, and trimethoprim-sulfamethoxazole, which are typically acquired within an IS*1*-mediated composite transposon, either on a plasmid or integrated into the chromosome as SGI11 (*Salmonella* genomic island) (27, 28, 34). XDR was defined as MDR with the addition of a ciprofloxacin resistance mechanism (QRDR mutation and/or plasmid-mediated quinolone resistance gene) and a ceftriaxone resistance gene (typically $bla_{\text{CTX-M-15}}$) (31, 59).

Chromosomal integration events were detected using the typing mode of ISMapper (60) to identify acquisition of an insertion sequence (IS) relative to a reference chromosome. To detect the integration sites of $bla_{\text{CTX-M-15}}$, its mobilizer, IS*Ecp1*, was used as a bait against a reference chromosome (Typhi 311189_291186, NZ_CP029894.1). Integration of SGI11 was detected using IS*1* as a bait element and Typhi CT18 as the reference chromosome (NC_003198).

Additional genomes

One-hundred twenty Typhi reference genomes from NCBI RefSeq200 database (accessed on 14 May 2020) were included in the analysis, collected between 1916 and 2019 (Table S1). A data set for comparative analysis against the U.S. data set was generated using all Typhi genomes isolated in the Indian subcontinent available in Pathogenwatch (61) (n = 1,606) (accessed on 22 March 2021). Specifically, this data set included genomes linked to Bangladesh (n = 637), India (n = 487), Nepal (n = 318), Pakistan (n = 158), and Sri Lanka (n = 3), or a combination of these countries (n = 3) (Table S2).

Thirty-eight Typhi genomes from the Murray collection (44) available at the European Nucleotide Archive at https://www.ebi.ac.uk/ena/browser/view/PRJEB3255 (Table S3) and a database of 1,804 globally representative *S.* Typhi genomes used to develop the GenoTyphi typing scheme (27) available at Pathogenwatch (https://pathogen.watch/genomes/all?collection=nti046ubbs7t-wong-et-al-2015&genusId=590) (Table S4) were included in a comparative analysis against the U.S. data set. Molecular subtyping and characterization of reference genomes was performed as above.

Jaccard index and genome length distance analysis

The exact JI was used as a measure of similarity between all genome pairs. First, the complete assembly of each genome was converted into a set of *k*-mers. JI was calculated as the ratio of shared *k*-mers over the total number of different *k*-mers between the two sets (shared *k*-mers, SNP *k*-mers [i.e., *k*-mers differing by just 1 bp], and indel *k*-mers [i.e., *k*-mers different in both data sets]). BinDash (19) was used to calculate JI, using parameters minhashtype = −1 (to compute the exact JI between highly similar genomes, that is, using the complete set of *k*-mers rather than calculating an estimated JI based on a subset of *k*-mers) and *k*-mer length (k) = 21 (this latter as previously defined as optimum in reference 20). The formula to calculate JI between genomes A and B (equation 1) is defined as the size of the intersection divided by the size of the union of the two *k*-mer sets of genomes A and B (see supporting information text in S1 Appendix).

$$\mathrm{JI}_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

Genome length was estimated from the number of unique *k*-mers in a genome (S). The upper *k*-mer length limit in Jellyfish v.2.2.6 (62) (k = 27) was used to generate *k*-mers from each genome sequence because with greater *k*-mer length, the probability

of having repeated *k*-mers by chance in a genome is lower and the genome length estimation is more accurate. *S* was computed by counting the occurrences of identical *k*-mers only once, that is, unique *k*-mers. To obtain a relative measure of genome size, the unique *k*-mer count is divided by 1 million base pairs (*S*/1,000,000). For every genome pair (A and B), the difference between their unique *k*-mer counts is recorded as a GLD value (equation 2).

$$\mathrm{GLD}_{(A,B)} = |S_A - S_B| \,/\, 1,000,000 \quad (2)$$

Taking into account that contig ends affect *k*-mer count, a correction described by equation 3 was applied in draft genomes, considering *S*, *k*, and the number of contigs of the assembly (*C*).

$$\mathrm{GLD}_{(A,B)} = |(S_A + (k-1)C_A) - (S_B + (k-1)C_B)| \,/\, 1,000,000 \quad (3)$$

Network visualization and community detection

The adjacency matrix of pairwise genome similarities generated by BinDash was used to build an undirected network. Gephi v.9 (https://gephi.org/ [63]) was used to visualize the network, applying the ForceAtlas2 algorithm for the layout. The network nodes, representing genomes, were colored according to metadata and genetic determinants of interest. Edges between nodes are represented whenever the corresponding JI or GLD value is equal to or higher than the user-defined threshold. A range of JI thresholds for a given application needs to be assessed to define the final components to study, referred to as JI groups. This depends on the specific study population and question pursued, but it is recommended to minimize complexity by setting a threshold that will result in a manageable number of JI groups [ideally, the number of clusters should not exceed the natural logarithm of the number of genomes (64)], to group the greatest number of genomes possible, and to factor in congruence with genetic determinants of interest, if available. In this data set, no distinct valley was observed in the distribution of JI values (Fig. S2 in S1 Appendix). We thus evaluated several statistics to optimize the network sparsification: transitivity (a measure indicating groups of nodes with strong internal connections), density, number of communities (smaller or larger than five members), and proportion of genomes in these two types of communities (Fig. S3 in S1 Appendix). For JI values greater than 0.97, we detected an increase in transitivity with a corresponding decrease in edge density (due to the removal of spurious edges linking communities). The number of small communities exponentially increased for JI ≥0.975. The presence of singletons and small communities may be influenced by incorrect assignment to the *Salmonella* serovar, sample bias, the thresholds applied, sequencing errors, and the intrinsic genetic diversity of the samples. Nevertheless, a plateau in the number of communities with five or more members was observed in the JI range between 0.980 and 0.984. In this range, 98.95% to 98.36% of genomes were assigned to a community with five or more members. The final JI threshold was set at 0.983. The Louvain method (65), implemented in Gephi, was used to define the JI groups by using resolution of 1.5. This community-finding algorithm aims to maximize the density of edges within communities while minimizing those between communities. Once the main JI groups are defined, they can be further dissected in several subgroups within the network using a more stringent JI and applying the same community detection algorithm.

Distinct differences in indels, including MGEs or accessory genome regions, between JI groups were detected using BLASTN (v.2.6.0+) (66) by comparing reference genomes from each JI group (Table S1). For those JI groups that did not contain a reference genome, a genome was reconstructed using PLACNETw (56). Plasmid presence was also detected using PlasmidSeeker (67). The BLAST searches between all possible reference pairs from different JI groups enable the detection of regions present in one genome of the pair and absent in the other. An estimation of their expected size can be obtained by clearing *L* (number of SNPs or the inserted region in size bp) from formula (Eq. 4), where

N corresponds to the total number of genome *k*-mers of the reference genome. If the size of the genome-specific regions targeted by BLAST is similar to *L*, it can be assumed that genome differences are due mainly to indels. Otherwise, SNPs or a mix of SNPs and indels account for the differences (supplemental information text in S1 Appendix).

$$\mathrm{JI} = N - (k-1) \,/\, N + L + k - 1 \quad (4)$$

MGE removal

To explore the contribution of MGEs in the JI group clustering, MGE (plasmids and/or SGI11) sequences were manually removed from the nucleotide fasta files of selected genomes. These "cured" sequences were then used to compute JI and generate networks as explained in the previous section.

Phylogeny reconstruction

kSNP 3.0 (68) was used to identify SNPs in WGS data (complete, assembled, and raw short-read data) using *k*-mers = 19. This optimal *k* was chosen with the kSNP tool Kchooser. SNP-based trees were reconstructed by maximum parsimony using the core-SNPs detected (option -core). All the trees generated in this study were visualized with iTol v.6 (69).

Plasmid copy number calculation

The presence of specific plasmids and their average plasmid copy number was estimated from the sequence read files of 1,836 Typhi isolates using PlasmidSeeker (67), including 1,157 genomes from JI-A, 97 from JI-B, 152 from JI-C, 5 from JI-D, 5 from JI-E, 7 from JI-F, 6 from JI-G, 158 from JI-H, 9 from JI-I, 10 from JI-J, 7 from JI-K, 4 from JI-L, 117 from JI-M, 75 from JI-N, and 28 singletons. PlasmidSeeker was executed with default parameters and a plasmid database of 1,064 plasmids from RefSeq200 (1,011 from *Salmonella* spp., and 153 belonging to PTU-E50, PTU-Y, PTU-E7, and PTU-E80 from Enterobacterales).

Statistical analysis

The varpart function implemented in the vegan Community Ecology R package (search.r-project.org/CRAN/refmans/vegan/html/varpart.html) was used to partition the variance in JI groups with respect to GenoTyphi lineages and MOB relaxase genes using an adjusted R^2 value. Chi-squared tests of independence were performed to examine geographic signals associated with JI groups.

ACKNOWLEDGMENTS

We acknowledge the state and local public health laboratories that participated in PulseNet and the National Antimicrobial Resistance Monitoring System (NARMS). We thank the laboratorians and epidemiologists that offered helpful review and critique.

This work was supported by the Centers for Disease Control and Prevention (contract no. 75D30119C06679 and 75D30121C11978 to F.D.L.C). This work was also supported by the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033 (PID2020-117923GB-I00 to F.D.L.C. and M.P.G.-B.), and the Spanish Ministry of Economy, Industry and Competitiveness (DI-17-09164 to S.R.-S.).

AUTHOR AFFILIATIONS

[1]Instituto de Biomedicina y Biotecnología de Cantabria, (CSIC, Universidad de Cantabria), Santander, Spain
[2]Division of Foodborne, Waterborne, and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
[3]Biomar Microbial Technologies, León, Spain

[4]Departamento de Ingeniería de las Comunicaciones, Universidad de Cantabria, Santander, Spain
[5]Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, Tennessee, USA
[6]ASRT, Inc., Suwanee, Georgia, USA

AUTHOR ORCIDs

Arancha Peñil-Celis http://orcid.org/0000-0003-2295-1563
Hattie E. Webb http://orcid.org/0000-0002-1190-7930
Jason P. Folster http://orcid.org/0000-0001-8514-5118
M. Pilar Garcillan-Barcia http://orcid.org/0000-0001-7058-5428
Fernando de la Cruz http://orcid.org/0000-0003-4758-6857

FUNDING

Funder	Grant(s)	Author(s)
Ministerio de Ciencia e Innovación (MCIN)	PID2020-117923GB-I00 MCIN/AEI/10.13039/501100011033	Fernando de la Cruz
Ministerio de Ciencia e Innovación (MCIN)	PID2020-117923GB-I00 MCIN/AEI/10.13039/501100011033	M. Pilar Garcillan-Barcia
HHS | Centers for Disease Control and Prevention (CDC)	Contracts No. 75D30119C06679 and 75D30121C11978	Fernando de la Cruz
Ministerio de Economía y Competitividad (MEC)	DI-17-09164	Santiago Redondo-Salvo

DATA AVAILABILITY

All data are available in the article and supplemental material. WGS data analyzed in this study are available at the Sequence Read Archive of NCBI (https://www.ncbi.nlm.nih.gov/sra/), PathogenWatch (https://pathogen.watch/), European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home), or Pathogenwatch (https://pathogen.watch). Illumina short-read data are available at the Sequence Read Archive of NCBI (https://www.ncbi.nlm.nih.gov/sra/) using the corresponding accession numbers in Tables S1 to S4. Oxford Nanopore long-read data generated in this study are available on NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) for the following genomes using the listed accession numbers: PNUSAS198714, SRR21971975; PNUSAS224101, SRR21971977; PNUSAS195139, SRR21971976. Custom Perl and Python scripts to implement BinDash and parse its results are available on GitHub (https://github.com/PenilCelis/Salmonella_Typhi_JINA).

ETHICS APPROVAL

This investigation was reviewed by the Centers for Disease Control and Prevention (CDC) and was conducted consistent with applicable federal law and CDC policy: 45 C.F.R. part 46, 21 C.F.R. part 56; 42 U.S.C. §241(d); 5 U.S.C. §552a; 44 U.S.C. §3501 et seq.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

S1 Appendix (mSystems00365-24-s0001.pdf). Supplemental information text, supplemental methods, and Figures S1 to S18.
S2 Appendix (mSystems00365-24-s0002.pdf). Gene annotation of MGEs.
Supplemental tables (mSystems00365-24-s0003.xlsx). Tables S1 to S4.

Open Peer Review

PEER REVIEW HISTORY (review-history.pdf). An accounting of the reviewer comments and feedback.

REFERENCES

1. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. 2019. Using genomics to track global antimicrobial resistance. Front Public Health 7:242. https://doi.org/10.3389/fpubh.2019.00242
2. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Van Der Vossen J, Tang S, Katase M, McClure P, Kimura B, Ching Chai L, Chapman J, Grant K. 2019. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol 79:96–115. https://doi.org/10.1016/j.fm.2018.11.005
3. Tolar B, Joseph LA, Schroeder MN, Stroika S, Ribot EM, Hise KB, Gerner-Smidt P. 2019. An overview of PulseNet USA databases. Foodborne Pathog Dis 16:457–462. https://doi.org/10.1089/fpd.2019.2637
4. Arnold BJ, Huang I-T, Hanage WP. 2022. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol 20:206–218. https://doi.org/10.1038/s41579-021-00650-4
5. Azarian T, Huang I-T, Hanage WP. 2020. Structure and dynamics of bacterial populations: pangenome ecology, p 115–128. In Tettelin H, Medini D (ed), The pangenome. Springer International Publishing.
6. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721. https://doi.org/10.1038/nrmicro1234
7. Brown EW, Bell R, Zhang G, Timme R, Zheng J, Hammack TS, Allard MW. 2021. *Salmonella* genomics in public health and food safety. EcoSal Plus 9:eESP00082020. https://doi.org/10.1128/ecosalplus.ESP-0008-2020
8. McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nat Microbiol 2:17040. https://doi.org/10.1038/nmicrobiol.2017.40
9. Whelan FJ, Hall RJ, McInerney JO. 2021. Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. Mol Biol Evol 38:3697–3708. https://doi.org/10.1093/molbev/msab139
10. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 29:304–316. https://doi.org/10.1101/gr.241455.118
11. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, Vehkala M, Välimäki N, Prentice MB, Ashour A, Avram O, Pupko T, Dobrindt U, Literak I, Guenther S, Schaufler K, Wieler LH, Zhiyong Z, Sheppard SK, McInerney JO, Corander J. 2016. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS Genet 12:e1006280. https://doi.org/10.1371/journal.pgen.1006280
12. Liu CM, Aziz M, Park DE, Wu Z, Stegger M, Li M, Wang Y, Schmidlin K, Johnson TJ, Koch BJ, Hungate BA, Nordstrom L, Gauld L, Weaver B, Rolland D, Statham S, Hall B, Sariya S, Davis GS, Keim PS, Johnson JR, Price LB. 2023. Using source-associated mobile genetic elements to identify zoonotic extraintestinal *E. coli* infections. One Health 16:100518. https://doi.org/10.1016/j.onehlt.2023.100518
13. Commichaux S, Rand H, Javkar K, Molloy EK, Pettengill JB, Pightling A, Hoffmann M, Pop M, Jayeola V, Foley S, Luo Y. 2023. Assessment of plasmids for relating the 2020 *Salmonella enterica* serovar Newport onion outbreak to farms implicated by the outbreak investigation. BMC Genomics 24:165. https://doi.org/10.1186/s12864-023-09245-0
14. Laing CR, Whiteside MD, Gannon VPJ. 2017. Pan-genome analyses of the species *Salmonella enterica*, and identification of genomic markers predictive for species, subspecies, and serovar. Front Microbiol 8:1345. https://doi.org/10.3389/fmicb.2017.01345
15. Eliades SJ, Brown JC, Colston TJ, Fisher RN, Niukula JB, Gray K, Vadada J, Rasalato S, Siler CD. 2021. Gut microbial ecology of the critically endangered Fijian crested iguana (*Brachylophus vitiensis*): effects of captivity status and host reintroduction on endogenous microbiomes. Ecol Evol 11:4731–4743. https://doi.org/10.1002/ece3.7373
16. Leclaire S, Nielsen JF, Drea CM. 2014. Bacterial communities in meerkat anal scent secretions vary with host sex, age, and group membership. Behav Ecol 25:996–1004. https://doi.org/10.1093/beheco/aru074
17. Puspaningrum EY, Nugroho B, Setiawan A, Hariyanti N. 2020. Detection of text similarity for indication plagiarism using winnowing algorithm based k-gram and Jaccard coefficient. J Phys Conf Ser 1569:022044. https://doi.org/10.1088/1742-6596/1569/2/022044
18. Temma S, Sugii M, Matsuno H. 2019. The document similarity index based on the Jaccard distance for mail filtering. 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC); JeJu, Korea (South). https://doi.org/10.1109/ITC-CSCC.2019.8793419
19. Zhao X. 2019. BinDash, software for fast genome distance estimation on a typical personal laptop. Bioinformatics 35:671–673. https://doi.org/10.1093/bioinformatics/bty651
20. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x
21. Acman M, van Dorp L, Santini JM, Balloux F. 2020. Large-scale network analysis captures biological features of bacterial plasmids. Nat Commun 11:2452. https://doi.org/10.1038/s41467-020-16282-w
22. Prokopenko D, Hecker J, Silverman EK, Pagano M, Nöthen MM, Dina C, Lange C, Fier HL. 2016. Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 genomes project. Bioinformatics 32:1366–1372. https://doi.org/10.1093/bioinformatics/btv752
23. Hancuh M, Walldorf J, Minta AA, Tevi-Benissan C, Christian KA, Nedelec Y, Heitzinger K, Mikoleit M, Tiffany A, Bentsi-Enchill AD, Breakwell L. 2023. Typhoid fever surveillance, incidence estimates, and progress toward typhoid conjugate vaccine introduction — worldwide, 2018–2022. MMWR Morb Mortal Wkly Rep 72:171–176. https://doi.org/10.15585/mmwr.mm7207a2
24. CDC. Typhoid fever and paratyphoid fever. Available from: https://www.cdc.gov/typhoid-fever/index.html
25. Gauld JS, Olgemoeller F, Heinz E, Nkhata R, Bilima S, Wailan AM, Kennedy N, Mallewa J, Gordon MA, Read JM, Heyderman RS, Thomson NR, Diggle PJ, Feasey NA. 2022. Spatial and genomic data to characterize endemic typhoid transmission. Clin Infect Dis 74:1993–2000. https://doi.org/10.1093/cid/ciab745
26. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, Garcillán-Barcia MP, de la Cruz F. 2020. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. Nat Commun 11:3602. https://doi.org/10.1038/s41467-020-17278-2
27. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, Murphy N, Holliman R, Sefton A, Millar M, Dyson ZA, Dougan G, Holt KE, International Typhoid Consortium. 2016. An extended genotyping framework for *Salmonella enterica* serovar Typhi, the cause of human typhoid. Nat Commun 7:12827. https://doi.org/10.1038/ncomms12827
28. Dyson ZA, Holt KE. 2021. Five years of GenoTyphi: updates to the global *Salmonella* Typhi genotyping framework. J Infect Dis 224:S775–S780. https://doi.org/10.1093/infdis/jiab414
29. François Watkins LK, Winstead A, Appiah GD, Friedman CR, Medalla F, Hughes MJ, Birhane MG, Schneider ZD, Marcenac P, Hanna SS, Godbole G, Walblay KA, Wiggington AE, Leeper M, Meservey EH, Tagg KA, Chen JC, Abubakar A, Lami F, Asaad AM, Sabaratnam V, Ikram A, Angelo KM, Walker A, Mintz E. 2020. Update on extensively drug-resistant *Salmonella* serotype Typhi infections among travelers to or from Pakistan and report of ceftriaxone-resistant *Salmonella* serotype Typhi infections among travelers to Iraq — United States, 2018–2019. MMWR Morb Mortal Wkly Rep 69:618–622. https://doi.org/10.15585/mmwr.mm6920a2
30. Nair S, Chattaway M, Langridge GC, Gentle A, Day M, Ainsworth EV, Mohamed I, Smith R, Jenkins C, Dallman TJ, Godbole G. 2021. ESBL-producing strains isolated from imported cases of enteric fever in England and Wales reveal multiple chromosomal integrations of $bla_{CTX\text{-}M\text{-}15}$ in XDR *Salmonella* Typhi. J Antimicrob Chemother 76:1459–1466. https://doi.org/10.1093/jac/dkab049

31. Klemm EJ, Shakoor S, Page AJ, Qamar FN, Judge K, Saeed DK, Wong VK, Dallman TJ, Nair S, Baker S, Shaheen G, Qureshi S, Yousafzai MT, Saleem MK, Hasan Z, Dougan G, Hasan R. 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:e00105-18. https://doi.org/10.1128/mBio.00105-18
32. Lima NCB, Tanmoy AM, Westeel E, de Almeida LGP, Rajoharison A, Islam M, Endtz HP, Saha SK, de Vasconcelos ATR, Komurian-Pradel F. 2019. Analysis of isolates from Bangladesh highlights multiple ways to carry resistance genes in *Salmonella* Typhi. BMC Genomics 20:530. https://doi.org/10.1186/s12864-019-5916-6
33. Holt KE, Phan MD, Baker S, Duy PT, Nga TVT, Nair S, Turner AK, Walsh C, Fanning S, Farrell-Ward S, Dutta S, Kariuki S, Weill F-X, Parkhill J, Dougan G, Wain J. 2011. Emergence of a globally dominant IncHI1 plasmid type associated with multiple drug resistant typhoid. PLoS Negl Trop Dis 5:e1245. https://doi.org/10.1371/journal.pntd.0001245
34. Chiou C-S, Alam M, Kuo J-C, Liu Y-Y, Wang P-J. 2015. Chromosome-mediated multidrug resistance in *Salmonella enterica* serovar Typhi. Antimicrob Agents Chemother 59:721–723. https://doi.org/10.1128/AAC.04081-14
35. Hooda Y, Sajib MSI, Rahman H, Luby SP, Bondy-Denomy J, Santosham M, Andrews JR, Saha SK, Saha S. 2019. Molecular mechanism of azithromycin resistance among typhoidal *Salmonella* strains in Bangladesh identified through passive pediatric surveillance. PLoS Negl Trop Dis 13:e0007868. https://doi.org/10.1371/journal.pntd.0007868
36. Carey ME, Dyson ZA, Ingle DJ, Amir A, Aworh MK, Chattaway MA, Chew KL, Crump JA, Feasey NA, Howden BP, et al. 2023. Global diversity and antimicrobial resistance of typhoid fever pathogens: insights from a meta-analysis of 13,000 *Salmonella* Typhi genomes. Elife 12:e85867. https://doi.org/10.7554/eLife.85867
37. da Silva KE, Tanmoy AM, Pragasam AK, Iqbal J, Sajib MSI, Mutreja A, Veeraraghavan B, Tamrakar D, Qamar FN, Dougan G, et al. 2022. The international and intercontinental spread and expansion of antimicrobial-resistant *Salmonella* Typhi: a genomic epidemiology study. Lancet Microbe 3:e567–e577. https://doi.org/10.1016/S2666-5247(22)00093-3
38. Pfeifer E, Bonnin RA, Rocha EPC. 2022. Phage-plasmids spread antibiotic resistance genes through infection and lysogenic conversion. mBio 13:e0185122. https://doi.org/10.1128/mbio.01851-22
39. Greig DR, Bird MT, Chattaway MA, Langridge GC, Waters EV, Ribeca P, Jenkins C, Nair S. 2022. Characterization of a P1-bacteriophage-like plasmid (phage-plasmid) harbouring $bla_{CTX\text{-}M\text{-}15}$ in *Salmonella enterica* serovar Typhi. Microb Genom 8. https://doi.org/10.1099/mgen.0.000913
40. Kidgell C, Pickard D, Wain J, James K, Diem Nga LT, Diep TS, Levine MM, O'Gaora P, Prentice MB, Parkhill J, Day N, Farrar J, Dougan G. 2002. Characterisation and distribution of a cryptic *Salmonella* Typhi plasmid pHCM2. Plasmid 47:159–171. https://doi.org/10.1016/s0147-619x(02)00013-6
41. Sikorski MJ, Hazen TH, Desai SN, Nimarota-Brown S, Tupua S, Sialeipata M, Rambocus S, Ingle DJ, Duchene S, Ballard SA, Valcanis M, Zufan S, Ma J, Sahl JW, Maes M, Dougan G, Thomsen RE, Robins-Browne RM, Howden BP, Naseri TK, Levine MM, Rasko DA. 2022. Persistence of rare *Salmonella* Typhi genotypes susceptible to first-line antibiotics in the remote Islands of Samoa. mBio 13:e0192022. https://doi.org/10.1128/mbio.01920-22
42. Pattenden T, Eagles C, Wahl LM. 2022. Host life-history traits influence the distribution of prophages and the genes they carry. Philos Trans R Soc Lond B Biol Sci 377:20200465. https://doi.org/10.1098/rstb.2020.0465
43. Nair S, Barker CR, Bird M, Greig DR, Collins C, Painset A, Chattaway M, Pickard D, Larkin L, Gharbia S, Didelot X, Ribeca P. 2024. Presence of phage-plasmids in multiple serovars of *Salmonella enterica*. Microb Genom 10:001247. https://doi.org/10.1099/mgen.0.001247
44. Baker KS, Burnett E, McGregor H, Deheer-Graham A, Boinett C, Langridge GC, Wailan AM, Cain AK, Thomson NR, Russell JE, Parkhill J. 2015. The Murray collection of pre-antibiotic era *Enterobacteriacae*: a unique research resource. Genome Med 7:97. https://doi.org/10.1186/s13073-015-0222-7
45. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. 2021. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nat Rev Microbiol 19:347–359. https://doi.org/10.1038/s41579-020-00497-1
46. Baker DN, Langmead B. 2023. Genomic sketching with multiplicities and locality-sensitive hashing using Dashing 2. Genome Res 33:1218–1227. https://doi.org/10.1101/gr.277655.123
47. Xu X, Yin Z, Yan L, Yi H, Wang H, Schmidt B, Liu W. 2023. RabbitKSSD: accelerating genome distance estimation on modern multi-core architectures. Bioinformatics 39:btad695. https://doi.org/10.1093/bioinformatics/btad695
48. Bonnici V, Cracco A, Franco G. 2022. A *k*-mer based sequence similarity for pangenomic analyses. In Nicosia G, et al (ed), Machine learning, optimization, and data science. Vol. 13164. Springer, Cham, Switzerland.
49. Zhang S, den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng X. 2019. SeqSero2: rapid and improved *Salmonella* serotype determination using whole-genome sequencing data. Appl Environ Microbiol 85:e01746-19. https://doi.org/10.1128/AEM.01746-19
50. Webb HE, Kim JY, Tagg KA, de la Cruz F, Peñil-Celis A, Tolar B, Ellison Z, Schwensohn C, Brandenburg J, Nichols M, Folster JP. 2022. Genome sequences of 18 *Salmonella enterica* serotype Hadar strains collected from patients in the United States. Microbiol Resour Announc 11:e00522-22. https://doi.org/10.1128/mra.00522-22
51. Bharat A, Petkau A, Avery BP, Chen JC, Folster JP, Carson CA, Kearney A, Nadon C, Mabon P, Thiessen J, Alexander DC, Allen V, El Bailey S, Bekal S, German GJ, Haldane D, Hoang L, Chui L, Minion J, Zahariadis G, Domselaar GV, Reid-Smith RJ, Mulvey MR. 2022. Correlation between phenotypic and *in silico* detection of antimicrobial resistance in *Salmonella enterica* in Canada using Staramr. Microorganisms 10:292. https://doi.org/10.3390/microorganisms10020292
52. Carattoli A, Hasman H. 2020. PlasmidFinder and in silico pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). Methods Mol Biol 2075:285–294. https://doi.org/10.1007/978-1-4939-9877-7_20
53. Garcillán-Barcia MP, Redondo-Salvo S, Vielva L, de la Cruz F. 2020. MOBscan: automated annotation of MOB relaxases. Methods Mol Biol 2075:295–308. https://doi.org/10.1007/978-1-4939-9877-7_21
54. Cury J, Abby SS, Doppelt-Azeroual O, Néron B, Rocha EPC. 2020. Identifying conjugative plasmids and integrative conjugative elements with CONJscan. Methods Mol Biol 2075:265–283. https://doi.org/10.1007/978-1-4939-9877-7_19
55. Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-López R, de la Cruz F. 2021. COPLA, a taxonomic classifier of plasmids. BMC Bioinformatics 22:390. https://doi.org/10.1186/s12859-021-04299-x
56. Vielva L, de Toro M, Lanza VF, de la Cruz F. 2017. PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. Bioinformatics 33:3796–3798. https://doi.org/10.1093/bioinformatics/btx462
57. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. 2021. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial Genomics 7:000685. https://doi.org/10.1099/mgen.0.000685
58. Wang RH, Yang S, Liu Z, Zhang Y, Wang X, Xu Z, Wang J, Li SC. 2024. PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations. Nucleic Acids Res 52:D756–D761. https://doi.org/10.1093/nar/gkad979
59. Hughes MJ, Birhane MG, Dorough L, Reynolds JL, Caidi H, Tagg KA, Snyder CM, Yu AT, Altman SM, Boyle MM, Thomas D, Robbins AE, Waechter HA, Cody I, Mintz ED, Gutelius B, Langley G, Francois Watkins LK. 2021. Extensively drug-resistant typhoid fever in the United States. Open Forum Infect Dis 8:fab572. https://doi.org/10.1093/ofid/ofab572
60. Hawkey J, Hamidian M, Wick RR, Edwards DJ, Billman-Jacobe H, Hall RM, Holt KE. 2015. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. BMC Genomics 16:667. https://doi.org/10.1186/s12864-015-1860-2
61. Argimón S, Yeats CA, Goater RJ, Abudahab K, Taylor B, Underwood A, Sánchez-Busó L, Wong VK, Dyson ZA, Nair S, Park SE, Marks F, Page AJ, Keane JA, Baker S, Holt KE, Dougan G, Aanensen DM. 2021. A global resource for genomic predictions of antimicrobial resistance and surveillance of *Salmonella* Typhi at pathogenwatch. Nat Commun 12:2879. https://doi.org/10.1038/s41467-021-23091-2
62. Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27:764–770. https://doi.org/10.1093/bioinformatics/btr011

63. Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. ICWSM 3:361–362. https://doi.org/10.1609/icwsm.v3i1.13937
64. Peixoto TP. 2014. Hierarchical block structures and high-resolution model selection in large networks. Phys Rev X 4:011047. https://doi.org/10.1103/PhysRevX.4.011047
65. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. J Stat Mech 2008:10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
66. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
67. Roosaare M, Puustusmaa M, Möls M, Vaher M, Remm M. 2018. PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. PeerJ 6:e4588. https://doi.org/10.7717/peerj.4588
68. Gardner SN, Slezak T, Hall BG. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics 31:2877–2878. https://doi.org/10.1093/bioinformatics/btv271
69. Letunic I, Bork P. 2021. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301

Journal of Global Antimicrobial Resistance 39 (2024) 69–72

Contents lists available at ScienceDirect

Journal of Global Antimicrobial Resistance

journal homepage: www.elsevier.com/locate/jgar

Short Communication

Azithromycin-resistant *mph*(A)-positive *Salmonella enterica* serovar Typhi in the United States

Kaitlin A. Tagg[a], Justin Y. Kim[a,b], Britton Henderson[a,b], Meseret G. Birhane[a], Caroline Snyder[a,c], Carla Boutwell[d], Abiye Iyo[d], Linlin Li[e], Eva Weinstein[e], Yvonne Mercado[f], Arancha Peñil-Celis[g], Matthew Mikoleit[h], Jason P. Folster[a], Louise K. Francois Watkins[a,*]

[a] *Division of Foodborne, Waterborne, and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA*
[b] *ASRT, Inc, Suwanee, GA, USA*
[c] *Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA*
[d] *Mississippi State Department of Health, Jackson, MS, USA*
[e] *California Department of Public Health, Richmond, CA, USA*
[f] *Madera County Department of Public Health, Madera, CA, USA*
[g] *Instituto de Biomedicina y Biotecnología de Cantabria, Universidad de Cantabria, Santander, Spain*
[h] *Division of Global Health Protection, Centers for Disease Control and Prevention, Atlanta, GA, USA*

ARTICLE INFO

Article history:
Received 12 February 2024
Revised 1 August 2024
Accepted 9 August 2024
Available online 20 August 2024

Editor: Stefania Stefani

Keywords:
S. Typhi
Azithromycin resistance

ABSTRACT

Objectives: The United States Centers for Disease Control and Prevention (CDC) conducts active surveillance for typhoid fever cases caused by *Salmonella enterica* serovar Typhi (Typhi). Here we describe the characteristics of the first two cases of *mph*(A)-positive azithromycin-resistant Typhi identified through US surveillance.

Methods: Isolates were submitted to public health laboratories, sequenced, and screened for antimicrobial resistance determinants and plasmids, as part of CDC PulseNet's routine genomic surveillance. Antimicrobial susceptibility testing and long-read sequencing were also performed. Basic case information (age, sex, travel, outcome) was collected through routine questionnaires; additional epidemiological data was requested through follow-up patient interviews.

Results: The patients are related and both reported travel to India (overlapping travel dates) before illness onset. Both Typhi genomes belong to the GenoTyphi lineage 4.3.1.1 and carry the azithromycin-resistance gene *mph*(A) on a PTU-FE (IncFIA/FIB/FII) plasmid. These strains differ genetically from *mph*(A)-positive Typhi genomes recently reported from Pakistan, suggesting independent emergence of azithromycin resistance in India.

Conclusions: Cases of typhoid fever caused by Typhi strains resistant to all available oral treatment options are cause for concern and support the need for vaccination of travellers to Typhi endemic regions. US genomic surveillance serves as an important global sentinel for detection of strains with known and emerging antimicrobial resistance profiles, including strains from areas where routine surveillance is not conducted.

1. Introduction

Salmonella enterica serovar Typhi (Typhi) is a globally important human pathogen, although disease burden falls predominantly in south and southeast Asia, and sub-Saharan Africa [2]. The United States is a low-incidence Typhi setting, with cases predominantly associated with international travel to endemic areas [3]. Thus, US Typhi surveillance serves as an important global sentinel for the surveillance of strains with known and emerging antimicrobial resistance (AR) profiles, including strains from areas where routine surveillance is not conducted.

Increasing multidrug resistance (MDR) and global spread of extensively drug resistant strains (XDR) [2] leaves dwindling options for oral antibiotic treatment of Typhi. While azithromycin

* Corresponding author at: 1600 Clifton Rd, NE, Atlanta, GA 30329, USA.
E-mail address: hvu9@cdc.gov (L.K. Francois Watkins).

https://doi.org/10.1016/j.jgar.2024.08.005

remains a viable oral treatment option for most cases of typhoid fever globally [2,4], azithromycin-resistant strains have recently been reported from India, Nepal and Singapore [5]. The mechanism of resistance is typically a single point mutation in the *acrB* gene, which has arisen in several independent Typhi lineages [6–8]. However, the acquired azithromycin-resistance gene *mph(*A) (encoding a macrolide phosphotransferase) has also been reported in Bangladesh [9], and most recently from a patient in Pakistan [10]. We describe here the first two cases of azithromycin-resistant *mph*(A)-positive Typhi identified in the United States through PulseNet, the US national network for molecular subtyping of enteric bacteria.

2. Material and methods

Isolates PNUSAS349400 and PNUSAS347532 were submitted by clinical diagnostic laboratories to public health laboratories and were sequenced as part of the Centers for Disease Control and Prevention (CDC) PulseNet national passive *Salmonella* surveillance network, per standardized methods (https://www.cdc.gov/nationalsurveillance/salmonella-surveillance.html; https://www.cdc.gov/pulsenet/pdf/PNL38-WGS-on-MiSeq-508.pdf); resistance determinants and plasmids were identified through routine screening, utilizing ResFinder and PlasmidFinder databases, respectively, as previously described [11]. Plasmids were further categorized into plasmid taxonomic units (PTU) using COPLA [12]. Sequenced reads were typed using GenoTyphi (https://github.com/katholt/genotyphi) [13].

For long-read sequencing of both isolates, genomic DNA was extracted (Wizard Genomic DNA Purification Kit, modified manufacturer's protocol, Promega, WI, USA) from cultures incubated on Tryptic Soy Agar-Sheep Blood overnight (37 °C). Libraries were prepared using the Rapid Barcoding Kit (SQK-RBK114.24; Oxford Nanopore Technologies [ONT], Oxford, UK) according to the manufacturer's protocol and sequenced for 72 h on a GridION sequencing platform (R10.4.1 flowcells, ONT). Reads were base-called using the SUP ("Super accurate") model of Guppy v6.5.7, filtered for quality using MinKNOW v23.04.5 (ONT), and filtered for minimum read length (>1000 bp) using Nanoq v0.10.0 [14]. Read sets were downsampled randomly using rasusa v0.7.1 [15] and assembled using flye v2.9 [16] (asm-coverage option set to 10 for PNUSAS349400). Assemblies were rotated to fix start positions of each contig using Circlator v1.5.5 [17], polished using Medaka v1.8.0 (https://github.com/nanoporetech/medaka), and screened for quality using socru v2.2.4 (to ensure expected genomic arrangement) and BUSCO v5.4.6 [18,19]. Long read data are deposited in National Center for Biotechnology Information (NCBI) under the BioSample IDs listed in Table 1.

Additionally, these two isolates underwent antimicrobial susceptibility testing (AST) according to CDC National Antimicrobial Resistance Monitoring System's protocol. Specifically, 14 antibiotics were tested (amoxicillin-clavulanic acid, ampicillin, azithromycin, cefoxitin, ceftriaxone, chloramphenicol, ciprofloxacin, colistin, gentamicin, meropenem, nalidixic acid, sulfamethoxazole, tetracycline, trimethoprim-sulfamethoxazole) and resistance was determined using CLSI breakpoints (https://www.cdc.gov/narms/antibiotics-tested.html).

Public health departments routinely submit epidemiologic information for all laboratory-confirmed cases of typhoid fever to CDC, including demographics, clinical outcome details, and travel history. We requested that public health officials perform a supplementary, second interview to collect additional epidemiologic and clinical information (including exposures, clinical course, and treatment information) from the two patients whose isolates carried the *mph*(A) gene.

This project was reviewed by CDC and deemed not to be research (IRB review was not required); the activity was conducted consistently with applicable federal law and CDC policy. Patients provided verbal consent for publication of their case information.

3. Results and discussion

Follow-up interviews revealed that patient one and patient two were related (daughter and mother, respectively). Patient one (with isolate PNUSAS349400) was a healthy woman in her 30 s who became ill in February 2023. She initially reported high fever and was hospitalized for her infection for six days. She failed to respond to multiple antibiotics (fever returned after a week), but subsequently recovered. Patient two (with isolate PNUSAS347532) was a woman in her 60 s with a history of type II diabetes mellitus who became ill in March 2023. Symptoms persisted for three weeks; she was initially evaluated in the emergency department and discharged home. After a second emergency department visit, she was hospitalized for four days and treated with multiple antibiotics over the course of her illness, including amoxicillin-clavulanic acid, azithromycin, ceftriaxone, ciprofloxacin, doxycycline, meropenem, minocycline, moxifloxacin; she ultimately recovered, and underwent laparoscopic cholecystectomy in April 2023.

Both patients spent a portion of their incubation period (defined as 6–30 days before illness onset) in New Delhi, India, where they attended the same wedding. Patient one stayed both at a hotel and with family; she reported consuming only Indian street food with no specific dietary restrictions, and consumed bottled water during her stay, but food may not have been prepared with bottled water. Patient two reported staying with friends and relatives and eating mostly meals prepared in the home; she reported consuming bottled water. Neither patient was vaccinated for typhoid fever before travel.

Isolates PNUSAS349400 (patient one) and PNUSAS347532 (patient two) belong to GenoTyphi lineage 4.3.1.1, a common genetic lineage in India [2,3]. They sit within a cluster of four genomes in a SNP-based phylogenetic tree and differ by a single SNP (Fig. 1) [1]. In addition to azithromycin resistance, they are multidrug resistant (MDR) displaying phenotypic and genotypic resistance to ampicillin, chloramphenicol, trimethoprim-sulfamethoxazole, and ceftriaxone, and decreased susceptibility to ciprofloxacin, due to

Table 1
Molecular and epidemiological characteristics of *mph*(A)-positive Typhi cases.

Patient	Strain ID	SAMN	GenoTyphi	AMR determinants	AST^	Plasmid type	Travel reported	Vaccinated
1	PNUSAS349400	SAMN35155331	4.3.1.1	*aac(6′)-Ib-cr, aadA5,* $bla_{CTX-M-15}$, bla_{OXA-1}, *catA1, dfrA17, dfrA7, gyrA*(S83Y), *mph*(A), *sul1*	ACSuCxCotAzm	PTU-FE (IncFIA/FIB/FII)	India	No
2	PNUSAS347532	SAMN35010716	4.3.1.1	*aac(3)-IIa, aac(6′)-Ib-cr, aadA5,* $bla_{CTX-M-15}$, bla_{OXA-1}, *catA1, dfrA17, dfrA7, gyrA*(S83Y), *mph*(A), *sul1*	ACSuCxGenCotAzm	PTU-FE (IncFIA/FIB/FII)	India	No

^ Antimicrobial susceptibility testing. Antimicrobial abbreviations are as follows: A, ampicillin; Azm, azithromycin; C, chloramphenicol; Cot, cotrimoxazole (trimethoprim-sulfamethoxazole); Cx, ceftriaxone; Gen, gentamicin; Su, sulfamethoxazole.

Fig. 1. SNP-based phylogenetic subtree of closely related genomes, exported directly from NCBI pathogen detection, (accessed 30.07.24) [1]. Scale bar indicates number of SNPs. PNUSAS349400 and PNUSAS347532 are highlighted in red.

Fig. 2. Pairwise comparison of multiresistance region (~27 kb) of plasmids pPNUSAS349400 and pPNUSAS347532, generated in Geneious Prime v2021.2.2. Plasmids were annotated using Prokka v1.14.6 and ISfinder (https://isfinder.biotoul.fr/blast.php). Complete and partial annotations are denoted in the following colours: coding sequences in yellow, resistance genes in teal, insertion sequences in purple, transposons in light green, and conserved segments of a class 1 integron in fuscia.

a combination of chromosomal and plasmid-mediated determinants (Table 1). Both isolates have a single *gyrA* mutation (S83Y); and a variant of SGI*11* inserted in the chromosome (between *cyaY* and *cyaA*), containing only *catA1, dfrA7* and *sul1* (similar to SGI*11b*, [20]). Resistance genes *catB3* (partial), bla_{OXA-1}, *aac(6′)-Ib-cr,* $bla_{CTX-M-15}$, *dfrA17, aadA5, sul1, qacE* and *mph*(A) were present on a PTU-FE (IncFIA/FIB/FII) plasmid in both genomes, within a large (~27 kb) multiresistance region (Fig. 2). While the multiresistance regions are structurally similar, plasmid pPNUSAS347532 carries *aac(3)-IIa* between two copies of IS*26*, a region that is not present in plasmid pPNUSAS349400 (Fig. 1). Thus, these two strains are closely genetically related, but not identical.

PTU-FE plasmids are common in *Escherichia coli* [21], but have not been reported from India, even in recent efforts to characterize ceftriaxone resistant strains [22], nor have they been detected before in Typhi surveillance in the United States ([23]; https://www.cdc.gov/typhoid-fever/surveillance.html). Since Typhi is prone to carriage of PTUs with the ability to colonize a wide range of bacterial genera [23], novel acquisition of an *E. coli*-associated PTU by a previously circulating Typhi lineage is plausible and, in fact, drove the recent emergence of XDR Typhi [24].

The *mph*(A)-positive Typhi reported here are of a different genotype and carry a different plasmid than recent *mph*(A)-positive genomes from Pakistan [10], indicating independent emergence of plasmid-mediated azithromycin resistance. The epidemiological and molecular evidence is suggestive of local transmission in India, either from a common exposure or person-to-person transmission. While we have not yet detected widespread circulation of this strain in India, the slight genetic variance (both in the chromosome and the plasmid) is worth further contemplation. It is certainly possible that the small variation in these strains occurred in each patient after initial infection; or perhaps there exists a common reservoir from which variants of the original *mph*(A)-positive strain are already evolving.

Strains resistant to penicillin, chloramphenicol, trimethoprim-sulfamethoxazole, fluoroquinolones and third-generation cephalosporins leave few treatment options. The emergence of azithromycin resistance, the only remaining oral treatment option, highlights the need for additional intervention and control measures, including vaccination for travellers and residents of endemic Typhi areas [7]. Because many typhoid infections likely go undetected, and treatment options are increasingly limited, ongoing US-based surveillance is important to detect *mph*(A)-positive Typhi in returning travellers and prevent transmission.

4. Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention, the Mississippi State Department of Health, the California Department of Public Health, or the California Health and Human Services Agency. Names of specific vendors, manufacturers or products are included for public health and informational purposes; inclusion does not imply endorsement of the vendors, manufacturers or products by the Centers for Disease Control and Prevention or the US Department of Health and Human Services.

Acknowledgements

We thank epidemiology and laboratory partners in state and local health departments.

Declarations

Funding: This work was supported by the Centers for Disease Control and Prevention.

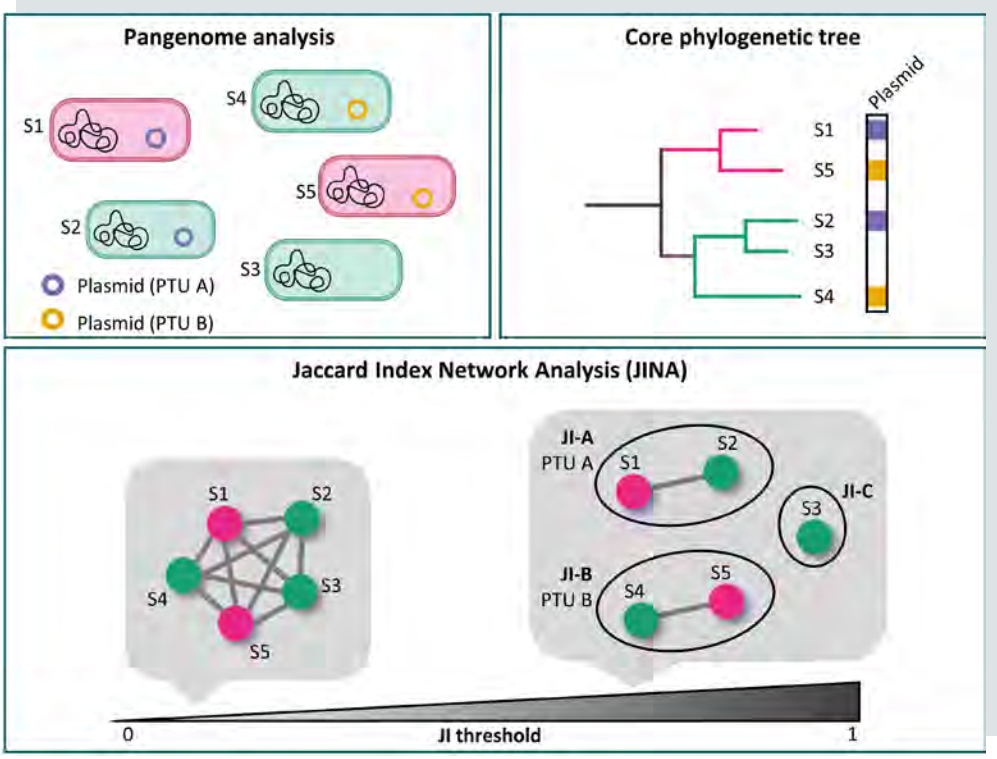
Declaration of competing interests: None declared.

Ethical approval: Not required.

References

[1] The NCBI Pathogen Detection Project [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 2016. https://www.ncbi.nlm.nih.gov/pathogens/ [accessed 30.07.24].
[2] da Silva KE, Tanmoy AM, Pragasam AK, Iqbal J, Islam Sajib MS, Mutreja A, et al. The international and intercontinental spread and expansion of antimicrobial-resistant *Salmonella* Typhi. Lancet Microbe 2022;3(8):E567–EE77.

[3] Carey ME, Dyson ZA, Ingle DJ, Amir A, Aworh MK, Chattaway MA, et al. Global diversity and antimicrobial resistance of typhoid fever pathogens: insights from a meta-analysis of 13,000 *Salmonella* Typhi genomes. Elife 2023;12:e85867.
[4] Hughes MJ, Birhane MG, Dorough L, Reynolds JL, Caidi H, Tagg KA, Snyder CM, Yu AT, Altman SM, Boyle MM, Thomas D, Robbins AE, Waechter HA, Cody I, Mintz ED, Gutelius B, Langley G, Francois Watkins LK. Extensively Drug-Resistant Typhoid Fever in the United States. Open Forum Infect Dis 2021;8(12):ofab572.
[5] Sajib MSI, Tanmoy AM, Hooda Y, Rahman H, Andrews JR, Garrett DO, et al. Tracking the emergence of azithromycin resistance in multiple genotypes of typhoidal *Salmonella*. MBio 2021;12(1):e03481-20.
[6] Octavia S, Chew KL, Lin RTP, Teo JWP. Azithromycin-resistant *Salmonella enterica* serovar Typhi AcrB-R717Q/L, Singapore. Emerging Infect Dis 2021;27(2):624–7.
[7] Carey ME, Jain R, Yousuf M, Maes M, Dyson ZA, Thu TNH, et al. Spontaneous emergence of azithromycin resistance in independent lineages of *Salmonella* Typhi in northern India. Clin Infect Dis 2021;72(5):e120–e1e7.
[8] Duy PT, Dongol S, Giri A, Nguyen To NT, Dan Thanh HN, Nhu Quynh NP, et al. The emergence of azithromycin-resistant *Salmonella* Typhi in Nepal. JAC Antimicrob Resist 2020;2(4):dlaa109.
[9] Dola NZ, Shamsuzzaman SM, Islam S, Rahman A, Mishu NJ, Nabonee MA. Distribution of ciprofloxacin- and azithromycin-resistant genes among *Salmonella* Typhi isolated from human blood. Int J Appl Basic Med Res 2022;12(4):254–9.
[10] Nizamuddin S, Khan EA, Chattaway MA, Godbole G. Case of carbapenem-resistant *Salmonella* Typhi infection, Pakistan, 2022. Emerging Infect Dis 2023;29(11):2395–7.
[11] Webb HE, Kim JY, Tagg KA, de la Cruz F, Peñil-Celis A, Tolar B, et al. Genome sequences of 18 *Salmonella enterica* serotype Hadar strains collected from patients in the United States. Microbiol Res Announc 2022;11(10):e0052222.
[12] Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-López R, et al. COPLA, a taxonomic classifier of plasmids. BMC Bioinform 2021;22:390.
[13] Dyson ZA, Holt KE. Five years of GenoTyphi: updates to the global *Salmonella* Typhi genotyping framework. J Infect Dis 2021;224(Supplement 7):S775–SS80.
[14] Steinig E, Coin L. Nanoq: ultra-fast quality control for nanopore reads (0.9.0). J Open Source Software 2022;7(69):2991.
[15] Hall MB. Rasusa: randomly subsample sequencing reads to a specified coverage. J Open Source Software 2022;7(69):3941.
[16] Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 2019;37(5):540–6.
[17] Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 2015;16(1):294.
[18] Page AJ, Ainsworth EV, Langridge GC. Socru: typing of genome-level order and orientation around ribosomal operons in bacteria. Microb Genom 2020;6(7):mgen000396.
[19] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31(19):3210–12.
[20] Lima NCB, Tanmoy AM, Westeel E, de Almeida LGP, Rajoharison A, Islam M, et al. Analysis of isolates from Bangladesh highlights multiple ways to carry resistance genes in *Salmonella* Typhi. BMC Genom 2019;20(1):530.
[21] Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. Nat Commun 2020;11(1):3602.
[22] Jacob JJ, Pragasam AK, Vasudevan K, Veeraraghavan B, Kang G, John J, et al. *Salmonella* Typhi acquires diverse plasmids from other Enterobacteriaceae to develop cephalosporin resistance. Genomics 2021;113(4):2171–6.
[23] Peñil-Celis A, Tagg KA, Webb HE, Redondo-Salvo S, Watkins LF, Vielva L, et al. Mobile genetic elements define the non-random structure of the Salmonella enterica serovar Typhi pangenome. mSystems 2024;6(8):e0036524.
[24] Klemm EJ, Shakoor S, Page AJ, Qamar FN, Judge K, Saeed DK, et al. Emergence of an extensively drug-resistant *Salmonella enterica* serovar Typhi clone harboring a promiscuous plasmid encoding resistance to fluoroquinolones and third–generation cephalosporins. MBio 2018;20(1):e00105–18.



El estudio de la relación genética entre bacterias mediante loci cromosómicos específicos constituye la base de la vigilancia genómica en salud pública. Sin embargo, este enfoque basado en la evolución vertical, deja de lado un aspecto crucial de la evolución bacteriana: la transferencia genética horizontal. Esta tesis evalúa la relación del pangenoma en *Salmonella* Typhi y Hadar mediante un enfoque basado en el Índice de Jaccard, capturando tanto las relaciones evolutivas verticales (homología por descendencia) como horizontales (homología por mezcla) en una red reticulada. El análisis revela grupos poblacionales estrechamente relacionados, linajes emergentes y el impacto de elementos genéticos móviles que impulsan cambios epidemiológicamente relevantes en períodos cortos de tiempo. Este enfoque de alta resolución no solo mejora la detección de brotes y la atribución de fuentes de contaminación, sino que también refuerza la vigilancia de la resistencia a antibióticos.

Bacterial relatedness measured using selected chromosomal loci forms the basis for public health genomic surveillance, yet this method primarily captures vertical evolution while neglecting horizontal gene transfer. This thesis evaluates pangenome relatedness in *Salmonella* Typhi and Hadar using a Jaccard Index approach, capturing both vertical (homology-by-descent) and horizontal (homology-by-admixture) evolutionary relationships within a reticulate network. The analysis revealed fine-scale population structures, emerging lineages, and the impact of mobile genetic elements driving epidemiologically relevant shifts over short periods of time. This high-resolution approach not only enhances outbreak detection and source attribution but also strengthens antimicrobial resistance surveillance.