Studies in Bacterial Genome Dynamics

a dissertation presented
by
Arya Kaul
to
The Division of Medical Sciences

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biomedical Informatics

Harvard University
Cambridge, Massachusetts
April 2026

Thesis advisor: Professor Michael Baym Arya Kaul

Studies in Bacterial Genome Dynamics

Abstract

Bacteria are among the most genetically dynamic organisms on Earth, continuously reshaping their genomes through gene gain, loss, and modification. Understanding how new genes arise, how they spread, and how selective pressures like antibiotic use reshape this diversity remain central questions in bacteriology. This thesis investigates these questions through three complementary studies unified by a shared interest in the fluid dynamics of bacterial genomes.

First, we explore a mechanism by which the streamlining bias that trims bacterial genomes can also serve as a source of evolutionary innovation. We propose and characterize a model of "deletion-born fusion genes," in which adaptive deletions fuse distant gene fragments into novel open reading frames. Unlike other gene birth mechanisms that begin with rare, neutral intermediates, these fusions reach high frequency by hitchhiking on the selective advantage of the deletion itself. We document examples in the Lenski Long-Term Evolution Experiment and in the *Mycobacterium tuberculosis–bovis* divergence. Finally, we develop a scalable screen to efficiently detect these genes across multi-million bacterial genome collections.

Second, we collate 1,817 high-quality genomes from the British National Collection of Type Cultures, spanning isolates collected from 1885 to the present, and trace how human antibiotic use reshaped the frequency and mobility of genomic resistance. We find that functional resistance genes circulated in clinically relevant isolates before the age of antibiotics, but were generally rare and chromosomally encoded. Following the clinical introduction of specific antibiotics, corresponding genomic resistance significantly increased in frequency and became progressively associated with multiple mobile genetic elements. These findings suggest that anthropogenic use of antibiotics did not generate resistance de novo, but both amplified and refined pre-existing genetic potential.

Third, we introduce the phylogeny-colored de Bruijn graph (pcDBG), a data structure that recolors each unitig in a compacted pangenome graph with the phylogenetic branch where its presence/absence pattern most parsimoniously changed state. Applied to over 1,000 *Staphylococcus aureus* ST8 genomes, we show that co-inherited sequence blocks extend roughly one kilobase before decaying, and that mapping evolutionary signals onto the USA300 reference recovers the known mobile element landscape without prior annotation. The pcDBG provides a general framework for characterizing the spatial organization of evolutionary history across bacterial pangenomes.

Together, these studies suggest a view of bacterial genomes as dynamic mosaics, continually being reshaped both by their environment and the pressures we impose on them. The genomes we sequence today are not endpoints, but fleeting snapshots of a continuous evolutionary process that long predates humanity and will long outlast us.

Contents

Listing of figures

For my Dadaji, for introducing me to the splendor of the natural world,
and my Nanuji, for showing me the value of hard work and effort.

Acknowledgments

This Ph.D. would simply not have been possible without the unconditional love and encouragement of my community.

To my thesis advisor, Michael Baym, thank you for providing me the space and freedom to learn, experiment, and grow. I began this rotation during the COVID-19 pandemic, and despite being mainly virtual, I remember being struck by the cohesion of the lab. To the members of the Baym lab, thank you for consistently being my greatest source of scientific inspiration and emotional support. All of you are deeply gifted scientists and my best friends, and my doctorate is immeasurably richer as a direct consequence of our coffee chats, late-night shenanigans, and whiteboard axes.

To my dissertation advisory committee — Shamil Sunyaev, Tami Lieberman, Aleksander Kostic, and Curtis Huttenhower — and to the program administrators — Lilen Uchima, Cathy Haskell, and Jamie Gunnerson — thank you for the guidance and support that carried me through.

To my French advisor, Karel Břinda, thank you for all your encouragement. Every meeting left me excited and energized, and you single-handedly rekindled my love for discovery. Je suis infiniment reconnaissant du temps que j'ai passé avec toi et toute l'équipe GenScale à Rennes, qui m'ont accueilli et fait sentir si bien chez moi. Pour mes amis Bretons, je n'oublierai jamais votre gentillesse. To the faculty and students of the University of Global Health Equity, thank you for demonstrating the positive difference each one of us has the power to make. To my colleagues at the Asian Development Bank, thank you for making me appreciate the infrastructure underpinning our daily lives and for always fighting the good fight. To all the people I met during my year abroad, thank you for reminding me that the beauty of this world is *infinite* in breadth and depth.

To my Boston community, you always reminded me that I am more than my work. Getting to know each of you has made Boston feel like home. A special thank you to some of the communities that made it so: Mission Hill Main Streets, Café Society, my Gator gang, Boston Boxing & Fitness, the Cambridge Crooners Collective, and the Ramakrishna Vedanta Society. To my Escondido friends, thank you for giving me buckets of lore to react to every morning and being my personal meme supply.

To my family, thank you for always being a safe haven. It is your love and encouragement that made my scientific career possible, and your support has made the Ph.D. feel easy even when it was not.

I am immensely grateful for all the individuals who have contributed to my growth over these past years. To anyone I have not had the space to mention — thank you for always being there for me, and know that no matter where I am, a piece of each of you is present throughout this work.

0

Introduction

Bacteria are both the oldest and most genetically diverse domain of known terrestrial life[76]. They thrive in habitats spanning the deep subterranean crust, the labyrinthine lining of the human gut, and even the clouds overhead; collectively their metabolic capabilities underpin diverse global biogeochemical cycles ranging from nitrogen fixation to oxygen respiration[12,156,81,56]. Belying their outsize impact, bacteria are predominantly composed of microscopic single-celled or-

ganisms; yet, all together bacteria make up an estimated 15% of Earth's biomass[63,10].

The evolutionary success and environmental ubiquity of bacteria may be partially attributed to their genetic diversity[175,103,174]. The hereditary material of terrestrial life is encoded in DNA, and the functional units of this DNA are called 'genes.' Bacteria, compared to their eukaryotic counterparts, retain the capacity to rapidly gain, lose, and modify their gene content[174]. However, this flexibility also raises a paradox. Bacterial genomes are extremely compact and shaped by a pervasive 'streamlining' pressure that favors the loss of non-essential DNA[121,116,179]. Given this bias towards minimalism, how do bacteria then exhibit such a cacophony of diverse genes? And how might external selection pressures, say the human introduction of antibiotics, reshape this diversity over time? In this thesis, we attempt to explore these central biological questions through three complementary investigations into the origins, spread, and the topology of bacterial genome dynamics.

On the Bacterial Genetic Landscape

The scale of bacterial genetic diversity is staggering.* Studies of well-characterized species routinely identify tens of thousands of distinct gene families, and metagenomic surveys of complex environments like the human gut have cataloged millions of coding sequences with no detectable similarity to known genes[174,21,97,135,9]. The realization that the genetic repertoire of a bacterial species can be much larger than the gene content of an individual strain has given rise to the concept of the bacterial "pangenome.†" The pangenome refers to the collective gene pool accessible to a bacterial species, and many have been demonstrated to be significantly larger than the total number of genes found in any one member of that species[174,97]. Together, these findings point to a view of the bacterial genome as less a static blueprint, and more a fluid mosaic. Where on the species level, bacterial genomes are continually reshaped by gene gain and gene loss.

*And the central reason I switched to computational microbiology!

†Note that this gene-centric definition is distinct from the newer sequence-centric definition[111].

The result is a mode of evolution that operates on gene content as much as gene sequence, a paradigm where the presence or absence of entire genes can rapidly shift[97]. Understanding the forces that govern this flux is a central goal of bacteriology.

Sources of Bacterial Genetic Novelty

Where do new bacterial genes come from? Several mechanisms have been characterized, each involving distinct tradeoffs between the generation of novelty and the preservation of function.

Gene duplication followed by diversification is perhaps the most intuitive route[4,165,144]. When a gene is duplicated, one copy can maintain its original function while the other is free to accumulate mutations, and potentially acquire new capabilities. This mechanism ensures that novel genes derive from functional templates, but the ancestral sequence constrains their evolutionary potential. Duplicated genes all begin from a common template and thus are constrained by the functional neighborhood of their progenitors.

Horizontal gene transfer allows bacteria to acquire genes wholesale from distantly related organisms[96]. This process can introduce entirely new functions in a single event. Yet the acquired genes must still integrate into the recipient's regulatory and metabolic networks, and many transferred sequences can be quickly lost if they fail to provide immediate benefit[127,176,181].

At the other extreme, de novo gene birth can generate proteins with no homology to any existing sequence. Overprinting, in which a new gene emerges from an alternative reading frame overlapping an existing coding sequence, and the expression of previously non-coding DNA both represent routes to genuinely novel proteins[136,131]. But proteins born from random or near-random sequence face steep challenges: they tend toward excessive hydrophobicity, aggregation, and disorder, making the path to stable function narrow[83,64].

Gene fusion occupies a middle ground. By joining fragments of distinct genes into a single new

open reading frame, fusion events can bring together existing protein domains in novel arrangements [172,134]. The resulting proteins inherit some functional potential from their parent sequences while gaining new capabilities from their new juxtaposition. Yet fusion genes face the same threat as any sequence of bacterial DNA, they must avoid being lost to drift or purged by selection before they can demonstrate their usefulness [181].

This is especially pertinent because bacterial genomes are shaped by a pervasive deletional bias that relentlessly trims non-essential DNA. Genomes that have undergone reductive evolution, like many intracellular obligate pathogens, illustrate the endpoint of this process, but even free-living bacteria with large genomes experience constant pressure toward genomic minimalism [68,67,116]. Deletions arise through diverse mechanisms, from homologous recombination to replication slippage to erroneous repair, and can be positively selected when they reduce the metabolic cost of DNA replication or eliminate genes whose products have become deleterious [117,179,68]. Against this backdrop, the emergence of new genes seems almost paradoxical: how can novelty arise when the genome is under constant pressure to contract?

One possibility, explored in Chapter 1 of this thesis, is the idea that this deletional bias need not be merely destructive. Deletions rearrange existing genetic material, and in doing so, they can create new genes even as they remove old ones. A deletion that fuses the upstream portion of one gene to the downstream portion of another generates a novel open reading frame. Unlike other mechanisms of gene birth, which produce neutral or deleterious intermediates that must survive drift before acquiring function, these deletion-born fusions can reach high frequency by hitchhiking *on the selective advantage of the deletion itself*. The very same destructive force that threatens to purge nascent genes may also serve as an engine of their creation.

Selection and the Spread of Genes

The generation of new genetic material is only half the story. Whether a gene persists depends on the selective forces acting upon it, and these forces can change dramatically in response to environmental shifts.

Antibiotic resistance provides a stark illustration. Nature has produced antimicrobial compounds for millennia, and resistance mechanisms are correspondingly ancient; functional resistance genes have been recovered from permafrost samples tens of thousands of years old[45,43]. Yet the clinical introduction of antibiotics in the twentieth century transformed the selective landscape. What may have once been rare curiosities, maintained at low frequency by the occasional encounter with a competitor's toxins, suddenly became essential for survival in clinical settings. The result was a proliferation of resistance genes, accompanied by their mobilization onto plasmids and other transferable elements that could readily spread these resistance genes across species boundaries[44].

Understanding this transformation requires examining the state of resistance genes before and after the antibiotic era[14]. This presents an obvious difficulty: how can we measure resistance to an antibiotic before that antibiotic was discovered? Historical culture collections offer a solution to this conundrum. The British National Collection of Type Cultures, founded in 1920, has preserved bacterial isolates stretching back over a century, and recent sequencing efforts have made thousands of these genomes available for analysis[49,3]. By matching isolates to their year of collection, we can trace the changing frequency and genomic context of resistance determinants across the advent of the antibiotic age.

Chapter 2 of this thesis undertakes this analysis. We find that resistance genes, while present in pre-antibiotic isolates, were generally rare and chromosomally encoded. With the introduction of specific antibiotics, the corresponding resistance genes significantly increased in frequency and became increasingly associated with mobile genetic elements. Over time, resistance determinants accumulated

on elements nested within other mobile elements, a genomic architecture that enables their rapid dissemination. Human antibiotic use did not generate resistance de novo; however, our analysis shows it significantly amplified both the prevalence and mobility of previously infrequent resistance genes.

Computational bacteriology at scale

The questions explored in this thesis: how new genes arise, how they spread, and how external pressures reshape genomes are increasingly tractable because of the veritable explosion in available bacterial genomic sequences. Public databases now contain petabytes of sequence data, and this bounty of data affords the depth to detect rare events, trace evolutionary trajectories, and identify patterns invisible at smaller scales.

But scale creates its own challenges. Traditional approaches to comparative genomics can struggle in the face of this data deluge. New methods capable of leveraging the size of modern genomic databases are increasingly needed. This need is particularly acute given the majority of bacteria remain unculturable[77]. As a result, techniques that can derive meaningful biological insight from genomic sequences are essential for cleverly navigating the jungle of bacterial diversity.

Chapter 3 develops a new framework for integrating evolutionary history into pangenome graphs. Colored de Bruijn graphs, which annotate graph structures with the genomes that contribute each sequence, have become a standard tool for pangenome analysis. Yet these graphs capture only the pattern of sequence sharing, not its evolutionary origin. We propose phylogeny-colored de Bruijn graphs, in which sequences are annotated not with the raw set of genomes containing them but with the phylogenetic branches where their distribution most parsimoniously changed state. This representation fuses graph adjacency with evolutionary history, enabling analyses of how co-inherited blocks are spatially organized and where evolutionary boundaries fall across the pangenome.

Overview of this thesis

This thesis investigates the dynamics of bacterial genomes through three complementary studies, unified by a shared interest in how genes arise, spread, and can be computationally analyzed at scale.

Chapter 1 asks whether the deletional bias that shapes bacterial genomes can also serve as a source of innovation. We first formalize our proposed model of "deletion-born fusion genes," then document these genes arising in both the Lenski Long-Term Evolution Experiment and in the divergence of *Mycobacterium tuberculosis* and *Mycobacterium bovis*. Finally, we develop a scalable computational screen to detect these genes across all 2.4 million publicly available bacterial genomes. Our findings reframe deletions as not merely destructive but as potential catalysts for evolutionary innovation.

Chapter 2 queries how human antibiotic use reshaped the landscape of resistance genes. Using 1,817 high-quality genomes from the National Collection of Type Cultures, matched to their years of isolation spanning 1885 to the present, we trace the frequency and genomic context of genomic resistance determinants before and after the introduction of major antibiotic classes. We find that resistance genes were present but rare in pre-antibiotic isolates, and that clinical introduction of antibiotics is associated with increases in both prevalence and mobility. Over time, resistance elements have become increasingly nested within multiple mobile elements, an architecture that helps facilitate their rapid dissemination. These findings suggest that human antibiotic use did not create resistance de novo, but instead significantly amplified pre-existing genetic potential.

Chapter 3 introduces phylogeny-colored de Bruijn graphs, a data structure that bridges the gap between pangenome graph representations and phylogenetic inference. Each sequence in the graph is annotated with the branch or branches of a phylogeny where its presence most parsimoniously changed state, fusing graph adjacency with evolutionary history. We define two complementary analyses enabled by this structure: evolutionary persistence, which measures the spatial extent of co-inherited blocks, and boundary severity, which characterizes the phylogenetic distance across evolutionary bound-

aries. We demonstrate that both recapitulate the known mobile element landscape of a diverse collection of methicillin-resistant *Staphylococcus aureus* genomes without prior annotation.

Together, these chapters aim to understand the fluid dynamism of bacterial genomes leveraging the power of computation. From the birth of new genes to the mobile architecture of antibiotic resistance, this thesis seeks to investigate a domain of life defined by ubiquitous and rapid transformation. Bacteria are among the most consequential organisms on the planet, and this work represents my humble offering to shed additional light on these tiny titans of terrestrial life.

Life and death appeared to me ideal bounds,
which I should first break through, and pour a
torrent of light into our dark world.

Mary Shelley, *Frankenstein* (1818)

1

Novel genes arise from genomic deletions across the bacterial tree of life

The following chapter has been preprinted, with DOI: 10.64898/2026.01.05.697752

Bacteria are the most genetically diverse domain of life on Earth[76,115]. This genetic diversity also translates to a profound diversity in their protein-coding genic repertoire. New genes

and gene families continually arise across the bacterial tree of life; pangenome studies of single species can identify tens of thousands of protein families, and metagenomic sequencing of environments like the human gut has revealed millions of protein-coding genes of unknown function [174,21,97,135,9]. This diversity raises the question of where new bacterial genes and their corresponding functions come from.

Different mechanisms for bacterial gene birth have been proposed and each uniquely constrains the generation of functional novelty. On the one hand, duplication and subsequent diversification ensures that novel genes derive from functional templates [4,165,144], but the preexisting functionality and domain structure constrain their evolutionary potential. On the other hand, overprinting [136,131], the expression of a novel gene overlapping a previously existing sequence but in a different reading frame, generates protein products without homology to existing proteins or prior functional constraints. Yet new proteins created from random peptides often suffer from excessive hydrophobicity, aggregation propensity, and intrinsic disorder [83,64]. Considering these tradeoffs, gene fusion, the stitching together of entire fragments of distinct genes, provides a gene birth mechanism that balances innovation with functionality by bringing together existing protein domains into new contexts [134], though such fusions must still avoid being lost to selection or genetic drift.

Random loss of newborn genes is compounded by a deletional bias that threatens to purge them from bacterial genomes before they can acquire beneficial function. Compact bacterial genomes, often less than fifteen percent non-coding [67], reflect a pervasive bias favoring loss of non-essential DNA [121,116,179]. This process, seen in natural [68], clinical [6,18], and experimental settings [105,139], is typically framed as a loss of genetic potential. These deletions may arise through diverse mechanisms, including homologous recombination [142], replication-associated slippage [16,35], erroneous repair [74], or site-specific recombinases [161]. Regardless of mechanism, deletions are a hallmark of bacterial genome evolution and can be positively selected by reducing the metabolic cost of replicating DNA or by tuning gene interaction networks [67,179,117].

Deletions can also generate new arrangements of existing material. In a recent analysis of the Lenski Long-Term Evolution Experiment (LTEE)[106], uz-Zaman et al. found that deletions moving regulatory elements contributed the largest share to the transcription and translation of previously non-coding DNA[169]. In other instances, deletions generated functional gene fusions that led to new anti-phage defense functions[170], to phenotypes with increased colony spread[57], and to potentially adaptive gene products in Mycobacterium tuberculosis lineages[66]. Despite the importance of gene fusions in generating novel phenotypes and individual cases linking deletions to the emergence of fusion genes, it remains unknown whether this represents a widespread mechanism for bacterial genome innovation.

Here, we explore a model of bacterial gene birth in which a deletion results in the fusion of the start of one open reading frame (ORF) and the end of another to create a novel ORF. Next, we identify the creation and maintenance of these deletion-born fusions in two densely sampled contexts: in the Lenski LTEE, and during mycobacterial speciation. Finally, we develop a computational technique to efficiently query putative deletion-born fusions across multi-million bacterial isolate genome collections and find evidence for this mechanism of gene birth across the bacterial tree of life. Our findings reframe the bacterial deletional bias as not merely destructive, but as a possible creative force.

1.1 Fusion genes can spread by hitchhiking on the fitness benefit of their causal deletion

Deletions in a bacterial genome result in the fusing of two distal sections of genetic material spanning the deletion junction. Either one or both deletion boundaries can fall inside an existing gene, thus generating a chimeric ORF at the junction. This is made more likely by the gene-dense architecture of bacterial genomes[17]. Both the novel ORF and the removal of intervening material can cause changes in relative fitness.

If a deletion confers a fitness benefit, any fusion ORF created as a by-product can increase in fre-

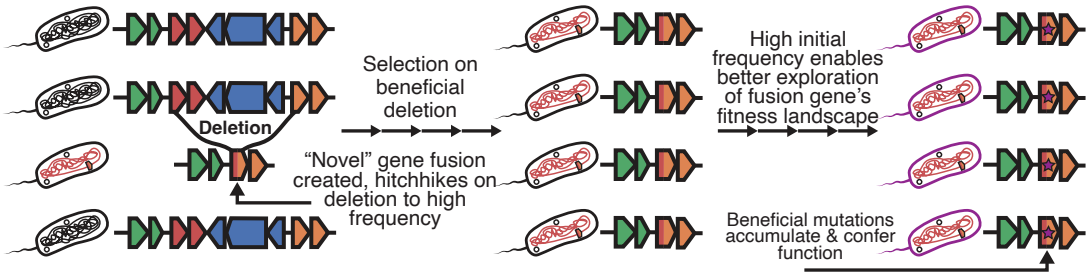


Figure 1.1: Proposed model of deletion-born fusion genes. (1) An initial deletion occurs in a subset of the population and spontaneously creates a deletion-born fusion gene, (2) The deletion is beneficial and rises in frequency with the fusion gene hitchhiking to high frequency, (3) fusion gene at high initial frequency explores fitness landscape and functionalizes.

quency through hitchhiking. Although most nascent genes are expected to be neutral or deleterious regardless of the mechanism by which they originated[88], deletion-born fusions may possess an advantage over other novel genes. Unlike mechanisms that depend on rare, initially neutral genes gradually drifting to appreciable frequency, deletion-born fusions can be quickly driven to elevated frequencies as direct correlates of positively selected structural changes. This hitchhiking may prolong their residence time in the population, increasing the likelihood that subsequent mutations will convert them into beneficial, functional alleles before being lost to drift or purifying selection (Figure 1.1). Moreover, because these fusions are assembled from pre-existing coding material, they are more likely to fold into active catalytic, structural, or regulatory domains. Overall, we expect that deletion-born fusion genes should be more likely to be functionalized than novel genes that emerge through different mechanisms.

To formalize this verbal model, we constructed a stochastic simulation that incorporates the role of genetic hitchhiking and allowed us to contrast the functionalization likelihood of novel genes that arose through different mechanisms. We found that starting novel gene frequency is the dominant driver of neofunctionalization, supporting the hypothesis that hitchhiking on a deletion is a means by which novel ORFs can functionalize (Figure A.1). We constructed a forward-time Wright-Fisher

simulation with 10^6 members, each with one of two states (no novel gene, and non-functional novel gene present with cost c). We assumed that each novel gene is initially non-functional, could be purged each generation with probability p_{purge}, and could functionalize with a probability μ. We also assumed that deletion-born fusion genes quickly increased in frequency to some frequency p_{init} (which should be reflective of the fitness advantage of the genomic deletion that gave rise to the gene) while genes born from other mechanisms always had $p_{init} = 10^{-6}$ (a lone member of the population begins with the gene). We demonstrated that across values for p_{purge} and μ the greatest determinant of at least one member of the population functionalizing the gene before it is purged from the population is p_{init}, the initial frequency of the gene. When $p_{init} = 10^{-6}$ the gene never functionalized consistently; however, as p_{init} increased, the probability of functionalizing increased substantially. This functionalization probability still fell below many values of μ; especially if the fusion is deleterious instead of neutral. It was only when p_{init} approached 1 that neofunctionalization occurred consistently under a sweep of parameters. These results suggested that young deletion-born fusion genes should be present in populations that are adapting to new environments, when genomic deletions are typically the most advantageous (i.e., p_{init} is the highest).

1.2 A novel deletion-born fusion fixes in the Lenski LTEE

To investigate the emergence of deletion-born fusions, we analyzed data from the Lenski LTEE. Briefly, the LTEE tracks the evolution of 12 replicate populations of *Escherichia coli* in minimal media since 1988, spanning over 82,000 generations as of 2025 [11,69]. By clustering predicted ORFs from sequenced clonal isolates against the ancestral genome, we identified a novel fusion gene created by a 57.5 kb deletion in the Ara+1 lineage. This deletion fused *yjcO* (a gene of unknown function) to *lysU* (lysyl-tRNA synthetase) (Figure 1.2A).

The fusion of *yjcO* and *lysU* is expressed and potentially folds but is likely non-enzymatic and func-

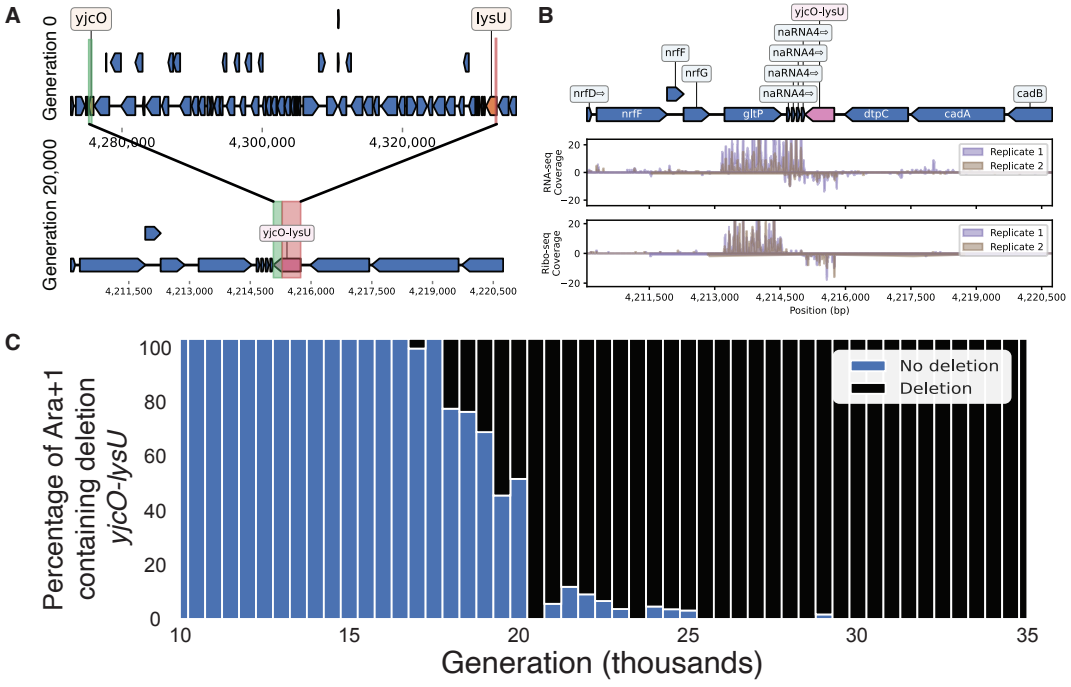


Figure 1.2: A 57.5 kb deletion creates a fusion gene in the Lenski LTEE Ara+1 lineage.
(**A**) **Top**: Genomic locus surrounding *yjcO-lysU* in REL606, **Bottom**: The same genomic locus of the clonal isolate from Generation 20,000 with the deletion and resulting *yjcO-lysU* fusion gene.
(**B**) Genomic locus of the *yjcO-lysU* region with coverage plots corresponding to RNA-seq and Ribo-profiling data from two clonal isolates picked at generation 50,000 in Ara+1.
(**C**) Metagenomic sequencing results comparing the fraction of reads supporting the deletion or not between Generation 10,000 and generation 35,000 in the Ara+1 lineage.

tionally inert. Re-analysis of existing RNA-seq and ribosome profiling data from clonal isolates in Ara+1 revealed that the *yjcO-lysU* fusion is both expressed and translated at generation 50,000 (Figure 1.2B)[58]. Domain annotation shows that the fusion retained multiple Sel1-like repeats from *yjcO*, but no catalytic domains were preserved from *lysU*. Across all clonal isolates sequenced after its appearance, the fusion gene shows no nucleotide substitutions. Based on the gene length and known LTEE mutation rates, we estimate the probability of at least one mutation occurring after its appearance to be $< 1\%$, consistent with the observed absence of variation and suggesting weak or no selection on the fusion.

Both the fusion and the deletion it arose from are first detected in metagenomic sequencing data at around generation 19,500 and appear to fix within $\approx$500 generations (Figure 1.2C). The speed of this selective sweep indicates that either the deletion itself or the resulting *yjcO-lysU* fusion likely conferred an approximate 6.5% fitness advantage under LTEE conditions (see Note A.1).

Supporting the hypothesis that the selective advantage lies in the deletion itself rather than the fusion gene, a parallel 43.4 kb deletion independently arose and fixed in the Ara-3 lineage at the same genomic locus, but did not produce a novel fusion (Figure A.2). Additional analyses of transposon sequencing data confirmed that disrupting the fusion gene had no deleterious effect[40], reinforcing the interpretation that the deletion, not the gene, is the beneficial variant. Genomic regions flanking deletions $\geq$1 kb exhibited significantly greater and more variable changes in expression and translation than randomly sampled regions (Figure A.3). These effects decayed with increasing window size, consistent with deletions inducing localized promoter capture and the disruption of transcriptional context. This supports a model in which deletions remodel local expression and translational landscapes, occasionally generating novel transcribed or translated products.

These results indicate that, even in the relatively constant conditions of the LTEE, new fusion genes can arise from large deletions and quickly increase in frequency in the population.

1.3 Deletion-born fusions emerge during speciation in the *Mycobacterium tuberculosis* complex

We next turned to a well-characterized and deeply sequenced bacterial speciation event, the *Mycobacterium tuberculosis-bovis* divergence, to investigate the potential for deletion-born fusions to arise in complex natural systems [118,31]. This divergence is characterized by multiple large deletions known as 'Regions of Difference' (RDs) that serve as phylogenetic markers for lineage classification within the MTBC [22].

We identified two putative deletion-born fusion genes: *acrR-glcD* (from a 2 kb deletion, corresponding to RD13) and *mlaE-htpX* (from a 12.5 kb deletion, corresponding to RD7) arising during this divergence (Figure 1.3). Using a collection of 47 long-read assemblies (37 *M. tb*, 10 *M. bovis*), we predicted all ORFs, clustered them into orthologous groups, and searched for novel ORFs consistent with deletions (see METHODS). Phylogenetic mapping showed that *acrR-glcD* occurred exclusively in *M. bovis*, while *mlaE-htpX* was found in both *M. bovis* and *M. tb* Lineage 6, the closest relative of *M. bovis* [15]. These monophyletic distributions suggest each fusion arose once, *acrR-glcD* in the *M. bovis* split and *mlaE-htpX* in the *M. bovis*/L6 divergence.

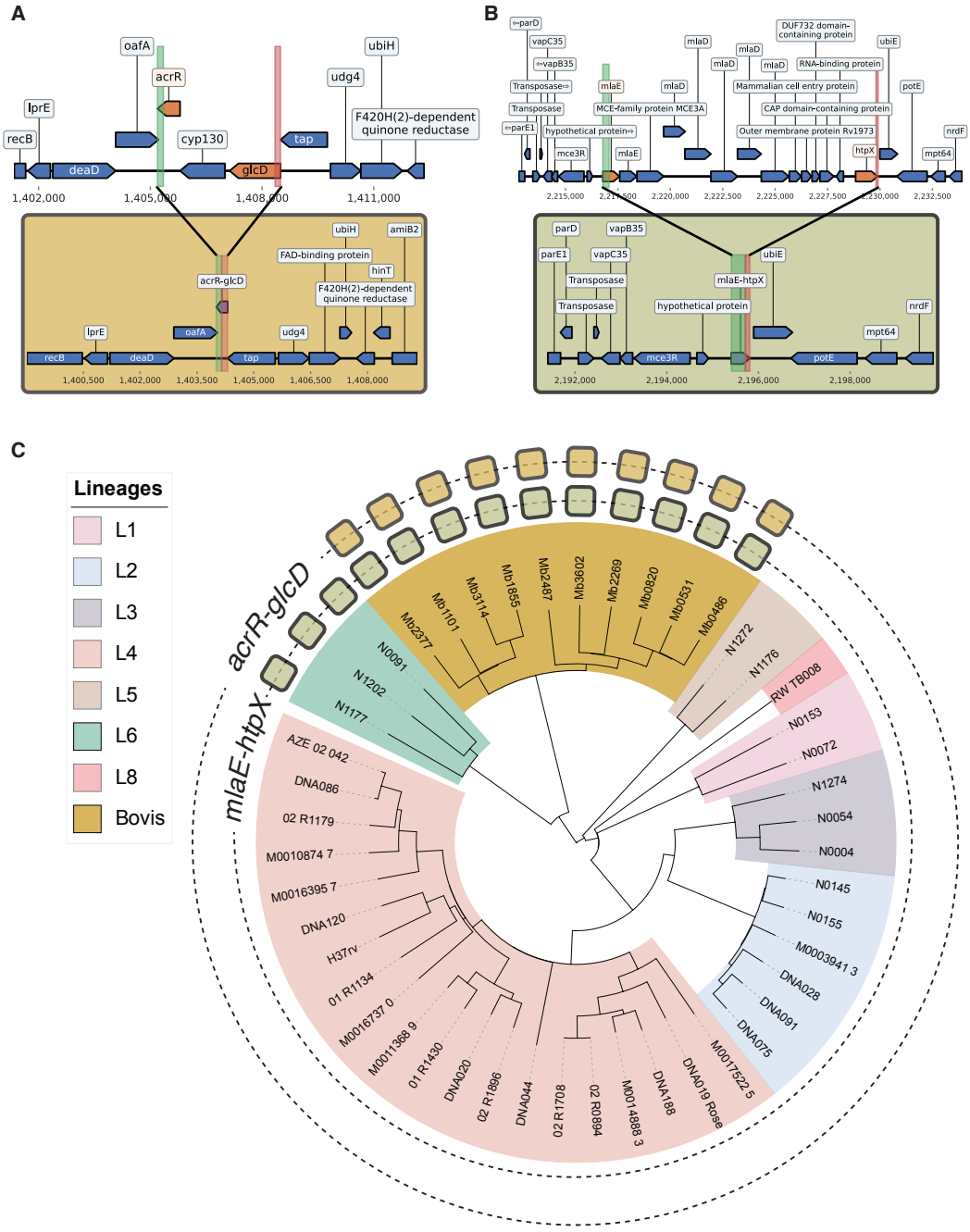
Neither fusion appears to be immediately functional. Structurally, *acrR-glcD* maintains reading frame continuity between the N-terminus of *glcD* and the C-terminus of *acrR*, whereas *mlaE-htpX* fuses *mlaE* in-frame to an out-of-frame fragment of *htpX*. Pfam domain searches revealed that both fusions lost the catalytic motifs of their ancestors and did not generate any new conserved domains. While neither *acrR-glcD* nor *mlaE-htpX* exhibit nucleotide variation in the genomes from Marin et al. and Charles et al.[118,31], a broader analysis (next section) revealed both variation and signatures of selection. The observation of deletion-born fusions in the twin contexts of laboratory evolution and natural speciation in evolutionarily distant bacteria implies that this process may be widespread across bacterial diversity.

1.4 An alignment-free approach to characterize structural variation across a multi-million bacterial genome collection

To query structural variation at the scale of the bacterial tree of life, our previous alignment-based techniques were computationally infeasible. We therefore developed an alignment-free approach that queries short k-mers from the beginning ("prefix") and the ending ("suffix") of candidate genes to infer structural rearrangements based upon the genomic distance between these sequences (Figure 1.4A). Because this method relies only on the location of exact k-mer matches, it can leverage FM-indices for sublinear query times[60], making searches across millions of genomes tractable. We applied this approach to AllTheBacteria (ATB), the largest current bacterial genome collection, comprising 2.4 million uniformly assembled single isolate genomes.

Figure 1.3 *(following page)*: Two deletion-born fusions arose in the *M. tb*/*M. bovis* speciation event.
(A) Top: representative pre-deletion *acrR-glcD* genomic neighborhood in *M. tb* (genome: RW-TB008). **Bottom**: representative post-deletion *acrR-glcD* genomic neighborhood in *M. bovis* (genome: Mb1855).
(B) Top: pre-deletion *mlaE-htpX* genomic neighborhood in *M. tb* (genome: DNA019_Rose). **Bottom**: post-deletion *mlaE-htpX* genomic neighborhood in *M. bovis* + L6 (genome: Mb3602).
(C) Phylogenetic tree of the 47 analyzed samples. Clade colors denote the lineage each isolate falls within, and the filled squares in each ring denote the genomes with *acrR-glcD* and *mlaE-htpX* respectively.

Figure 1.3: *(Continued)*



The distribution of observed prefix-suffix distances reflects structural variation within a given gene. Nearly all insertion, deletion, or partial inversion events will change the distribution of distances. For deletions, the ancestral pre-deletion sequence will have a larger prefix-suffix distance than a fused post-deletion sequence. We used multi-modality in the prefix-suffix distance distribution as an indicator of structural variation within a locus (see METHODS).

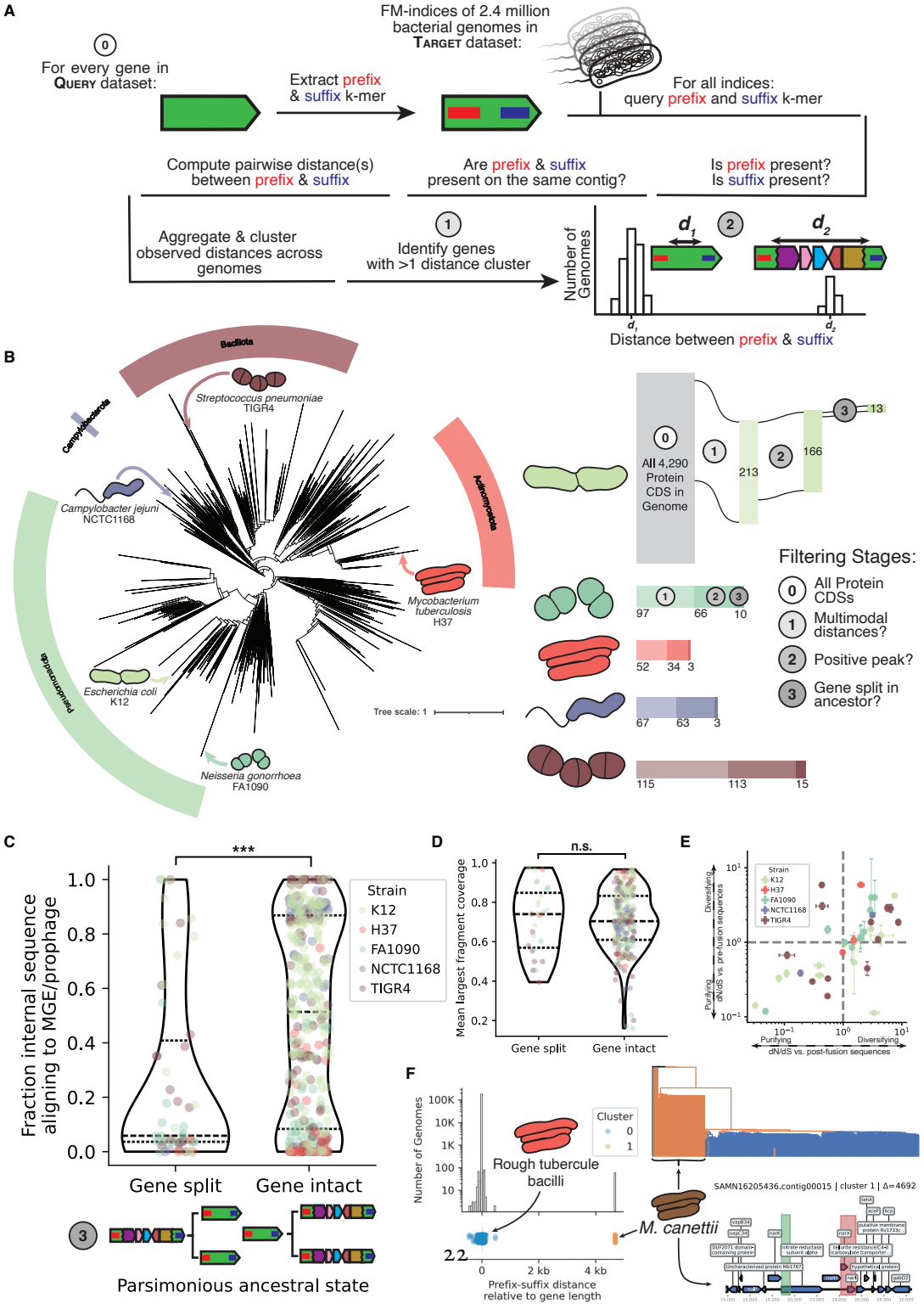
We first validated this prefix-suffix approach on the three previously identified fusion genes and observed distinct peaks at the ancestral and fused distances in all cases (Figure A.4). Notably, the scale of ATB revealed additional structural variation at these loci that was invisible in our earlier, smaller-scale analyses. At the *yjcO-lysU* locus, we identified nine distinct distance clusters spanning a range of structural variants. The 80 genomes in cluster 0 (prefix-suffix distance matching the fusion length) formed a monophyletic clade restricted to LTEE-derived isolates. The remaining 116,892 genomes showed extended prefix-suffix distances ranging from 29 to 96 kb, consistent with independent insertions and deletions at this locus. Cluster 3, comprising 48,892 genomes with a prefix-suffix distance centered at 57 kb, matches the ancestral LTEE length; the remaining 68,000 genomes exhibit distinct structural configurations, revealing additional diversity at this locus (Figure A.4 A).

In Mycobacteria, the deeper sampling from ATB uncovered 4 distinct *acrR-glcD* alleles and 17 *mlaE-htpX* alleles. Alignment-based Bayesian selection analysis showed strong evidence for diversifying selection at codon 50 in *acrR-glcD* (posterior >0.9) and purifying selection at five positions (87, 91, 92, 108, 125) in *mlaE-htpX* (see METHODS). These patterns indicate that, even absent recognizable catalytic domains, these fusions are experiencing selective pressure, consistent with emerging or maintained function.

While we developed the prefix-suffix method to identify deletion-born fusions, the approach detects any recurrent structural variation at a locus. In the next section we filter candidates to deletion-born fusions, but many of the events we filter out are themselves biologically interesting. We note the identification of internal deletions in the *mngB* gene in *E. coli* (Figure A.5A), repeat prophage inser-

Figure 1.4 *(following page)*: Deletion-born fusions are present across the bacterial tree of life.
(A) Schematic of the prefix-suffix k-mer approach.
(B) Left: bacterial tree of life condensed to the genus level showing the five type strains queried. **Right**: filtering cascade with counts of genes passing each stage: (0) all protein-coding ORFs, (1) multimodal prefix-suffix distances, (2) positive distance peak present, (3) gene inferred as split in ancestor. Bacterial illustrations were traced from SEM images: *E. coli* [125], *N. gonorrhoeae* [46], *M. tuberculosis* [42], *C. jejuni* [73], *S. pneumoniae* [159]
(C) Fraction of intervening sequence aligning to MGEs/prophages in genes with split versus intact ancestral states (Mann-Whitney U test, $p < 10^{-6}$).
(D) Mean coverage of the largest alignment block between prefix and suffix k-mers; no significant difference between distributions.
(E) dN/dS estimates for 44 candidate fusions. X-axis: dN/dS versus intact alleles; Y-axis: dN/dS versus reconstructed pre-deletion sequences. Error bars show 95% confidence intervals.
(F) *narX* from *M. tuberculosis* H37Rv. Histogram and stripplot show prefix-suffix distances; Cluster 0 = rough tubercle bacilli (including the fusion), Cluster 1 = *M. canettii* (putative ancestral split state). Phylogeny colored by cluster assignment; bottom panel shows the 10 kb genomic neighborhood of the ancestral *narX* locus in *M. canettii*.

Figure 1.4: *(Continued)*



tions into *rep13e12* gene in *M. tuberculosis* (Figure A.5B), and variable gene cargo in an uncharacterized mobile element disrupting the *pdp* gene in *S. pneumoniae* (Figure A.5C). Thus, the prefix-suffix signal is generalizable to surveying structural variation beyond our specific application and, because of its computational efficiency, scales readily to ever-expanding genome collections.

1.5 Putative deletion-born fusions are found across the bacterial tree of life

We next asked how pervasive deletion-born fusions are across diverse bacterial phyla. We applied the prefix-suffix k-mer screen to all annotated protein-coding sequences from five type strains spanning well-studied and diverse bacterial clades (*Escherichia coli* K12, *Mycobacterium tuberculosis* H37Rv, *Neisseria gonorrhoeae* FA1090, *Campylobacter jejuni* NCTC1168, and *Streptococcus pneumoniae* TIGR4). We selected these species since they span phylogenetically distinct clades, and because their clinical importance has made them among the most deeply sequenced bacteria, with each represented by at least 60,000 genomes in the ATB.

Across the strains, our pipeline resolved 44 putative deletion-born fusion candidates inferred as "split" at the MRCA of the genomes sampled: 13 in *E. coli* K12, 3 in *M. tuberculosis* H37Rv, 10 in *N. gonorrhoeae* FA1090, 3 in *C. jejuni* NCTC1168, and 15 in *S. pneumoniae* TIGR4 (Figure 1.4B).

We identified these putative deletion-born fusions via a sequence of consecutive filters: We first selected genes whose prefix-suffix distances were multimodal, indicative of structural variation. We then retained candidates that showed one cluster centered near the length of the gene (a difference of 0 implying an intact gene) and at least one more cluster at a larger, positive distance (putative pre-deletion sequence) (see METHODS). By solely relying on the distance between the prefix and suffix k-mers, we are unable to distinguish between insertion into a gene and a deletion that led to the formation of that gene. To distinguish between these possibilities, for each candidate gene, we extracted com-

plete genomes with equal representation from each cluster, added outgroups, built a k-mer–based distance tree, and performed parsimony-based ancestral state reconstruction using cluster labels as discrete states (see METHODS). Candidates were retained when the most parsimonious cluster assignment for the most recent common ancestor (MRCA) of all sampled genomes was a cluster with a positive distance length (split gene) implying that the ancestral state was an unfused gene that later experienced a deletion and gene formation.

In genes whose parsimonious ancestral state is predicted to be intact, the observed positive prefix-suffix distance peaks are significantly explained by the insertion of foreign elements, particularly mobile genetic elements (MGEs) and prophages (Figure 1.4C). By contrast, at loci whose MRCA is inferred not to have carried the intact gene, the intervening segments have significantly fewer matches to MGEs or prophages, consistent with separation/fusion via deletion or recombination rather than insertion.

To ensure the distance signals identified were not derived from spurious k-mer matches, we sequence-aligned the putative fusion gene to genomes where it was predicted to be split. We found that all putative deletion-born fusions have at least 80% gap-excluded identity to their split ancestors, implying the prefix-suffix approach detected true sequence homology. Further, the distribution of the relative contributions of the largest alignment fragment in putative deletion-born fusions (MRCA = Gene split) matched that of disrupted genes (MRCA = Gene intact, which we expect to be fully random), implying that spurious k-mer matches to either the prefix or suffix alone are undetectably rare (Figure 1.4D).

The lack of spurious k-mer matches is likely attributed to the significant requirements we enforce: at least 54 base pairs of exact nucleotides matching on the same contig separated by roughly the same amount across numerous genomes. To test how robust this approach was to varying lengths of k we also tested numerous values for three sets of 1,000 random bacterial RefSeq proteins. We found that below $k = 20$, the number of genes with multimodal distributions rises sharply (Figure A.6). Though

some of these signals might be real, we elected to continue with the more stringent value of $k = 27$ to ensure high confidence matches.

1.6 Detection of both deletion-born fusions and broader structural variation is dependent on sampling depth

We next applied the prefix-suffix approach to subsampled collections of 100,000 protein coding ORFs extracted from all 54,630 "complete" RefSeq genomes available at the time of analysis (Figure 1.5A). These genomes contained a total of 219,648,508 protein coding ORFs, these ORFs were clustered into 23,126,961 protein families (see METHODS). Concordant with prior results, we found that most (60.78%) of the protein families have only one sequence within them, so-called "singletons" (Figure A.7A)[54,182,155]. This observation can partially be explained by the bias of which bacterial genomes are selected for sequencing, assembled with high-confidence, and then annotated. Indeed, we find that the top 20 represented species make up about 35% of the 16,774 species sampled across the dataset (Figure A.7B).

We sampled 100,000 protein families and selected representatives from each in three distinct ways. In the first, all protein families were given an equal chance of being drawn ("All Families"). In the second, only those families with at least one other sampled member were sampled ("Non-singletons"). In the third, only families with more than 20 sequences in their family were considered ("Large Families").

Applying the prefix-suffix approach to these 300,000 proteins, we identify increasing deletion-born fusions when moving to more represented protein families (see METHODS). Only 1 gene was identified in "All Families", 8 genes in "Non-singletons" and 44 genes in "Large Families" (Figure 1.5B). Importantly, these same trends are also observed for the intermediate filtering stages: "Large Families" has not only the most putative deletion-born fusions, but also the most proteins where some struc-

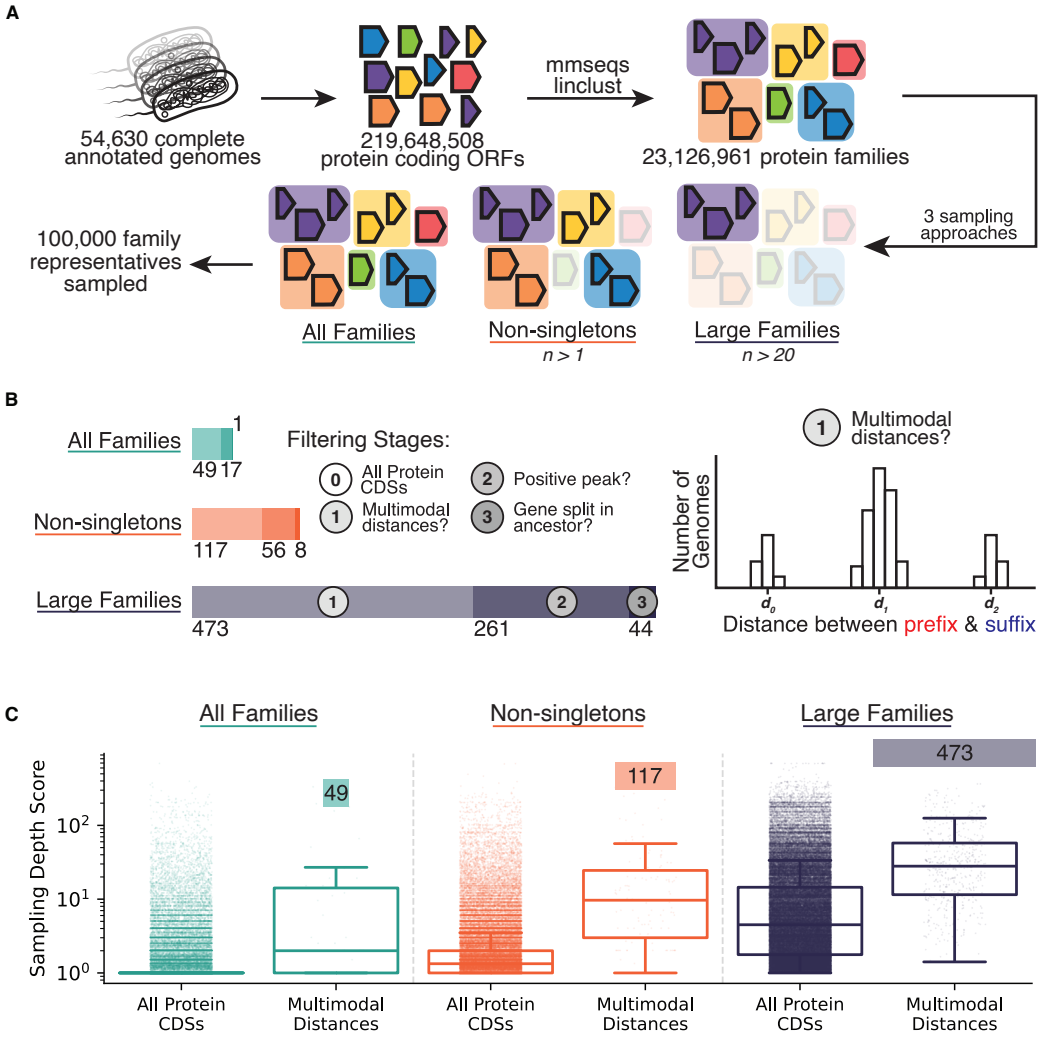


Figure 1.5: Identification of deletion-born fusions scales with genomic sampling depth.
(A) Sampling strategy schematic. Protein-coding ORFs from 54,630 complete RefSeq genomes were clustered into families, and 100,000 families were sampled under three strategies: "All Families" (uniform), "Non-singletons" (≥2 members), and "Large Families" (>20 members).
(B) Filtering cascade showing genes passing each stage. "Large Families" yields 44 deletion-born fusions versus 8 ("Non-singletons") and 1 ("All Families"). Cartoon schematic of multimodal distance filtering stage is depicted on the right.
(C) Sampling depth score (total genomes / total unique species) for all sampled proteins and multimodal proteins in each sampling strategy.

tural rearrangement was observed (473 multimodal distances vs. 49 in "All Families"; 217 putative gene insertions vs. 16 in "All Families").

We sought to examine the extent to which this disparity reflects sampling depth rather than biological distribution by defining a sampling depth score as the total number of genomes containing members of that family divided by the number of unique species represented. A score of 1 indicates a protein family sampled either once or broadly across many species, whereas higher scores indicate repeated sampling within the same species and thus deeper lineage-specific coverage.

Larger protein families were enriched for more deeply sampled species. Protein families in the "Large Families" category exhibited a mean sampling depth score of 15.64, indicating that a typical protein in this set is represented, on average, by 15–16 genomes from the same species (Figure 1.5C). This depth was significantly greater than that of the "Non-singletons" (mean = 3.31) and "All Families" (mean = 1.95) categories (Mann–Whitney U test, $p < 10^{-30}$ for all comparisons), explaining the greater power to detect structural variation in more deeply sampled protein families.

Proteins exhibiting multimodal prefix-suffix distance distributions showed significantly higher sampling depth than the full set of proteins drawn with that strategy ("All Families" mean = 24.15, "Non-singletons" mean = 28.64, "Large Families" mean = 50.53; Mann–Whitney U test, $p < 10^{-12}$ for all comparisons). This pattern indicates that, even controlling for sampling strategy, the detection of structural variation is biased toward protein families with deeper within-species sampling.

Together, these results demonstrate that the apparent enrichment of deletion-born fusions in certain datasets is driven by sampling depth rather than biological prevalence. Detecting these events requires observing both pre-deletion and post-deletion states within the same lineage, which in turn demands not only broad phylogenetic sampling but also dense sampling within species. As genomic databases continue to expand and sampling becomes deeper across the bacterial tree of life, we anticipate that deletion-born fusions will be revealed as more widespread than currently detectable.

1.7 Discussion

In this work, we describe a distinct route to gene birth in which adaptive deletions generate fusion ORFs as by-products. These "deletion-born fusion genes" inherit their initial frequency from a beneficial structural change rather than drifting from rarity and are assembled from pre-existing coding material rather than arising de novo.

This mechanism complements existing models while occupying a distinct niche. Duplication and diversification require maintaining redundant copies in genomes under strong streamlining pressure, overprinting must preserve the ancestral reading frame; and horizontal gene transfer introduces novelty but defers origin to an external donor. Classical gene fusions generally persist only when the fusion is beneficial[181]. By contrast, deletion-born fusions arise as incidental consequences of selection acting at the level of genome architecture rather than protein function. They require no external material, no long-term maintenance of redundancy, and no immediate benefit of the fused product. Since these fusions are assembled from existing sequences, they are more likely to produce biophysically viable proteins than sequences emerging de novo from non-coding DNA, as proposed for novel genes arising in eukaryotes[28,173]. Our simulations formalize the hitchhiking advantage showing that elevated starting frequency is the dominant determinant of whether functionalization occurs before loss.

We document this process across multiple evolutionary timescales. In the Lenski LTEE, we observe a large deletion that rapidly sweeps to fixation, generating a novel fusion gene in the process; a convergent deletion at the same locus suggests selection acted on the deletion rather than the gene. At a longer timescale, we identify deletion-born fusions arising during the *Mycobacterium tuberculosis-M. bovis* divergence, demonstrating that such fusions can persist across speciation events. Finally, by screening millions of bacterial genomes, we identify putative deletion-born fusions across diverse bacterial clades, indicating that this mechanism reflects a general opportunity for bacterial genome

innovation.

We have not, however, identified any clear novel beneficial function to an identified deletion-born fusion. The *yjcO-lysU* fusion in the LTEE shows no nucleotide variation and the acrR-glcD and mlaE-htpX fusions in the MTBC lack catalytic domains and show minimal variation. Most candidates from our tree-of-life screen exhibit dN/dS ratios consistent with diversifying rather than strong purifying selection, suggesting they remain in the early stages of sequence exploration. The fusions identified here likely represent snapshots of this process, in which proteins persist long enough to sample sequence space but have not yet undergone strong functional refinement. Furthermore, while the genome streamlining literature provides strong indirect evidence that deletions are often beneficial, we have not proven any specific deletion was advantageous and its accompanying fusion was neutral or deleterious.

The absence of clearly functional fusions may reflect biology, but it may also reflect the conservatism of our approach. Our survey provides a lower bound on the prevalence of deletion-born fusions. A single nucleotide mutation in either the prefix or suffix k-mer is sufficient to exclude a genome from the search, so fusions that have accumulated terminal mutations are systematically missed. The problem is compounded by the biology of the events we are seeking deletions often arise through recombination between repetitive sequences, yet repeats are precisely where short-read assemblies tend to break, placing true deletion junctions on contig edges and preventing us from measuring the distance between them. Future studies employing more sensitive indices could relax the exact-match requirement, potentially revealing functionalized fusions that have since diverged at their termini.

Sampling bias further constrains detection. We show that detection scales with sampling depth: the "Large Families" category, drawn from deeply sampled protein families, yielded 44 deletion-born fusion candidates compared to just 1 in the uniformly sampled "All Families" category. This disparity likely reflects the requirement to observe both pre- and post-deletion states within the same dataset, rather than true biological enrichment in well-studied species. Current genomic databases are dom-

inated by culturable, human-associated pathogens; environmental lineages remain sparsely sampled. This creates a double bind: we both lack the breadth to detect fusions deep in the bacterial tree, and the depth to detect recent fusions in poorly sampled lineages. We expect organisms whose lifestyles predispose them to deletions (i.e. intracellular pathogens) to be enriched for deletion-born fusions, yet many remain too sparsely sequenced to test this hypothesis.

This is not merely a limitation of our approach. Any broad computational analysis of existing databases will inherit these gaps, a fact particularly pertinent given the recent proliferation of protein language models. Our protein family analysis demonstrated that most bacterial proteins in high-quality genome collections remain singletons, solely sampled once. We attribute this to the same sampling bias: we simply have not conducted deep, high-quality sampling across bacterial diversity. Targeted sequencing of underrepresented taxa will be necessary to close these gaps, underscoring that large-scale computational analyses are only as powerful as the underlying biological data relied upon.

The prefix-suffix k-mer approach developed here has utility beyond deletion-born fusions. It enables detection of recurrent structural variation at specific loci across multi-million genome collections neither relying on alignments nor prior knowledge of ancestral states. In this study alone, we characterized internal deletions, repeat prophage insertions, and variable gene cargo in an uncharacterized mobile element. More broadly, the method should be applicable to any process that alters the genomic distance between conserved flanking sequences: novel bacterial or phage introns [101,120], integron cassette expansion [100], variable plasmid cargo [150], or structural variation in complex metagenomic communities [32]. We have not yet characterized which protein domains or functional categories are enriched among structurally variable loci; such meta-analyses are a natural extension. As sequence databases expand in scale and complexity, alignment-free approaches may become the only tractable means of extracting biological signal from the noisy chorus of bacterial diversity.

Our characterization of deletion-born fusions carries implications beyond evolutionary biology. In clinical microbiology, large deletions are frequently observed during adaptation to host environ-

ments[6,18]; our results suggest that some of these events may generate novel proteins with unpredictable properties. In synthetic biology, rational protein design often proceeds by domain shuffling fusing functional modules from different proteins to create chimeras with new activities[90]. Understanding how nature assembles and filters such chimeras may inform future efforts to engineer functional novelty.

1.8 Methods

1.8.1 Simulation Framework

We simulated the fate of a novel gene using a minimal forward-time Wright–Fisher model with a fixed haploid population size $N = 10^6$ and two initial states: 0 (no fusion), and 1 (non-functional novel gene). State 1 carried a fitness cost c (relative fitness $1 - c$), while states 0 had fitness 1. State 2 is when the novel gene functionalizes, and no individual started at this point. Simulations were initialized with the fusion in state 1 at frequency p_{init} and the rest at state 0. Populations were updated each generation by fitness-proportionate multinomial sampling (selection + drift). After "reproduction", individuals in state 1 could functionalize (transition from state 1 to state 2) with probability p_{func} per generation, and novel gene-bearing individuals in state 1 can lose the gene (transition to state 0) with probability p_{purge} per generation. Each replicate was run until the novel gene was either lost or functionalized (defined as the first appearance of any state 2 individual, with a maximum of 100,000 generations.

For parameter sweeps (Figure A.1), we evaluated a log-spaced grid (20 points) of p_{func} from 10^{-16} to 10^{-2} and p_{purge} from 10^{-8} to 10^{-2} for each combination of c and p_{init} running 100 replicate simulations per grid point with a fixed seed of 42. Simulations were implemented in Python using numpy and results were saved as matrices of functionalization probabilities[72]. Visualizations provided by seaborn[171]. Code for simulations is available in the project repository.

1.8.2 LTEE Data Analysis

We obtained mutation call files for LTEE clonal isolates from the Barrick lab LTEE-Ecoli repository (https://github.com/barricklab/LTEE-Ecoli). For each .gd file, we used gdtools APPLY to generate an isolate-specific genome sequence by applying the curated variants to REL606. We produced per-isolate FASTA outputs. These isolate-specific assemblies were used for consistent gene calling and pangenome construction across all isolates.

To provide standardized gene models for pangenome inference, we annotated each isolate-specific FASTA with Prokka[149], producing per-isolate GFF files. These GFFs were used as the input to Panaroo[166]. For each of the twelve lineages, we ran Panaroo on the set of clonal isolate annotations for that population together with the corresponding ancestral annotation (Anc+ or Anc−). Panaroo was run with –merge_paralogs enabled and with stringent similarity thresholds (e.g., –len_dif_percent 0.98, –threshold 0.98, –family_threshold 0.7), using the "moderate" clean mode.

To identify candidate novel genes arising during evolution, we analyzed each population-specific Panaroo run using a custom script. Briefly, isolate identifiers were mapped to LTEE metadata (population and generation), and Panaroo's gene_presence_absence.csv and struct_presence_absence.Rtab were scanned to identify gene clusters whose first appearance occurred after the ancestor and increased in presence among later isolates ("appearing genes").

To distinguish gene families plausibly arising from structural rearrangement from those reflecting annotation noise or near-identical ancestral sequence, we filtered candidate appearing genes by sequence similarity to the REL606 ancestor. For each candidate gene cluster, we extracted its representative nucleotide sequence from pan_genome_reference.fa and aligned it to the ancestral REL606 genome with BLASTN[26]. We computed the fraction of the candidate gene covered by its single largest BLAST match to REL606 and removed candidates with $> 85\%$ coverage by the best match, consistent with those being largely ancestral sequence rather than novel junction-derived sequence.

yjcO-lysU was identified in this manner and its genomic context from the ancestor was visualized using DNA Features Viewer[183]. RNA-sequencing and Ribosome profiling were downloaded from Favate et al. and processed according with the same approach as published[58]. DNA Features Viewer was again used to visualize the reads to the relevant clonal genome.

For candidates that passed the REL606 similarity filter, we extracted a short sequence spanning the putative novel junction. When BLAST produced two distinct hits to REL606, we converted the two hit intervals to BED format and used bedops with a 30 bp range to identify the local overlap/adjacent junction region[128]; when BLAST produced a single hit, we extracted a 30 bp window around the inferred boundary of the match (depending on hit orientation/endpoint). The corresponding subsequence was extracted from the candidate gene sequence and retained as a "junction" FASTA. To query the relative fraction of the population with a specific variant, both the junction FASTA and the original sequence were aligned with minimap278 to all metagenomic reads from Good et al[108,69]. The relative fraction of reads supporting each variant was computed and results were visualized with seaborn. Domain annotation was performed by using hmmsearch (version 3.4) on the translated protein sequence against Pfam-A database (version 38)[53,122].

To analyze the RNA-seq and Ribo-seq coverage changes around large structural variants, we downloaded a LTEE structural variant table from https://barricklab.org/shiny/LTEE-Ecoli/ on 2025.06.30. We first restricted to the 12 sequenced endpoint clones sequenced in the RNA/Ribo dataset and selected events annotated as deletions or substitutions (DEL or SUB) larger than 1 kb. Because the deletion coordinates were reported in REL606 reference coordinates, we transferred breakpoints into each evolved clone's coordinate system using whole-genome alignment block coordinates (*.coords, generated by nucmer)[99]. For each clone, we parsed alignment blocks describing reference-to-query interval mappings and mapped each breakpoint by locating the corresponding block (allowing a 50 bp tolerance) and applying the within-block offset; events with neither breakpoint mappable were excluded, while events with only one breakpoint mappable were evaluated using a one-sided window

around the mapped breakpoint.

Per-base RNA-seq and Ribo-seq coverage was provided as strand-specific depth files for two replicates per clone and for ancestral controls. For each deletion and each window size, we extracted coverage across the mapped interval, averaged coverage across all available replicate and strands to obtain a single mean coverage value and then computed the log_2 fold change of evolved versus ancestral mean coverage in the same window. To assess spatial scale, we repeated this analysis across multiple window sizes (100 bp to 50 kb). As a matched control, for each clone and window size, we sampled the same number of random genomic windows, excluding regions overlapping deletion windows (including flanks), mapped those windows into clone coordinates using the same alignment-based procedure, and computed coverage log_2 fold changes identically. Distributions of absolute log_2 fold change values were then compared between deletion-flanking windows and matched random windows for RNA-seq and Ribo-seq.

1.8.3 Mycobacterium tuberculosis Complex Analysis

We first downloaded the 36 clinical *M. tb* genomes assembled in Marin et al. 2022 from NCBI using BioProjects PRJNA719670, PRJNA480888, PRJNA436997 and PRJNA421446[118]. The 10 *M. bovis* genomes used were also downloaded from NCBI using BioProjects PRJNA832544[31]. Genomes were retrieved in FASTA format and organized per isolate. The *M. tuberculosis* H37Rv reference genome (accession AL123456) was also downloaded in GenBank format and used as a reference for downstream analyses.

To enable consistent gene-content comparisons across isolates, we predicted protein-coding genes de novo for all genomes using pyrodigal and exported predicted proteins as amino acid FASTA files[102]. All predicted proteins across genomes were then clustered into gene families using MMseqs2[158]. We created a single MMseqs2 sequence database from all proteins, clustered sequences using identity-based clustering (minimum sequence identity 0.7), and exported cluster assignments as a tab-delimited

table. These clusters were used to define gene families and identify accessory families whose presence varied across *M. tb* and *M. bovis* isolates.

To prioritize candidate deletion-born fusions, we examined accessory gene families with lineage-restricted presence patterns and then performed nucleotide-level mapping in genomes lacking the gene family. For each candidate family, we extracted representative nucleotide sequences and aligned them to the corresponding genome FASTA files using BLASTN. BLAST hits were converted to genomic intervals, merged to identify discrete matching regions, and filtered to enrich for signatures consistent with deletion-born fusion rather than simple absence, fragmentation, or duplication. Specifically, retained candidates required at least 80% total aligned coverage across merged hits, a best single-hit length not exceeding 80% of the gene length, and multiple hits separated by at least 1 kb in the target genome. Candidates passing these criteria were treated as putative deletion-born fusion genes. Domain analysis was done as before for the *yjcO-lysU* fusion.

To place candidate events in an evolutionary context, we constructed a core-genome alignment and phylogeny using Parsnp with H37Rv as the reference, using the set of *M. tb* and *M. bovis* assemblies analyzed above[94]. The resulting phylogeny was visualized in iTOL and BLAST results were visualized using DNA Features Viewer[107,26,183].

1.8.4 Prefix-Suffix K-mer Screen

We extracted prefix and suffix k-mers from each query gene using coding nucleotide FASTA inputs. For each gene sequence, we took a k-mer of length k ($k = 27$) from near the 5′ end and a second k-mer of length k from near the 3′ end, excluding a small buffer region from each terminus to avoid start and stop codons and to shift the k-mers out of frame relative to the annotated coding sequence. Specifically, we used a fixed gap distance $g = 4$ bp: the prefix k-mer was taken from positions g to $g+k$ (4 to 31). Assume the length of the gene is L, then the suffix k-mer would be taken from $L-g-k$ to $L-g$.

We queried these k-mers against the AllTheBacteria collection of 2.4 million bacterial isolate assemblies. Assemblies were processed in batches according to their Miniphy'd output[20]: genome FASTA files were stored in compressed tar.xz archives, and for each archive we concatenated subsets of 500 genomes into temporary multi-FASTA files and built BWA indices on these batches [109]. Indexing was performed with bwa index, producing FM-indices for each genome batch.

Exact k-mer placements were then obtained using bwa fastmap, run with the query k-mer FASTA and a matching k-mer length equal to k (the same value used in extraction). We used a large maximum hit window (-w 99999) to retain all exact match locations reported by fastmap. Fastmap outputs were gzip-compressed and parsed to recover, for each genome, all contig-level match positions for both prefix and suffix k-mers.

For each genome and each query gene, we computed a prefix-suffix distance only when at least one prefix match and one suffix match occurred on the same contig. Distances were computed from the genomic coordinates of the matched k-mers on that contig. Matches split across contigs were ignored, and genomes lacking a same-contig prefix-suffix pair were treated as missing for that gene.

Supplementary tables can be found at the online preprint and describe the passing putative deletion-born fusion genes. Each gene had their amino acid translated sequence uploaded to SeqHub from Tatta Bio[85]. Export tables were downloaded as CSVs and converted to Excel spreadsheets for dissemination.

1.8.5 Multimodal Distance Detection and Clustering

We identified structurally variable genes by clustering the per-genome prefix-suffix distances for each query gene and testing whether the resulting distance distribution was multimodal. For each gene, we aggregated the set of observed "Difference" values across genomes (the prefix-suffix distance expressed relative to the expected gene length) along with their multiplicities, and clustered these one-dimensional values using a density-based algorithm (DBSCAN1D)[30]. Clustering was run with ep-

silon = 800 (bp) and min_samples = 25, and we treated DBSCAN noise points as outliers (removed from downstream summaries). A gene was called "multimodal" only if DBSCAN identified at least two non-noise clusters, corresponding to two tight peaks in the distance distribution.

To enrich specifically for candidates consistent with deletion-born fusion genes, we applied additional filters to multimodal loci. We required one cluster centered near 0 (the intact allele, where the observed prefix-suffix distance matches the reference gene length) and at least one additional cluster centered at a positive distance greater than 800 bp, consistent with a split ancestral state in which the prefix and suffix k-mers are separated by a substantial intervening segment. Genes with only small positive shifts (e.g., internal deletions) or without an intact-like cluster near 0 were excluded at this stage. The set of genes passing these clustering and distance-peak criteria was carried forward for phylogenetic filtering and downstream analyses.

1.8.6 Phylogenetic Reconstruction and Ancestral State Inference

To distinguish deletion-born fusions from gene disruption by insertion (Figure 1.4B, filter 3), we performed phylogenetic reconstruction and ancestral state inference using the DBSCAN cluster assignments from the prefix-suffix distance analysis. For each candidate gene passing the multimodality and peak-shape filters, we downsampled genomes to obtain a tractable but representative set for tree building by sampling equal numbers of genomes from each distance cluster (300 total genomes per gene, split evenly across clusters). Sampling was restricted to a high-quality genome set from the ATB to reduce artifacts from fragmented assemblies. For each sampled genome, we extracted the identified prefix and suffix k-mer, as well as the entire interleaving sequence to a new FASTA file and masked that same region in the whole genome sequence.

Trees were constructed from masked sequences using attotree (an optimized version of Mashtree) with default options on the masked whole genome sequences [19,92]. To enable rooting, we also added two outgroup genomes per focal species, chosen as close relatives outside the focal clade: for *E. coli*

K-12, *Klebsiella pneumoniae* MGH78578 (GCF_000016305.1) and *Salmonella enterica* serovar Typhimurium LT2 (GCF_000006945.2); for *M. tb.* H37Rv, *M. canettii* (NC_015848.1) and *M. caprae* (CP016401.1); for *N. gonorrhoeae* FA1090, *N. lactamica* (NC_014752.1) and *N. meningitidis* MC58 (NC_003112.2); for *C. jejuni* NCTC1168, *C. coli* (NC_022660.1) and *C. lari* (NC_012039.1); and for *S. pneumoniae* TIGR4, *S. mitis* (FN568063.1) and *S. oralis* (FR720602.1). For each of the two possible outgroups included, we identified the genome with the largest mean distance between it and the other leaves and chose that as the outgroup to root the resulting tree at.

For each candidate gene passing the distance-based filters, we inferred whether the ancestral state was "split" or "intact" using a tree-based parsimony approach. Cluster labels were treated as discrete character states, and we reconstructed internal node states by multi-state Fitch parsimony[61]. The ancestral state for each gene was defined as the parsimony assignment at the most recent common ancestor (MRCA) of the ingroup genomes. Genes were classified as consistent with deletion-born fusions when the MRCA state corresponded to a "split" cluster (the cluster with a positive prefix-suffix difference) and at least one descendant clade carried an "intact" cluster (the cluster with mean difference near 0). Conversely, genes whose MRCA state was "intact" were interpreted as cases where the intact gene was ancestral, and the positive-distance cluster reflects disruption (often by insertion) and were excluded from the deletion-born fusion set.

1.8.7 Mobile Genetic Element and Prophage Detection

To assess whether structurally variable loci were dominated by insertions of mobile genetic elements (MGEs) or prophages (Figure 1.4C), we aligned the sequence between the matched prefix and suffix k-mers to curated MGE and prophage databases. The MGE database was taken from MGEdb and the prophage database from Prophage-DB (bacterial host prophages)[87,51]; the two FASTA sets were concatenated into a single nucleotide BLAST database. For each genome sampled, we extracted the intervening sequence between the matched prefix and suffix k-mers on the same contig (i.e., the

sequence whose length drives the positive prefix-suffix distance signal) and queried it against the combined database using BLASTN. For each intervening sequence, we merged overlapping BLAST hit intervals along the query and computed the total number of query bases covered by any hit; the fraction of the intervening sequence explained by MGEs/prophages was then calculated as covered bases divided by query length. These per-genome fractions were summarized by locus and compared between loci whose inferred MRCA state was "split" versus "intact" using a two-sided Mann–Whitney U test.

1.8.8 Selection Analysis

For the selection analysis on *mlaE-htpX* and *acrR-glcD*, genomes with the intact ORF were identified based on their membership to the cluster with a relative prefix-suffix "Distance" of 0. The prefix-suffix k-mer and intervening nucleotides were extracted, deduplicated, and codon-aligned using MACSE with default options [140]. A phylogeny of the masked whole genome sequences was built using Parsnp and selection analysis was performed with FUBAR, implemented in the HyPhy package with default options [124].

We assessed signatures of selection on candidate deletion-born fusions using two complementary dN/dS-style comparisons (Figure 1.4E). First, to estimate selection acting on the observed "intact" allele, we focused on genomes assigned to the cluster closest to zero difference (the intact-like cluster). For each genome we used the extracted locus sequence and deduplicated identical sequences by hashing, retaining a count of how many genomes shared each unique sequence. Each unique sequence was then codon-aligned to the canonical query CDS using MACSE [140], and we counted synonymous and nonsynonymous differences across aligned codons, ignoring codons overlapping gaps or ambiguous bases and tracking premature stop gains separately. Per-cluster estimates were computed by summing synonymous and nonsynonymous counts across unique sequences and weighting each sequence by the number of genomes in which it occurred; dN/dS was then calculated as weighted nonsynonymous

divided by weighted synonymous changes.

Second, to estimate the degree of divergence expected from the pre-deletion ("split") state, we constructed "surrogate" genes for genomes assigned to the inferred ancestral cluster (the MRCA cluster from parsimony). For each genome we used the extracted intervening locus sequence and flanking sequence, deduplicated both sequence sets, and mapped each unique removed-region sequence into its parent flank sequence by exact substring matching (allowing reverse complement). We then aligned the canonical query CDS to each unique removed-region sequence using BLASTN and selected a set of high-scoring segment pairs that approximately tiled the query. Each BLAST segment was projected back onto the parent flank coordinates and intersected with ORFs predicted on the flank sequence using pyrodigal; for each segment we selected the ORF with the greatest overlap and stitched the resulting ORF nucleotide sequences together in query order, joining adjacent blocks with "NNN" to preserve codon-phase ambiguity. These stitched constructs were treated as surrogate pre-deletion sequences and were codon-aligned to the canonical query CDS again using MACSE, after which synonymous and nonsynonymous differences were counted as above and aggregated into a weighted dN/dS estimate.

1.8.9 Protein Family Clustering and Sampling

We constructed a large-scale bacterial protein family catalogue from complete RefSeq genomes to quantify how sampling depth affects detection of deletion-born fusion genes (Figure 1.5). All complete bacterial genomes available in RefSeq at the time of analysis (54,630 assemblies) were downloaded using ncbi-genome-download in GenBank format. From each genome, we extracted all annotated protein-coding sequences by parsing the GenBank feature tables. For each CDS, we extracted both the nucleotide sequence and the corresponding protein sequence. When a translation was provided in the GenBank record it was used directly; otherwise, the nucleotide sequence was translated in-frame. Protein sequences were written to per-genome FASTA files and then concatenated into a

single combined protein FASTA for clustering.

Protein sequences were clustered into gene families using MMseqs2 Linclust[158]. Clustering was performed with a minimum pairwise sequence identity of 80%, a minimum target coverage of 80% (coverage mode 1), and 80 k-mers per sequence, using mmseqs easy-linclust. This procedure yielded 23,126,961 protein families, spanning both singleton and multi-member families. The resulting cluster table and representative sequences were used for downstream sampling and analysis.

To evaluate how database sampling affects detection of structural variation, we defined three complementary protein-family sampling strategies. In the "All Families" strategy, we sampled protein families uniformly at random from the full set of MMseqs2 clusters, including singletons. In the "Non-singletons" strategy, we excluded singleton families and sampled uniformly from families containing at least two members. In the "Large Families" strategy, we restricted sampling to families with more than 20 members, enriching for deeply sampled lineages. For each strategy, we analyzed an equal number of protein families (100,000).

For each protein family, we quantified sampling depth using a sampling depth score defined as the total number of genomes containing members of that family divided by the number of unique species represented among those genomes. A score near 1 indicates a family sampled broadly but shallowly across species, whereas higher values indicate repeated sampling of the same species (e.g., many isolates of a single lineage). Sampling depth scores were used to compare the effective detectability of deletion-born fusions across the three sampling strategies.

1.9 Acknowledgments

I would like to warmly thank all past and present members of the Baym and GenScale team for both illuminating conversations and additional scientific insight. I would also like to thank Dr. Luke Barrett Harrison for helping to catch a misprint in our preprint. This work was supported by NIGMS

of the National Institutes of Health (R35GM133700 and R35GM156320), the David and Lucile Packard Foundation, the Pew Charitable Trusts, and the Alfred P. Sloan Foundation. The prefix-suffix approach is based upon research performed in France within the GenScale team at the Inria Center at Rennes University. The work was supported by a Chateaubriand Fellowship of the Office for Science & Technology of the Embassy of France in the United States and a mobility grant from the Collège doctoral de Bretagne. This research was additionally supported by the French National Research Agency (ANR) under Grant ANR-24-CE45-1226 for the REALL project (KB). Portions of this research were conducted on the O2 High Performance Compute Cluster, supported by the Research Computing Group, at Harvard Medical School.

Philip Larkin famously proposed that what will survive of us is love.

Wrong. What will survive of us is plastic, swine bones and lead-207,

the stable isotope at the end of the uranium-235 decay chain.

Robert Macfarlane, *Underland: A Deep Time Journey* (2019)

2

Genomic resistance in clinical bacteria increased in frequency and mobility after the age of antibiotics

The following chapter has been published, with DOI: 10.1099/mgen.0.001474

Since the earliest days of modern antibiotics, their use has been threatened by resistance[44]. Bacteria acquire antibiotic resistance, the ability to survive a higher dose than can be safely given to a patient, both by mutating their genomes and acquiring new genes. The use of antibiotics then selects for resistant mutants, allowing them to preferentially replicate and spread. Indeed, for every antibiotic introduced, resistance soon follows[44].

To understand the evolutionary history of clinical resistance, we not only need to understand the selective pressures on resistant mutants, but also what the frequency and prevalence of resistance alleles were before the widespread use of antibiotics. This presents a problem: how can we know the rate of resistance to an antibiotic before it was even discovered? Contemporaneous with early antibiotic discovery, freeze-dried collections of bacterial isolates began to be collected and preserved to this day. In 1920, the UK's National Collection of Type Cultures (NCTC) was founded, preserving both clinical and environmental isolates for long-term storage. Prior studies have leveraged these collections to gain insights from the phylogeography of individual strains to retrospective diagnoses of ancient disease outbreaks[14,130,47]. More recently, large-scale sequencing efforts have produced many publicly available, high-quality genomic assemblies from the NCTC, which we use here to study resistance alleles around the advent of antibiotics[49].

Most explanations for antibiotic resistance attribute the problem to human use of antibiotics, suggesting that consumption and production of these drugs have created selective pressure that drives bacteria to develop and retain resistance[44]. However, recent work studying microbes isolated before the widespread use of antibiotics has found antibiotic resistance even in clinically relevant isolates. The first NCTC isolate accessioned, NCTC 1, a *Shigella flexneri* isolated in 1915, is resistant to penicillin and erythromycin, well before the discovery of either drug[8,7]. Phylogenetic reconstruction has predicted that methicillin-resistant *Staphylococcus aureus* first emerged in the mid-1940s, about 14 years before the introduction of methicillin in the clinic[71]. In the environment, genomic studies of microbial isolates from permafrost have identified functional antibiotic resistance genes well

over 30,000 years old[45,137]. While this natural occurrence is unsurprising, given that microbial antibiotic production far predates human use, the frequency, type and context of resistance mutations in human-associated bacteria before the use of antibiotics remain unknown[45].

Here, we sought to identify and quantify trends in resistance genes contemporaneous with antibiotic introduction in genomes from the NCTC. We note that the present study, by virtue of being solely computational, only considers the presence of genomic resistance determinants. Carriage of a known resistance gene or allele does not necessarily imply resistance to that antibiotic. For example, if a resistance gene is unexpressed, the microbe may still be antibiotic sensitive. However, carriage of a resistance element enables a larger pool of possible mutations that induce phenotypic resistance[143].

We measured and characterized the genomic resistance that existed before the age of antibiotics to better understand how the clinical introduction of a given antibiotic correlates with differences in the frequency and features of resistance. Through the construction and analysis of a time-matched database of microbial genomes, we show that human introduction of an antibiotic is associated with changing both the prevalence and mobility of genomic resistance elements within isolates deposited in the NCTC.

2.1 Methods

2.1.1 Creation of a time-matched database of microbial genomes

To generate a historical database of microbial genomes, we turned to the NCTC, administered by the UK Health Security Agency. We chose the NCTC due to its status as the oldest strain collection in the world, and as a result, it is one of the most diverse collections of isolates cultured before the age of antibiotics[3]. We note that although the database has a predominantly clinical focus, there are also various environmental isolates deposited.

To generate the database, we systematically analyzed the web metadata associated with each isolate

from the NCTC web server, as no complete database of isolates is publicly available (https://culturecollections.org.uk). Upon downloading each isolate's metadata, we aggregated the metadata and then performed a regex search for a putative year of isolation. Due to the NCTC's long history, standardized data entry fields and data logging practices are not common throughout the database. If a year of isolation is recorded, it can appear in several different fields. From the regex search, we take all prospective years of isolation and rank order the priority of each year based on a heuristic determined from direct communication with the NCTC. Although a 'collection date' field exists for assemblies deposited in GenBank, we found it to be inaccurate: out of 2,958 assemblies in the NCTC3000 project, 1,845 have non-empty collection dates and only 590 have a single year listed in the collection_date entry. We also found several strains with demonstrably incorrect collection_date entries. For example, NCTC 11317 is listed as being isolated in 1969, while the GenBank entry for it lists the collection_date as 2015. In another instance, NCTC 13760 is only listed as being 'pre-1989' but has a collection_date entry of 1949. For this reason, we neglected to use the collection_date entry and relied on the NCTC metadata associated with each to determine a year of isolation.

For all strains for which a year of isolation was determined, we searched for complete genome assemblies within the European Nucleotide Archive (ENA), RefSeq and GenBank. We found that if a given accession had a genome assembly in either RefSeq or GenBank, then there was a copy also deposited within the ENA. For consistency, we only moved forward with assemblies downloaded from the ENA. This output was then merged and analyzed (Figure 2.1). Out of the 1,919 genomes in our database, 1,752 are from the NCTC3000 collection (91%). The NCTC3000 was generated from a collaboration between the NCTC, Pacific Biosciences and the Wellcome Sanger Institute, which sequenced and assembled around 3,000 NCTC isolates using long-read technology[49].

To evaluate the quality of our genomic dataset, we calculated three metrics: CheckM2 completeness and contamination scores, as well as N50[34]. We defined a genome as low quality if it had a completeness score below 90%, a contamination score above 5% or an N50 value less than 5 kb. After

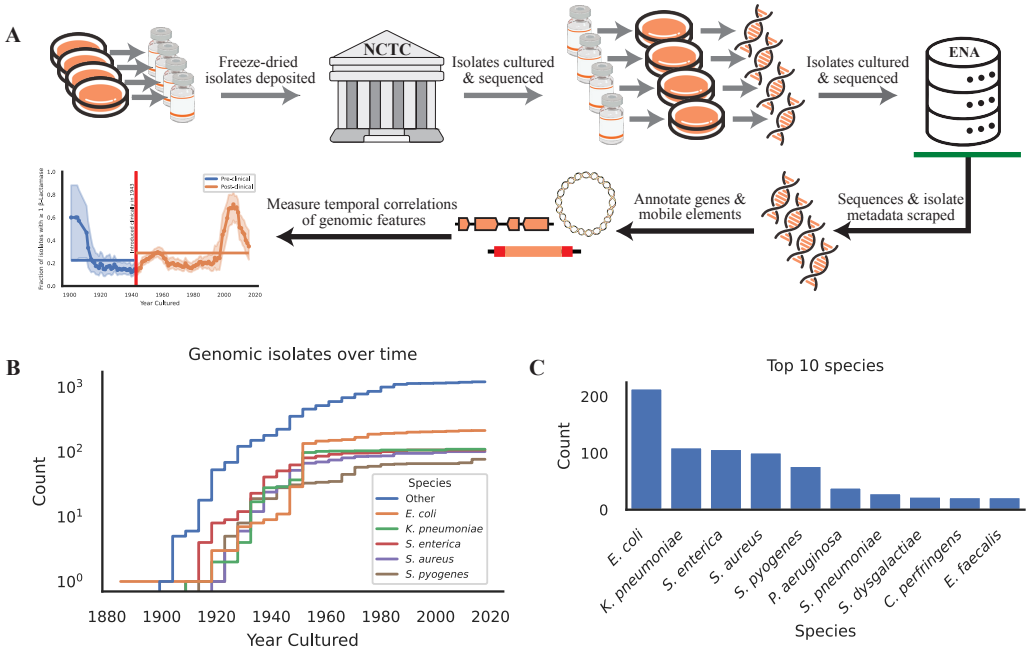


Figure 2.1: Overview of the collated database.

(A) Schematic of the workflow: arrows after the green bar indicate novel work done in this paper.

(B) Cumulative count of the genomic isolates captured over time, coloured with the top five species represented in the database.

(C) Raw counts of the top 10 species in the database.

applying these quality control criteria, we filtered out 102 genomes from our initial dataset of 1,919 and were left with a final set of 1,817 high-quality genomes suitable for downstream analysis.

2.1.2 Identification of genomic resistance elements

Here, we use the term 'genomic resistance elements' to mean loci on a genome (either SNPs or complete genes), which have been associated with resistance to antibiotics. While the presence of these genomic resistance elements is not a direct predictor of resistance, their presence does correlate with phenotypic resistance[39]. To identify genomic resistance elements within each isolate, we took two orthogonal approaches. First, we used the Resistance Genome Identifier and version 3.2.2 of the Comprehensive Antibiotic Resistance Database (CARD) with the parameters (–input_type contig –exclude_nudge)[2]. CARD is an extensive database of curated antimicrobial resistance gene sequences and resistance-conferring mutations and can provide a complete view of genomic resistance elements present within a given genome. Additionally, we used AMRFinderPlus on isolates from one of the following taxa using the –organism flag (*Klebsiella pneumoniae*, *Escherichia coli*, *Salmonella enterica*, *Acinetobacter baumannii*, *Enterococcus faecium*, *Staphylococcus aureus*, *Streptococcus pyogenes*, *Streptococcus pneumoniae*, *Enterobacter cloacae* and *Pseudomonas aeruginosa*)[59]. AMRFinderPlus similarly performs alignment searches in a large database of known resistance-associated alleles. When combined with the –organism flag, it also incorporates prior knowledge about which resistance alleles are commonly found in each taxon and filters those results out. As a result, it should be viewed as a refinement tool, allowing for the identification of resistance alleles specific to each taxon. Results from both approaches are described, and it is noted when a given analysis uses one set of results over the other.

2.1.3 Identification of mobility

To classify plasmids present in the genomic assemblies analyzed, we used MOB-suite's MOB-recon and MOB-typer tools with default options, and every contig analyzed was classified as deriving from the chromosome or a plasmid[38,141]. MOB-recon identifies plasmids based on the presence of known replicons and relaxases, and MOB-typer further characterizes them using additional features such as *oriT* sequences and mate-pair formation markers. While this approach primarily identifies plasmids, it is possible that a small number of integrative and conjugative elements (ICEs) or integrative and mobilizable elements, which share some conjugation-related genes with plasmids, may be included in the classification. To identify integrons, we used IntegronFinder with default options[129]. To identify transposons (TN) and insertion sequences (IS) present in each isolate, we used MobileElementFinder with default options[87]. To identify prophages, we used phigaro with default options[157].

For all genomic resistance elements, we classified them with one or more of the following labels. If the element was present on a contig classified as a plasmid by MOBTyper, then it was determined to be 'plasmid associated'. If 95% or more of the element sequence was encompassed by an integron gene cassette, then it was determined to be 'integron associated'. If the element was flanked by two TN or two IS elements within 10 kb of each other and of the same type, or if the element overlapped with a TN or IS, we classified it as 'TN/IS associated'. If a resistance element overlapped a predicted prophage, we labelled it 'prophage-associated'. To account for the observation that nearby genes can be erroneously packaged inside nearby prophages, we also investigated genomic resistance elements within 1 kb of predicted prophages[27,24]. However, we found minimal changes to the results and elected to move forward with the more stringent overlap requirement. We have kept the annotation of genomic resistance elements within 1 kb distance to prophages in our online tables. If the element met none of these requirements, then it was classified as 'Immobile'.

We saw a need for a standardized computational pipeline to predict genomic element mobility in

bacterial genomes and have packaged this pipeline for general use into the tool, 'CallMeMobile'. This pipeline is publicly available for download and installation on GitHub (https://github.com/aryakaul/callmemobile).

2.1.4 Null distribution for genomic resistance prevalence

The year in which a given antibiotic was clinically introduced was determined from [80]. To determine the significance of the observed difference between the fraction of isolates containing a genomic resistance element against an antibiotic before and after its clinical introduction, we generated a null distribution using two shuffling methods. In the first method, all years of isolation were collated, and each isolate was randomly assigned a year of isolation from the pool of possible values without replacement. This unstructured shuffling allowed for isolates from different years to be reassigned to any other year; for example, some isolates from 1958 could be assigned to 2009, while others could be assigned to 1930. To account for possible structure in the year of isolation, we employed a second method: structured shuffling. In this approach, all isolates from a given year were randomly assigned to the same new year. For example, all isolates cultured in 1958 would be randomly assigned to the same year, 1930.

We performed both shuffling methods 10,000 times to generate a null distribution of differences in resistance prevalence before and after antibiotic introduction. A Gaussian distribution was fitted to the created null distribution, and empirical p-values were calculated based on the observed distance. Because the unstructured method consistently yielded artificially low p-values due to sample density imbalances across years, we report only the p-values derived from the structured shuffling, which better accounts for these temporal sampling biases. To correct for multiple hypothesis testing across antibiotic categories, we applied the Benjamini–Hochberg procedure to control the false discovery rate. Adjusted p-values were computed using the multipletests function in the statsmodel Python package (method='fdr_bh'), and results were considered significant at a corrected p-value threshold of 0.05 [148]. This correction minimizes the proportion of false positives among significant results while retaining

statistical power[13].

2.1.5 Genomic annotation, data processing and visualizations

Genomes were uniformly annotated with Bakta using version 5 of their database and default options[147]. DNA Features Viewer was used to visualize resulting annotations[183]. Dataframes were processed with Pandas, and all plots were created with Seaborn[119,162,171].

2.1.6 Phylogenetic tree of beta-lactamases

To construct a phylogeny from the beta-lactamases in our dataset, we selected all beta-lactamase genes from families with at least 20 members, based on their AMR gene family annotations. Nucleotide sequences for each resistance gene were extracted from the 'AMR Gene Family' field and aligned using MAFFT v7.505 with default parameters[91]. A representative *blaNDM-1* sequence from CARD was included as an outgroup. The resulting multiple sequence alignment was used to generate a maximum likelihood phylogenetic tree using FastTree v2.1.11 with default settings[138]. Tree visualization and metadata annotation, including species, year of isolation and beta-lactamase family, were performed using iTOL (Interactive Tree Of Life) v6[107].

2.1.7 Sampling Bias

As with any analysis of a historical database, sampling bias is a significant consideration. This analysis using the NCTC is no exception.

First, the NCTC is based in the UK and administered by the UK Health Security Agency. As a result, most isolates deposited are either from the UK or Europe, and most isolates are clinically relevant. Thus, the isolates included in our study are not necessarily representative of the global bacterial population present during those periods.

In addition, the availability and preservation of isolates from different periods are not uniform, with some periods being more highly sampled than others (Figure 2.1B). This temporal bias could impact our ability to accurately assess the prevalence and characteristics of antibiotic resistance alleles throughout history. In addition, the isolates deposited are reflective of researchers' priorities at given points in time. For example, there is a noticeable spike in the fraction of isolates with genomic resistance elements deposited after the year 2000; this likely coincides with increased interest in clinical resistance and its markers.

Though we strove to mitigate these biases through random sampling and a focus on species with large sample sizes, we note that the complete elimination of this sampling bias is impossible.

2.2 Creation of a time-matched database of bacterial genomes

At the time of download, 5,665 isolates were present in the NCTC. Of these, 1,919 had both a predicted year of isolation from the metadata and publicly available assembled genomes. These 1,919 isolates represent 605 species with the majority (651 isolates) being either *Escherichia coli* (228), *Klebsiella pneumoniae* (122), *Salmonella enterica* (119), *Staphylococcus aureus* (105), or *Streptococcus pyogenes* (77).

We present an interactive Jupyter notebook in the project Github repository to make the database easily accessible for other researchers (https://github.com/baymlab/MicroTrawler). Both the complete data of the 5,665 isolates scraped from the NCTC, the 1,919 genomes scraped, and the 1,817 high quality genomes used for the rest of this chapter's analysis can be found in the published work.

2.3 Genomic resistance was present in clinical isolates before the age of antibiotics

We first measured the prevalence of resistance to a given antibiotic before the clinical introduction of that antibiotic. To focus our analysis on the most informative genomic resistance elements likely to lead to a phenotypic change in resistance profile, we employed AMRFinderPlus and focused on a subset of the species in our database. These included the 274 isolates belonging to one of the ESKAPE pathogens (*Enterococcus faecium* (6), *Staphylococcus aureus* (101), *Klebsiella pneumoniae* (110), *Acinetobacter baumannii* (9), *Pseudomonas aeruginosa* (39), and *Enterobacter* sp. (9)), the 212 isolates from *Escherichia coli*, the 106 isolates of *Salmonella enterica*, the 77 isolates of *Streptococcus pyogenes*, and the 29 isolates of *Streptococcus pneumoniae* bringing the total to 698 isolates. Each of the 698 assemblies and their species information was input to AMRFinderPlus to query for resistance elements not commonly found within these taxa (see METHODS).

To investigate if these isolates harbored genomic resistance elements before the clinical introduction of an antibiotic class, we computed the fraction of isolates before anthropogenic introduction of a given antibiotic that contain alleles associated with resistance to that antibiotic.

We note that the antibiotic categories used in this analysis are defined by the genomic resistance elements detected, not strictly by chemical class or clinical usage. In cases where a single resistance mechanism confers cross-resistance to multiple antibiotic classes, such as the *oqxAB* efflux pump, which affects both phenicols and quinolones, we grouped these together under a combined category. Conversely, when distinct mechanisms target a single antibiotic (i.e. acetyltransferases specific to chloramphenicol), we treated them separately to reflect their genomic specificity. This gene-first approach allows for more accurate interpretation of resistance dynamics based on what is mechanistically and detectably encoded in the genomes.

We find that although uncommon, it is possible to detect resistance elements in the genomes of clin-

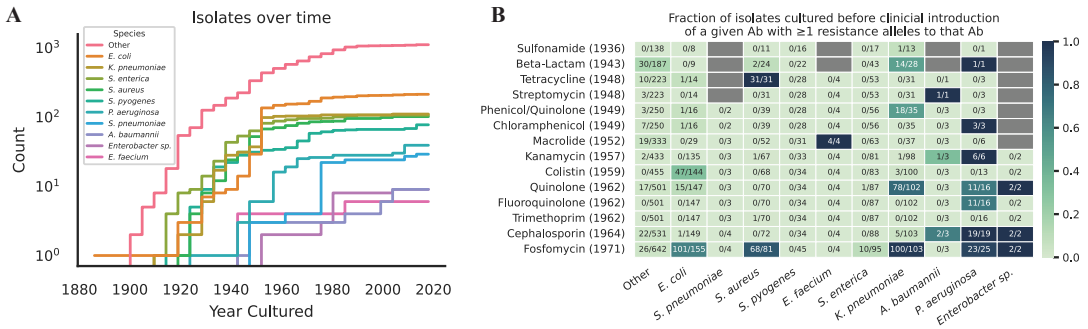


Figure 2.2: Resistance alleles were present but uncommon across species and drug classes before the introduction of antibiotics.

(A) A cumulative histogram of each of the genomic isolates over time in our dataset was analyzed with AMRFinderPlus. **(B)** Heatmap representing the fraction of isolates harboring resistance alleles before the introduction of a given antibiotic. The denominator in each cell is the number of isolates of a given species cultured before the clinical introduction of the antibiotics in the corresponding row (represented by the year next to the antibiotic). The numerator is the number of isolates cultured before clinical introduction of the antibiotic containing resistance alleles to that antibiotic.

ically relevant isolates cultured before the clinical introduction of a given antibiotic (Figure 2.2B, Note B.1). Moreover, our analysis reveals differential rates of resistance for different antibiotics and microbial species. To explore this further, we re-ran the analysis using CARD, this time without applying any taxonomic constraints to filter intrinsic resistance genes (Figure 2.2). This approach includes all resistance elements regardless of whether they are commonly found in the taxon. As a result, it captures many genes that are currently considered part of the normal genomic background of a species. The widespread prevalence of genomic resistance elements observed therefore likely reflect intrinsic resistance elements rather than acquired or emergent resistance mechanisms. This comparison illustrates the importance of taxon-aware filtering when estimating resistance prevalence likely to be clinically or phenotypically significant (Figure 2.2B, Figure B.2).

To better understand these isolates, we selected a variety of pre-antibiotic era isolates with genomic resistance and analyzed their isolation source and the genomic context of the resistance elements (Note B.1, Figure B.3). Overall, we show that while rare, historical isolates deposited in the context of both clinical and environmental settings contain genomic resistance elements to a given antibiotic, even

before clinical usage of that antibiotic.

2.4 Genomic resistance against most antibiotics significantly rose after clinical introduction of that antibiotic

Though we observed some genomic resistance in isolates before the age of antibiotics, we were interested in determining if the clinical introduction of an antibiotic is significantly associated with an increase in the prevalence of genomic resistance to that antibiotic. We began with beta-lactamases, as their prevalence and diversity are well-associated with rising resistance to beta-lactams in the clinic and beta-lactamases are thought to have minimal pleiotropic influences [52,113].

For each year we had isolates from, we measured the fraction of isolates in that year with one or more beta-lactamases. We then took the mean of the fractions before the clinical introduction of penicillin in 1943 and after and measured the difference between the means, denoted by delta. To measure the significance of delta, we shuffled the year of isolation for each isolate and recomputed the delta as before. This null distribution was fitted to a normal distribution, and the observed delta's significance was computed ('Methods'). We find that the increased fraction of isolates containing beta-lactamases is significantly correlated with the clinical introduction of penicillin in 1943 (Figure 2.3A). Subsetting the data to only ESKAPE pathogens provides an even more significant difference (Figure 2.3B). To determine if these results held to broader classes of antibiotics, we expanded this analysis to include all antibiotics for which we had at least 30 isolates with genomic resistance elements.

We also note a peak in resistance prevalence shortly after 2000, followed by a decline beginning around 2010. This trend corresponds to a shift in the species composition of isolates deposited in the NCTC during that period. Specifically, there is an increase in the diversity of species and a relative decrease in the number of deposited ESKAPE pathogens, which contributed to the observed decline in resistance prevalence post-2010 (Figure 2.1B, Figure B.1).

To further explore the temporal dynamics of beta-lactam resistance, we analyzed both the gene family and functional class distributions of beta-lactamases (Figure B.4). Family-level analysis highlights the dominance of SHV and TEM enzymes in the mid-twentieth century (Figure B.4A). Functional reclassification based on CARD annotations and Bush and Jacoby's updated scheme reveals the early prevalence of penicillinases and cephalosporinases, followed by a delayed but notable appearance of extended spectrum beta-lactamases (ESBLs) and carbapenemases in later decades (Figure B.4B)[25]. We note the presence of an *SHV-206* beta-lactamase in a 1937 *Klebsiella pneumoniae* isolate (NCTC 5048); however, this allele lacks the key active site mutations typically required for extended-spectrum activity in SHV beta-lactamases and was, therefore, not classified as an ESBL[168]. To explore the evolutionary origins of beta-lactamase genes in our dataset, we constructed a maximum likelihood phylogenetic tree of all identified beta-lactamase sequences (Figure B.5). The data are consistent with multiple independent origins of beta-lactamase gene families across both time and bacterial taxa, consistent with the parallel emergence and dissemination of resistance via distinct mechanisms.

We find that for most drug classes, the clinical introduction of antibiotics is significantly associated with a rise in resistance alleles present in the NCTC, with the notable exceptions of tetracycline, phenicol/quinolones and fosfomycin (Figure 2.3C, Figure B.6). Pre-1948 tetracycline resistance is driven by the chromosomally encoded efflux pump, *tet*[25], in S. aureus[167]. Phenicol/quinolone resistance before 1949 is largely a result of the *oqxAB* efflux pumps in *Klebsiella pneumoniae*[110]. Though both genomic elements are thought to only induce low- to medium-resistance to these drug classes, they still constitute clinically relevant genotypes that provide evidence that the pangenome of these

isolates often contained genomic resistance elements even before the selective pressure induced by human introduction of antibiotics.

In contrast, fosfomycin resistance pre-dating its clinical introduction in the early 1970s exhibits a broader range of mechanisms. Fosfomycin's mechanism of action involves inhibiting the enzyme *murA*, which catalyses an early step in cell wall synthesis. In some instances, point mutations in *glpT* and *uhpT* were observed that decrease the uptake of fosfomycin into the bacterial cell[180]. In other instances, *murA* exhibited mutations known to cause resistance to fosfomycin[65,163]. Additionally, the presence of *fosA* and *fosB* was noted, which encode enzymes capable of deactivating fosfomycin entirely[55]. These genomic determinants of resistance were spread across the species analyzed. We found no association with the change in genomic resistance prevalence when we further stratified the antibiotics by mechanism of action (Figure B.7).

Together, these analyses show that, for most antibiotic classes, the prevalence of corresponding genomic resistance elements among NCTC isolates increases in the years following their clinical introduction. We were next interested in investigating how genomic mobility and the capacity for horizontal gene transfer have changed for resistance elements over time.

2.5 Genomic resistance exhibits increasing mobility over time

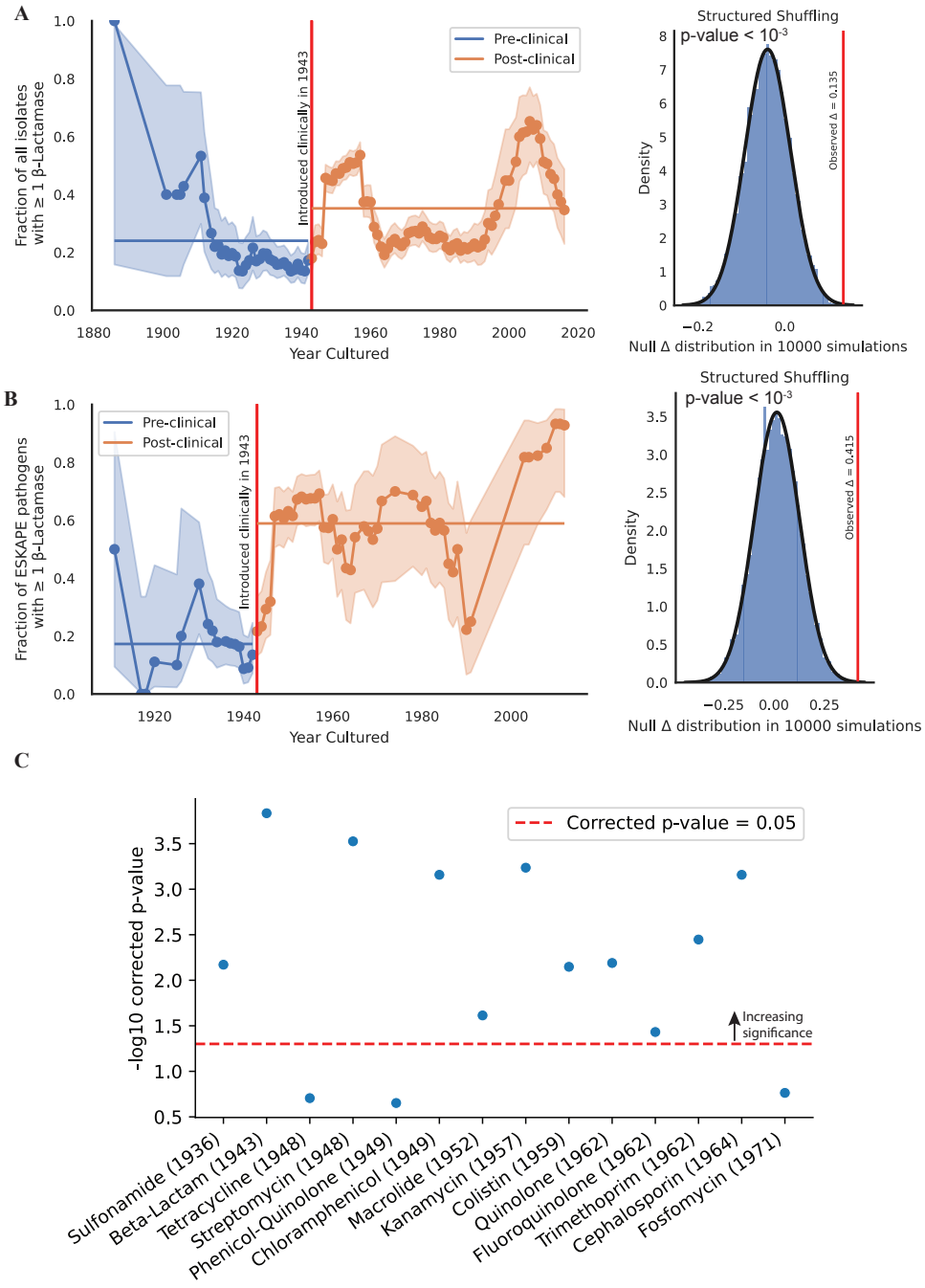
To investigate the trends in the association of antibiotic resistance elements with mobile elements over time, we classified each resistance element as either 'Immobile' or some combination of 'Plasmid associated', 'Integron associated', 'TN/IS associated' and 'Prophage associated' (see 'Methods').

For those resistance elements that were mobile, we investigated the relative fractions of each of the different mobility types over time. We observe that TN-/IS-associated resistance elements are the most common, followed by plasmid associated, integron associated and prophage associated. Importantly, all four seem to show an increase over time, suggesting enhanced potential for horizontal spread of

genomic resistance elements among isolates deposited throughout the NCTC (Figure 2.4A).

Figure 2.3 ***(following page)*****: Clinical introduction of antibiotics is significantly associated with an increase in the fraction of isolates containing resistance to that antibiotic across drugs and mechanisms of action.**
(A) Left: Fraction of isolates with ≥ 1 beta-lactamase over time. Error bars represent the beta distribution 95% confidence estimate given the number of isolates in that year. **Right**: the significance of the observed difference in the mean fraction of isolates with ≥ 1 beta-lactamase after 10,000 simulations of randomly shuffling the year of isolation for the data.
(B) Same as in **(A)** but restricted to ESKAPE pathogens.
(C) The same analysis was performed across all isolates and over all drug classes with 10,000 shuffles. Each category reflects gene-defined resistance rather than strict drug classes. Multiple hypotheses corrected using Benjamini–Hochberg correction.

Figure 2.3: *(Continued)*



To investigate whether the cause for this increase in resistance mobility was driven by an increase in mobility overall, we investigated the prevalence of different mobile elements over time, irrespective of whether they carried resistance elements. We found that the fraction of isolates containing one or more plasmids is relatively constant throughout the NCTC (Figure 2.4B). However, when we then look at the fraction of those plasmids that contain one or more resistance genes, we find a steady increase over time, indicating that throughout the NCTC, it becomes more likely to identify isolates with resistance elements carried on plasmids over time (Figure 2.4B). We performed the same analysis for integrons and prophages. Integrons tended to be rarer in our dataset but did exhibit the same association with resistance elements. We note that since we have fewer integrons in our dataset, these results should be interpreted with caution (Figure 2.4C). Prophages tended to be abundant across our dataset and only began to be associated with resistance elements for more recent isolates (Figure 2.4D). Similar to the resistance prevalence trends, the decline in mobility-associated resistance elements after 2010 reflects changes in isolate composition. The post-2010 submissions included a wider range of taxa with fewer resistant ESKAPE pathogens, whose resistance tends to be more broadly associated with mobility.

The previous analysis was done across all genomic resistance elements as identified by CARD. To determine if these same results held for a more focused subset of resistance elements, we again turned to beta-lactamases. Analyzing the mobility of beta-lactamases over time, we find both an overall increase in at least one class of mobility and an increase in plasmid-associated beta-lactamases as time goes on (Figure B.8).

Finally, we observe a progressive rise in the proportion of resistance genes associated with more than one possible mechanism of mobility (transposons, plasmids, integrons or prophages) (Figure 2.4E, Figure B.9). One copy of the *dfrA14* gene was found to be associated with three of the four mobility mechanisms tested; this variant of dihydrofolate reductase (*dfrA*) confers resistance to trimethoprim and is commonly linked with class 1 integrons[70]. This *dfrA14* was found on NCTC 13440, a *Klebsiella pneumoniae* strain isolated from the Bolzano Regional Hospital in Italy[5]. We find that it is si-

multaneously a cassette in a class 1 integron, within the insertion sequence IS6100, located on a 61 kb contig predicted to belong to the conjugative AA552 (*IncN*) plasmid cluster, and although not located within a prophage, it does reside 200 bp from a predicted prophage. We note an observed increase in the cumulative mobility mechanisms of genomic resistance elements as time goes on. We interpret this increase in mobility as supporting the 'nesting doll' pattern of antibiotic resistance, wherein resistance genes become more frequently integrated into multiple mobile genetic elements over time[154].

2.6 Discussion

In Alexander Fleming's Nobel Prize acceptance speech, he described a model where sub-lethal doses would gradually select for resistance: 'Mr. X. has a sore throat. He buys some penicillin and gives himself, not enough to kill the streptococci but enough to educate them to resist penicillin'; in contrast, our study further reinforces that even in the pre-antibiotic era fully resistant pathogens were already likely circulating[62]. Rather than de novo resistance emerging in response to antibiotic exposure, we found that antibiotic resistance alleles were already present at low levels in clinically relevant isolates prior to the widespread use of antibiotics. In the metaphor, most likely some subset of streptococci either infecting Mr. X. or in the environment around Mr. X. had already been 'educated' to resist the antibiotic. Mr. X.'s use of penicillin merely selects for that subset to proliferate further. This 'pre-existing resistance' is consistent with the idea that clinically relevant populations naturally contain genomic resistance elements even before clinical usage of antibiotics[84,37,48]. Though these resistance elements were indeed rare, we note their existence and subsequent rise in frequency after human usage of a given antibiotic.

Furthermore, over time, genomic resistance elements are more likely to be associated with mobile genetic elements. Mobile genetic elements play a critical role in the horizontal transfer of antibiotic resistance genes between bacterial strains[1,164]. The increasing mobility of genomic resistance elements

over time suggests the accumulation and dissemination of resistance determinants within bacterial populations through horizontal gene transfer mechanisms. These findings align with the results from prior work investigating resistance carried by pre-antibiotic era conjugative plasmids and another more recent study sequencing pre-antibiotic era plasmids and finding an absence of resistance elements[78,29].

This phenomenon has significant implications for the rapid spread of antibiotic resistance, as mobile elements can facilitate the transfer of resistance genes between different bacterial species and genera. In addition, we find evidence for the 'nesting doll' theory of antibiotic resistance, where resistance elements become increasingly nested within multiple mobility mechanisms as time has gone on[154,86]. This accrual of different mobility mechanisms could be due to homologous recombination between mobile elements and the selection of those mobile elements that carry resistance elements as cargo.

We note some drawbacks to the present study. First, by virtue of being solely computational, we cannot make claims about the phenotypic resistance of any of the isolates analyzed. Even though our analysis is relegated to genomic resistance, we argue that genomic resistance serves as the raw material for evolution to act upon and is necessarily correlated with phenotypic resistance. Second, all historical analyses of archived isolates are biased by what samples were chosen to be significant enough to add to that archive. We do not believe this bias impacts our first main finding that genomic resistance was present at low levels in clinically relevant isolates even before the introduction of those antibiotics, since pre-discovery, there was no reason to preferentially deposit resistant isolates. However, we do acknowledge this bias likely skews the further results of changes in resistance prevalence and mobility over time. Though the specific magnitudes and proportions may be influenced by sampling bias, we believe the overall trends observed likely still hold true despite this bias. Third, we did not incorporate antibiotic usage data into our resistance analysis due to the persistent global challenge of sparse, inconsistent and often proprietary consumption data. A recent global estimate highlights the scarcity of rigorous usage data, especially pre-1980; available data are often modeled or survey-based, exclude informal markets and are preferentially sampled in wealthy countries[23]. Additionally, the

NCTC dataset lacks precise geographic metadata for many isolates, limiting our ability to align resistance patterns with antibiotic usage. If more comprehensive usage statistics and improved isolate metadata become available, linking antibiotic usage to resistance emergence would be a compelling direction for future work.

More broadly, what we call a resistance gene is itself teleological: while to us their primary role is to resist the antibiotics we use to treat patients, they may have evolved for an entirely different purpose[43]. Efflux pumps, for example, confer resistance to a range of natural stressors, and the presence of these genes in bacterial populations prior to the widespread use of antibiotics suggests they were already fulfilling important physiological roles.

The presence of resistance elements in a subsample of the older isolates in the NCTC indicates that they were likely a regular, if infrequent, occurrence in infections before the use of antibiotics. Consistent with our common understanding, the introduction of antibiotics into clinical environments accelerated the prevalence and mobility of these once-rare elements, contributing to the widespread prevalence of resistance observed today.

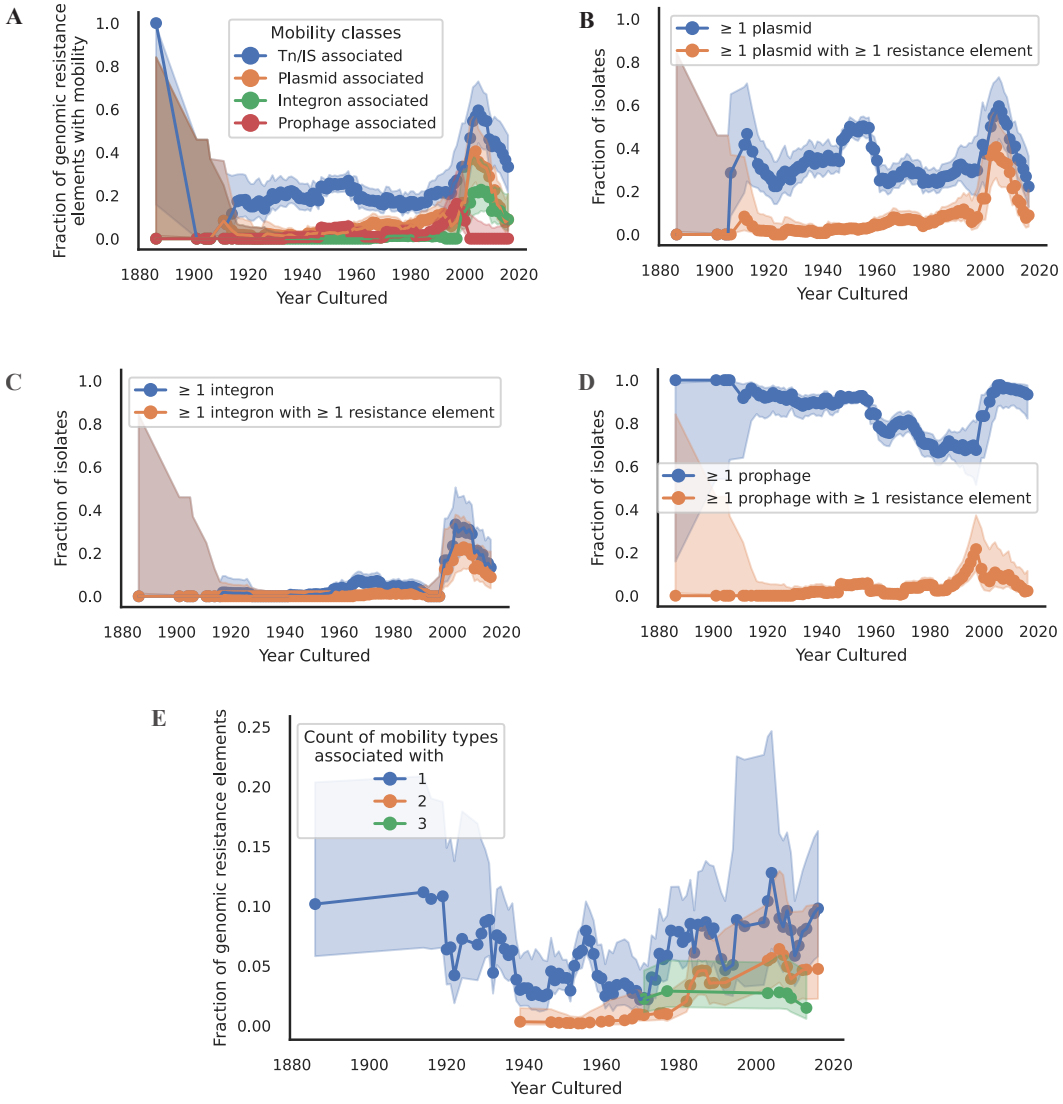
Our findings add to the growing literature showing that the genetic foundations of antibiotic resistance did not arise solely in response to human antibiotic use[8,45,137]. Rather, they were drawn from a long-standing repertoire of resistance alleles already present in the pangenome of clinically relevant isolates. Nevertheless, the introduction of antibiotics into clinical environments does appear to be followed by an acceleration of the prevalence and mobility of these once-rare elements in the NCTC. Recognizing this broader context reinforces that the emergence of clinically relevant antibiotic resistance is not simply a recent phenomenon triggered by modern drug use; rather, it is rooted in ancient genomic adaptations that have been amplified by the selective pressures introduced through widespread antibiotic deployment.

2.7 Acknowledgments

We thank the UK Health Security Agency and the National Collection of Type Cultures for providing access to their data and expertise, which guided our research. We thank Fernando Rossine and the entirety of the Baym Lab, whose insights and thoughtful commentary were instrumental in shaping our findings. Portions of this research were conducted on the O2 High Performance Compute Cluster, supported by the Research Computing Group at Harvard Medical School. Elements of Figure 2.1A adapted from CC-BY artwork by the Database Center for Life Science, Firefox, Matanbz, Zahra Ibrahem and Laboratoires Servier.

Figure 2.4 *(following page)*: Over time, resistance alleles in the NCTC exhibit increasing mobility. Error bars represent the beta distribution 95% confidence estimate given the number of isolates in that year.
(A) The fraction of resistance elements associated with a given mobility type. Note that a given resistance allele might have more than one form of mobility.
(B) The fraction of isolates over time that contain one or more plasmids, and the fraction that contain one or more plasmids predicted to harbor one or more resistance elements.
(C) Same as **(B)** but considering integrons.
(D) Same as **(B)** but considering prophages.
(E) The fraction of resistance elements with one or more independent mobility types associated with it.

Figure 2.4: *(Continued)*



Πάντα ἀλλήλοις ἐπιπλέκεται

καὶ ἡ σύνδεσις ἱερά

All things are interwoven,

and the Web is holy.

Marcus Aurelius, *Meditations*

3

Phylogeny-colored de Bruijn graphs enable the spatial analysis of pangenome evolution

Bacterial pangenomes are shaped by the interplay of vertical inheritance, recombination, and horizontal gene transfer. Understanding how these processes distribute sequence across a population requires data structures that represent both the *content* of shared and vari-

able genomic regions and the *evolutionary history* that generated them. De Bruijn graph-based approaches have become central to pangenome representation, allowing reference-free, scalable encodings of sequence variation across large collections of genomes[112,178,36,153]. However, a fundamental gap remains between the information encoded in these graphs and the evolutionary history yielding their topology.

A compacted de Bruijn graph decomposes a set of input genomes into unitigs (maximal non-branching paths in the graph) and records their adjacency[33]. When augmented with sample membership information, the resulting colored de Bruijn graph encodes which input genomes contain which unitig[82,123]. This representation has proven powerful for tasks such as classifying core and accessory genome components[146], detecting genomic variants[126], and performing genome-wide epistasis and coselection analyses[98]. Tools including Cuttlefish[93], Bifrost[75], GGCAT[41], and others have made construction of colored de Bruijn graphs tractable for datasets comprising thousands of bacterial genomes.

Yet the colored de Bruijn graph only captures the pattern of sequence sharing, *not* its evolutionary origin. A unitig present in a subset of genomes may have been vertically inherited from a common ancestor, acquired by a single horizontal transfer event and then inherited clonally, or gained independently in multiple lineages. These scenarios produce identical entries in the color vector but imply distinct biological events. Distinguishing between them requires phylogenetic inference; specifically, ancestral state reconstruction on a tree relating the input genomes. Given a phylogeny and a presence/absence vector, parsimony or likelihood-based methods can identify the branches on which state changes most plausibly occurred, thus localizing each unitig's distribution to specific points in the evolutionary history of the population.

Ancestral state reconstruction and pangenome graph analysis have developed largely independently. Phylogenetic methods for gene gain and loss typically operate on gene presence/absence tables derived from annotation pipelines, abstracting away the fine-grained sequence structure that de Bruijn graphs

preserve. Conversely, graph-based pangenome analyses exploit adjacency and sequence content but lack an evolutionary dimension. They can identify that two adjacent unitigs differ in their sample membership, but cannot determine whether those different patterns arose from the same evolutionary event, from closely related events on neighboring branches, or from entirely distinct events on opposite sides of the phylogeny. This distinction matters because it can separate local refinement of a single genomic variation from genuine boundaries between independently inherited regions.

In this work, we introduce the phylogeny-colored de Bruijn graph (pcDBG), a data structure that bridges this gap by recoloring each unitig in a compacted colored de Bruijn graph with its most parsimonious phylogenetic assignment (Figure 3.1). This fuses graph adjacency with evolutionary history into a single structure, thereby enabling analyses that are inaccessible to either component alone. We describe two complementary analyses that exploit the joint availability of graph adjacency and phylogenetic labels and apply them to a diverse collection of *Staphylococcus aureus* ST8 genomes.

3.1 Construction of phylogeny-colored de Bruijn graphs

In a classical colored de Bruijn graph, each unitig carries a color indicating which input genomes contain it (Figure 3.1A). The pcDBG replaces these membership colors with phylogenetic colors: each unitig is instead colored by the branch or branches of the phylogeny where its presence/absence pattern most parsimoniously changes state (Figure 3.1B). Two adjacent unitigs sharing a phylogenetic color were gained or lost at the same point in evolutionary history; two adjacent unitigs with different colors represent a boundary between differently inherited genomic regions.

The phylogenetic assignment of each unitig is encoded as a 'cut-string', a compact representation of the branch or branches where the state changed. A single-node cut-string implies one evolutionary event, at a particular node in the branch and affects all the descendants of that node. A multi-node cut-string implies multiple independent changes, and in the scenario where multiple scenarios are equally

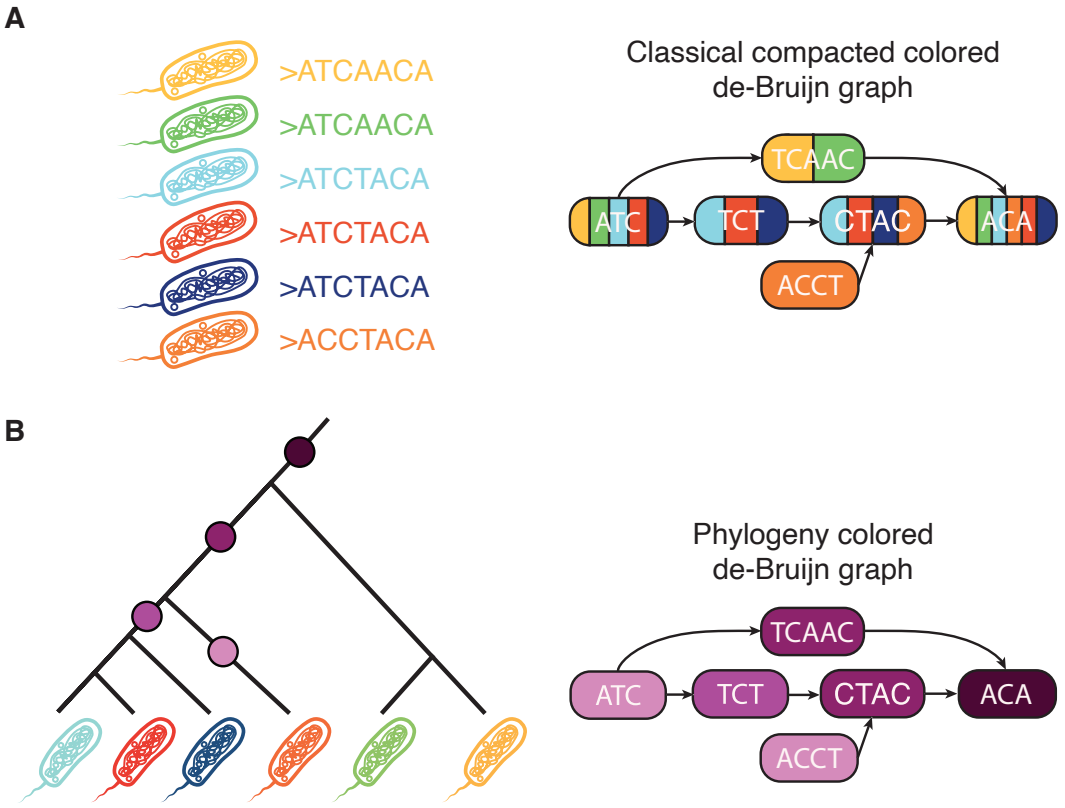


Figure 3.1: Construction of a phylogeny-colored de Bruijn graph.
(A) A classical compacted colored de Bruijn graph built from a set of input sequences. Each unitig is colored by which input genomes contain it.
(B) The corresponding phylogeny-colored de Bruijn graph. Each unitig's color is replaced by the branch/branches of the tree where its presence/absence pattern most parsimoniously changes state. Adjacent unitigs sharing a color were gained or lost at the same point in evolutionary history; adjacent unitigs with different colors represent a boundary between differently inherited regions.

parsimonious we store all possible solutions. A key question for downstream analysis is whether two adjacent unitigs with distinct cut-strings represent the same evolutionary event or genuinely different ones. To answer this, we construct an ancestor-descendant concordance criterion: two cut-strings are 'concordant' if the nodes in one are ancestral to the nodes in the other on the phylogeny. This captures the case where a unitig assigned to a deep clade-defining branch sits adjacent to a unitig assigned to a more recent branch within that same clade. The latter scenario is merely a refinement of the first, deeper evolutionary event. Edges where neither flanking assignment is ancestral to the other are classified as 'discordant'.

3.2 Evolutionary persistence and boundary severity in *S. aureus* ST8

Using this criteria, we have implemented two complementary analyses on the pcDBG: evolutionary persistence and boundary severity. Evolutionary persistence measures the spatial extent of concordant blocks. Beginning from a random seed unitig, how quickly does evolutionary history (concordance) decay as one traverses outwards (Figure 3.2A)? Boundary severity asks at those edges where the history does change (discordant), how different are the two flanking histories (Figure 3.2B)?

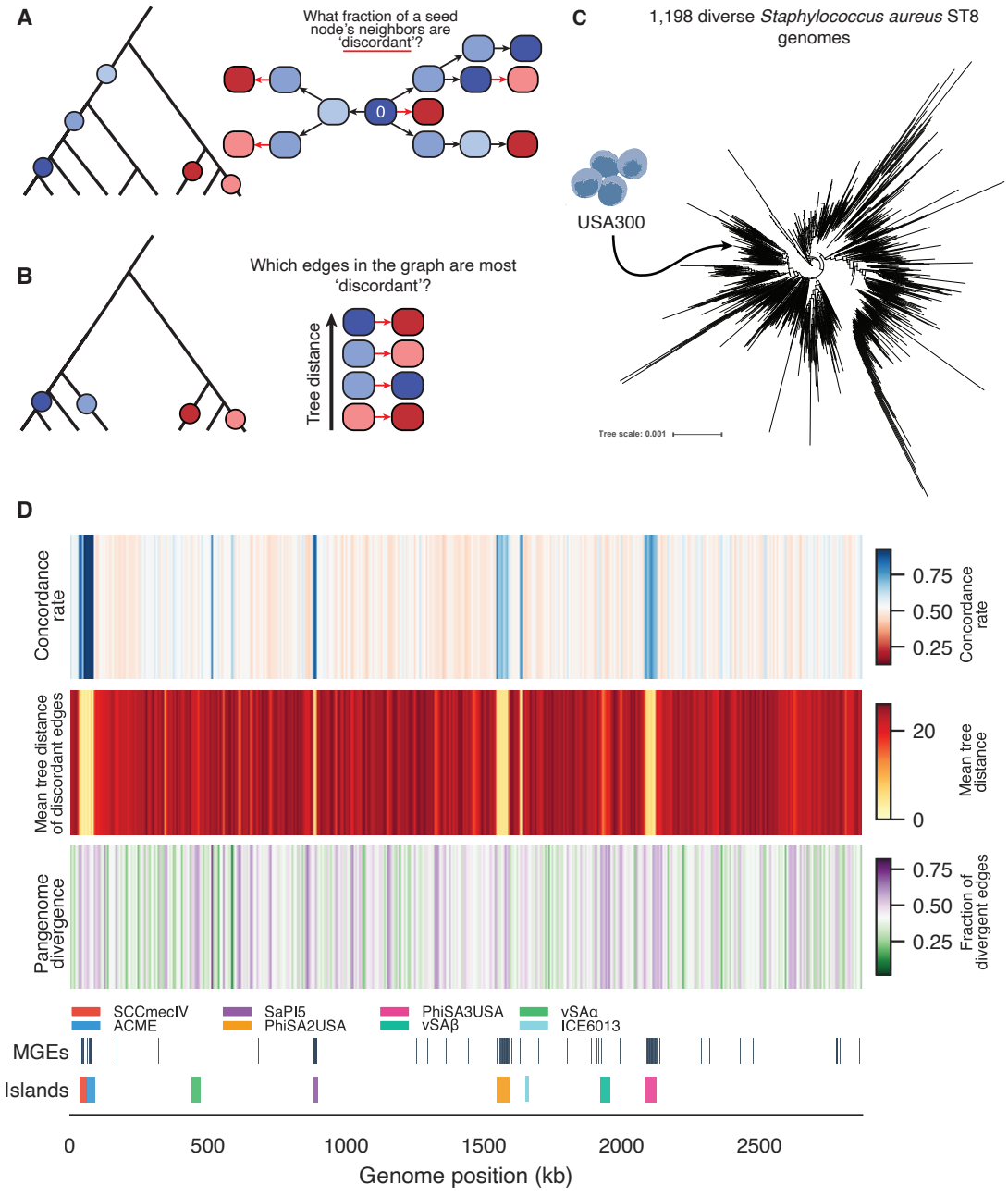
We applied both analyses to a pcDBG constructed from 1,198 diverse *S. aureus* ST8 genomes (Figure 3.2C, see METHODS). ST8 encompasses the globally prevalent methicillin-resistant USA300 lineage and is an ideal test case since it carries well-characterized mobile elements (SCCmec, the ACME element, and Staphylococcal Pathogenicity Islands) that are known to appear across ST8 genomes in diverse contexts [160]. The resulting pcDBG contained 1,316,848 unitigs and 1,727,459 edges.

The evolutionary persistence analysis on this set of genomes revealed that co-inherited blocks extend well beyond the immediate graph neighborhood (Figure C.1A). The fraction of evolutionarily concordant neighbors was strongly enriched relative to null expectation at short hop distances and decayed steadily, crossing the null only around the 25th hop. Given the mean unitig length in this graph (47 bp, Figure C.1B), this corresponds to co-inherited blocks on the order of a kilobase; fairly small relative to the 2.87 Mb *S. aureus* chromosome. The shape of the decay curve, a steep initial drop followed by a gradual approach to null, indicates a mixture of block sizes, as expected for a genome carrying MGEs of varying sizes.

The boundary severity analysis revealed that among the edges that are truly discordant, the observed mean tree distance (19.08) was lower than the null expectation (20.42), with a marked spike of mild transitions at low tree distances (Figure C.2). This spike corresponds to discordant edges where flanking unitigs were gained or lost on immediately adjacent branches of the phylogeny, indicating that many evolutionary boundaries in the ST8 pangenome separate blocks that differ only in fine-grained sub-lineage membership rather than in deep phylogenetic origin. More broadly, the leftward shift of the observed distribution relative to the null is consistent with a clonal collection of genomes whose pangenomic variation is driven by differential MGE acquisition from within the lineage: when co-inherited blocks end, the flanking histories tend to involve closely related branches rather than distant parts of the phylogeny.

Figure 3.2 ***(following page)*****: Analysis of the S. aureus ST8 pangenome using the pcDBG.**
(A) Schematic of evolutionary persistence: starting from a seed node, the fraction of neighbors that are evolutionarily concordant is measured at increasing graph distances.
(B) Schematic of boundary severity: for each discordant edge, the phylogenetic tree distance between flanking assignments is computed.
(C) Phylogeny of 1,198 diverse S. aureus ST8 genomes, with the closest relative to USA300 highlighted. **(D)** Evolutionary persistence and boundary severity mapped onto the USA300 FPR3757 reference genome. Top track: concordance rate per sliding window (blue = concordant, red = discordant, centered on the genome-wide mean of 0.53). Second track: mean tree distance of discordant edges per window (yellow = closely related boundaries, dark red = deeply divergent boundaries). Third track: pangenome divergence, the fraction of edge activity in each window arising from structural or unmapped edges (green = reference-conforming, purple = elevated divergence, centered on the genome-wide mean of 0.40). Bottom track: annotated mobile genetic elements (top row) and genomic islands (bottom row) from the USA300 reference, including the integrative conjugative element ICE6013.

Figure 3.2: *(Continued)*



3.3 Reference genome mapping localizes boundaries to known genomic islands

To determine whether evolutionary boundaries cluster at specific genomic positions, we mapped the results from both analyses onto the USA300 FPR3757 reference genome (Figure 3.2D). Of the unitigs in the graph, only 643,817 (48.8% of all the unitigs in the pcDBG) passed alignment filters and mapped to the reference. Of the 1,727,459 total edges, 591,937 had both flanking unitigs map to the reference, 467,375 were partially mapped, and 668,147 were unmapped. The large fraction of unitigs and edges without sufficient alignment to USA300 is a consequence of the high diversity present within *S. aureus*, even within a single sequence type. Concordance and tree distance statistics were computed in 2,868 sliding windows (5 kb windows, 1 kb steps) spanning the reference and compared against annotated mobile elements and genomic islands.

The concordance rate across the sections of the pcDBG aligning to USA300 was non-uniform, with a genome-wide mean of 0.527 (Figure 3.2D, top track). Regions of high concordance correspond to stretches of conserved core backbone shared across ST8 sub-lineages, while regions of low concordance mark evolutionary boundaries where the graph transitions between differently inherited content. These low-concordance regions co-localized with annotated MGEs, including the SCCmec type IV cassette near the chromosomal origin, the vSAα and vSAβ genomic islands, SaPI5, the ACME element, the PhiSA2USA and PhiSA3USA prophages, and the integrated conjugative element ICE6013.

The mean tree distance of discordant edges was generally high and uniform across the genome (Figure 3.2D, middle track), indicating that when evolutionary boundaries do occur, they tend to separate deeply divergent branches of the phylogeny rather than closely related sub-lineages. This is consistent with the boundary severity analysis (Figure C.2), which showed that the observed distribution of tree distances is shifted toward lower values relative to the null but retains a high overall mean. The spatial

uniformity of tree distance across the genome, punctuated by localized variation at MGE loci, suggests that the recombination events shaping this population draw from a broad range of phylogenetic donors rather than being dominated by transfers between closely related strains.

We additionally computed a pangenome divergence fraction for each window, defined as the proportion of all edge activity arising from structural rearrangements or edges where one or both unitigs lacked a reference alignment (Figure 3.2D, third track). The genome-wide mean divergence fraction was 0.40, indicating that a little under half of all edge activity across the genome involves content *not* represented in the USA300 reference. This fraction appeared elevated at known MGE loci and broadly uniform elsewhere, consistent with accessory content concentrating at established integration sites while the core backbone remains represented within the reference. Of the divergent edges, the vast majority (467,375 of 468,308) arose from partially unmapped edges rather than structural rearrangements (933), indicating that the USA300 reference is collinear with the mapped portion of the pangenome but that the ST8 population carries substantial accessory content absent from this single reference.

As the mapped concordance rate and evolutionary persistence measure the spatial extent of co-inherited blocks, one against the USA300 reference, and the other across the pangenomic graph, we wondered if those two signals agreed. Of the 5,000 seeds used in the persistence analysis, 2,411 (48.2%) mapped to USA300 and could be assigned a sliding window across USA300. We find that the two metrics were significantly positively correlated (Spearman $\rho = 0.16$, $p < 10^{-15}$). Windows with higher concordance rates indeed harbored seeds with longer co-inherited blocks in the graph. The correlation is modest, consistent with the two metrics integrating over largely non-overlapping subsets of the pangenome, but confirms that the spatial patterns in the reference mapping reflect genuine block structure rather than alignment artifacts.

Importantly, the spatial distribution of all three tracks recapitulated known features of *S. aureus* USA300 biology without any prior knowledge or annotations provided. Each of the annotated ge-

nomic islands and prophages in the reference corresponded to a visible perturbation in concordance, tree distance, or divergence fraction.

3.4 Discussion

We have introduced the phylogeny-colored de Bruijn graph, a data structure that annotates each unitig in a pangenome graph with the phylogenetic branch or branches where its presence/absence pattern most parsimoniously changed state. By fusing graph adjacency with evolutionary history, the pcDBG enables analyses that are inaccessible to either a colored de Bruijn graph or ancestral state reconstruction alone.

The two analyses we defined, evolutionary persistence and boundary severity, characterize complementary aspects of pangenome architecture. Persistence measures how far one can walk before the evolutionary history changes; severity measures what that change looks like when it occurs. When mapped to the USA300 reference genome, both analyses recapitulated the known mobile element landscape of *S. aureus* ST8, with concordance, tree distance, and divergence signals co-localizing with annotated mobile elements. The pangenome divergence fraction further quantified the extent to which the ST8 accessory genome exceeds any single reference, with 40% of edge activity involving content absent from USA300.

Several extensions of the pcDBG remain open for exploration. One immediate analysis is to apply the data structure to diverse bacterial clades from highly clonal to highly transformable. We expect markedly different graph structures when one compares a clonal species like *Mycobacterium tuberculosis* to a more recombinant species such as *Helicobacter pylori*. Systematic comparison across species would provide a quantitative basis for classifying recombination regimes across bacterial genomes. Though we have only built pcDBGs on whole genomes, one can imagine a graph built on the variation present in a particular gene across many bacterial species, thus allowing an analysis of the history

of diversification within a particular coding region. Additionally, one could use the depth of phylogenetic assignments to provide a further temporal dimension to the analysis. By sweeping a depth threshold from root to tips while measuring graph properties, one could decompose the pangenome into temporal strata, revealing when in the evolutionary history different layers of sequence variation were generated.

The pcDBG adds a phylogenetic layer to the pangenome graph without discarding any information already present in the colored de Bruijn graph. The color vectors, graph topology, and sequence content remain available; the phylogenetic assignment is simply an additional annotation. Any existing analysis that operates on colored de Bruijn graphs can therefore be augmented with evolutionary context. We anticipate that this integration of graph topology and phylogenetic inference will prove useful across the range of organisms and questions for which pangenome graphs are routinely constructed.

3.5 Methods

3.5.1 pcDBG construction

A compacted colored de Bruijn graph was constructed using Cuttlefish 1 with $k = 31$, producing a GFA1 file with unitig sequences and sample membership tags. The color matrix was extracted and deduplicated to identify unique presence/absence vectors. Fitch parsimony was applied to each unique vector against the phylogenetic tree: the bottom-up pass computed the intersection (or union, if empty) of children's state sets at each internal node; the top-down pass resolved ambiguities by inheriting the parent's state where possible[61]. Cuts were identified at every branch where the assigned state changed between parent and child. Annotations were stored in a SQLite database alongside unitig sequences and graph adjacency.

3.5.2 Ancestor-descendant concordance

Two adjacent unitigs were classified as concordant if their phylogenetic assignments are compatible under an ancestor-descendant criterion, and discordant otherwise. The intuition is that a unitig assigned to a deep clade-defining branch and an adjacent unitig assigned to a more recent branch within that same clade share a common evolutionary origin; the second assignment is merely a refinement of the first. Only edges where the flanking assignments cannot be related by ancestry indicate boundaries between independently inherited regions.

Formally, ancestor-descendant relationships are evaluated efficiently per node pair using an Euler-tour traversal of the phylogeny. Entry and exit times are recorded for each node during a depth-first traversal; node a is an ancestor of node b if and only if $\text{entry}(a) \leq \text{entry}(b)$ and $\text{exit}(b) \leq \text{exit}(a)$. For two single-node cut-strings, concordance requires that one node is an ancestor or descendant of the other. For multi-node cut-strings, concordance requires that every node in the smaller set be paired with a distinct node in the larger set such that *each pair* satisfies the ancestor-descendant relationship. When one or both cut-strings have multiple alternative solutions, concordance is satisfied if any pair of alternatives (one from each cut-string) allows such a matching. Edges where no combination of alternatives are concordant yield a 'discordant' matching.

3.5.3 Evolutionary persistence & boundary severity analysis

Both analyses operate on the concordance classifications defined above. Evolutionary persistence measures the spatial decay of concordance along graph walks; boundary severity characterizes the phylogenetic distance across discordant edges.

For evolutionary persistence, breadth-first search was performed from 5,000 randomly sampled seed unitigs to a maximum depth of 100 hops. At each hop distance, the fraction of newly reached unitigs concordant with the seed was recorded. A null expectation was estimated by sampling 1,000,000

random cut-string pairs weighted by unitig frequency and computing the fraction satisfying the concordance criterion. The persistence length of the aggregate decay curve was defined as the hop distance at which concordance first dropped below this null. Per-seed persistence lengths were computed analogously for each individual seed's decay curve and used in downstream correlation analyses.

For boundary severity, every edge in the pcDBG was classified as concordant or discordant. For discordant edges, the phylogenetic distance between flanking cut-strings was computed from a precomputed pairwise distance matrix over all named tree nodes. For single-node cut-strings, the distance is a direct lookup. For multi-node cut-strings, the distance between two solutions is the average of the two directed minimum distances. When one or both cut-strings have alternative solutions, the minimum distance across all pairs of alternatives is taken. A null distribution was estimated by drawing 5,000,000 random cut-string pairs weighted by frequency.

3.5.4 *Staphylococcus aureus* ST8 genome collection

All annotated 17,385 *Staphylococcus aureus* ST8 genomes were downloaded from the AllTheBacteria (ATB) collection[79]. Mash was used with a sketch size of 10,000 and a k-mer length of 21 to compute a pairwise distance matrix across all samples[133]. Single-linkage hierarchical clustering was performed on this matrix, and a distance threshold was selected via binary search to yield approximately 1,200 clusters. For each cluster, the medoid genome (the genome with the lowest mean distance to all other cluster members) was chosen as the representative, producing a final set of 1,198 diverse ST8 genomes.

To generate a phylogeny for use, we used attotree (an optimized version of Mashtree) with default options on the contigs downloaded from the ATB[19,92].

3.5.5 Reference genome mapping

Unitig sequences were aligned to the USA300 FPR3757 reference genome (GenBank accession CP000255.1 [50]) using BLAST (e-value $< 10^{-10}$, single best hit per unitig) [26]. Hits with percent identity below 90% or query coverage below 50% were discarded; 643,817 of 673,957 total hits (95.5%) passed these filters. Edges were classified as local (both unitigs mapped within 10 kb on the reference), structural (both mapped but >10 kb apart), or unmapped (one or both unitigs without a passing hit). Local edges were assigned a reference position at the midpoint of their flanking unitig positions. Sliding window statistics were computed using 5 kb windows with 1 kb steps, yielding 2,868 windows across the reference. For each window, the concordance rate was computed as the fraction of local edges that were concordant, and the pangenome divergence fraction was computed as the number of structural and unmapped edges divided by the total edge activity (local plus structural plus unmapped) in that window. Concordance patterns were compared against annotated mobile elements and genomic islands parsed from the USA300 GenBank annotation.

3.5.6 Software and data availability

The pcDBG construction pipeline is available at https://github.com/aryakaul/phylogeny-colored-dbg.

4

Conclusion

I began this thesis with a paradox. Bacterial genomes are shaped by a pervasive streamlining pressure that favors the loss of non-essential DNA; however, bacteria exhibit a cacophony of diverse genes unmatched by any other domain of terrestrial life. How then does such diversity flourish in genomes under constant pressure to contract? We framed the bacterial genome as less a static blueprint and more a fluid mosaic, continually reshaped by gene gain and loss. Each of the three chapters that

followed investigated a different facet of that dynamism. Together, they suggest that the forces shaping bacterial genomes are richer, and more complex, than their conventional descriptions imply.

In Chapter 1, we asked whether the deletional bias that contracts bacterial genomes might simultaneously serve as an engine of innovation. We formalized the model of deletion-born fusion genes, documented them arising in both the Lenski Long-Term Evolution Experiment and in the divergence of *Mycobacterium tuberculosis* and *Mycobacterium bovis*, and developed a computational screen that identified these events across 2.4 million publicly available bacterial genomes. The deletional bias is real and relentless, but it is not purely subtractive. The same molecular event that contracts a genome can generate a novel open reading frame, one capable of reaching high frequency by hitchhiking on the selective advantage of the deletion itself. Streamlining, it turns out, is *also* a source of the very diversity it appears to oppose. The force that trims can also create.

In Chapter 2, we traced how human antibiotic use reshaped a pre-existing genetic landscape. Using 1,817 genomes from the National Collection of Type Cultures, matched to their years of isolation spanning 1885 to the present, we showed that genomic resistance determinants were present in pre-antibiotic isolates, albeit at lower frequencies and largely immobile. Anthropogenic antibiotic use changed not the existence of these genes but their prevalence and mobility. Over time, resistance elements became increasingly nested within multiple layers of mobile genetic elements, an architecture that facilitates rapid dissemination across species boundaries. The antibiotic era did not conjure resistance from nothing; it simply amplified and mobilized pre-existing genetic potential. Our findings align with the growing consensus that human antibiotic selective pressure did not add new tiles to the fluid mosaic, it rearranged the ones already there and gave them the means to mobilize.

In Chapter 3, we developed a computational framework for integrating evolutionary history into pangenome graphs. Colored de Bruijn graphs encode which genomes share which sequences, but this representation captures only the pattern of sharing, not its evolutionary origin. We introduced phylogeny-colored de Bruijn graphs, in which each sequence is annotated with the phylogenetic branch

or branches where its distribution most parsimoniously changed state. This recoloring fused graph adjacency with evolutionary history, revealing spatial structure invisible to either representation alone. In a diverse collection of *Staphylococcus aureus* ST8 genomes, co-inherited blocks extended over kilobases before their evolutionary history decayed, and the boundaries between differently inherited regions co-localized with known mobile genetic elements, all without prior annotation. Pangenome graph topology was never only structural, it always encoded evolutionary history; all that was required was the right lens to read it.

A thread connects these three investigations beyond their shared subject matter. In each case, a familiar force (deletion, selection, sequence sharing) turned out to behave differently than its name alone would predict. Deletion creates as it destroys. Selection amplifies and mobilizes rather than invent. Graph structure records history rather than depicting content. Importantly, these are not refutations of the conventional understanding behind these concepts, simply additional refinement. Acknowledging this nuance is important, since it opens our minds to additional scientific questions we would never think to ask without it. If deletion only destroyed, we would never look for the genes it births. If resistance only emerged in response to antibiotics, we would not search for it in century-old isolates. If pangenome graphs only recorded sequence content, we would not query them for evolutionary signal.

But asking these new questions required new kinds of evidence, and in each case the evidence was unlocked by scale. Deletion-born fusion genes are rare events; identifying them required screening millions of genomes. The temporal arc of resistance became visible only through a historical collection spanning more than a century. The spatial organization of evolutionary boundaries across a pangenome demanded a graph built from large collections of diverse genomes. Yet scale also exposes the limits of what we currently know. Our protein family analysis in Chapter 1 demonstrated that most bacterial proteins in high-quality genome collections remain singletons, sampled only once. Current databases are dominated by culturable, human-associated pathogens; environmental lineages

remain sparsely represented. Any broad computational analysis of existing databases will inherit these gaps. An important takeaway from this work is that large-scale computational analyses are *only as powerful* as the underlying biological data they rely upon.

Each chapter also opens directions I did not have the time nor tools to fully explore. The prefix-suffix k-mer approach developed in Chapter 1 was built to detect deletion-born fusions, but it generalizes to any process that alters the genomic distance between conserved flanking sequences: novel introns, integron cassette expansions, variable plasmid cargo, or structural variation in complex metagenomic communities. We have not yet characterized which protein domains or functional categories are enriched among structurally variable loci; such meta-analyses are a natural extension. More pressingly, we identified no clear novel function for any deletion-born fusion. The candidates we found appear to be snapshots of the earliest stages of sequence exploration, proteins that persist long enough to sample sequence space but have not yet undergone functional refinement. Whether some fraction of these eventually acquire beneficial roles, and how one might detect such functionalization computationally, remains an open question.

The historical trajectory described in Chapter 2 ends at the present, but its logic extends forward. If the nesting of resistance elements within mobile genetic elements continues to deepen, predicting the dissemination potential of any given resistance gene will require understanding not just its own mobility but the mobility of the elements that carry it. Surveillance efforts that track only resistance gene presence without characterizing their genomic architecture may miss the difference between a chromosomally embedded determinant unlikely to spread and the same gene perched atop a stack of mobile elements poised for lateral transfer. Integrating mobility context into genomic surveillance pipelines would provide a more complete picture of outbreak risk than presence/absence calls alone.

The phylogeny-colored de Bruijn graph presented in Chapter 3 was run on a single clonal lineage. An immediate extension is to apply it across species with markedly different recombination regimes, from the highly clonal *Mycobacterium tuberculosis* to the highly transformable *Helicobacter pylori*.

We expect fundamentally different graph topologies, and systematic comparison across species might provide a quantitative basis for classifying distinct genomic strategies. One could also use the depth of phylogenetic assignments to add a temporal dimension: by sweeping a depth threshold from root to tips while measuring graph properties, the pangenome could be decomposed into temporal strata, revealing when in evolutionary history different layers of sequence variation were generated.

Bacterial genomes are not static blueprints, but fluid mosaics, continually reshaped by the interplay of forces whose effects we are only beginning to appreciate in their full complexity. My thesis has offered three humble contributions toward that appreciation: a mechanism of gene birth hiding inside a destructive process, a historical record of selection reshaping a pre-existing genetic landscape, and a computational framework for merging evolutionary history with the topology of a genome graph. As bacterial sequence data continues to expand in size and scope, the questions posed here become increasingly tractable. Together, these chapters represent my attempt to understand a domain of life defined by ubiquitous and rapid transformation, and they suggest that some of the richest novel biology can be found in those places our intuition is most surprised.

A

Supplementary Material of Chapter 1

A.1 Estimation of *yjcO-lysU*/deletion selection coefficient

For a haploid population in which a novel beneficial allele has fitness $1 + s$, and the wildtype has a fitness of 1, the allele-frequency dynamics are given by:

$$p_{t+1} = \frac{p_t (1 + s)}{1 + s p_t}$$

Where p_t is the frequency of the novel allele at generation t. This can be approximated to the differential equation, a classical result from Kimura[95]:

$$\frac{dp}{dt} = sp (1 - p)$$

Integrating over this equation and solving for t gives:

$$t \approx \frac{1}{s} ln \left(\frac{p_1 (1 - p_0)}{p_0 (1 - p_1)} \right)$$

Conditioning on the allele fixing, we can set $p_0 = \frac{1}{N_e}$ and $p_1 = 1 - \frac{1}{N_e}$. For large effective populations, this yields the commonly used approximation:

$$t \approx \frac{2}{s} ln (N_e)$$

Solving for s gives:

$$s \approx \frac{2ln(N_e)}{t}$$

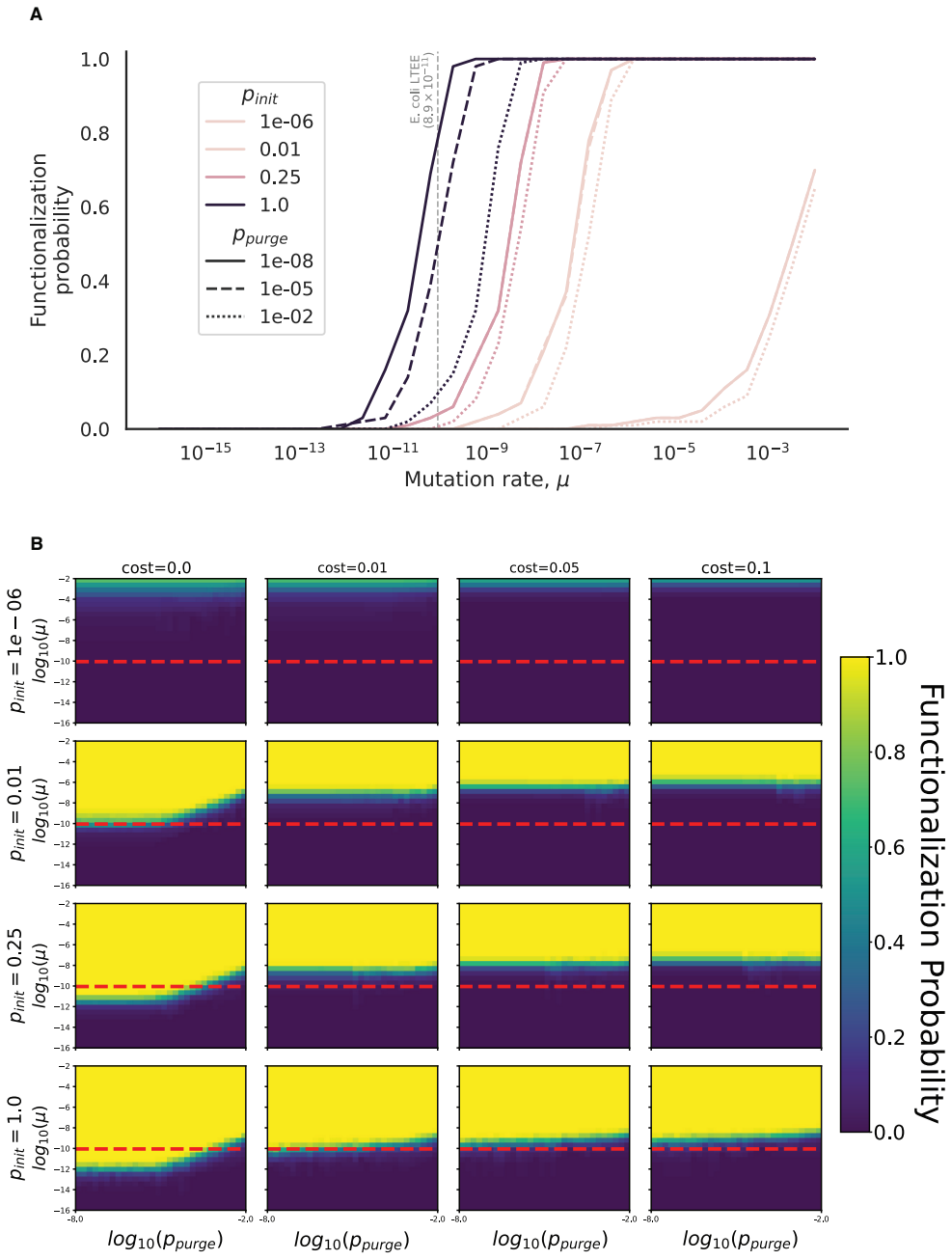
From Good et al.[69], we estimate $N_e = 10^7$, from our metagenomic analysis, we estimate $t = 500$:

$$s \approx \frac{2ln\left(10^7\right)}{500} \approx 0.065$$

Yielding the estimated selection coefficient of 6.5% for the *yjcO-lysU* fusion/deletion.

Figure A.1 ***(following page)*****: Forward simulations quantify the hitchhiking advantage.**
(A) Probability that at least one lineage functionalizes as a function of the mutation rate μ; colors denote four starting frequencies of the nascent fusion (p_{init}) and line styles three purge probabilities (p_{purge}). The dashed vertical line marks the LTEE point-mutation rate. All curves are shown for a fusion gene with a pre-functionalized fitness cost of 0.01.
(B) Heat-maps show the same probability across grids of p_{purge} (x-axis, log_{10} scale) and fitness cost of the unfixed fusion (c, four columns) for the four p_{init} values (rows); the color of the cell indicates the proportion of times the fusion functionalized before being purged. The dashed red line denotes the LTEE point-mutation rate. Each cell summarizes 100 Wright–Fisher runs of 10^6 haploids.

Figure A.1: *(Continued)*



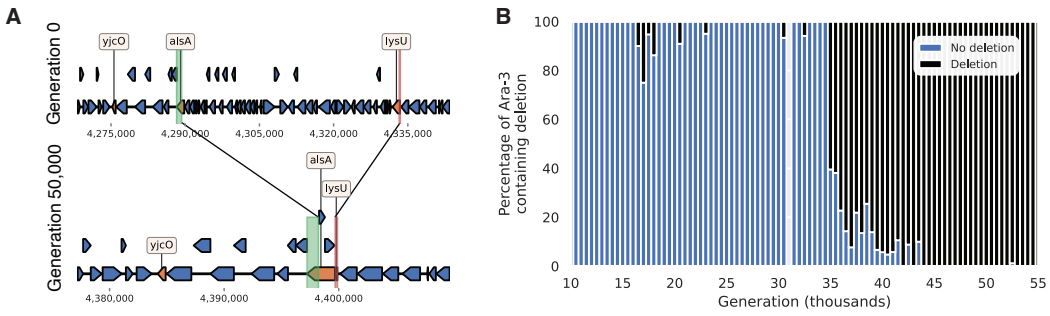


Figure A.2: A convergent 43.4 kb deletion sweeps at the same locus in Ara-3.
(A) Schematic of the deletion in Ara-3, orange genes highlight the resulting prior genes involved in the deletion. Green and red bars correspond to BLAST alignments to this region.
(B) Metagenomic sequencing results showing the fraction of reads supporting the deletion or not.

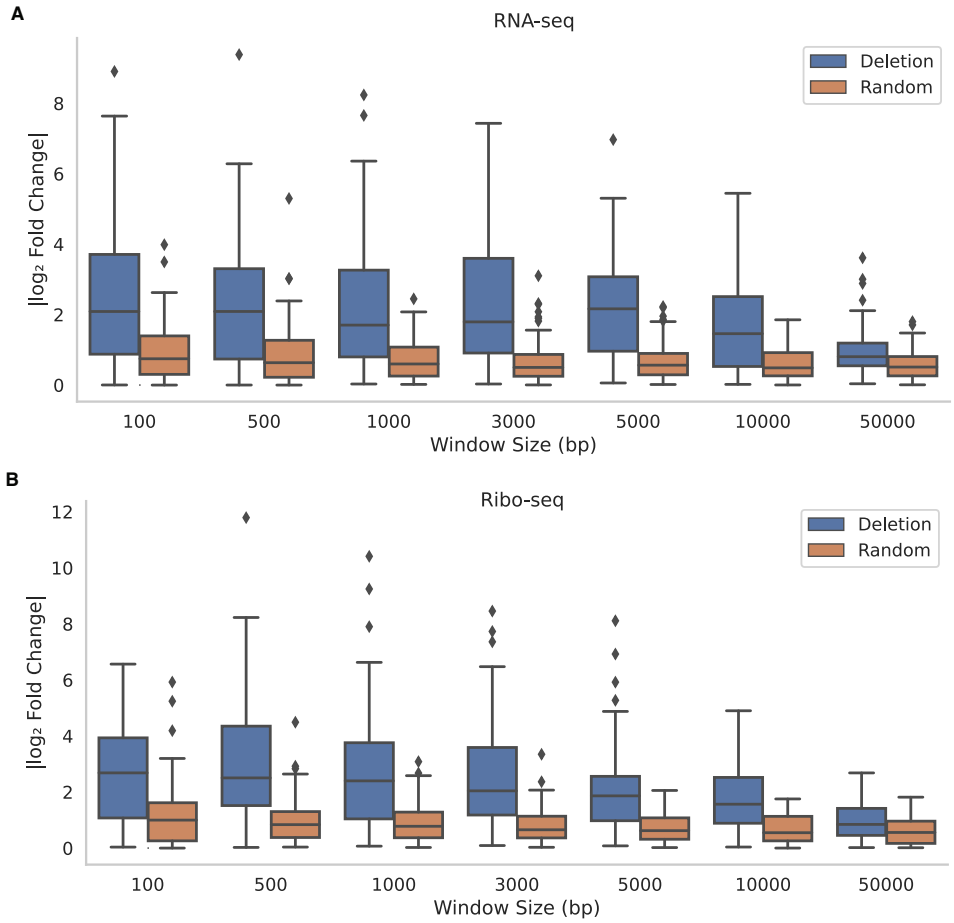


Figure A.3: Large LTEE deletions are associated with significant changes to local transcription and translation.
(A) RNA-seq log_2 fold changes for windows of varying size flanking $\geq$ 1 kb deletions (blue) and for randomly sampled windows (orange). Fold changes are calculated between the ancestral strain and the evolved population at generation 50,000. Data downloaded from Favate et al. 2022 [58].
(B) As in (**A**) but analyzing Ribo-seq data.

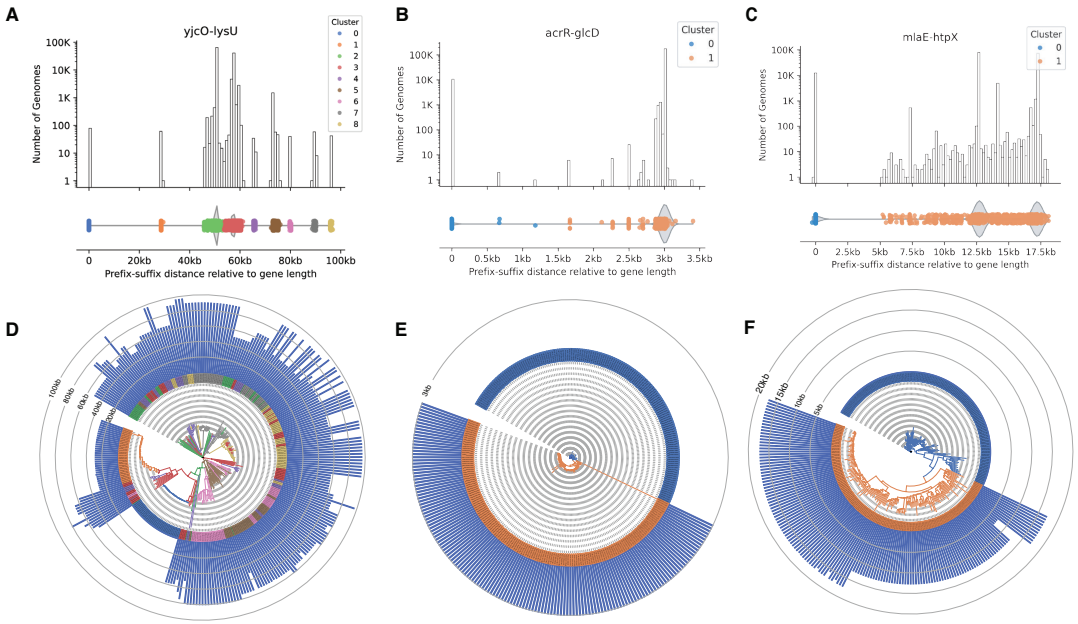


Figure A.4: The prefix-suffix approach captures previously identified deletion-born fusions and reveals additional structural variation.

(A) Top: Log-scaled histogram of prefix-suffix distances relative to the length of the original *yjcO-lysU* fusion. **Bottom**: underlying point distribution for histogram. Every point represents a single genome in the ATB dataset; colors correspond to DBSCAN-derived clusters.

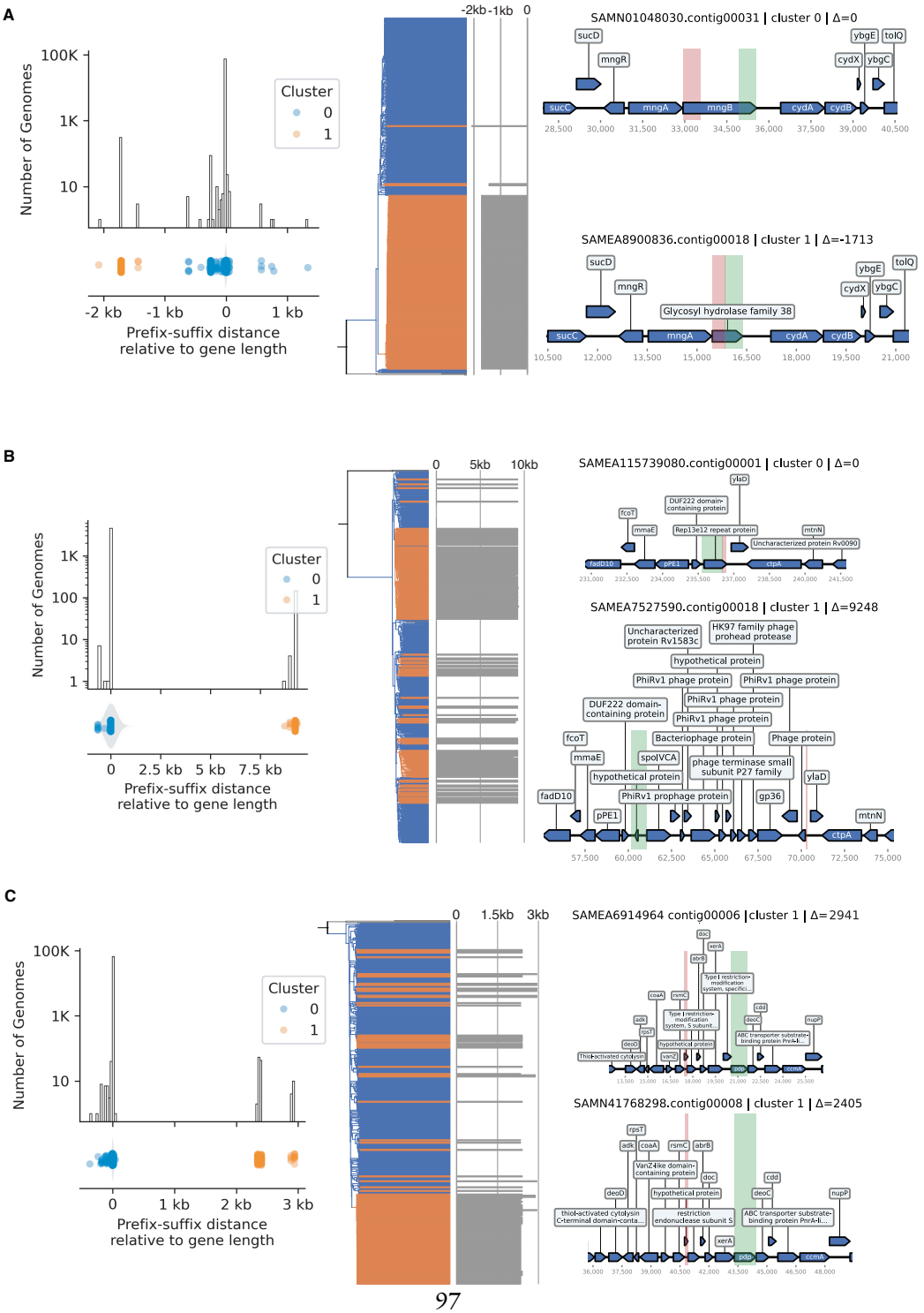
(B), **(C)** are the same as in **(A)** but with *acrR-glcD* and *mlaE-htpX* respectively.

(D) Distance-metric based phylogeny of 300 randomly sampled genomes with equal genomes sampled across the number of clusters identified. Clades are colored by cluster membership, and blue vertical bars off leaves represent the distance between the prefix-suffix found in that genome relative to the *yjcO-lysU* gene.

(E), **(F)** are the same as in **(D)** but with *acrR-glcD* and *mlaE-htpX* respectively.

Figure A.5 ***(following page):*** **Prefix-suffix approach captures diverse structural variation beyond deletion-born fusions.** Each row shows: prefix-suffix distance distribution (left), rooted phylogeny of 300 sampled genomes (middle), and representative genomic contexts (right).
(A) Internal deletion in *mngB* (*E. coli* K12).
(B) Repeat prophage insertions in *rep13e12* (*M. tuberculosis* H37Rv).
(C) Variable gene cargo disrupting *pdp* (*S. pneumoniae* TIGR4); the two Cluster 1 representatives differ by ≈500 bp due to the presence/absence of a Type I restriction system protein downstream of *xerA*.

Figure A.5: *(Continued)*



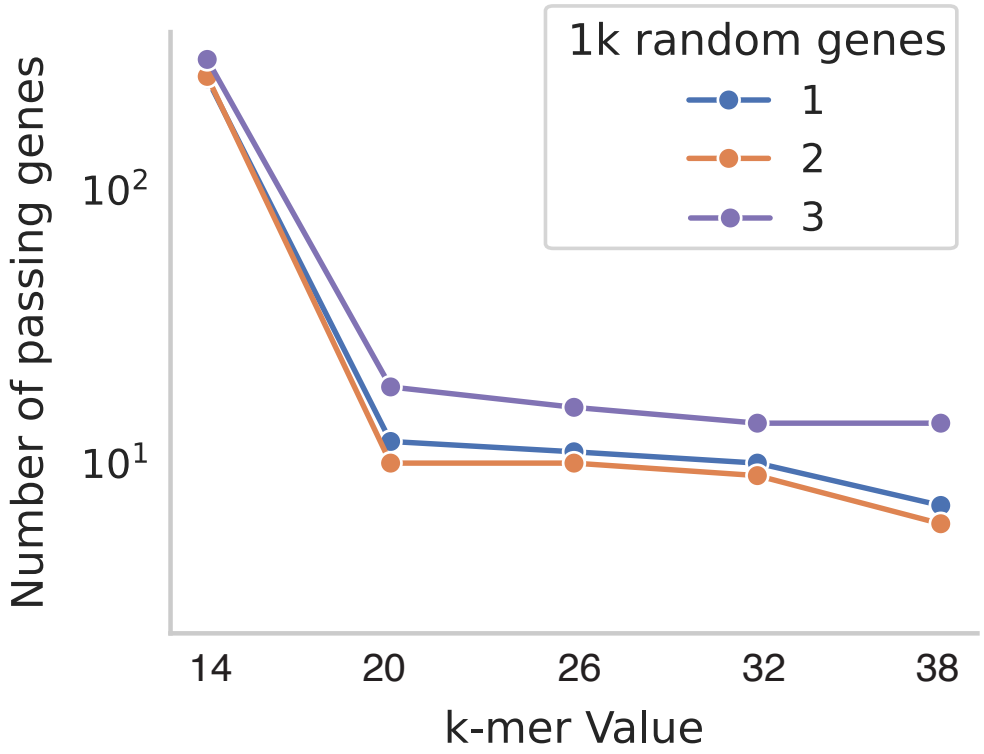


Figure A.6: Prefix-suffix approach is robust to values of k ≥ 20.
Number of genes with multimodal prefix-suffix distances detected across three random samples of 1,000 RefSeq genes at varying k-mer lengths.

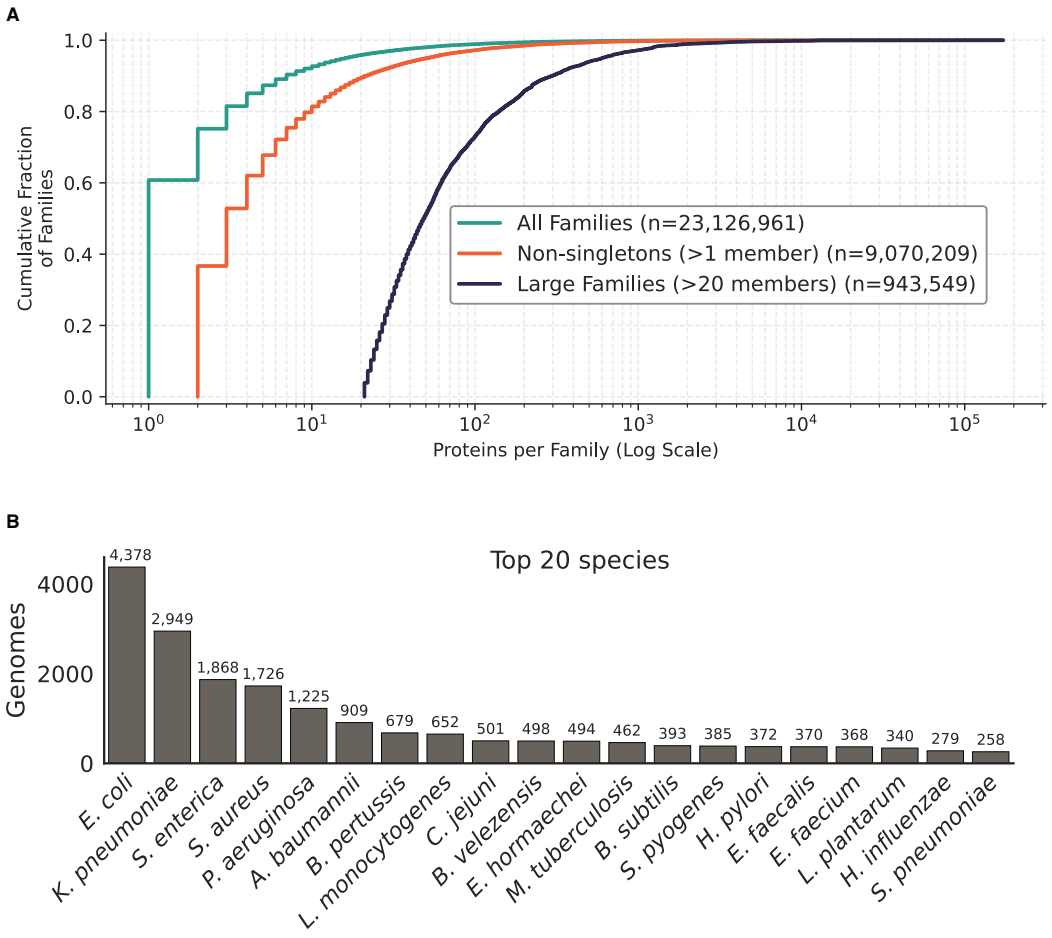


Figure A.7: Representation of both protein families and bacterial species are highly skewed in RefSeq complete genomes. **(A)** Cumulative distribution function of the number of protein members per family. Most represented proteins are singletons.

(B) Top 20 species represented in the 54,630 complete genomes proteins were pulled from. Count of each is displayed above each bar.

B

Supplementary Material of Chapter 2

B.1 Pre-antibiotic era isolates & their genomic resistance determinants

B.1.1 The *mdtM* gene in a clinically relevant *Salmonella enterica* strain isolated in 1911.

NCTC 74 is an isolate of *Salmonella enterica* serotype Typhi isolated from a case of food poisoning in 1911 [104]. *mdtM* is a multi-drug transporter efflux pump associated with increased resistance to ciprofloxacin, norfloxacin, levofloxacin, kanamycin, streptomycin, gentamicin, nalidixic acid, chloramphenicol, ethidium bromide and acriflavine, including fluoroquinolone antibiotics, which are common drugs to treat *Salmonella enterica* Typhi infections [152,151].

B.1.2 *oqxAB* efflux pump in a *K. pneumoniae* strain isolated in 1920.

NCTC 418 is an isolate of *K. pneumoniae* deposited within Johns Hopkins Hospital since at least 1911 [177]. The only accompanying metadata associated with it is that it is 'probably a descendant of the original capsule bacillus of Pfeiffer'. Pfeiffer's capsule bacillus is now known as *Haemophilus influenzae*; however, comparing k-mer sketching distance of the genome against RefSeq genomes shows strong similarity to *K. pneumoniae* isolates, indicating that this classification is likely incorrect. NCTC 418 carries the *oqxAB* operon encoding a small drug efflux pump enabling the bacterium to expel a broad range of toxic compounds. The *oqxAB* efflux pump has been shown to confer resistance to various antimicrobial agents, including quinolones and biocides [110].

B.1.3 Intact plasmid-borne beta-lactam resistance system in a clinically relevant *Staphylococcus aureus* strain isolated in 1932

NCTC 4136 is an isolate of *Staphylococcus aureus* isolated from a case of food poisoning in 1932 [89]. Genomic analysis of its sequence demonstrates the presence of intact *blaR1*, *blaZ* and *blaI* genes in its

genome. In addition, both PlasmidFinder and mob_recon predict that the contig where these genes are found is a plasmid. Annotations on the contig indicate the presence of replication proteins and an origin of replication (Figure B.3C). *BlaI* serves as the repressor of the operon, while *blaR1* senses beta-lactams and *blaZ* is a beta-lactamase[114].

B.1.4 Mobile *sul2* gene in an environmental *K. pneumoniae* strain isolated in 1933

NCTC 9127 is derived from a study that analyzed the contamination of ice cream in Iowa[132]. Later identified as *K. pneumoniae*, we find 100% sequence homology to the reference *sul2* gene deposited in the CARD database. On the bacterial chromosome (contig of 5.3 Mb length), we identify *sul2*, a sulfonamide resistant analogue for dihydropteroate synthase. In addition, we identify the typical dihydropteroate synthase gene in *K. pneumoniae*, *folP*, located upstream of *sul2*. Consistent with prior literature, we found *glmM*, a phosphoglucosamine mutase, to be downstream of *sul2*[145]. We additionally found that both *sul2* and the *glmM* protein sit inside an ICE known to typically carry both genes, indicating that the horizontal transfer of sulfonamide resistance likely predates the clinical usage of sulfonamide (Figure B.3D).

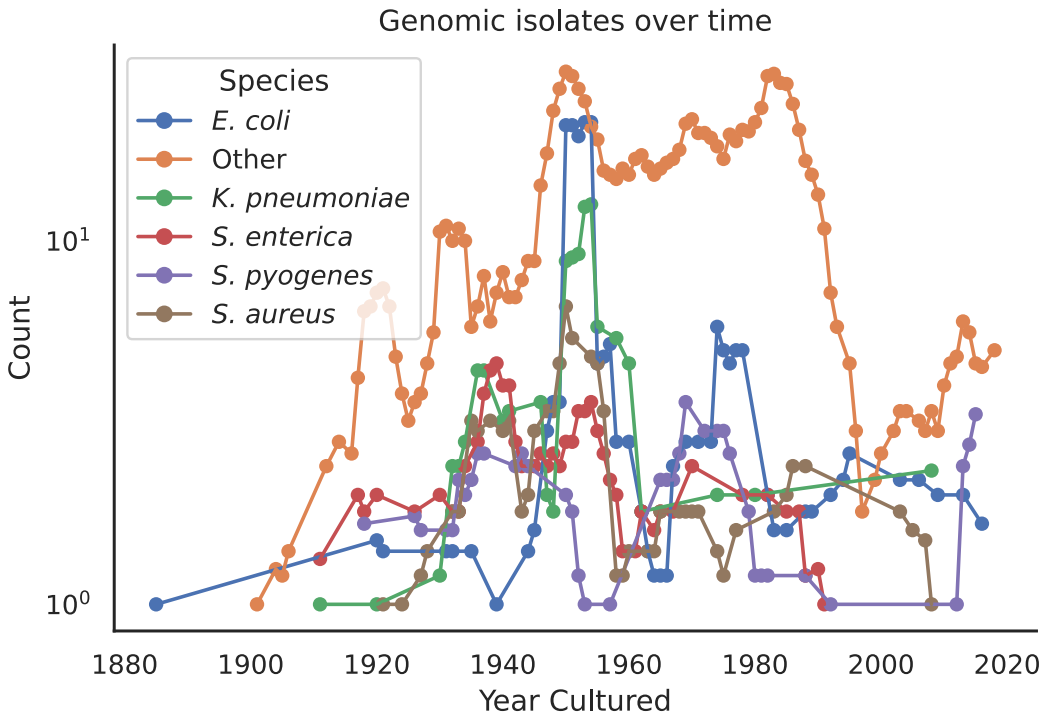


Figure B.1: Rolling count of isolates in the dataset over time.
Rolling average of number of isolates per specie in the data analyzed. Points represent the average number of isolates isolated within a window 5 years.

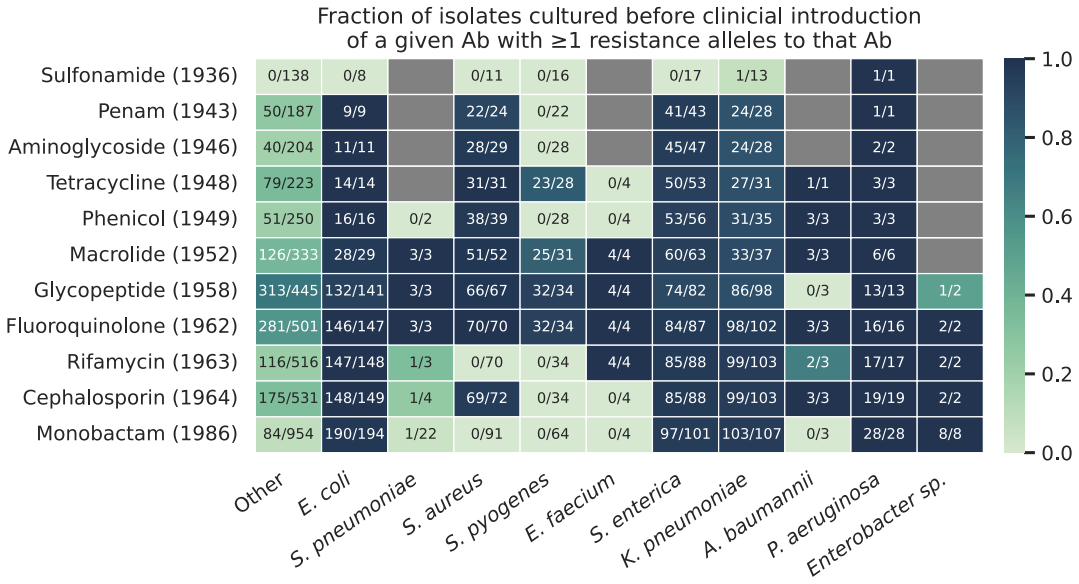


Figure B.2: Resistance-associated genomic elements known to exist within specific species were also ubiquitous before clinical introduction of given antibiotics.

Heatmap representing the fraction of isolates harboring resistance alleles before the introduction of a given antibiotic. Denominator in each cell is the number of isolates of a given species cultured before the clinical introduction of the antibiotics in the corresponding row (represented by the (Year) next to the antibiotic). The numerator is the number of isolates cultured before clinical introduction of the antibiotic containing resistance alleles to that antibiotic. Resistance allele prediction and classification done from CARD without prior species knowledge.

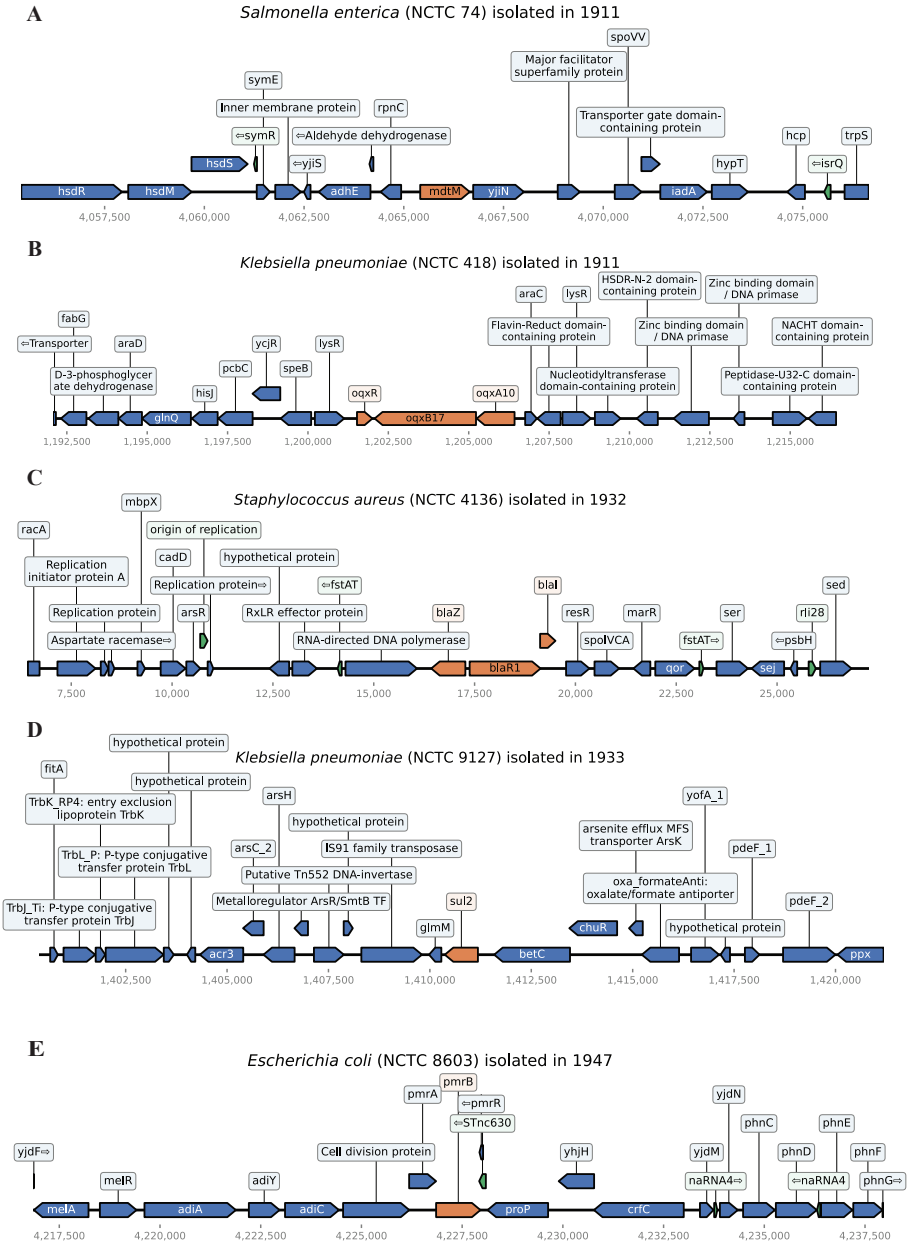


Figure B.3: Genomic neighborhood of pre-antibiotic era genomic resistance elements reveals similar genomic structures observed for these resistance alleles.

10kb genomic window surrounding each of the pre-antibiotic resistance elements in the isolates described in Note B.1

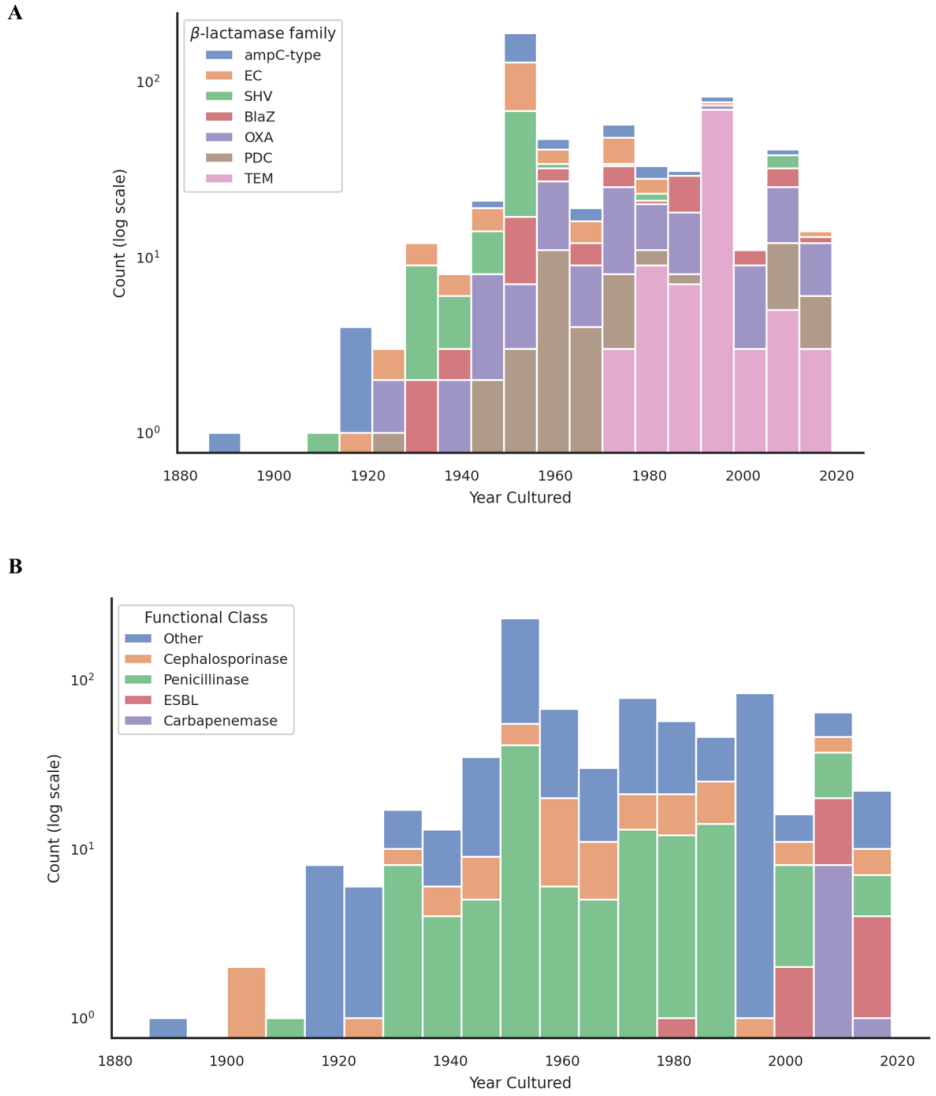


Figure B.4: Beta-lactamase diversity increased over time, with delayed emergence of ESBLs and carbapenemases.
(A) Stacked histogram showing the count of beta-lactamase genes identified in NCTC isolates, grouped by gene family (e.g., TEM, SHV, OXA) and binned by year of isolate.
(B) Same data as in **(A)**, but functionally classified into penicillinases, cephalosporinases, extended-spectrum beta-lactamases (ESBLs), and carbapenemases. "Other" includes beta-lactamases not fitting into the four major functional groups or lacking sufficient experimental evidence for classification.

Figure B.5 *(following page)*: Phylogeny of beta-lactamase genes reveals independent origins across families and time. A maximum-likelihood phylogenetic tree of all beta-lactamase gene sequences identified in our dataset, rooted using a representative *blaNDM-1* outgroup. Concentric annotation rings indicate the year of isolation (inner gradient, 1920–2016), host species (middle ring), and beta-lactamase family (outer ring).

Figure B.5: *(Continued)*

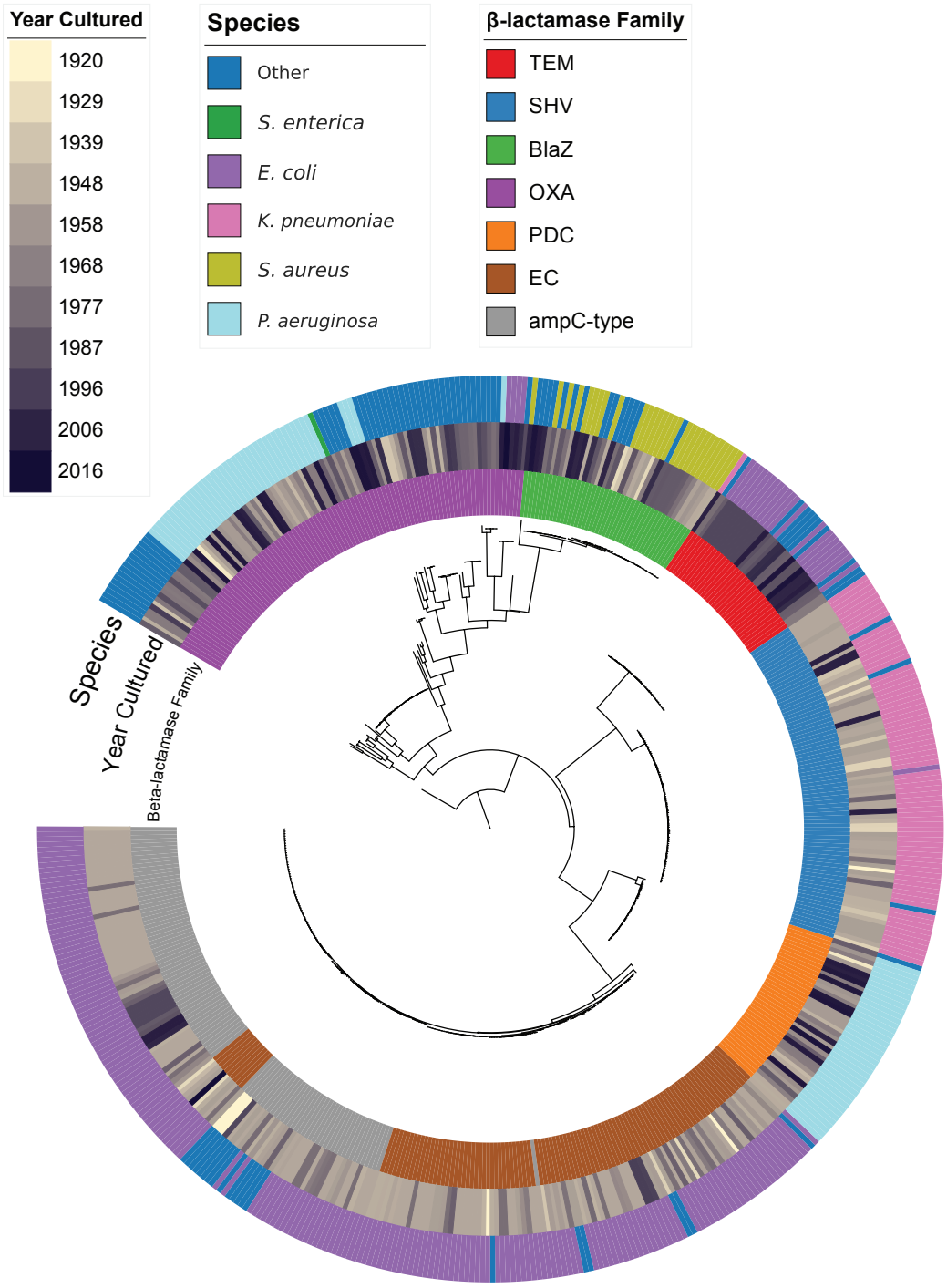
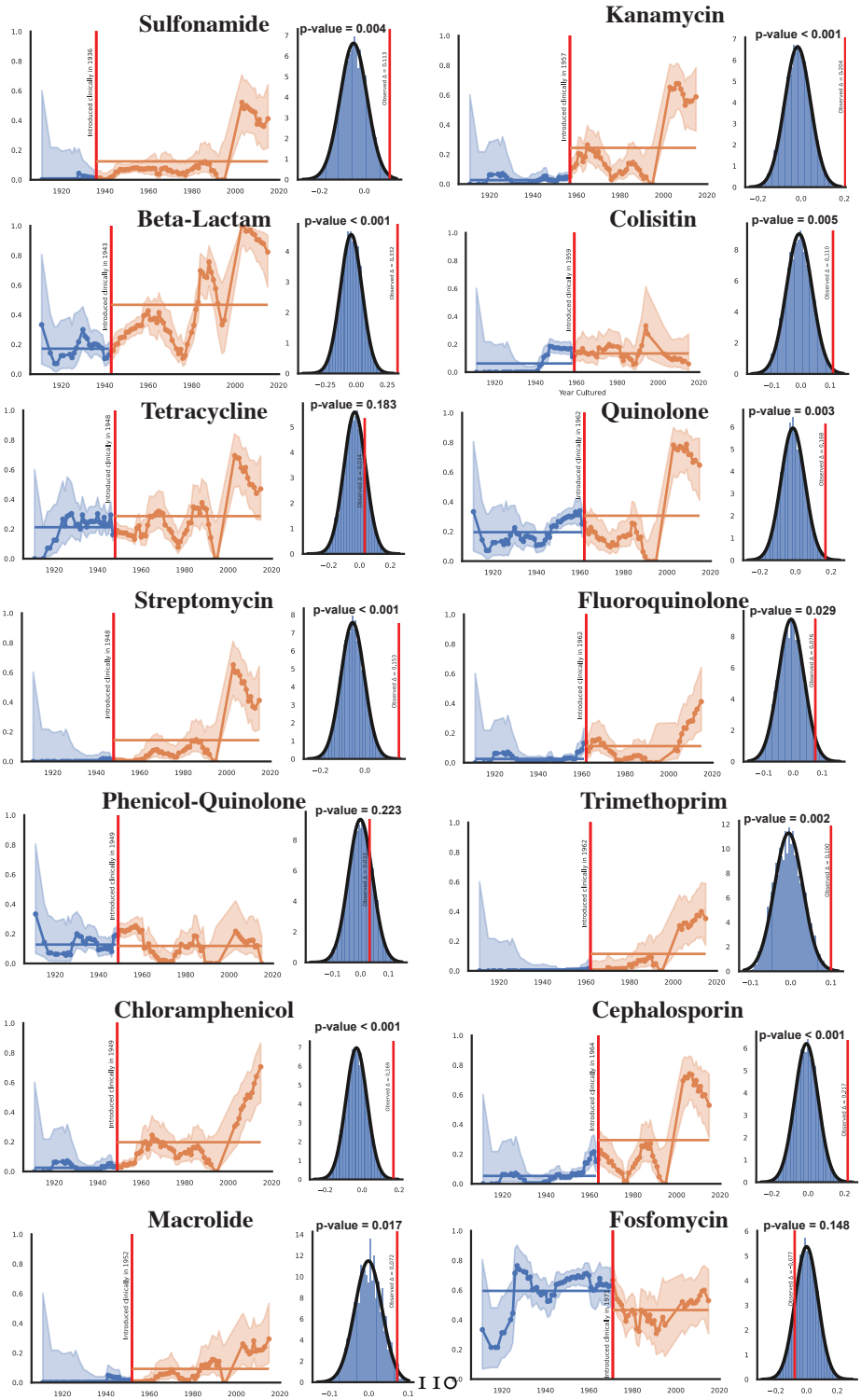


Figure B.6 *(following page)*: Significant increase in observed resistance frequencies after clinical introduction of an antibiotic across most drug types.
As in Figure 2.3 but across all drug types.

Figure B.6: *(Continued)*



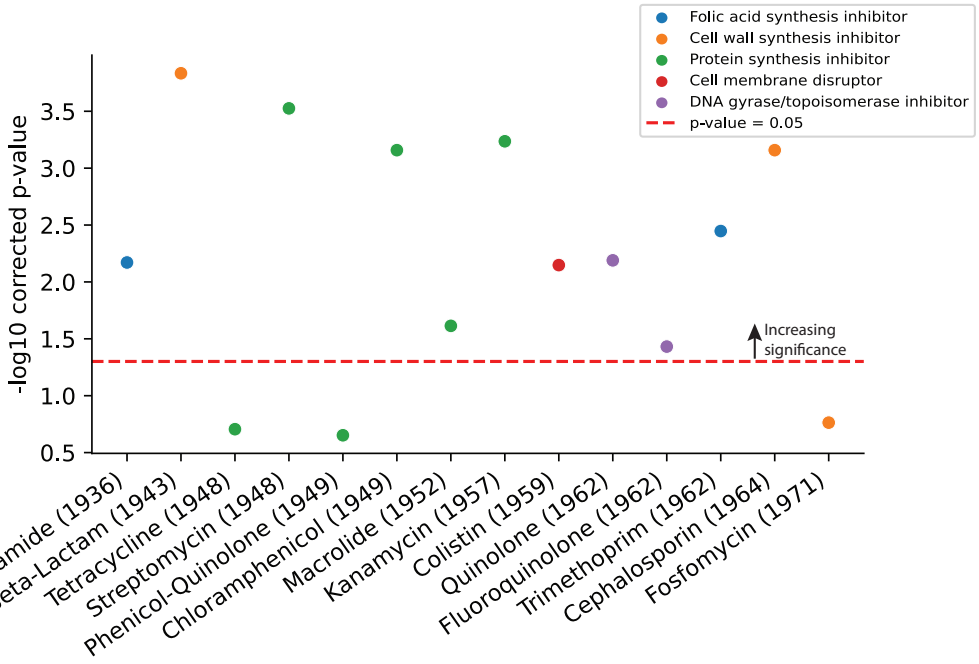


Figure B.7: Significant increase in observed resistance rates after clinical introduction are not associated with antibiotic mechanism.

Significance of change in resistance prevalence over all isolates and drug classes with 10,000 shuffles. Multiple hypotheses corrected with Benjamini-Hochberg. Individual points colored by antibiotic mechanism.

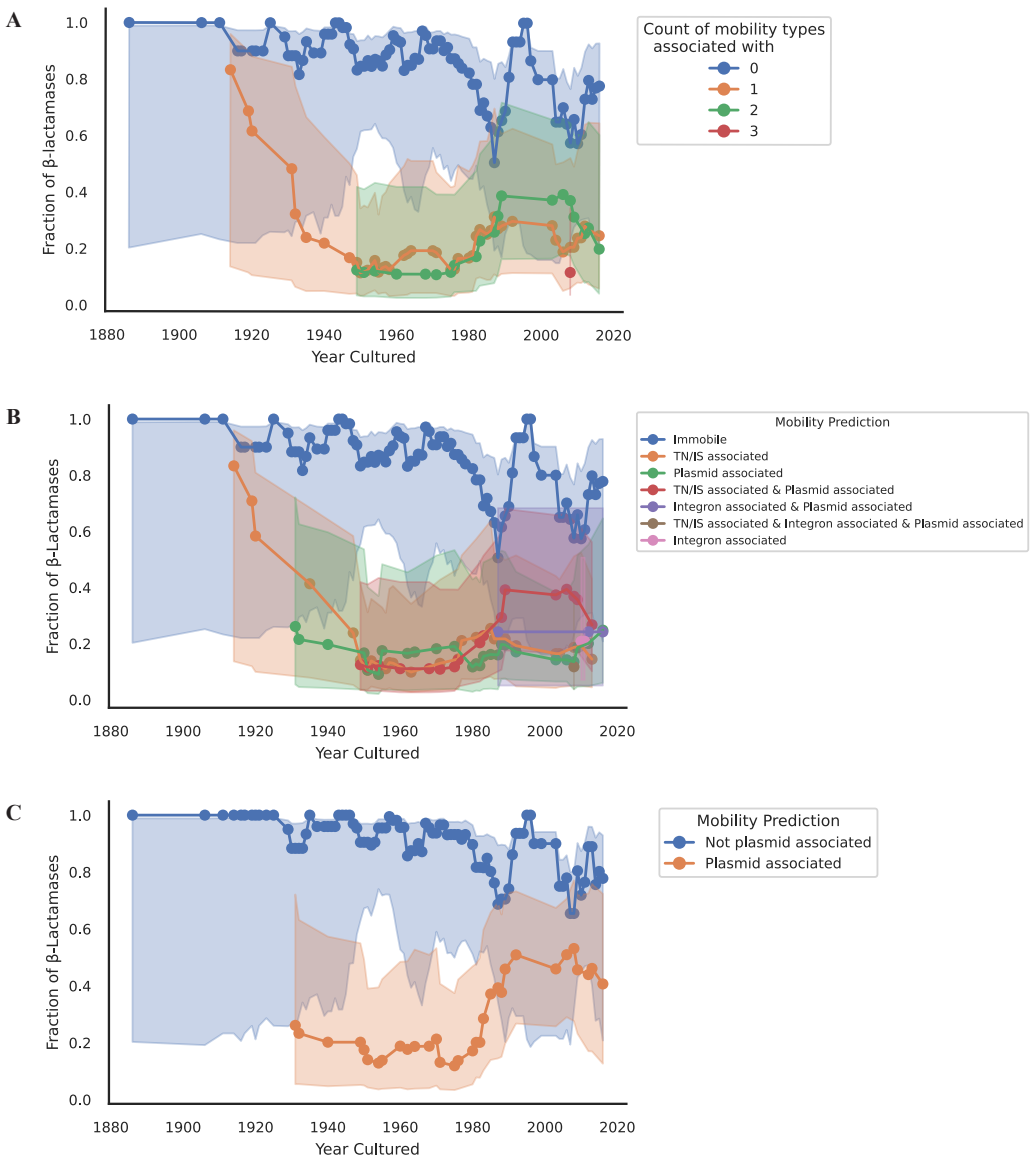


Figure B.8: Beta-lactamases experience increasing mobility over the years, with most mobility driven by plasmids. Fraction of beta-lactamases classified as mobile over the course of the NCTC, shaded regions represent 95% confidence intervals based on sampling size.

(A) Number of mobility types associated with each beta-lactamase in the NCTC.

(B) Breakdown of which combination of mobility types cause the counts in **(A)**.

(C) Fraction of beta-lactamases with some or no plasmid association.

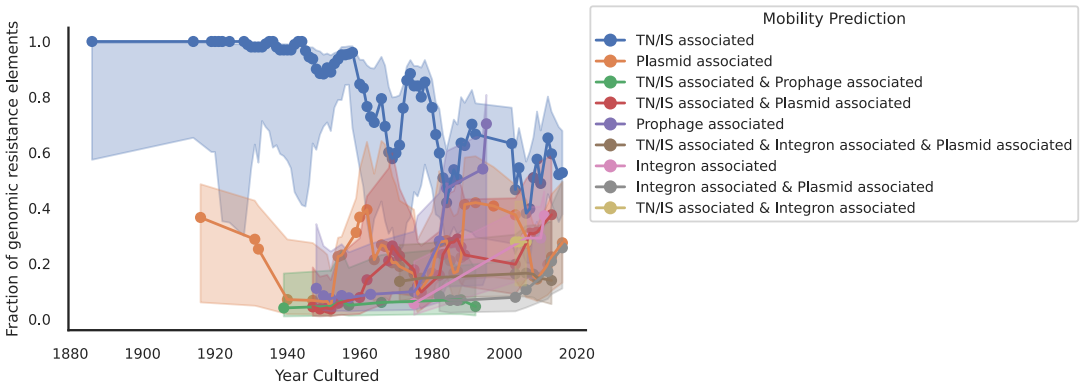


Figure B.9: Increase in mobility access of mobile resistance elements appears to be driven by an increase in Plasmid-associated resistance elements.

Fraction of mobile resistance elements as a function of time and the specific combination of mobility types observed. The shaded regions denote 95% confidence intervals from a beta distribution.

C

Supplementary Material of Chapter 3

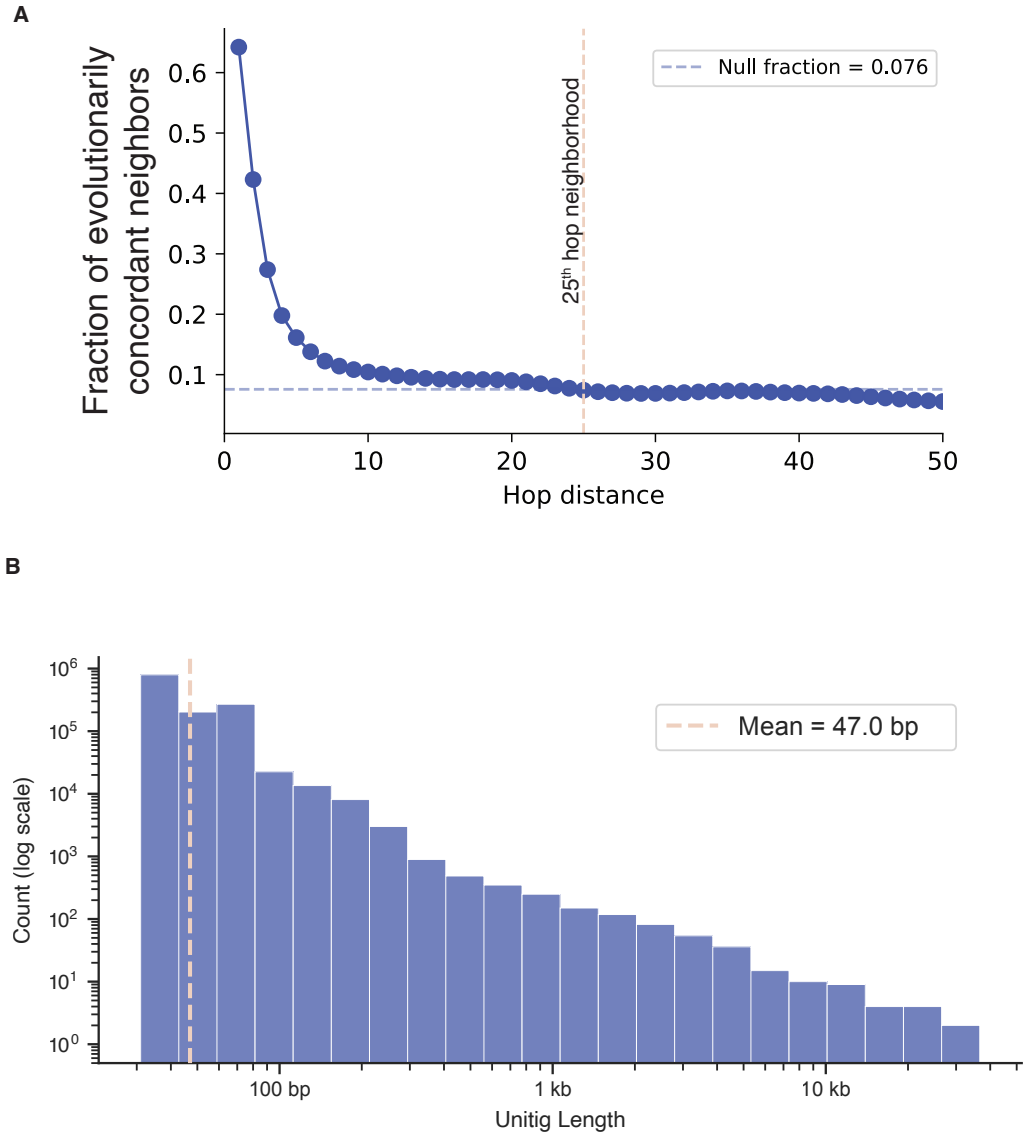


Figure C.1: Evolutionary persistence and unitig length distribution in 1,198 *Staphylococcus aureus* ST8 genomes.
(A) Fraction of evolutionarily concordant neighbors as a function of hop distance from seed unitigs. Dashed horizontal line: null expectation. Vertical dashed line: 25th hop, where the curve crosses null.
(B) Distribution of unitig lengths in the ST8 pcDBG (log scale). Dashed line: mean unitig length.

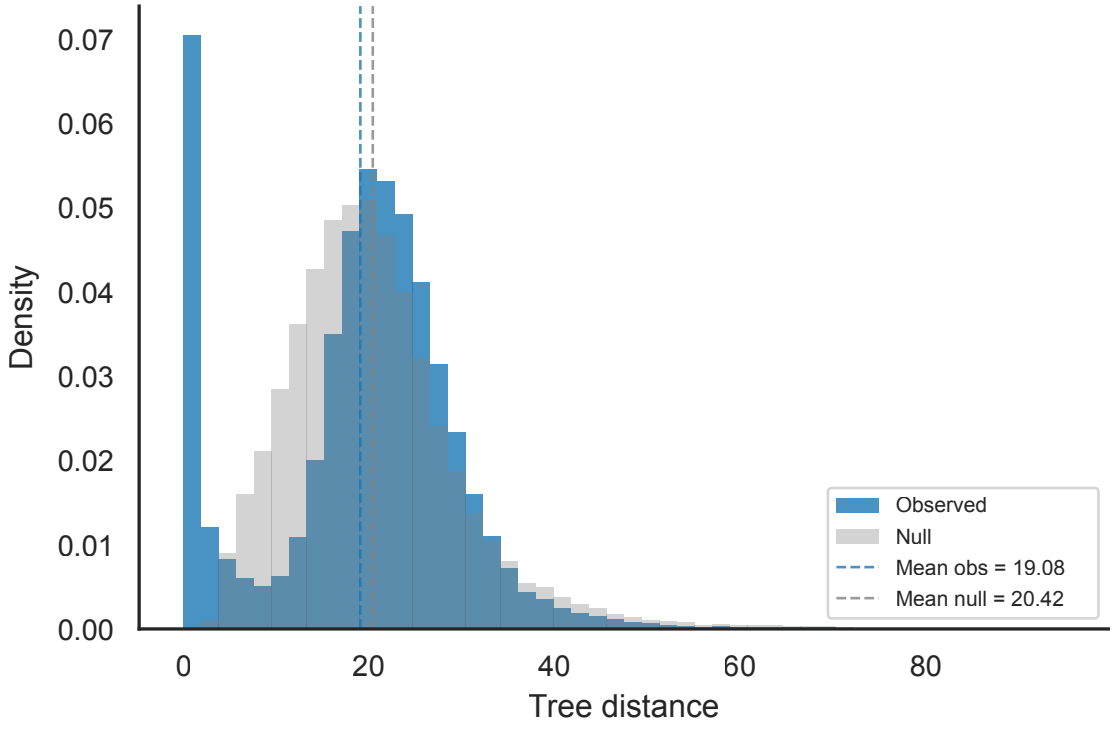


Figure C.2: Boundary severity distribution in 1,198 *Staphylococcus aureus* ST8 genomes.
Distribution of tree distances at discordant edges compared to a null model based on random pairing of phylogenetic assignments weighted by edge-endpoint frequency. The observed distribution is shifted toward lower tree distances relative to null, indicating that evolutionary boundaries in the ST8 pangenome are milder than expected by chance.

References

[1] Acman, M., Wang, R., van Dorp, L., Shaw, L. P., Wang, Q., Luhmann, N., Yin, Y., Sun, S., Chen, H., Wang, H., & Balloux, F. (2022). Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene blaNDM. *Nature Communications*, 13(1), 1131. number: 1 publisher: Nature Publishing Group.

[2] Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H.-K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., & McArthur, A. G. (2020). Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 48(D1), D517–D525. PMID: 31665441 PMCID: PMC7145624.

[3] Alexander, S. & Gurung, S. (2020). 100 years and counting: The National Collection of Type Cultures centenary. *The Lancet Microbe*, 1(6), e235–e236.

[4] Álvarez-Lugo, A. & Becerra, A. (2021). The Role of Gene Duplication in the Divergence of Enzyme Function: A Comparative Approach. *Frontiers in Genetics*, 12. publisher: Frontiers.

[5] Aschbacher, R., Doumith, M., Livermore, D. M., Larcher, C., & Woodford, N. (2008). Linkage of acquired quinolone resistance (qnrS1) and metallo-beta-lactamase (blaVIM-1) genes in multiple species of Enterobacteriaceae from Bolzano, Italy. *The Journal of Antimicrobial Chemotherapy*, 61(3), 515–523. PMID: 18184647.

[6] Ashish, A., Paterson, S., Mowat, E., Fothergill, J. L., Walshaw, M. J., & Winstanley, C. (2013). Extensive diversification is a common feature of Pseudomonas aeruginosa populations during respiratory infections in cystic fibrosis. *Journal of Cystic Fibrosis*, 12(6), 790–793.

[7] Baker, K. S., Burnett, E., McGregor, H., Deheer-Graham, A., Boinett, C., Langridge, G. C., Wailan, A. M., Cain, A. K., Thomson, N. R., Russell, J. E., & Parkhill, J. (2015). The Murray collection of pre-antibiotic era Enterobacteriacae: a unique research resource. *Genome Medicine*, 7(1), 97.

[8] Baker, K. S., Mather, A. E., McGregor, H., Coupland, P., Langridge, G. C., Day, M., Deheer-Graham, A., Parkhill, J., Russell, J. E., & Thomson, N. R. (2014). The extant World War 1 dysentery bacillus NCTC1: A genomic analysis. *The Lancet*, 384(9955), 1691–1697.

[9] Baltoumas, F. A., Karatzas, E., Paez-Espino, D., Venetsianou, N. K., Aplakidou, E., Oulas, A., Finn, R. D., Ovchinnikov, S., Pafilis, E., Kyrpides, N. C., & Pavlopoulos, G. A. (2023). Exploring microbial functional biodiversity at the protein family level—From metagenomic sequence reads to annotated protein clusters. *Frontiers in Bioinformatics*, 3. publisher: Frontiers.

[10] Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, 115(25), 6506–6511.

[11] Barrick, J. E. & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12), 827–839. number: 12 publisher: Nature Publishing Group.

[12] Beaver, R. C. & Neufeld, J. D. (2024). Microbial ecology of the deep terrestrial subsurface. *The ISME Journal*, 18(1), wrae091.

[13] Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

[14] Bennett, R. J. & Baker, K. S. (2019). Looking Backward To Move Forward: The Utility of Sequencing Historical Bacterial Genomes. *Journal of Clinical Microbiology*, 57(8), 10.1128/jcm.00100–19.

[15] Bespiatykh, D., Bespyatykh, J., Mokrousov, I., & Shitikov, E. (2021). A Comprehensive Map of Mycobacterium tuberculosis Complex Regions of Difference. *mSphere*, 6(4), e00535–21.

[16] Bichara, M., Wagner, J., & Lambert, I. B. (2006). Mechanisms of tandem repeat instability in bacteria. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 598(1), 144–163.

[17] Bobay, L.-M. & Ochman, H. (2017). The Evolution of Bacterial Genome Architecture. *Frontiers in Genetics*, 8. publisher: Frontiers.

[18] Bottai, D., Frigui, W., Sayes, F., Di Luca, M., Spadoni, D., Pawlik, A., Zoppo, M., Orgeur, M., Khanna, V., Hardy, D., Mangenot, S., Barbe, V., Medigue, C., Ma, L., Bouchier, C., Tavanti, A., Larrouy-Maumus, G., & Brosch, R. (2020). Tbd1 deletion as a driver of the evolutionary success of modern epidemic Mycobacterium tuberculosis lineages. *Nature Communications*, 11(1), 684. publisher: Nature Publishing Group.

[19] Břinda, K. (2025). karel-brinda/attotree. https://github.com/karel-brinda/attotree. original-date: 2023-06-13T09:24:45Z.

[20] Břinda, K., Lima, L., Pignotti, S., Quinones-Olvera, N., Salikhov, K., Chikhi, R., Kucherov, G., Iqbal, Z., & Baym, M. (2025). Efficient and robust search of microbial genomes via phylogenetic compression. *Nature Methods*, 22(4), 692–697. publisher: Nature Publishing Group.

[21] Brockhurst, M. A., Harrison, E., Hall, J. P. J., Richards, T., McNally, A., & MacLean, C. (2019). The Ecology and Evolution of Pangenomes. *Current Biology*, 29(20), R1094–R1103.

[22] Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L. M., Pym, A. S., Samper, S., van Soolingen, D., & Cole, S. T. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6), 3684–3689. PMID: 11891304 PMCID: PMC122584.

[23] Browne, A. J., Chipeta, M. G., Haines-Woodhouse, G., Kumaran, E. P. A., Hamadani, B. H. K., Zaraa, S., Henry, N. J., Deshpande, A., Reiner, R. C., Day, N. P. J., Lopez, A. D., Dunachie, S., Moore, C. E., Stergachis, A., Hay, S. I., & Dolecek, C. (2021). Global antibiotic consumption and usage in humans, 2000–18: a spatial modelling study. *The Lancet. Planetary Health*, 5(12), e893–e904. PMID: 34774223 PMCID: PMC8654683.

[24] Budzik, J. M., Rosche, W. A., Rietsch, A., & O'Toole, G. A. (2004). Isolation and Characterization of a Generalized Transducing Phage for Pseudomonas aeruginosa Strains PAO1 and PA14. *Journal of Bacteriology*, 186(10), 3270–3273. publisher: American Society for Microbiology.

[25] Bush, K. & Jacoby, G. A. (2010). Updated Functional Classification of β-Lactamases. *Antimicrobial Agents and Chemotherapy*, 54(3), 969–976.

[26] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). Blast+: architecture and applications. *BMC bioinformatics*, 10, 421. PMID: 20003500 PMCID: PMC2803857.

[27] Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.-L., & Brüssow, H. (2003). Phage as agents of lateral gene transfer. *Current Opinion in Microbiology*, 6(4), 417–424.

[28] Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, 487(7407), 370–374. number: 7407 publisher: Nature Publishing Group.

[29] Cazares, A., Figueroa, W., Cazares, D., Lima, L., Turnbull, J. D., McGregor, H., Dicks, J., Alexander, S., Iqbal, Z., & Thomson, N. R. (2024). Pre and Post antibiotic epoch: insights into the historical spread of antimicrobial resistance. page: 2024.09.03.610986 section: New Results.

[30] Chambers, D. (2025). d-chambers/dbscan1d. https://github.com/d-chambers/dbscan1d. original-date: 2019-10-07T05:04:44Z.

[31] Charles, C., Conde, C., Vorimore, F., Cochard, T., Michelet, L., Boschiroli, M. L., & Biet, F. (2023). Features of Mycobacterium bovis Complete Genomes Belonging to 5 Different Lineages. *Microorganisms*, 11(1), 177. number: 1 publisher: Multidisciplinary Digital Publishing Institute.

[32] Chen, L., Zhao, N., Cao, J., Liu, X., Xu, J., Ma, Y., Yu, Y., Zhang, X., Zhang, W., Guan, X., Yu, X., Liu, Z., Fan, Y., Wang, Y., Liang, F., Wang, D., Zhao, L., Song, M., & Wang, J. (2022). Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nature Communications*, 13(1), 3175. publisher: Nature Publishing Group.

[33] Chikhi, R., Limasset, A., & Medvedev, P. (2016). Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, 32(12), i201–i208.

[34] Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). Checkm2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, 20(8), 1203–1212. publisher: Nature Publishing Group.

[35] Clayton, A. L., Jackson, D. G., Weiss, R. B., & Dale, C. (2016). Adaptation by Deletogenic Replication Slippage in a Nascent Symbiont. *Molecular Biology and Evolution*, 33(8), 1957–1966.

[36] Cleary, A., Ramaraj, T., Kahanda, I., Mudge, J., & Mumey, B. (2019). Exploring Frequented Regions in Pan-Genomic Graphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(5), 1424–1435.

[37] Collignon, P. (2015). Antibiotic resistance: are we all doomed? *Internal Medicine Journal*, 45(11), 1109–1115. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/imj.12902.

[38] Coluzzi, C., Garcillán-Barcia, M. P., de la Cruz, F., & Rocha, E. P. (2022). Evolution of Plasmid Mobility: Origin and Fate of Conjugative and Nonconjugative Plasmids. *Molecular Biology and Evolution*, 39(6), msac115.

[39] Corona, F. & Martinez, J. (2013). Phenotypic Resistance to Antibiotics. *Antibiotics*, 2(2), 237–255.

[40] Couce, A., Limdi, A., Magnan, M., Owen, S. V., Herren, C. M., Lenski, R. E., Tenaillon, O., & Baym, M. (2024). Changing fitness effects of mutations through long-term bacterial evolution. *Science*, 383(6681), eadd1417. publisher: American Association for the Advancement of Science.

[41] Cracco, A. & Tomescu, A. I. (2023). Extremely fast construction and querying of compacted and colored de Bruijn graphs with GGCAT. *Genome Research*, (pp. genome;gr.277615.122v2).

[42] Dahl, J. L. (2004). Electron microscopy analysis of *Mycobacterium tuberculosis* cell division. *FEMS Microbiology Letters*, 240(1), 15–20.

[43] Davies, J. (1990). What are antibiotics? Archaic functions for modern activities. *Molecular Microbiology*, 4(8), 1227–1232. PMID: 2280684.

[44] Davies, J. & Davies, D. (2010). Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews*, 74(3), 417–433. publisher: American Society for Microbiology.

[45] D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W. L., Schwarz, C., Froese, D., Zazula, G., Calmels, F., Debruyne, R., Golding, G. B., Poinar, H. N., & Wright, G. D. (2011). Antibiotic resistance is ancient. *Nature*, 477(7365), 457–461.

[46] Dekker, N. P., Lammel, C. J., & Brooks, G. F. (1991). Scanning electron microscopy of piliated *Neisseria gonorrhoeae* processed with hexamethyldisilazane. *Journal of Electron Microscopy Technique*, 19(4), 461–467.

[47] Devault, A. M., Golding, G. B., Waglechner, N., Enk, J. M., Kuch, M., Tien, J. H., Shi, M., Fisman, D. N., Dhody, A. N., Forrest, S., Bos, K. I., Earn, D. J. D., Holmes, E. C., & Poinar, H. N. (2014). Second-Pandemic Strain of Vibrio cholerae from the Philadelphia Cholera Outbreak of 1849. *New England Journal of Medicine*, 370(4), 334–340. publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1308663.

[48] Diaz Caballero, J., Wheatley, R. M., Kapel, N., López-Causapé, C., Van der Schalk, T., Quinn, A., Shaw, L. P., Ogunlana, L., Recanatini, C., Xavier, B. B., Timbermont, L., Kluytmans, J., Ruzin, A., Esser, M., Malhotra-Kumar, S., Oliver, A., & MacLean, R. C. (2023). Mixed strain pathogen populations accelerate the evolution of antibiotic resistance in patients. *Nature Communications*, 14, 4083. PMID: 37438338 PMCID: PMC10338428.

[49] Dicks, J., Fazal, M.-A., Oliver, K., Grayson, N. E., Turnbull, J. D., Bane, E., Burnett, E., Deheer-Graham, A., Holroyd, N., Kaushal, D., Keane, J., Langridge, G., Lomax, J., McGregor, H., Picton, S., Quail, M., Singh, D., Tracey, A., Korlach, J., Russell, J. E., Alexander, S., & Parkhill, J. (2023). NCTC3000: A century of bacterial strain collecting leads to a rich genomic data resource. *Microbial Genomics*, 9(5), 000976.

[50] Diep, B. A., Gill, S. R., Chang, R. F., Phan, T. H., Chen, J. H., Davidson, M. G., Lin, F., Lin, J., Carleton, H. A., Mongodin, E. F., Sensabaugh, G. F., & Perdreau-Remington, F. (2006). Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. *The Lancet*, 367(9512), 731–739.

[51] Dieppa-Colón, E., Martin, C., Kosmopoulos, J. C., & Anantharaman, K. (2025). ProphageDB: a comprehensive database to explore diversity, distribution, and ecology of prophages. *Environmental Microbiome*, 20(1), 5.

[52] D'Andrea, M. M., Arena, F., Pallecchi, L., & Rossolini, G. M. (2013). Ctx-M-type β-lactamases: A successful story of antibiotic resistance. *International Journal of Medical Microbiology*, 303(6), 305–317.

[53] Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. In *Genome Informatics 2009*: PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.

[54] Ellrott, K., Jaroszewski, L., Li, W., Wooley, J. C., & Godzik, A. (2010). Expansion of the Protein Repertoire in Newly Explored Environments: Human Gut Microbiome Specific Protein Families. *PLOS Computational Biology*, 6(6), e1000798. publisher: Public Library of Science.

[55] Falagas, M. E., Vouloumanou, E. K., Samonis, G., & Vardakas, K. Z. (2016). Fosfomycin. *Clinical Microbiology Reviews*, 29(2), 321–347. PMID: 26960938 PMCID: PMC4786888.

[56] Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879), 1034–1039.

[57] Farr, A. D., Remigi, P., & Rainey, P. B. (2017). Adaptive evolution by spontaneous domain fusion and protein relocalization. *Nature Ecology & Evolution*, 1(10), 1562–1568. publisher: Nature Publishing Group.

[58] Favate, J. S., Liang, S., Cope, A. L., Yadavalli, S. S., & Shah, P. (2022). The landscape of transcriptional and translational changes over 22 years of bacterial adaptation. *eLife*, 11. publisher: eLife Sciences Publications, Ltd.

[59] Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H., & Klimke, W. (2021). Amrfinderplus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports*, 11(1), 12728. number: 1 publisher: Nature Publishing Group.

[60] Ferragina, P. & Manzini, G. (2000). Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science* (pp. 390–398). ISSN: 0272-5428.

[61] Fitch, W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology*, 20(4), 406–416.

[62] Fleming, A. (1945). Penicillin. Nobel Prize in Physiology or Medicine 1945.

[63] Flemming, H.-C. & Wuertz, S. (2019). Bacteria and archaea on Earth and their abundance in biofilms. *Nature Reviews Microbiology*, 17(4), 247–260.

[64] Foy, S. G., Wilson, B. A., Bertram, J., Cordes, M. H. J., & Masel, J. (2019). A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution. *Genetics*, 211(4), 1345–1355.

[65] Fu, Z., Ma, Y., Chen, C., Guo, Y., Hu, F., Liu, Y., Xu, X., & Wang, M. (2016). Prevalence of Fosfomycin Resistance and Mutations in murA, glpT, and uhpT in Methicillin-Resistant Staphylococcus aureus Strains Isolated from Blood and Cerebrospinal Fluid Samples. *Frontiers in Microbiology*, 6, 1544. PMID: 26793179 PMCID: PMC4707275.

[66] Gallant, J., Mouton, J., Ummels, R., ten Hagen-Jongman, C., Kriel, N., Pain, A., Warren, R. M., Bitter, W., Heunis, T., & Sampson, S. L. (2020). Identification of gene fusion events in Mycobacterium tuberculosis that encode chimeric proteins. *NAR Genomics and Bioinformatics*, 2(2), lqaa033.

[67] Giovannoni, S. J., Cameron Thrash, J., & Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *The ISME Journal*, 8(8), 1553–1565. publisher: Nature Publishing Group.

[68] Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappé, M. S., Short, J. M., Carrington, J. C., & Mathur, E. J. (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, 309(5738), 1242–1245. publisher: American Association for the Advancement of Science.

[69] Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., & Desai, M. M. (2017). The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678), 45–50. number: 7678 publisher: Nature Publishing Group.

[70] Guerra, B., Soto, S., Cal, S., & Mendoza, M. C. (2000). Antimicrobial Resistance and Spread of Class 1 Integrons among Salmonella Serotypes. *Antimicrobial Agents and Chemotherapy*, 44(8), 2166–2169. publisher: American Society for Microbiology.

[71] Harkins, C. P., Pichon, B., Doumith, M., Parkhill, J., Westh, H., Tomasz, A., de Lencastre, H., Bentley, S. D., Kearns, A. M., & Holden, M. T. G. (2017). Methicillin-resistant Staphylococcus aureus emerged long before the introduction of methicillin into clinical practice. *Genome Biology*, 18, 130. PMID: 28724393 PMCID: PMC5517843.

[72] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. publisher: Nature Publishing Group.

[73] Hilbert, F., Scherwitzel, M., Paulsen, P., & Szostak, M. P. (2010). Survival of Campylobacter jejuni under conditions of atmospheric oxygen tension with the support of Pseudomonas spp. *Applied and Environmental Microbiology*, 76(17), 5911–5917. PMID: 20639377 PMCID: PMC2935043.

[74] Hoff, G., Bertrand, C., Piotrowski, E., Thibessard, A., & Leblond, P. (2018). Genome plasticity is governed by double strand break DNA repair in Streptomyces. *Scientific Reports*, 8(1), 5272. publisher: Nature Publishing Group.

[75] Holley, G. & Melsted, P. (2020). Bifrost: Highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biology*, 21(1), 249.

[76] Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1(5), 1–6. publisher: Nature Publishing Group.

[77] Hughes, J. B., Hellmann, J. J., Ricketts, T. H., & Bohannan, B. J. M. (2001). Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. *Applied and Environmental Microbiology*, 67(10), 4399–4406.

[78] Hughes, V. M. & Datta, N. (1983). Conjugative plasmids in bacteria of the 'pre-antibiotic' era. *Nature*, 302(5910), 725–726. publisher: Nature Publishing Group.

[79] Hunt, M., Lima, L., Anderson, D., Bouras, G., Hall, M., Hawkey, J., Schwengers, O., Shen, W., Lees, J. A., & Iqbal, Z. (2025). Allthebacteria – all bacterial genomes assembled, available, and searchable. ISSN: 2692-8205 page: 2024.03.08.584059 section: New Results.

[80] Hutchings, M. I., Truman, A. W., & Wilkinson, B. (2019). Antibiotics: past, present and future. *Current Opinion in Microbiology*, 51, 72–80.

[81] Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., Versalovic, J., Wollam, A. M., Worley, K. C., Wortman, J. R., Young, S. K., Zeng, Q., Aagaard, K. M., Abolude, O. O., Allen-Vercoe, E., Alm, E. J., Alvarado, L., Andersen, G. L., Anderson, S., Appelbaum, E., Arachchi, H. M., Armitage, G., Arze, C. A., Ayvaz, T., Baker, C. C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M. J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G. A., Buhay, C. J., Busam, D. A., Campbell, J. L., Canon, S. R., Cantarel, B. L., Chain, P. S. G., Chen, I.-M. A., Chen, L., Chhibba, S., Chu, K., Ciulla, D. M., Clemente, J. C., Clifton, S. W., Conlan, S., Crabtree, J., Cutting, M. A., Davidovics, N. J., Davis, C. C., DeSantis, T. Z., Deal, C., Delehaunty, K. D., Dewhirst, F. E., Deych, E., Ding, Y., Dooling, D. J., Dugan, S. P., Michael Dunne, W., Scott Durkin, A., Edgar, R. C., Erlich, R. L., Farmer, C. N., Farrell, R. M., Faust, K., Feldgarden, M., Felix, V. M., Fisher, S., Fodor,

A. A., Forney, L. J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D. C., Fronick, C. C., Fulton, L. L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M. Y., Goldberg, J. M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Kinder Haake, S., Haas, B. J., Hamilton, H. A., Harris, E. L., Hepburn, T. A., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth, C., Huang, K. H., Huse, S. M., Izard, J., Jansson, J. K., Jiang, H., Jordan, C., Joshi, V., Katancik, J. A., Keitel, W. A., Kelley, S. T., Kells, C., King, N. B., Knights, D., Kong, H. H., Koren, O., Koren, S., Kota, K. C., Kovar, C. L., Kyrpides, N. C., La Rosa, P. S., Lee, S. L., Lemon, K. P., Lennon, N., Lewis, C. M., Lewis, L., Ley, R. E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C. A., Dwayne Lunsford, R., Madden, T., Mahurkar, A. A., Mannon, P. J., Mardis, E. R., Markowitz, V. M., Mavromatis, K., McCorrison, J. M., McDonald, D., McEwen, J., McGuire, A. L., McInnes, P., Mehta, T., Mihindukulasuriya, K. A., Miller, J. R., Minx, P. J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S. M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K. S., Pop, M., Priest, M. E., Proctor, L. M., Qin, X., Raes, J., Ravel, J., Reid, J. G., Rho, M., Rhodes, R., Riehle, K. P., Rivera, M. C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M. C., Russ, C., Sanka, R. K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J. A., Schloss, P. D., Schmidt, T. M., Scholz, M., Schriml, L., Schubert, A. M., Segata, N., Segre, J. A., Shannon, W. D., Sharp, R. R., Sharpton, T. J., Shenoy, N., Sheth, N. U., Simone, G. A., Singh, I., Smillie, C. S., Sobel, J. D., Sommer, D. D., Spicer, P., Sutton, G. G., Sykes, S. M., Tabbaa, D. G., Thiagarajan, M., Tomlinson, C. M., Torralba, M., Treangen, T. J., Truty, R. M., Vishnivetskaya, T. A., Walker, J., Wang, L., Wang, Z., Ward, D. V., Warren, W., Watson, M. A., Wellington, C., Wetterstrand, K. A., White, J. R., Wilczek-Boney, K., Wu, Y., Wylie, K. M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B. P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J. D., Birren, B. W., Gibbs, R. A., Highlander, S. K., Methé, B. A., Nelson, K. E., Petrosino, J. F., Weinstock, G. M., Wilson, R. K., White, O., & The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207–214.

[82] Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2), 226–232.

[83] Irbäck, A., Peterson, C., & Potthast, F. (1996). Evidence for Non-Random Hydrophobicity Structures in Protein Chains. arXiv:chem-ph/9512004.

[84] Jangir, P. K., Yang, Q., Shaw, L. P., Caballero, J. D., Ogunlana, L., Wheatley, R., Walsh, T., & MacLean, R. C. (2022). Pre-existing chromosomal polymorphisms in pathogenic E. coli potentiate the evolution of resistance to a last-resort antibiotic. *eLife*, 11, e78834. publisher: eLife Sciences Publications, Ltd.

[85] Jha, N., Kravitz, J., West-Roberts, J., Lu, C., Camargo, A. P., Roux, S., Cornman, A., & Hwang, Y. (2025). Gaia: An AI-enabled genomic context–aware platform for protein sequence annotation. *Science Advances*, 11(25), eadv5109.

[86] Jia, C., Wang, Z., Huang, C., Teng, L., Zhou, H., An, H., Liao, S., Liu, Y., Huang, L., Tang, B., & Yue, M. (2023). Mobilome-driven partitions of the resistome in *Salmonella*. *MSYSTEMS*, 8(6), e00883–23. number-of-pages: 16 publisher-place: Washington publisher: Amer Soc Microbiology Web of Science ID: WOS:001143818300042.

[87] Johansson, M. H. K., Bortolaia, V., Tansirichaiya, S., Aarestrup, F. M., Roberts, A. P., & Petersen, T. N. (2021). Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: Mobileelementfinder. *Journal of Antimicrobial Chemotherapy*, 76(1), 101–109.

[88] Johnson, M. S., Martsul, A., Kryazhimskiy, S., & Desai, M. M. (2019). Higher-fitness yeast genotypes are less robust to deleterious mutations. *Science*, 366(6464), 490–493. publisher: American Association for the Advancement of Science.

[89] Jordan, E. O. (1931). Staphylococcus food poisoning. *Journal of the American Medical Association*, 97(23), 1704–1707.

[90] Kadelka, C., Wheeler, M., Veliz-Cuba, A., Murrugarra, D., & Laubenbacher, R. (2023). Modularity of biological systems: a link between structure and function. *Journal of The Royal Society Interface*, 20(207), 20230505.

[91] Katoh, K., Misawa, K., Kuma, K.-i., & Miyata, T. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–3066. PMID: 12136088 PMCID: PMC135756.

[92] Katz, L., Griswold, T., Morrison, S., Caravas, J., Zhang, S., Bakker, H., Deng, X., & Carleton, H. (2019). Mashtree: a rapid comparison of whole genome sequence files. *Journal of Open Source Software*, 4(44), 1762.

[93] Khan, J. & Patro, R. (2021). Cuttlefish: Fast, parallel and low-memory compaction of de Bruijn graphs from large-scale genome collections. *Bioinformatics*, 37(Supplement_1), i177–i186.

[94] Kille, B., Nute, M. G., Huang, V., Kim, E., Phillippy, A. M., & Treangen, T. J. (2024). Parsnp 2.0: scalable core-genome alignment for massive microbial datasets. *Bioinformatics*, 40(5), btae311.

[95] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1 edition. DOI: 10.1017/CBO9780511623486.

[96] Koonin, E. V., Makarova, K. S., & Aravind, L. (2001). Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annual Review of Microbiology*, 55(1), 709–742.

[97] Koonin, E. V., Makarova, K. S., & Wolf, Y. I. (2021). Evolution of Microbial Genomics: Conceptual Shifts over a Quarter Century. *Trends in Microbiology*, 29(7), 582–592.

[98] Kuronen, J., Horsfield, S. T., Pöntinen, A. K., Mallawaarachchi, S., Arredondo-Alonso, S., Thorpe, H., Gladstone, R. A., Willems, R. J., Bentley, S. D., Croucher, N. J., Pensar, J., Lees, J. A., Tonkin-Hill, G., & Corander, J. (2024). Pangenome-spanning epistasis and coselection analysis via de Bruijn graphs. *Genome Research*, 34(7), 1081–1088.

[99] Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2), R12. PMID: 14759262 PMCID: PMC395750.

[100] Labbate, M., Case, R. J., & Stokes, H. W. (2009). *The Integron/Gene Cassette System: An Active Player in Bacterial Adaptation*, (pp. 103–125). Humana Press: Totowa, NJ. DOI: 10.1007/978-1-60327-853-9_6.

[101] Lambowitz, A. M. & Zimmerly, S. (2004). Mobile Group II Introns. *Annual Review of Genetics*, 38(Volume 38, 2004), 1–35. publisher: Annual Reviews.

[102] Larralde, M. (2022). Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, 7(72), 4296.

[103] Lawrence, J. G. & Roth, J. R. (1999). Genomic Flux: Genome Evolution by Gene Loss and Acquisition. In *Organization of the Prokaryotic Genome* chapter 15, (pp. 263–289). John Wiley & Sons, Ltd.

[104] Le Minor, L., Véron, M., & Popoff, M. (1982). [A proposal for Salmonella nomenclature]. *Annales De Microbiologie*, 133(2), 245–254. PMID: 7149525.

[105] Lee, M.-C. & Marx, C. J. (2012). Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations. *PLOS Genetics*, 8(5), e1002651. publisher: Public Library of Science.

[106] Lenski, R. E., Rose, M. R., Simpson, S. C., & Tadler, S. C. (1991). Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *The American Naturalist*, 138(6), 1315–1341. publisher: The University of Chicago Press.

[107] Letunic, I. & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, 52(W1), W78–W82.

[108] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100.

[109] Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760.

[110] Li, J., Zhang, H., Ning, J., Sajid, A., Cheng, G., Yuan, Z., & Hao, H. (2019). The nature and epidemiology of OqxAB, a multidrug efflux pump. *Antimicrobial Resistance & Infection Control*, 8(1), 44.

[111] Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., Garg, S., Groza, C., Guarracino, A., Harvey, W. T., Heumos, S., Howe, K., Jain, M., Lu, T.-Y., Markello, C., Martin, F. J., Mitchell, M. W., Munson, K. M., Mwaniki, M. N., Novak, A. M., Olsen, H. E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J. A., Sirén, J., Tomlinson, C., Villani, F., Vollger, M. R., Antonacci-Fulton, L. L., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Carroll, A., Chang, P.-C., Cody, S., Cook, D. E., Cook-Deegan, R. M., Cornejo, O. E., Diekhans, M., Ebert, P., Fairley, S., Fedrigo, O., Felsenfeld, A. L., Formenti, G., Frankish, A., Gao, Y., Garrison, N. A., Giron, C. G., Green, R. E., Haggerty, L., Hoekzema, K., Hourlier, T., Ji, H. P., Kenny, E. E., Koenig, B. A., Kolesnikov, A., Korbel, J. O., Kordosky, J., Koren, S., Lee, H., Lewis, A. P., Magalhães, H., Marco-Sola, S., Marijon, P., McCartney, A., McDaniel, J., Mountcastle, J., Nattestad, M., Nurk, S., Olson, N. D., Popejoy, A. B., Puiu, D., Rautiainen, M., Regier, A. A., Rhie, A., Sacco, S., Sanders, A. D., Schneider, V. A., Schultz, B. I., Shafin, K., Smith, M. W., Sofia, H. J., Abou Tayoun, A. N., Thibaud-Nissen, F., Tricomi, F. F., Wagner, J., Walenz, B., Wood, J. M. D., Zimin, A. V., Bourque, G., Chaisson, M. J. P., Flicek, P., Phillippy, A. M., Zook, J. M., Eichler, E. E., Haussler, D., Wang, T., Jarvis, E. D., Miga, K. H., Garrison, E., Marschall, T., Hall, I. M., Li, H., & Paten, B. (2023). A draft human pangenome reference. *Nature*, 617(7960), 312–324.

[112] Liu, B., Guo, H., Brudno, M., & Wang, Y. (2016). deBGA: Read alignment with de Bruijn graph-based seed and extension. *Bioinformatics*, 32(21), 3224–3232.

[113] Livermore, D. M. (2003). Bacterial Resistance: Origins, Epidemiology, and Impact. *Clinical Infectious Diseases*, 36(Supplement_1), S11–S23.

[114] Llarrull, L. I., Toth, M., Champion, M. M., & Mobashery, S. (2011). Activation of BlaR1 protein of methicillin-resistant Staphylococcus aureus, its proteolytic processing, and recovery from induction of resistance. *The Journal of Biological Chemistry*, 286(44), 38148–38158. PMID: 21896485 PMCID: PMC3207446.

[115] Louca, S., Mazel, F., Doebeli, M., & Parfrey, L. W. (2019). A census-based estimate of Earth's bacterial and archaeal diversity. *PLOS Biology*, 17(2), e3000106. publisher: Public Library of Science.

[116] Lynch, M. (2006). Streamlining and Simplification of Microbial Genome Architecture. *Annual Review of Microbiology*, 60(1), 327–349.

[117] Lynch, M. & Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences*, 112(51), 15690–15695. publisher: Proceedings of the National Academy of Sciences.

[118] Marin, M., Vargas, R., Harris, M., Jeffrey, B., Epperson, L. E., Durbin, D., Strong, M., Salfinger, M., Iqbal, Z., Akhundova, I., Vashakidze, S., Crudu, V., Rosenthal, A., & Farhat, M. R.

(2022). Benchmarking the empirical accuracy of short-read sequencing across the M. tuberculosis genome. *Bioinformatics (Oxford, England)*, 38(7), 1781–1787. PMID: 35020793 PMCID: PMC8963317.

[119] McKinney, W. (2010). : (pp. 56–61). Austin, Texas. [Online; accessed 2025-01-14].

[120] Merk, L. N., Jones, T. A., & Eddy, S. R. (2025). Presence of group II introns in phage genomes. *Nucleic Acids Research*, 53(15), gkaf761.

[121] Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, 17(10), 589–596. publisher: Elsevier PMID: 11585665.

[122] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1), D412–D419.

[123] Muggli, M. D., Bowe, A., Noyes, N. R., Morley, P. S., Belk, K. E., Raymond, R., Gagie, T., Puglisi, S. J., & Boucher, C. (2017). Succinct colored de Bruijn graphs. *Bioinformatics*, 33(20), 3181–3187.

[124] Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). Fubar: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Molecular Biology and Evolution*, 30(5), 1196–1205.

[125] Nagarajan, D., Nagarajan, T., Roy, N., Kulkarni, O., Ravichandran, S., Mishra, M., Chakravortty, D., & Chandra, N. (2018). Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *The Journal of Biological Chemistry*, 293(10), 3492–3509. PMID: 29259134 PMCID: PMC5846155.

[126] Narzisi, G., Corvelo, A., Arora, K., Bergmann, E. A., Shah, M., Musunuri, R., Emde, A.-K., Robine, N., Vacic, V., & Zody, M. C. (2018). Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Communications Biology*, 1(1), 20.

[127] Navarre, W. (2016). The Impact of Gene Silencing on Horizontal Gene Transfer and Bacterial Evolution. In *Advances in Microbial Physiology*, volume 69 (pp. 157–186). Elsevier.

[128] Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28(14), 1919–1920.

[129] Néron, B., Littner, E., Haudiquet, M., Perrin, A., Cury, J., & Rocha, E. P. C. (2022). Integronfinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in Klebsiella. *Microorganisms*, 10(4), 700. number: 4 publisher: Multidisciplinary Digital Publishing Institute.

[130] Njamkepo, E., Fawal, N., Tran-Dien, A., Hawkey, J., Strockbine, N., Jenkins, C., Talukder, K. A., Bercion, R., Kuleshov, K., Kolínská, R., Russell, J. E., Kaftyreva, L., Accou-Demartin, M., Karas, A., Vandenberg, O., Mather, A. E., Mason, C. J., Page, A. J., Ramamurthy, T., Bizet, C., Gamian, A., Carle, I., Sow, A. G., Bouchier, C., Wester, A. L., Lejay-Collin, M., Fonkoua, M.-C., Le Hello, S., Blaser, M. J., Jernberg, C., Ruckly, C., Mérens, A., Page, A.-L., Aslett, M., Roggentin, P., Fruth, A., Denamur, E., Venkatesan, M., Bercovier, H., Bodhidatta, L., Chiou, C.-S., Clermont, D., Colonna, B., Egorova, S., Pazhani, G. P., Ezernitchi, A. V., Guigon, G., Harris, S. R., Izumiya, H., Korzeniowska-Kowal, A., Lutyńska, A., Gouali, M., Grimont, F., Langendorf, C., Marejková, M., Peterson, L. A. M., Perez-Perez, G., Ngandjio, A., Podkolzin, A., Souche, E., Makarova, M., Shipulin, G. A., Ye, C., Žemličková, H., Herpay, M., Grimont, P. A. D., Parkhill, J., Sansonetti, P., Holt, K. E., Brisse, S., Thomson, N. R., & Weill, F.-X. (2016). Global phylogeography and evolutionary history of Shigella dysenteriae type 1. *Nature Microbiology*, 1, 16027. PMID: 27572446.

[131] Ohno, S. (1984). Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proceedings of the National Academy of Sciences*, 81(8), 2421–2425.

[132] Olson, H. C. & Hammer, B. W. (1933). The bacteriology of butter. *Agricultural Experiment Station Iowa State College of Agriculture and Mechanic Arts*, 159, 64.

[133] Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1), 132.

[134] Pasek, S., Risler, J.-L., & Brézellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics (Oxford, England)*, 22(12), 1418–1423. PMID: 16601004.

[135] Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., & Segata, N. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3), 649–662.e20.

[136] Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., Firth, A., & Karlin, D. (2018). Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLOS ONE*, 13(10), e0202513. publisher: Public Library of Science.

[137] Perron, G. G., Whyte, L., Turnbaugh, P. J., Goordial, J., Hanage, W. P., Dantas, G., & Desai, M. M. (2015). Functional Characterization of Bacteria Isolated from Ancient Arctic Soil Exposes Diverse Resistance Mechanisms to Modern Antibiotics. *PLOS ONE*, 10(3), e0069533. publisher: Public Library of Science.

[138] Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). Fasttree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3), e9490.

[139] Raeside, C., Gaffé, J., Deatherage, D. E., Tenaillon, O., Briska, A. M., Ptashkin, R. N., Cruveiller, S., Médigue, C., Lenski, R. E., Barrick, J. E., & Schneider, D. (2014). Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with Escherichia coli. *mBio*, 5(5), 10.1128/mbio.01377–14. publisher: American Society for Microbiology.

[140] Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). Macse: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE*, 6(9), e22594. publisher: Public Library of Science.

[141] Robertson, J. & Nash, J. H. E. (2018). Mob-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics*, 4(8). [Online; accessed 2022-08-25].

[142] Rocha, E. P. C., Cornet, E., & Michel, B. (2005). Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems. *PLOS Genetics*, 1(2), e15. publisher: Public Library of Science.

[143] Ruppé, E., Cherkaoui, A., Charretier, Y., Girard, M., Schicklin, S., Lazarevic, V., & Schrenzel, J. (2020). From genotype to antibiotic susceptibility phenotype in the order Enterobacterales: a clinical perspective. *Clinical Microbiology and Infection*, 26(5), 643.e1–643.e7.

[144] Sanchez-Herrero, J. F., Bernabeu, M., Prieto, A., Hüttener, M., & Juárez, A. (2020). Gene Duplications in the Genomes of Staphylococci and Enterococci. *Frontiers in Molecular Biosciences*, 7. publisher: Frontiers.

[145] Sánchez-Osuna, M., Cortés, P., Barbé, J., & Erill, I. (2019). Origin of the Mobile Di-Hydro-Pteroate Synthase Gene Determining Sulfonamide Resistance in Clinical Isolates. *Frontiers in Microbiology*, 9. [Online; accessed 2022-04-15].

[146] Schulz, T., Wittler, R., & Stoye, J. (2022). Sequence-based pangenomic core detection. *iScience*, 25(6), 104413.

[147] Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., & Goesmann, A. (2021). Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, 7(11), 000685.

[148] Seabold, S. & Perktold, J. (2010). : (pp. 92–96). Austin, Texas. [Online; accessed 2025-06-17].

[149] Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14), 2068–2069. PMID: 24642063.

[150] Sereika, M., Mussig, A. J., Jiang, C., Knudsen, K. S., Jensen, T. B. N., Petriglieri, F., Yang, Y., Jørgensen, V. R., Delogu, F., Sørensen, E. A., Nielsen, P. H., Singleton, C. M., Hugenholtz, P., & Albertsen, M. (2025). Genome-resolved long-read sequencing expands known microbial diversity across terrestrial habitats. *Nature Microbiology*, 10(8), 2018–2030. publisher: Nature Publishing Group.

[151] Shaheen, A., Ismat, F., Iqbal, M., Haque, A., De Zorzi, R., Mirza, O., Walz, T., & Rahman, M. (2015). Characterization of putative multidrug resistance transporters of the major facilitator-superfamily expressed in Salmonella Typhi. *Journal of Infection and Chemotherapy*, 21(5), 357–362.

[152] Shaheen, A., Ismat, F., Iqbal, M., Haque, A., Ul-Haq, Z., Mirza, O., De Zorzi, R., Walz, T., & Rahman, M. (2021). Characterization of the multidrug efflux transporter styMdtM from Salmonella enterica serovar Typhi. *Proteins*, 89(9), 1193–1204. PMID: 33983672 PMCID: PMC8338744.

[153] Sheikhizadeh, S., Schranz, M. E., Akdel, M., De Ridder, D., & Smit, S. (2016). PanTools: Representation, storage and exploration of pan-genomic data. *Bioinformatics*, 32(17), i487–i493.

[154] Sheppard, A. E., Stoesser, N., Wilson, D. J., Sebra, R., Kasarskis, A., Anson, L. W., Giess, A., Pankhurst, L. J., Vaughan, A., Grim, C. J., Cox, H. L., Yeh, A. J., the Modernising Medical Microbiology (MMM) Informatics Group, Sifri, C. D., Walker, A. S., Peto, T. E., Crook, D. W., & Mathers, A. J. (2016). Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. *Antimicrobial Agents and Chemotherapy*, 60(6), 3767–3778. publisher: American Society for Microbiology.

[155] Siew, N. & Fischer, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins*, 53(2), 241–251. PMID: 14517975.

[156] Smets, W., Moretti, S., Denys, S., & Lebeer, S. (2016). Airborne bacteria in the atmosphere: Presence, purpose, and potential. *Atmospheric Environment*, 139, 214–221.

[157] Starikova, E. V., Tikhonova, P. O., Prianichnikov, N. A., Rands, C. M., Zdobnov, E. M., Ilina, E. N., & Govorun, V. M. (2020). Phigaro: high-throughput prophage sequence annotation. *Bioinformatics*, 36(12), 3882–3884.

[158] Steinegger, M. & Söding, J. (2017). Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11), 1026–1028. publisher: Nature Publishing Group.

[159] Straume, D., Piechowiak, K. W., Olsen, S., Stamsås, G. A., Berg, K. H., Kjos, M., Heggenhougen, M. V., Alcorlo, M., Hermoso, J. A., & Håvarstein, L. S. (2020). Class A PBPs have a distinct and unique role in the construction of the pneumococcal cell wall. *Proceedings of the*

National Academy of Sciences, 117(11), 6129–6138. publisher: Proceedings of the National Academy of Sciences.

[160] Strauß, L., Stegger, M., Akpaka, P. E., Alabi, A., Breurec, S., Coombs, G., Egyir, B., Larsen, A. R., Laurent, F., Monecke, S., Peters, G., Skov, R., Strommenger, B., Vandenesch, F., Schaumburg, F., & Mellmann, A. (2017). Origin, evolution, and global transmission of community-acquired *Staphylococcus aureus* ST8. *Proceedings of the National Academy of Sciences*, 114(49).

[161] Suzuki, N., Inui, M., & Yukawa, H. (2008). Random genome deletion methods applicable to prokaryotes. *Applied Microbiology and Biotechnology*, 79(4), 519–526.

[162] The pandas development team (2024). pandas-dev/pandas: Pandas. DOI: 10.5281/ZENODO.3509134.

[163] Thompson, M. K., Keithly, M. E., Sulikowski, G. A., & Armstrong, R. N. (2015). Diversity in fosfomycin resistance proteins. *Perspectives in Science*, 4, 17–23.

[164] Tokuda, M. & Shintani, M. (2024). Microbial evolution through horizontal gene transfer by mobile genetic elements. *Microbial Biotechnology*, 17(1), e14408. publisher: John Wiley & Sons, Ltd.

[165] Toll-Riera, M., Millan, A. S., Wagner, A., & MacLean, R. C. (2016). The Genomic Basis of Evolutionary Innovation in Pseudomonas aeruginosa. *PLOS Genetics*, 12(5), e1006005. publisher: Public Library of Science.

[166] Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, 21(1), 180.

[167] Truong-Bolduc, Q. C., Dunman, P. M., Strahilevitz, J., Projan, S. J., & Hooper, D. C. (2005). Mgra is a multiple regulator of two new efflux pumps in Staphylococcus aureus. *Journal of Bacteriology*, 187(7), 2395–2405. PMID: 15774883 PMCID: PMC1065235.

[168] Tsang, K. K., Lam, M. M. C., Wick, R. R., Wyres, K. L., Bachman, M., Baker, S., Barry, K., Brisse, S., Campino, S., Chiaverini, A., Cirillo, D. M., Clark, T., Corander, J., Corbella, M., Cornacchia, A., Cuénod, A., D'Alterio, N., Di Marco, F., Donado-Godoy, P., Egli, A., Farzana, R., Feil, E. J., Fostervold, A., Gorrie, C. L., Hassan, B., Hetland, M. A. K., Hoa, L. N. M., Hoi, L. T., Howden, B., Ikhimiukor, O. O., Jenney, A. W. J., Kaspersen, H., Khokhar, F., Leangapichart, T., Ligowska-Marzęta, M., Löhr, I. H., Long, S. W., Mathers, A. J., McArthur, A. G., Nagaraj, G., Oaikhena, A. O., Okeke, I. N., Perdigão, J., Parikh, H., Pham, M. H., Pomilio, F., Raffelsberger, N., Rakotondrasoa, A., Kumar, K. L. R., Roberts, L. W., Rodrigues, C., Samuelsen, O., Sands, K., Sassera, D., Seth-Smith, H., Shamanna, V., Sherry, N. L.,

Sia, S., Spadar, A., Stoesser, N., Sunde, M., Sundsfjord, A., Thach, P. N., Thomson, N. R., Thorpe, H. A., Torok, M. E., Trang, V. D., Trung, N. V., Vornhagen, J., Walsh, T., Warne, B., Wilson, H., Wright, G. D., Holt, K. E., & KlebNET-GSP AMR Genotype-Phenotype Group (2024). Diversity, functional classification and genotyping of SHV β-lactamases in Klebsiella pneumoniae. *Microbial Genomics*, 10(10). [Online; accessed 2025-06-19].

[169] uz Zaman, M. H., D'Alton, S., Barrick, J. E., & Ochman, H. (2024). Promoter recruitment drives the emergence of proto-genes in a long-term evolution experiment with Escherichia coli. *PLOS Biology*, 22(5), e3002418. publisher: Public Library of Science.

[170] Warsi, O., Knopp, M., Surkov, S., Jerlström Hultqvist, J., & Andersson, D. I. (2020). Evolution of a New Function by Fusion between Phage DNA and a Bacterial Gene. *Molecular Biology and Evolution*, 37(5), 1329–1341.

[171] Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.

[172] Watson, A. K., Lopez, P., & Bapteste, E. (2022). Hundreds of Out-of-Frame Remodeled Gene Families in the Escherichia coli Pangenome. *Molecular Biology and Evolution*, 39(1), msab329.

[173] Weisman, C. M., Murray, A. W., & Eddy, S. R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLOS Biology*, 18(11), e3000862. publisher: Public Library of Science.

[174] Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S.-R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G. F., Rose, D. J., Zhou, S., Schwartz, D. C., Perna, N. T., Mobley, H. L. T., Donnenberg, M. S., & Blattner, F. R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences*, 99(26), 17020–17024. publisher: Proceedings of the National Academy of Sciences.

[175] Wiedenbeck, J. & Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews*, 35(5), 957–976.

[176] Will, W. R., Navarre, W. W., & Fang, F. C. (2015). Integrated circuits: How transcriptional silencing and counter-silencing facilitate bacterial evolution. *Current Opinion in Microbiology*, 23, 8–13.

[177] Winslow, C.-E. A., Kligler, I. J., & Rothberg, W. (1919). Studies on the classification of the colon-typhoid group of bacteria with special reference to their fermentative reactions. *Journal of Bacteriology*, 4(5), 429–503. publisher: American Society for Microbiology.

[178] Wittler, R. (2020). Alignment- and reference-free phylogenomics with colored de Bruijn graphs. *Algorithms for Molecular Biology*, 15(1), 4.

[179] Wolf, Y. I. & Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 35(9), 829–837. PMID: 23801028 PMCID: PMC3840695.

[180] Xu, S., Fu, Z., Zhou, Y., Liu, Y., Xu, X., & Wang, M. (2017). Mutations of the Transporter Proteins GlpT and UhpT Confer Fosfomycin Resistance in Staphylococcus aureus. *Frontiers in Microbiology*, 8, 914. PMID: 28579984 PMCID: PMC5437707.

[181] Yanai, I., Wolf, Y. I., & Koonin, E. V. (2002). Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biology*, 3(5), research0024.1.

[182] Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., Heidelberg, K. B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C. S., Li, H., Mashiyama, S. T., Joachimiak, M. P., Belle, C., Chandonia, J.-M., Soergel, D. A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B. J., Bafna, V., Friedman, R., Brenner, S. E., Godzik, A., Eisenberg, D., Dixon, J. E., Taylor, S. S., Strausberg, R. L., Frazier, M., & Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLOS Biology*, 5(3), e16. publisher: Public Library of Science.

[183] Zulkower, V. & Rosser, S. (2020). Dna Features Viewer: a sequence annotation formatting and plotting library for Python. *Bioinformatics*, 36(15), 4350–4352.