Atlas Journal of Biology 2025, pp. 846–855
https://doi.org/10.5147/ajb.262

Published in the United States of America
https://www.atlas-publishing.org

Atlas Journal of Biology - ISSN 2158-9151. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

Comparative Genomics of Foodborne Pathogens: Diversity, Virulence, and Epidemiological Relevance

Bradford Dreshawn[1], Khalid Lodhi[2], Jiazheng Yuan[2], Danielle Graham[2], Justin Graham[2], Mohamed Maldani[2], Erin White[2], Afua Arhin[3], and My Abdelmajid Kassem[1*]

[1] Plant Genomics and Bioinformatics Lab, Department of Biological and Forensic Sciences, Fayetteville State University, Fayetteville, NC 28301, USA; [2] Department of Biological and Forensic Sciences, Fayetteville State University, Fayetteville, NC 28301, USA; [3] School of Nursing, Fayetteville State University, Fayetteville, NC 28301, USA

Received: June 6, 2025 / Accepted: July 4, 2025

Abstract

Foodborne bacterial infections are a major global health concern, causing millions of illnesses and deaths annually. Advances in microbial genomics have improved pathogen characterization, yet the relationship between genomic traits and public health outcomes remains unclear. This study investigates 50 foodborne bacterial species by analyzing genome size, GC content, virulence gene count, and antimicrobial resistance (AMR) gene presence in relation to global infection rates and mortality. Our findings reveal substantial genomic diversity, with genome sizes ranging from 1.2 Mb to 9.0 Mb and virulence gene counts from 2 to 312. Genome size, gene number, and GC content are strongly correlated, but neither virulence nor AMR gene counts consistently predict mortality or global case numbers. These weak associations suggest that host susceptibility, ecological adaptation, and gene expression contribute significantly to pathogenicity. This study also highlights the value of microbial forensics in foodborne outbreak investigations. Integrating whole-genome sequencing (WGS), comparative genomics, and phylogenetic analysis allows for tracing pathogen origins during contamination events. Bacteria such as *Salmonella enterica*, *Escherichia coli*, and *Listeria monocytogenes* frequently feature in forensic cases due to their high public health impact. The use of machine learning (ML) and Artificial Intelligence (AI) enhanced genomic surveillance holds promise for improving pathogen source attribution and biosecurity. These results highlight the complexity of bacterial virulence and call for integrated approaches combining genomic, epidemiological, and forensic data. Future work should emphasize functional genomics, host-pathogen interactions, and predictive modeling to enhance foodborne disease prevention and outbreak response strategies.

Keywords: Microbial forensics, Foodborne pathogens, Genomic epidemiology, Antimicrobial resistance, Virulence factors, Whole-genome sequencing.

[*] Corresponding author: mkassem@uncfsu.edu

1. Introduction

Foodborne bacterial infections are a major global health concern, causing millions of illnesses and deaths annually. According to the World Health Organization (WHO), foodborne diseases affect 600 million people each year, leading to approximately 420,000 deaths, with the highest burden observed in low- and middle-income countries (WHO, 2025). These infections are commonly associated with contaminated food, water, improper hygiene practices, within children under five years old, immunocompromised individuals, and the elderly being the most vulnerable (Kirk et al., 2015). The economic impact is also severe, with billions of dollars lost annually due to healthcare costs, loss of productivity, and food recalls (Scallan et al., 2011). Despite decades of research, foodborne bacterial infections remain a persistent challenge due to the emergence of antimicrobial resistance (AMR), environmental adaptability, and complex pathogen-host interactions (Rocourt et al., 2003).

The field of microbial genomics has revolutionized our understanding of bacterial pathogens, providing valuable insights into the genetic basis of virulence, antimicrobial resistance, and epidemiological trends (Didelot et al., 2017). The advent of whole-genome sequencing (WGS) and comparative genomics has enabled researchers to characterize the genetic diversity of bacterial species, facilitating the identification of key virulence determinants, mobile genetic elements, and resistance mechanisms (Oniciuc et al., 2018; Saini et al., 2024). However, despite these advances, the relationship between genomic features and disease severity remains poorly understood. While some pathogens exhibit large, complex genomes with numerous virulence factors, others have streamlined genomes yet cause severe disease, suggesting that genomic complexity alone may not fully explain pathogenic potential (Merhej and Raoult, 2011).

Several studies have explored genome size, GC content, and gene composition as potential indicators of bacterial adaptability, virulence, and resistance capacity (Ochman and Davalos, 2006; Bobay and Ochman, 2017). Genome size has been linked to metabolic versatility and environmental adaptability, with larger genomes often associated with higher numbers of virulence and resistance genes (Toft and Andersson, 2010). However, this trend does not hold universally, as some highly virulent pathogens, such as Clostridium botulinum and Helicobacter pylori, have relatively small genomes, yet produce potent toxins or have evolved mechanisms to evade host immunity (Rossetto et al., 2014). Conversely, opportunistic pathogens like Pseudomonas aeruginosa and Burkholderia species have large genomes, allowing for adaptation to diverse environments but not necessarily higher virulence (Stover et al., 2000).

Another important genomic feature is GC content, which varies widely among bacteria and may influence DNA stability, mutation rates, and gene regulation (Hildebrand et al., 2010). Some studies suggest that higher GC content correlates with greater environmental persistence, as observed in soil-dwelling and free-living bacteria, whereas host-associated pathogens often exhibit lower GC content, possibly due to genome reduction and specialization (Moran, 2002). However, the direct impact of GC content on bacterial virulence and epidemiological success remains debated (Bentley and Parkhill, 2004).

In addition to genome size and GC content, virulence gene content and antimicrobial resistance (AMR) gene presence are critical factors influencing pathogenic potential. Bacteria encode a wide array of virulence factors, including toxins, adhesins, secretion systems, and immune evasion proteins, which collectively determine their ability to infect, survive, and proliferate within hosts (Ribet and Cossart, 2015). Similarly, AMR genes allow bacteria to withstand antibiotic treatment, complicating infection management and increasing mortality risks (Martínez et al., 2009). The global rise of antibiotic-resistant foodborne pathogens, such as multidrug-resistant Salmonella and extended-spectrum beta-lactamase (ESBL)-producing Escherichia coli, has been a growing concern in clinical and food safety settings (Djordjevic et al., 2024). However, while AMR genes contribute to treatment efficacy, their direct relationship with mortality rates is not well established, as some highly resistant bacteria cause mild infections, whereas others with few resistance genes can be highly lethal (Baker et al., 2018).

Given these uncertainties, this study aims to systematically analyze the genomic features of 50 foodborne bacterial species and assess their correlation with public health metrics, including mortality rate and annual global infection cases. By evaluating genome size, GC content, virulence gene count, and AMR gene presence, we seek to determine whether genomic complexity influences pathogenic severity. Understanding these relationships could improve risk assessment, surveillance, and intervention strategies for foodborne diseases. The findings from this study may also contribute to genome-based predictive models, helping identify high-risk pathogens and guiding public health policies on food safety and infectious disease control.

2. Materials and Methods

2.1. Data Collection and Selection Criteria

The genomic and epidemiological data for this study were obtained from well-established public databases. Bacterial genome sequences were retrieved from NCBI and RefSeq databases (O'Leary et al., 2016; Sayers et al., 2024). In parallel, epidemiological data, including global infection cases and mortality rates, were collected from the Global Burden of Disease (GBD) study and World Health Organization (WHO) reports. These sources were chosen for their comprehensive surveillance of foodborne diseases across different geographical regions and population groups.

To ensure the reliability of our dataset, bacterial species were included only if they had well-documented genomic data and available epidemiological statistics related to human infections. Species with incomplete genome sequences, low-quality assemblies, or insufficient epidemiological records were excluded from the study.

2.2. Genomic Feature Analysis

To explore the genetic diversity of foodborne bacteria, several key genomic features were extracted from assembled genome sequences. Genome size (Mb), gene number, and GC content (%) were directly obtained from RefSeq annotations. These fundamental metrics provide insights into bacterial genome organization, metabolic potential, and evolutionary adaptation.

To assess virulence potential, the number of virulence genes in each bacterial species was identified using the Virulence Factor Database (VFDB) (Chen et al., 2016). This database contains curated information on known bacterial virulence determinants, including toxins, adhesion proteins, and immune evasion mechanisms. The presence of antimicrobial resistance (AMR) genes was determined using ResFinder and the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2020). These databases enable the detection of genetic determinants associated with antibiotic resistance, including genes encoding beta-lactamases, efflux pumps, and ribosomal modifications.

To ensure data accuracy, we applied quality control filters during data extraction. Only complete or high-quality draft genomes were included, and duplicate entries and plasmid-only assemblies were excluded. Virulence and AMR genes were identified using standardized thresholds of >90% nucleotide identity and >80% coverage to minimize false positives and ensure consistency across species. The wide ranges observed in virulence and AMR gene counts reflect true biological variation among species and are not due to redundancy or assembly artifacts. Cases with zero gene counts were verified as true negatives, typically occurring in species with reduced genomes (e.g., Mycoplasma spp.), rather than data omissions.

2.3. Statistical Data Analysis

Descriptive statistics were calculated to summarize the genomic and epidemiological characteristics of the dataset. For each genomic feature, mean, standard deviation, quartiles, and range were computed to assess the overall distribution and variability among bacterial species. All data analyses have been conducted using Jupyter Notebook (Kluyver et al., 2016) with Python (van Rossum and Drake, 2009) and its libraries: NumPy (Harris et al., 2020) for numerical operations, Pandas (McKinney, 2010) for data handling, Matplotlib (Hunter, 2007) for plotting, and Seaborn (Waskom, 2021) for enhanced visualization.

To evaluate potential relationships between genomic traits and public health impact, Pearson and Spearman correlation coefficients were used. Pearson correlation was applied to assess linear relationships between continuous variables, while Spearman correlation was used to capture potential non-linear associations. These analyses aimed to determine whether genome size, gene number, GC content, virulence gene count, or AMR gene presence were predictive of mortality rate and global infection cases.

To facilitate interpretation, multiple data visualization techniques were employed. Histograms and boxplots were generated to illustrate distribution patterns and outliers in genomic features, while heatmaps were used to depict correlation matrices. Additionally, scatterplots were created to examine potential associations between genomic complexity and pathogenic severity. These visual representations provided an intuitive means of identifying trends, clusters, and potential outliers within the dataset.

3. Results

3.1. Genomic Diversity Among Foodborne Bacteria

The comprehensive dataset summarizing the species, family, genome size (Mb), gene number, GC content (%), virulence gene count, antimicrobial resistance (AMR) gene count, annual global infection cases, and reported mortality rates (%) for 50 foodborne bacterial species analyzed in this study are shown in Table 1. The summary statistics of the 50 foodborne bacteria (Table 2) reveal substantial variability in genome size, gene number, GC content, virulence factors, AMR genes, and their associated public health impact. The average genome size is 3.97 Mb, with a wide range (1.2–9.0 Mb), mirroring the diversity in gene number (0–18,000, mean 3,847). GC content varies from 27% to 68% (mean 44.28%), indicating different evolutionary adaptations. The number of virulence genes also shows high dispersion (2–312, mean 47.4), as does AMR gene count (0–7,000, mean 152.4), suggesting that some pathogens possess extensive resistance mechanisms while others have none (Table 1). The distribution of annual global cases (1–2.8 billion, median 141,000) and mortality rates (0.1–93%) highlights the uneven burden of these bacteria, with a few species causing massive outbreaks (Table 1). The high standard deviations across most variables emphasize the heterogeneity in genome characteristics and public

Table 1. . Genomic, Virulence, Antimicrobial Resistance, and Epidemiological Profiles of 50 Foodborne Bacterial Species. Comprehensive dataset summarizing the species, family, genome size (Mb), gene number, GC content (%), virulence gene count, antimicrobial resistance (AMR) gene count, annual global infection cases, and reported mortality rates (%) for 50 foodborne bacterial species analyzed in this study. This detailed table supports comparative analyses of genomic features and their potential associations with public health impact and microbial forensic investigations.

Species	Family	Genome Size (Mb)	Gene Number	GC Cont. (%)	Vir Genes Number	AMR Genes	Annual Cases Worldwide	Mortality Rate (%)
Salmonella enterica	*Enterobacteriaceae*	4.6	4600	52	50	10	78000000	1
Listeria monocytogenes	*Listeriaceae*	3	3000	38	30	5	23000000	30
Escherichia coli	*Enterobacteriaceae*	5	4700	50	40	10	2000000	5
Shigella_sonnei	*Enterobacteriaceae*	4.8	4500	51	35	5	80000000	0.1
Clostridium botulinum	*Clostridiaceae*	3.9	3600	28	20	3	1000	70
Klebsiella pneumoniae	*Enterobacteriaceae*	5.5	5500	57	60	7	790000	41
Proteus mirabilis	*Morganellaceae*	4.1	3700	39	9	3	1,500,000	9
Enterobacter cloacae	*Enterobacteriaceae*	5.6	4,000	55	4	10	5,600,000	30
Serratia marcescens	*Enterobacteriaceae*	5	5,000	58	21	10	4,000	30
Citrobacter freundii	*Enterobacteriaceae*	5	5,000	52	12	7	290,000	10
Yersinia enterocolitica	*Yersiniaceae*	4.6	4,000	47	57	10	117,000	34
Yersinia pseudotuberculosis	*Yersiniaceae*	4.6	260	48	67	8	90,000	5
Edwardsiella tarda	*Hafniaceae*	3.8	3,700	57	16	9	10,000	8.6
Morganella morganii	*Morganellaceae*	4	3,500	51	30	19	73,000	21
Pseudomonas aeruginosa	*Pseudomonadaceae*	7	5,000	67	20	6	1,000,000	18
Burkholderia cepacia	*Burkholderiaceae*	9	8,100	67		7,000	1	9
Burkholderia pseudomallei	*Burkholderiaceae*	7.4	6,500	68	146	100	165,000	10
Legionella pneumophila	*Legionellaceae*	4.8	3,000	38	300	1	18,000	10
Aeromonas hydrophila	*Aeromonadaceae*	5.5	18,000	60	312	8	40,000	15
Vibrio cholerae	*Vibrionaceae*	4	3,900	47	10	40	2,000,000	50
Vibrio parahaemolyticus	*Vibrionaceae*	5.2	4,400	53	300	5	200,000	18
Vibrio_vulnificus	*Vibrionaceae*	7.5	4,000	46	160	7	1,000	33
Campylobacter jejuni	*Campylobacteraceae*	1.8	1,700	30	158	30	1,500,000	1
Campylobacter coli	*Campylobacteraceae*	1.8	2,000	31	11	12	1,500,000	1
Helicobacter pylori	*Helicobacteraceae*	1.7	1,200	39	6	42	1,200,000	4
Bacillus cereus	*Bacillaceae*	6.4	6,800	44	11	13	60,000	1
Bacillus anthracis	*Bacillaceae*	5.3	5,000	37	2	10	100,000	20
Bacillus thuringiensis	*Bacillaceae*	5.5	6,000	39	8	100	10	93
Clostridium difficile	*Clostridiaceae*	4.3	4,000	29	2	6	1,500,000	6
Clostridium perfringens	*Clostridiaceae*	3.3	2,500	30	20	7	1,000,000	30
Clostridium tetani	*Clostridiaceae*	2.8	2,700	28	100	3	1,000,000	10
Clostridium septicum	*Clostridiaceae*	3.4	3,400	27		13	1,000	50
Staphylococcus aureus	*Staphylococcaceae*	2.8	2,000	33	9	20	1,600,000	10
Staphylococcus epidermidis	*Staphylococcaceae*	2.5	2,500	32	63	8	1,000	40
Staphylococcus saprophyticus	*Staphylococcaceae*	2.5	2,500	33	3	5	1,500,000	1
Listeria_ivanovii	*Listeriaceae*	2.9	2,800	38	4	1	100	20
Streptococcus pyogenes	*Streptococcaceae*	1.9	1,500	39	46	6	18,000,000	10
Streptococcus agalactiae	*Streptococcaceae*	2	2,000	35	13	7	28,000	8
Streptococcus pneumoniae	*Streptococcaceae*	2	2,500	39	36	5	500	5
Streptococcus mutans	*Streptococcaceae*	2	2,000	36	9	4	2,800,000,000	5
Streptococcus suis	*Streptococcaceae*	2	8,000	40	5	6	240,000	7
Enterococcus faecalis	*Enterococcaceae*	3.6	3,000	37	10	10	100,000	10
Enterococcus faecium	*Enterococcaceae*	3.2	0	38	7	10	50,000	13
Mycobacterium tuberculosis	*Mycobacteriaceae*	4.4	4,000	65	15	10	9,000	50
Mycobacterium avium	*Mycobacteriaceae*	5	4,000	68	7	3	200,000	25
Francisella tularensis	*Francisellaceae*	1.8	1,800	32	6	2	205	5
Brucella abortus	*Brucellaceae*	3.3	4,200	57	8	0	100	1
Brucella melitensis	*Brucellaceae*	3.3	3,000	57	10	4	500,000	2
Coxiella burnetii	*Coxiellaceae*	1.9	2,000	42	2	0	1,000	1
Rickettsia rickettsii	*Rickettsiaceae*	1.2	1,300	30	5	0	250	5

Table 2. Summary statistics of genome size, gene number, GC content, virulence genes, AMR genes, annual cases worldwide, and mortality rates among 50 foodborne bacteria.

	Genome Size (Mb)	Gene Number	GC Content (%)	Virulence Genes	AMR Genes	Annual Cases World	Mortality Rate (%)
Count	50.00	50.00	50.00	48.00	50.00	50.00	50.00
Mean	3.97	3847.20	44.28	47.40	152.40	60499803.32	17.83
Std	1.73	2679.74	12.06	77.30	988.36	395649413.95	19.40
Min	1.20	0.00	27.00	2.00	0.00	1.00	0.10
25%	2.58	2500.00	35.25	7.75	5.00	5250.00	5.00
50%	3.95	3650.00	39.50	14.00	7.00	141000.00	10.00
75%	5.00	4575.00	52.75	47.00	10.00	1500000.00	28.75
Max	9.00	18000.00	68.00	312.00	7000.00	2800000000.00	93.00

Atlas Journal of Biology - ISSN 2158-9151. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

health risks associated with these pathogens.

The genomic features of the 50 foodborne bacterial species exhibit substantial variability, as seen in the distribution plots and boxplots. Genome size ranges from 1.2 Mb to 9.0 Mb (mean: 3.97 Mb), with most bacteria clustering around 2–5 Mb (Figure 1). The gene count follows a similar trend, varying from 0 to 18,000 genes (mean: 3,847), with a few species having exceptionally large genomes and gene content (Figure 2). GC content shows a wide distribution (27%–68%) (Figure 3), indicating diverse evolutionary strategies across species. Boxplots reveal that some species exhibit outliers in genome size, gene number, and GC content (Figure 4), suggesting the influence of horizontal gene transfer, environmental adaptation, or pathogenic specialization.

A barplot ranking bacterial species by genome size (Figure 5) highlights Burkholderia cepacia, Vibrio vulnificus, and Burkholderia pseudomallei as the species with the largest genomes (>7 Mb). In contrast, Rickettsia rickettsii, Helicobacter pylori, and Campylobacter jejuni have the smallest genomes (<2 Mb), suggesting a more specialized or host-dependent lifestyle.

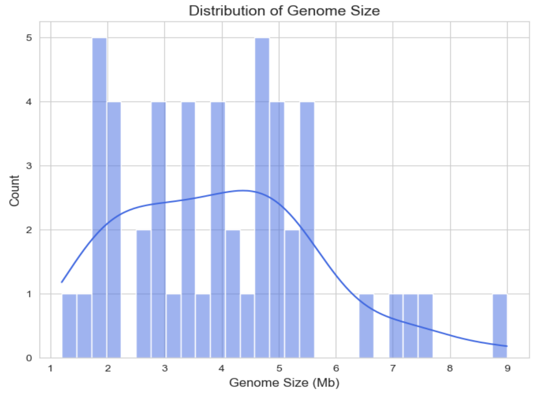

Figure 1. Distribution of genome size among 50 foodborne bacteria species.

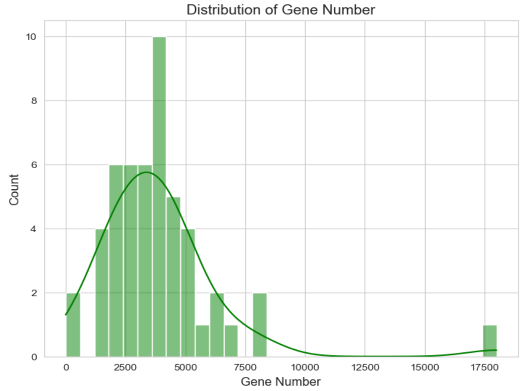

Figure 2. Distribution of gene number among 50 foodborne bacteria species.

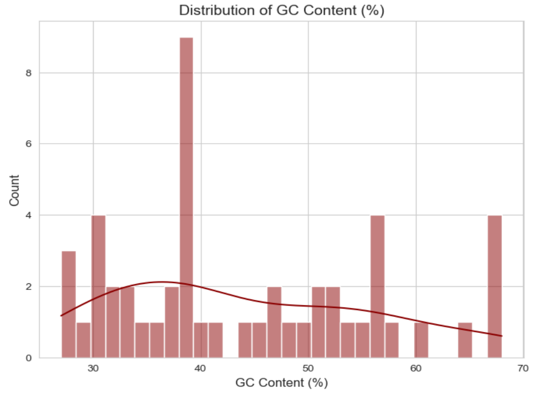

Figure 3. Distribution of GC content among 50 foodborne bacteria species.

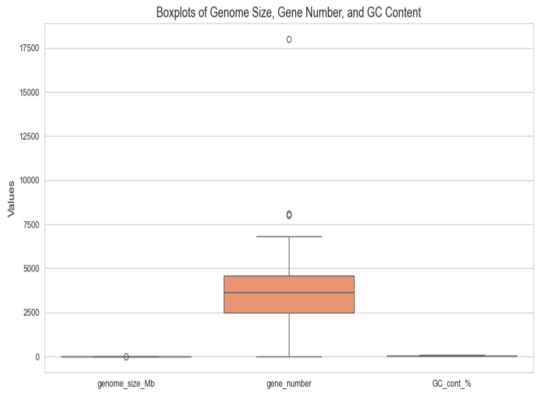

Figure 4. Boxplots of genome size, gene number, and GC content among 50 foodborne bacteria species.

3.2. Virulence and Antimicrobial Resistance (AMR) Gene Distributions

The number of virulence genes varies widely across species, ranging from 2 to 312 (mean: 47.4) (Figure 6). Most bacteria harbor relatively few virulence genes, but a few species possess over 100, suggesting enhanced pathogenic potential. The boxplot (Figure 7) confirms this trend, with several outliers exhibiting exceptionally high virulence gene counts.

Similarly, AMR gene distribution is highly skewed, with most bacteria containing few resistance genes (mean: 152.4), while a select few harbor thousands of AMR genes (Figure 8). The presence of extreme outliers in the boxplot (Figure 9) suggests that some species have undergone extensive resistance acquisition, likely due to antibiotic pressure or horizontal gene transfer.

3.3. Correlation Between Genomic Features and Public Health Impact

A heatmap (Figure 10) shows strong positive correlations between genome size, gene number (0.55), and GC content (0.67), indicating that larger genomes tend to have higher GC content and more genes. Virulence gene count exhibits moderate correlations with genome size (0.32) and gene number (0.42), suggesting that more complex genomes may contain more virulence factors. AMR gene count, annual cases, and mortality rates show weak or no correlation with genomic traits, implying that pathogenicity and public health burden are influenced by additional ecological or host factors.

3.4. Scatterplot Analysis of Mortality Rate vs. Genomic Traits

Scatterplots comparing mortality rates with genomic features (Figure 11) indicate that genome size, gene number, and GC content do not strongly correlate with mortality, suggesting that larger genomes or higher gene counts do not predict increased lethality. Virulence gene count shows a slight upward trend, with some species possessing >100 virulence genes exhibiting higher mortality; however, the overall pattern remains dispersed, indicating the influence of additional factors.

Similarly, AMR gene counts and annual global case numbers do not show clear relationships with mortality, as bacteria with high resistance gene counts or widespread prevalence often exhibit low mortality rates. These patterns emphasize the complexity of bacterial pathogenicity and the role of host, ecological, and environmental factors beyond genomic content.

3.5. Pairwise Comparisons of Genomic and Epidemiological Features

A comprehensive pairplot (Figure 12) visualizes relationships among numerical features across species. Genome size, gene number, and GC content exhibit strong correlations, consistent with observations in the correlation heatmap, reflecting genome structural relationships across foodborne bacteria. In contrast, relationships between genomic features and public health metrics

Atlas Journal of Biology - ISSN 2158-9151. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

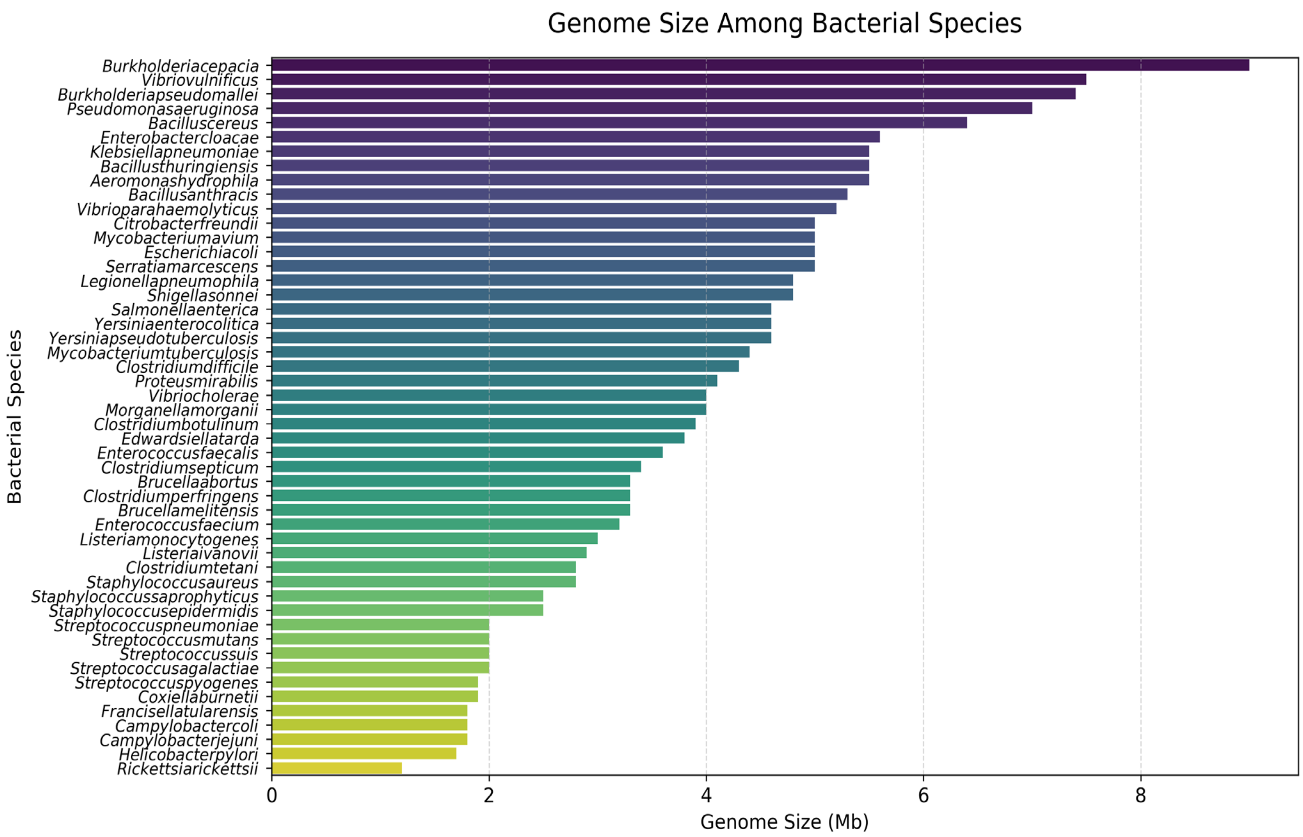

Figure 5. Ranked barplot of genome size by species among 50 foodborne bacteria species.

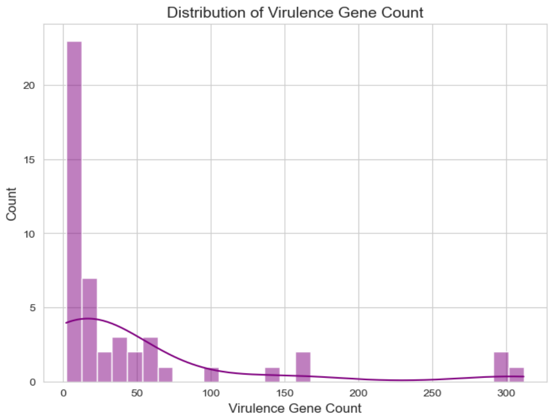

Figure 6. Distribution of virulence gene count among 50 foodborne bacteria species.

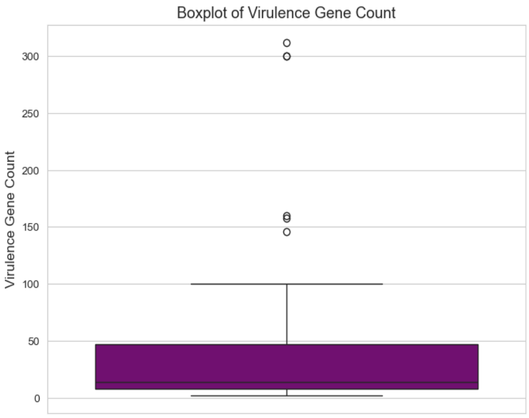

Figure 7. Boxplot of virulence gene count among 50 foodborne bacteria species.

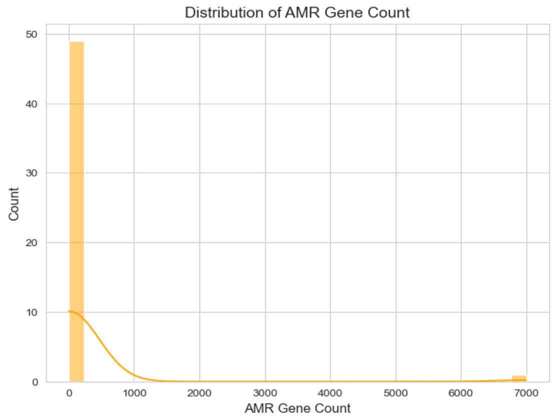

Figure 8. Distribution of AMR gene count among 50 foodborne bacteria species.

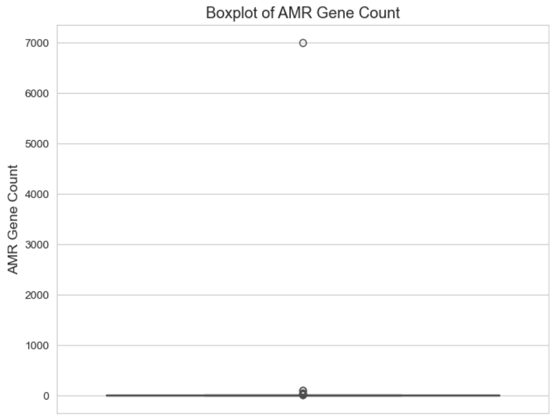

Figure 9. Boxplot of AMR gene count among 50 foodborne bacteria species.

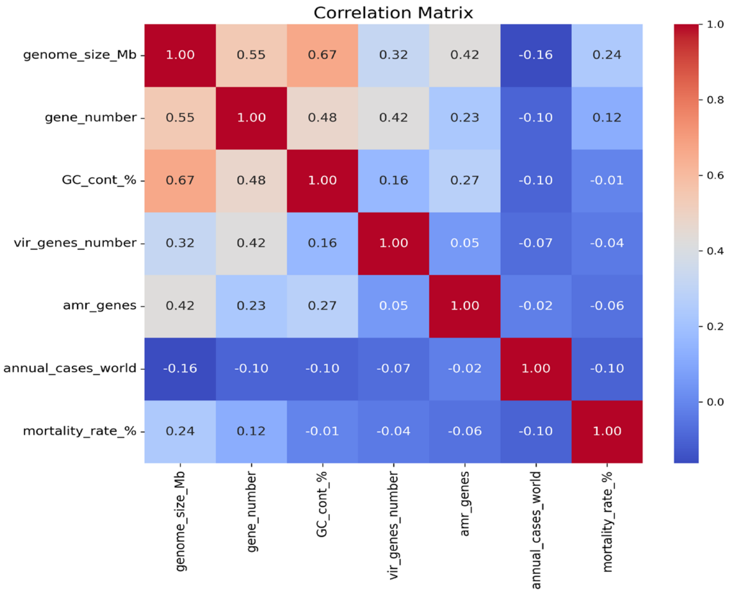

Figure 10. Correlation heatmap of genomic features among 50 foodborne bacteria species.

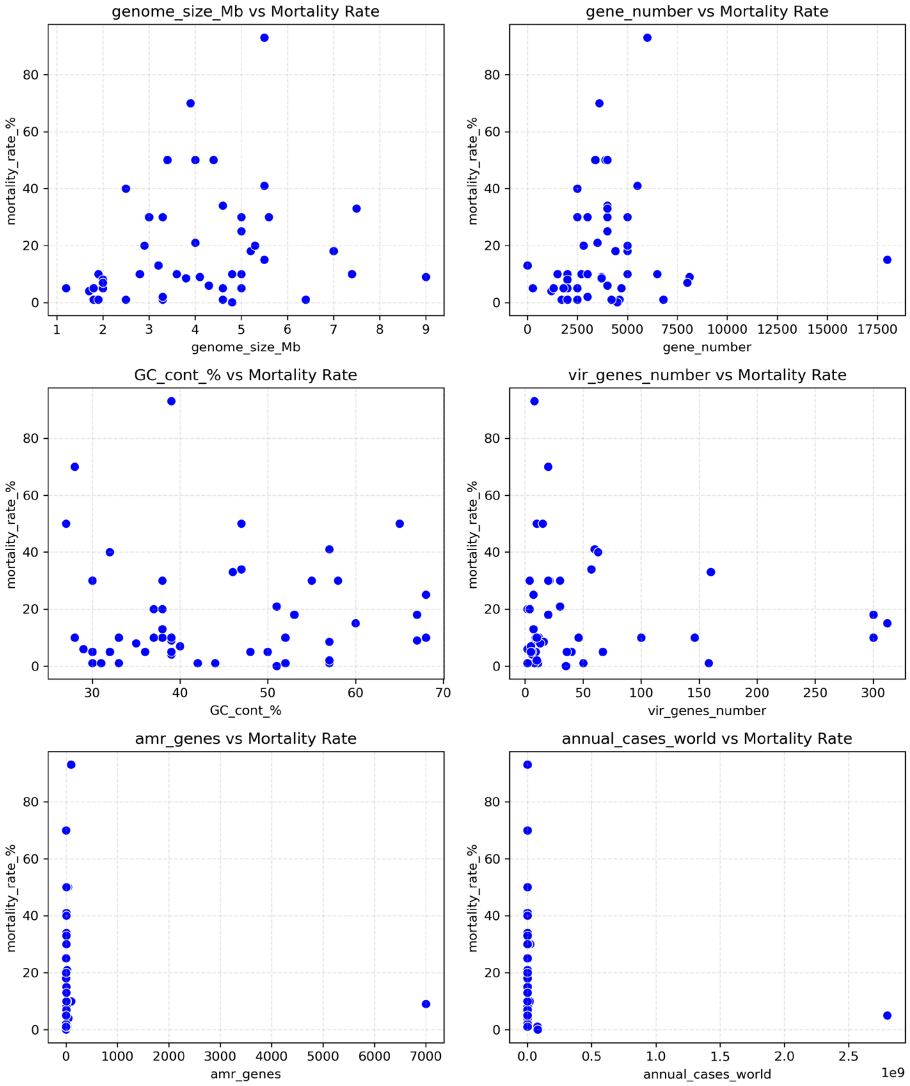

Figure 11. Scatterplots of mortality rate vs. genomic traits among 50 foodborne bacteria species.

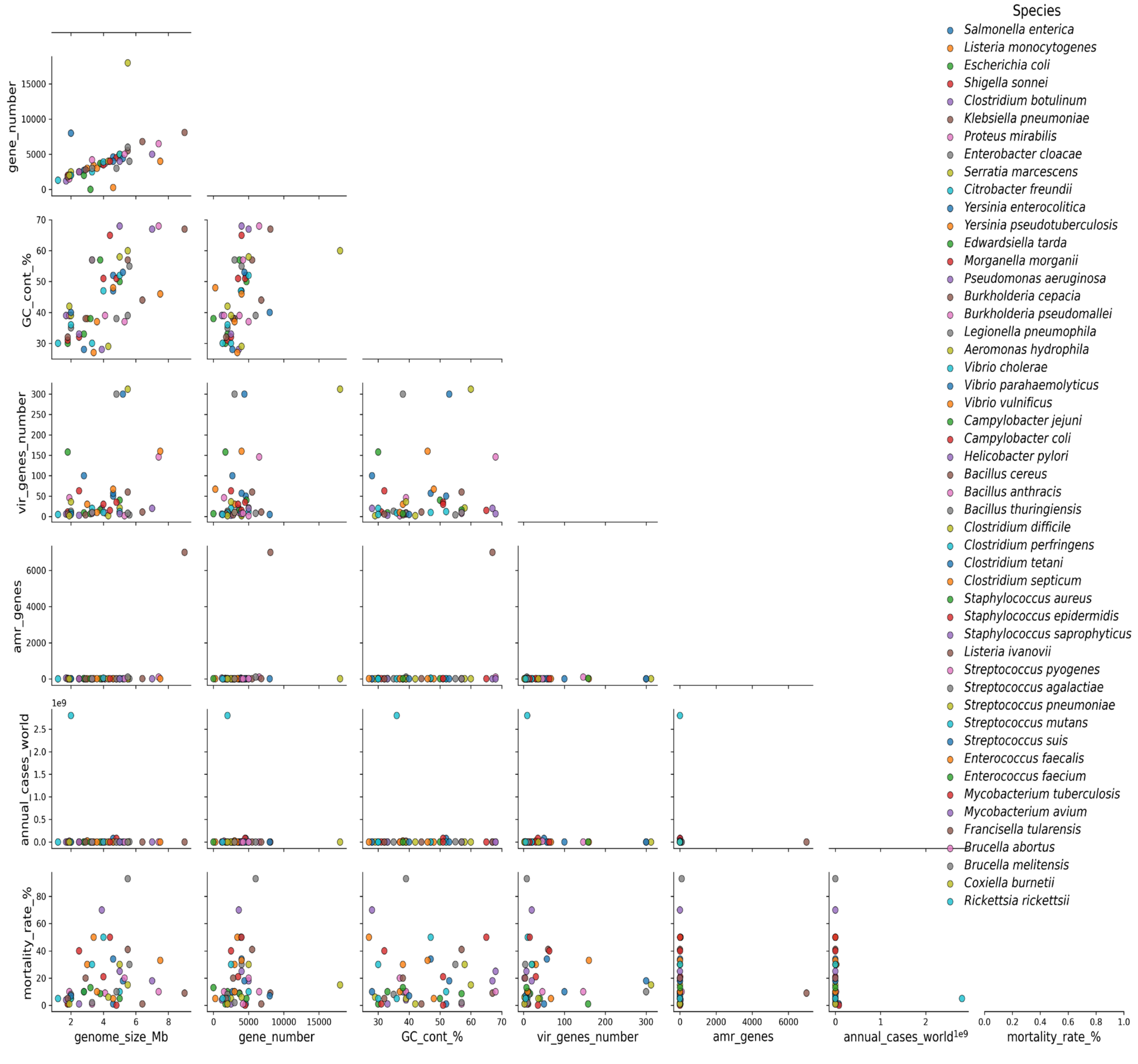

Figure 12. Pairplot of all numerical features among 50 foodborne bacteria species.

(mortality rates, annual cases) remain weak or absent, highlighting that genome characteristics alone are insufficient to predict disease severity or burden.
Virulence and AMR gene counts display skewed distributions, with a few species harboring high counts while most remain low, without clear clustering patterns based on pathogenic severity. These findings underscore the multifactorial nature of foodborne bacterial pathogenicity and the necessity of integrating genomic, epidemiological, and ecological data to improve risk assessment, surveillance, and disease prevention strategies.

4. Discussion

This study provides a comprehensive analysis of the genomic diversity, virulence potential, and antimicrobial resistance (AMR) profiles of 50 foodborne bacterial species, linking these features to their public health impact. Our findings reveal substantial variability in genome size, gene content, virulence factors, and resistance genes, but weak correlations between genomic traits and epidemiological severity. These results highlight the multifactorial nature of bacterial pathogenicity, emphasizing the need for integrated genomic, clinical, ecological, and host-related assessments to better predict and mitigate foodborne disease risks (Pightling et al., 2021; Hendriksen et al., 2019).

4.1. Genomic Complexity and its Relationship to Pathogenic Potential

Our results confirm that genome size is strongly correlated with gene number ($r = 0.55$) and GC content ($r = 0.67$), consistent with previous findings that larger bacterial genomes often encode more genes, allowing for greater metabolic versatility and adaptability (Ochman and Davalos, 2006). However, genome size alone does not dictate pathogenicity, as some highly virulent bacteria, such as Helicobacter pylori and Campylobacter jejuni, have small genomes (<2 Mb), while others with larger genomes, such as Burkholderia species, are opportunistic rather than obligate pathogens. This suggests that pathogenicity is more reliant on specific virulence factors than on genome size alone.
Furthermore, virulence gene count exhibits only moderate correlations with genome size ($r = 0.32$) and gene number ($r = 0.42$), reinforcing that pathogenic potential is not solely determined by genomic expansion. Some pathogens with relatively low gene counts, such as Clostridium botulinum, produce highly potent toxins, making them lethal despite their small genomes (Rossetto et al., 2014). Conversely, some species with large genomes, such as Bacillus cereus, are not always highly virulent, highlighting the role of regulatory mechanisms, environmental conditions, and host interactions in determining disease severity (Ehling-Schulz et al., 2019).

4.2. Antimicrobial Resistance and Public Health Risk

The distribution of AMR genes across bacterial species is highly skewed, with some species possessing thousands of resistance genes, while others have none. This variability aligns with prior studies showing that horizontal gene transfer (HGT) plays a crucial role in the acquisition of resistance determinants, particularly in species frequently exposed to antibiotic pressure (Frost et al., 2005; von Wintersdorff et al., 2016). Global surveillance efforts using metagenomics and WGS have highlighted the spread of AMR determinants across environmental and clinical settings, underscoring the need for integrated monitoring frameworks (Hendriksen et al., 2019; Collignon and McEwen, 2019).

Despite the presence of high AMR gene counts in some bacteria, no strong correlation was found between AMR genes and mortality rate, suggesting that antibiotic resistance alone does not directly predict disease severity. While resistance may contribute to treatment failure and prolonged infections, it does not necessarily increase intrinsic virulence (Martinez et al., 2009). Additionally, some highly resistant bacteria, such as Enterococcus faecium, typically cause chronic but non-lethal infections, whereas low-resistance pathogens like Listeria monocytogenes can be fatal, particularly in immunocompromised individuals (Radoshevich and Cossart, 2018).

4.3. Weak Correlations Between Genomic Features and Mortality Rate

A key finding of this study is the lack of strong correlations between genome size, gene number, GC content, virulence genes, and mortality rate. Scatterplots show that some highly lethal bacteria have relatively simple genomes, while others with large genomes and numerous virulence factors cause only mild infections. This reinforces the idea that mortality rate is shaped by a complex interplay of factors, including host susceptibility, transmission dynamics, and immune evasion strategies (Casadevall and Pirofski, 2018).

For instance, pathogens such as Vibrio vulnificus and Yersinia pseudotuberculosis exhibit high fatality rates despite moderate genome sizes, whereas Escherichia coli and Salmonella enterica, which have relatively large genomes, cause a spectrum of diseases ranging from mild gastroenteritis to severe systemic infections (Jolley and Maiden, 2010; Baker et al., 2018). This suggests that key virulence determinants, such as toxin production, immune evasion mechanisms, and host-pathogen interactions, may be better predictors of disease severity than genome size alone (Pightling et al., 2021).

4.5. Evolutionary and Ecological Influences on Pathogenicity

The pairwise comparison of genomic and epidemiological features indicates that certain bacterial species cluster together based on genome size and GC content, but not necessarily on virulence or resistance patterns. This aligns with previous research showing that bacterial genome evolution is driven by niche adaptation, environmental pressures, and host interactions rather than a simple accumulation of virulence factors (Merhej and Raoult, 2011).

For example, intracellular pathogens like Rickettsia and Mycobacterium exhibit genome reduction, reflecting their reliance on host cellular machinery, while environmentally persistent bacteria like Pseudomonas aeruginosa maintain large, flexible genomes to survive diverse conditions (Toft and Andersson, 2010). Machine learning and AI approaches are increasingly being used to analyze these complex genomic and ecological patterns, enhancing our understanding of pathogen evolution and spread (Jiang et al., 2022; Danko et al., 2021; Libbrecht and Noble, 2015). This suggests that pathogen success is dictated by ecological fitness rather than genome complexity alone, underscoring the need for contextual analysis of bacterial pathogenicity.

4.6. Implications for Food Safety and Disease Prevention

These findings have important implications for foodborne disease surveillance and risk assessment. Given the weak correlations between genomic traits and public health burden, a multifactorial approach incorporating genomic, epidemiological, and host-pathogen interaction data is essential for improving predictive models of foodborne disease risk (Scallan et al., 2011; Collignon and McEwen, 2019).

Genomic screening alone may not be sufficient to assess virulence risk; functional studies on toxin production, immune evasion, and transmission dynamics are critical. AMR monitoring should focus on clinically relevant resistance genes and their impact on treatment outcomes rather than total gene counts. Surveillance strategies should prioritize high-risk species with both high virulence and frequent foodborne transmission, such as Listeria monocytogenes, Vibrio vulnificus, and Salmonella enterica.

Future research should explore machine learning approaches to integrate genomic, epidemiological, and clinical data for more accurate pathogenicity risk assessments (Jiang et al., 2022; Libbrecht and Noble, 2015). Additionally, functional genomics studies could help determine which virulence and resistance factors are most predictive of severe disease outcomes.

4.7. Foodborne Bacteria and Their Role in Microbial Forensics

The genomic analysis of foodborne bacteria not only enhances our understanding of pathogenicity and public health risks but also plays a crucial role in microbial forensics—a field dedicated to identifying and tracking microbial agents in criminal, bioterrorism, and foodborne outbreak investigations. By leveraging whole-genome sequencing (WGS), phylogenetic analysis, and comparative genomics, microbial forensics can trace the origin, evolution, and transmission routes of foodborne pathogens, providing critical evidence in cases of food contamination, bioterrorism, and intentional adulteration (Thirunavukkarasu et al., 2018).

Today, advancements in microbial forensics enable real-time genomic surveillance, allowing authorities to rapidly identify specific bacterial strains, their virulence factors, and antimicrobial resistance genes, aiding in forensic investigations and outbreak control (Oliveira et al., 2024; Pightling et al., 2021). The integration of ML and AI models with WGS data can enhance source attribution, evolutionary tracking, and prediction of outbreak dynamics in microbial forensics (Jiang et al., 2022; Danko et al., 2021).

The bacterial species analyzed in this study, including Salmonella enterica, Escherichia coli, Listeria monocytogenes, and Vibrio cholerae, are among the most common culprits in foodborne outbreaks and forensic investigations (Todd, 2017; Baliyan et al., 2025). As global food supply chains become increasingly complex, the combination of pathogen genomics and forensic tools will be essential for ensuring food safety, tracing sources of contamination, and mitigating biosecurity threats.

4.8. Limitations and Future Directions

While this study provides valuable insights into the genomic and epidemiological diversity of foodborne bacteria, several limitations remain. Key factors such as host susceptibility, infection dose, and environmental conditions were not included, yet they play crucial roles in disease severity and transmission. Future research should incorporate functional genomics approaches to analyze how gene expression influences virulence and antimicrobial resistance, rather than relying solely on gene presence. Additionally, ML techniques could significantly improve predictive modeling by identifying genomic signatures linked to high-risk pathogens, enabling more precise risk assessment and targeted public health interventions.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflicts of interest.

Authors Contributions

Conceptualization: MAK
Data collection: DB
Formal Analysis: BD, MAK
Writing—original draft preparation: MAK, KL
Writing—review and editing: AA, DG, JG, JY, MD, KL, EW
Supervision: MAK
Project administration: MAK

References

American Academy of Microbiology. (2003). Microbial forensics: A scientific assessment (Colloquium held June 7–9, 2002, Burlington, Vermont). American Society for Microbiology. https://www.ncbi.nlm.nih.gov/books/NBK560476. https://doi.org/10.1128/AAMCol.7June.2002

Baker, S., N. Thomson, FX. Weill., and K. Holt (2018) Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens, Science, vol. 360, no. 6390, pp. 733-738. https://doi.org/10.1126/science.aar3777.

Baliyan, N., R. Kumari, A. Kumari, and A. Kumar (2025). Food Safety and Detection of Bacterial Foodborne Pathogens in India, HPn, 2025, pp 1–24. https://doi.org/10.1007/978-3-031-32047-7_7-1.

Bentley, S.D., and J. Parkhill (2004) Comparative genomic structure of prokaryotes, Annu. Rev. Genet. vol. 38, pp. 771-791. https://doi.org/10.1146/annurev.genet.38.072902.094318.

Bobay, L.M., H. Ochman (2017) The evolution of bacterial genome architecture. Front. Genet., vol. 8, 72. https://doi.org/10.3389/fgene.2017.00072.

Casadevall, A., and L. Pirofski (2018) What Is a Host? Attributes of Individual Susceptibility, Infect. Immun., vol. 86 no.2, pp. 1-12. https://doi.org/10.1128/iai.00636-17.

Chen, L., J. Yang, J. Yu, Z. Yao, L. Sun, Y. Shen, and Q. Jin (2016) VFDB: A reference database for bacterial virulence factors, NAR, vol. 33, no. 1, pp.325-328. https://doi.org/10.1093/nar/gki008.

Collignon PJ and SA McEwen (2019). One Health—Its Importance in Helping to Better Control Antimicrobial Resistance. Tropical Medicine and Infectious Disease. 2019; 4(1):22. https://doi.org/10.3390/tropicalmed4010022.

Danko D, D Bezdan, EE Afshin, S Ahsanuddin, C Bhattacharya, DJ Butler, KR Daisy, et al. (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell 184(13), 3376-3393.e17. https://doi.org/10.1016/j.cell.2021.05.002.

Didelot, X., R. Bowden, D.J. Wilson, T.E.A. Peto, and D.W. Crook (2017) Transforming clinical microbiology with bacterial genome sequencing, Nat. Rev. Genet., vol. 13, no. 9, pp. 601-612. https://doi.org/10.1038/nrg3226.

Djordjevic, S.P., V. M. Jarocki, T. Seemann, M. L. Cummins, A. E. Watt, and B. Drigo et al. (2024) Genomic surveillance for antimicrobial resistance — a One Health perspective, Nat. Rev. Genet., Vol. 25, pp. 142–157. https://doi.org/10.1038/s41576-023-00649-y.

Ehling-Schulz, M., Lereclus, D., and Koehler, T.M., The Bacillus cereus Group: Bacillus Species with Pathogenic Potential, Microbiol. Spectr., 2019, vol. 7, no. 3, pp. 1-35. https://doi.org/10.1128/microbiolspec.gpp3-0032-2018.

Frost, L., R. Leplae, A. O. Summers, and A. Toussaint (2005) Mobile genetic elements: the agents of open-source evolution, Nat. Rev. Microbiol., vol. 3, pp. 722–732. https://doi.org/10.1038/nrmicro1235.

Gogarten, J.F., S. Calvignac-Spencer, C. L. Nunn, M. Ulrich, N. Saiepour, and H. V. Nielsen et al (2020) Metabarcoding of eukaryotic parasite communities describes diverse parasite assemblages spanning the primate phylogeny, Mol. Ecol. Resour., vol. 20, pp. 204–215. https://doi.org/10.1111/1755-0998.13101.

Harris, C.R., K. J. Millman, S. T. van der Walt, R. Gommers, P. Virtanen, and D. Cournapeau et al (2020) Array programming with NumPy, Nature, vol. 585, no. 7825, pp. 357-362. https://doi.org/10.1038/s41586-020-2649-2.

Hendriksen RS, P Munk, P Njage, B van Bunnik, L McNally, O Lukjancenko, et al. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nature Communications, 10(1), 1124. https://doi.org/10.1038/s41467-019-08853-3.

Hildebrand, F., A. Meyer, and A. Eyre-Walker (2010) Evidence of selection upon genomic GC-content in bacteria, PLoS Genet., vol. 6, no. 9, pp. 1-9. https://doi.org/10.1371/journal.pgen.1001107.

Hunter, J.D., and A. Matplotlib (2007) 2D graphics environment, CiSE, vol. 9, no. 3, pp. 90-95. https://doi.org/10.1109/MCSE.2007.55.

Jiang Y, J Luo, D Huang, Y Liu, and DD Li (2022). Machine Learning Advances in Microbiology: A Review of Methods and Applications. Frontiers in Microbiology, 13, 2022. https://doi.org/10.3389/fmicb.2022.925454.

Jolley, K.A., and M. C. Maiden (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics, vol. 11, no. 1, pp. 595. https://doi.org/10.1186/1471-2105-11-595.

Kirk, M.D., S. M. Pires, R. E. Black, M. Caipo, J. A. Crump, B. Devleesschauwer et al (2015) WHO Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis, PLoS Med., vol. 12, no. 12, pp. e1001921. https://doi.org/10.1371/journal.pmed.1001921.

Libbrecht MW and WS Noble (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics 16, 321–332 (2015). https://doi.org/10.1038/nrg3920.

Martinez, J.L., (2009) Environmental pollution by antibiotics and by antibiotic resistance determinants, Environmental Pollution, vol.157, no. 11, pp. 2893-2902. https://doi.org/10.1016/j.envpol.2009.05.051.

McKinney, W., (2010) Data structures for statistical computing in Python, Proceedings of the 9th Python in Science Conference, 56-61. https://doi.org/10.25080/Majora-92bf1922-00a.

Merhej, V., D. and Raoult (2011) Rickettsial evolution in the light of comparative genomics, Biol. Rev., vol. 86, no. 2, pp. 379-405. https://doi.org/10.1111/j.1469-185X.2010.00151.x.

Moran, N.A., (2002) Microbial minimalism: Genome reduction in bacterial pathogens, Cell, vol. 108, no. 5, pp. 583-586. https://doi.org/10.1016/S0092-8674(02)00665-7.

O'Leary, N.A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, NAR, vol. 44, no. D1, pp. D733–D745. https://doi.org/10.1093/nar/gkv1189.

Ochman, H., and L.M., Davalos (2006) The nature and dynamics of bacterial genomes. Science, vol. 311, no. 5768, pp. 1730-1733. https://doi.org/10.1126/science.1119966.

Oliveira, M., K. Marszalek, M. Kowalski, A. Frolova, P. P. Labaj, W. Branicki et al (2024) Sequencing Technologies in Forensic Microbiology: Current Trends and Advancements, Forensic Sci. vol. 4, no. 4, pp. 523-545. https://doi.org/10.3390/forensicsci4040035.

Oniciuc, A.E., E. Likotrafiti, A. Alvarez-Molina, M. Prieto, J.A. Santos, and A. Alvarez-Ordóñez (2018) The Present and Future of Whole Genome Sequencing (WGS) and Whole Metagenome Sequencing (WMS) for Surveillance of Antimicrobial Resistant Microorganisms and Antimicrobial Resistance Genes across the Food Chain. Genes, vol. 9, no.5, pp.268. https://doi.org/10.3390/genes9050268.

Pightling AW, JB Pettengill, Y Luo, JD Baugher, H Rand, and E Strain (2021). Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. Frontiers in Microbiology, 12, 616970. https://doi.org/10.3389/fmicb.2021.616970.

Radoshevich, L., and P. Cossart (2018) Listeria monocytogenes: towards a complete picture of its physiology and pathogenesis, Nat. Rev. Microbiol., vol.16, pp. 32–46. https://doi.org/10.1038/nrmicro.2017.126.

Ribet, D., and P. Cosset (2015) How bacterial pathogens colonize their hosts and invade deeper tissues, Microbes and Infect., 2015, vol. 17, no. 3, pp. 173-183. https://doi.org/10.1016/j.micinf.2015.01.004.

Rocourt, J., G., K. Moy, Vierk, and J. Schlundt (2003) The present state of foodborne disease in OECD countries. World Health Organization, ISBN-92-4-159109-9.

Rossetto, O., M. Pirazzini, and C. Montecucco, (2014). Botulinum neurotoxins: Genetic, structural and mechanistic insights, Nat. Rev. Microbiol., vol. 12, no. 8, pp. 535-549. https://doi.org/10.1038/nrmicro3295.

Saini, P., V. Bandsode, A. Singh, S. K. Mendem, T. Semmler, M. Alam, and N. Ahmed (2024) Genomic insights into virulence, antimicrobial resistance, and adaptation acumen of Escherichia coli isolated from an urban environment, mBio, vol. 15, no. 3, pp. e03545-23. https://doi.org/10.1128/mbio.03545-23.

Sayers, E.W., Beck, J., Bolton, E.E., Brister, J.R., Chan, J., Comeau, D.C., et al., Database resources of the National Center for Biotechnology Information, Nucleic Acids Res., 2024, Jan 5; vol. 52, no. 1, pp. 33-43. https://doi.org/10.1093/nar/gkad1044.

Scallan, E., R. M. Hoekstra, F. J. Angulo, R. V. Tauxe, M. A. Widdowson, S. L. Roy, J. L. Jones., and M. P. Griffin (2011) Foodborne illness acquired in the United States—major pathogens, Emerging Infectious Diseases, vol. 17, no. 1, pp. 7-15. https://doi.org/10.3201/eid1701.P11101.

Stover, C.K., X. Q. Pham, A. L. Erwin., S. D. Mizoguchi, P. Warrener, M. J. Hickey et al (2000) Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen, Nature, vol. 406, pp. 959–964. https://doi.org/10.1038/35023079.

Thirunavukkarasu, N., E. Johnson, S. Pillai, D. Hodge, L. Stanker, T. Wentz et al (2018) Botulinum Neurotoxin Detection Methods for Public Health Response and Surveillance, Front. Bioeng. Biotechnol., Vol. 6, pp. 80. https://doi.org/10.3389/fbioe.2018.00080.

Todd, E.C.D. (2017) Foodborne Disease in the Middle East. In: Murad, S., Baydoun, E., Daghir, N. (eds) Water, Energy & Food Sustainability in the Middle East, Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-48920-9_17.

Toft, C., and S.G. E. Andersson (2010) Evolutionary microbial genomics: insights into bacterial host adaptation, Nat. Rev. Genet., vol. 2898, pp 1-11. https://doi.org/10.1038/nrg2798.

Torok, T.J., R. V. Tauxe, R. P. Wise, J. R. Livengood, R. Sokolow, S. Mauvais et al (1997) A Large Community Outbreak of Salmonellosis Caused by Intentional Contamination of Restaurant Salad Bars, JAMA., vol. 278, no. 5, pp. 389–395. https://doi.org/10.1001/jama.1997.03550050051033.

van Rossum, G., and F. L. Drake (2009) Python 3 Reference Manual. CreateSpace

von Wintersdorff, C.J.H., J., Penders, J.M., van Niekerk, N.D., Mills, S., Majumder, L.B., van Alphen, et al (2016) Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer, Front. Microbiol. vol. 7, pp.173. https://doi.org/10.3389/fmicb.2016.00173.

Waskom, M.L. (2021) Seaborn: Statistical data visualization, JOSS., vol. 6, no. 60, pp. 3021. https://doi.org/10.21105/joss.03021.

Atlas Journal of Biology - ISSN 2158-9151. Published By Atlas Publishing, LLC (www.atlas-publishing.org)