MANSOURA JOURNAL OF BIOLOGY
Official Journal of Faculty of Science, Mansoura University, Egypt
E-mail: scimag@mans.edu.eg ISSN: 2974-492X

Genomic Insights into *Enterobacteriaceae*: Pan-Genome Analysis, and Functional Profiling

***Magda M. Awad*[1], *Mohamed Abdelmoteleb*[1], *Yehia A. Osman*[1, *]**

[1]*Botany Department, Faculty of Science, Mansoura University. Elgomhouria St., Mansoura, 35516, Egypt)*
[*] ***Correspondence to:*** *Yehia A. Osman yaolazeik@mans.edu.eg - 01144746346)*

Received:1/2/2025
Accepted: 9/2/2025

Abstract: *Enterobacteriaceae* is a diverse family of Gram-negative bacteria that includes clinically significant pathogens such as *Enterobacter* spp., *Escherichia coli*, *Klebsiella pneumoniae*, *Salmonella enterica*, and *Shigella* spp. These bacteria exhibited remarkable genomic plasticity, facilitating adaptation to various environments, acquisition of antimicrobial resistance (AMR) genes, and the evolution of virulence traits. This study conducted a comprehensive pan-genome analysis of 22 *Enterobacteriaceae* strains to investigate their genetic diversity, with a focus on multidrug resistance (MDR) genes and functional classification using the Clusters of Orthologous Groups (COG) database. The pan-genome analysis revealed a completely open genomic structure, with no core genes shared among all strains. Instead, the accessory genome dominated, comprising genes associated with virulence, host adaptation, and antimicrobial resistance. MDR genes were exclusively found in the accessory genome, highlighting their variable distribution and potential for horizontal transfer. Functional annotation using the COG database showed a consistent distribution of genes related to essential cellular functions, with carbohydrate and amino acid metabolism being the most represented categories. However, variations in certain functional groups indicated strain-specific adaptations and genomic plasticity. These findings underscored the evolutionary dynamics of *Enterobacteriaceae*, where the accessory genome drives genetic diversity and adaptation. The exclusive presence of MDR genes in the accessory genome emphasized the role of horizontal gene transfer in antimicrobial resistance dissemination. Understanding these genomic patterns is crucial for tracking the emergence of resistant strains and developing effective therapeutic and public health strategies.

keywords: *Enterobacteriaceae*, pan-genome, accessory genome, Clusters of Orthologous Groups (COG), genomic diversity, functional annotation, bacterial evolution.

1.Introduction

Enterobacteriaceae comprises a diverse group of Gram-negative bacteria, including major human pathogens such as *Escherichia coli*, *Klebsiella*, *Salmonella*, and *Shigella*. These organisms are responsible for many infections, from urinary tract infections to life-threatening sepsis. A major concern with *Enterobacteriaceae* is the increasing prevalence of multidrug resistance (MDR), driven by the acquisition of antimicrobial resistance (AMR) genes through horizontal gene transfer (HGT). Mobile genetic elements, such as plasmids, transposons, and integrons, facilitated the dissemination of resistance determinants, including extended-spectrum β-lactamases (ESBLs) and carbapenemases, which severely limit treatment options [1].

The pan-genomic analysis provided a comprehensive framework for studying the genetic diversity of *Enterobacteriaceae*, revealing the distribution of core, accessory, and unique genes across multiple strains. This approach is advantageous in identifying genetic determinants associated with MDR, as resistance genes often reside within the accessory genome. Comparative analysis of the pan-genome allowed for the identification of key resistance genes, their genomic contexts, and potential mechanisms of dissemination [2].

A functional classification of genes using Clusters of Orthologous Groups (COG) analysis further enhanced our understanding of MDR by categorizing genes based on their predicted functions. COG analysis enabled the identification of genes involved in crucial cellular processes, such as membrane transport, stress response, and antibiotic resistance mechanisms. Specifically, genes classified under COG categories related to defense mechanisms (e.g., antibiotic efflux pumps and β-lactamases) and genetic information processing (e.g., recombination and repair systems) play a significant role in the evolution of MDR phenotypes. By integrating COG analysis with pan-genomic data, we could identify functional adaptations that contribute to the persistence and spread of MDR traits within *Enterobacteriaceae* [3].

In this study, we performed a pan-genomic analysis of multiple *Enterobacteriaceae* strains, focusing on genes linked to MDR and their functional classification using COG. By characterizing the core, accessory, and unique genomes, we aimed to uncover the genetic factors contributing to antimicrobial resistance. The findings provided valuable insights into the molecular mechanisms underlying MDR in *Enterobacteriaceae* and highlighted potential targets for novel therapeutic interventions.

2. Materials and methods

2.1. Strain Selection

A total of 22 strains representing diverse *Enterobacteriaceae* species, including *Enterobacter*, *Escherichia coli*, *Klebsiella*, *Salmonella*, and *Shigella* spp., were selected for this study (**Table 1**).

Genomic data were obtained from publicly accessible repositories, with a preference for complete and high-quality draft genome sequences. The selected databases included NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank, ensuring reliable and well-annotated genome assemblies [4].

2.2. Genome Annotation

Genome annotation for the 22 complete *Enterobacteriaceae* strains was carried out using the Prokaryotic Genome Annotation System (Prokka) v1.14.5. This tool facilitated the identification of various genomic features, including coding sequences (CDS), rRNA, tRNA, and other elements. The annotation process produced a GFF3 file for each genome, containing comprehensive details about the positions and characteristics of genes and other functional elements within the genomic sequences [5].

Table 1: List of Enterobacteriaceae Strains Used in This Study

Organism	Accession No.	Code
Enterobacter cloacae	NZ_CP053568.1	En1
Enterobacter cloacae	NZ_CP092042.1	En2
Enterobacter cloacae	CP109676.1	En3
Enterobacter cloacae	CP109679.1	En4
Enterobacter aerogenes	CP002824.1	En5
Enterobacter aerogenes	FO203355.1	En6
Escherichia coli	NZ_CP007136.1	En7
Escherichia coli	AE005174.2	En8
Escherichia coli	NC_012947.1	En9
Escherichia coli	NZ_CP010444.1	En10
Shigella flexneri	AE014073.1	En11
Shigella sonnei	CP037997.1	En12
Klebsiella pneumoniae	CP028915.1	En13
Klebsiella pneumoniae	NZ_CP052761.1	En14
Klebsiella pneumoniae	NZ_CP009461.1	En15
Klebsiella pneumoniae	NZ_CP136385.1	En16
Klebsiella pneumoniae	NZ_CP096810.1	En17
Klebsiella quasipneumoniae	NZ_CP066173.1	En18
Salmonella enterica	NZ_CP136141.1	En19
Salmonella enterica	NZ_CP060508.1	En20
Salmonella enterica	NZ_CP060512.1	En21
Salmonella enterica	NZ_CP060522.1	En22

2.3. Pan-genome Construction

The pan-genome analysis was conducted using Roary (v3.13.3), a high-throughput tool designed for rapid clustering of homologous genes from multiple bacterial genomes. The software processes annotated genome assemblies in GFF3 format and identifies orthologous genes by performing pairwise comparisons. A sequence identity threshold of 95% was set to define orthologous gene clusters, ensuring that only highly similar sequences were grouped. The resulting pan-genome was categorized into three components: **Core genome** (Genes shared by all analyzed strains, representing the conserved and essential functions within *Enterobacteriaceae*), **Accessory genome** (Genes present in two or more strains but not in all, contributing to

strain-specific adaptations such as antimicrobial resistance and virulence), and **Unique genome** (Genes found in only one strain, often associated with niche-specific traits or recent horizontal gene transfer events). This analysis provided insights into the genomic diversity among the 22 *Enterobacteriaceae* strains, highlighting the genetic elements linked to antimicrobial resistance and functional adaptations [6].

To explore the evolutionary relationships among *Enterobacteriaceae* strains, a heatmap was generated to visualize the gene presence-absence patterns derived from the pan-genome, highlighting both shared and strain-specific genes [7]. In addition, a phylogenetic tree was constructed using core genes, with sequence alignment performed via MUSCLE and tree construction using the neighbor-joining method in MEGAX. The tree's reliability was assessed with 100 bootstrap replicates. This combined approach provided a comprehensive view of the core genome conservation and the variability of accessory genes across the strains [8].

2.4. Functional Analysis of Genes

The genes identified in the pan-genome were functionally annotated through an extensive rpsBLASTp search against the NCBI Clusters of Orthologous Groups (COG) database (https://www.ncbi.nlm.nih.gov/research/cog/) [9]. This approach aimed to assign functional categories to the genes based on their sequence similarity to known orthologous groups. A stringent E-value threshold of $1\times10-5$ was applied during the rpsBLASTp search to ensure high specificity and minimize false-positive results. This cutoff allowed only significant matches to be included in the functional annotation, ensuring that the assigned functions accurately reflected the roles of the genes within the pan-genome [10].

To visualize the distribution of COG categories across the strains, a bar plot was generated that displayed the percentage of genes assigned to each COG category. This was done using ggplot2 in R, where each bar represented a specific COG category and its relative abundance in terms of the percentage of genes in that category for each strain. The bars were grouped by strain to show the variation in COG category distribution across the different strains [11].

3. Results

3.1. Pan-genome Construction

The pan-genome analysis of 22 *Enterobacteriaceae* genomes revealed a total of 23,343 predicted protein-coding genes, which were categorized into four groups: core (0 genes, present in 99-100% of strains), softcore (0 genes, present in 95-99%), shell (6,333 genes, present in 15-95%), and cloud (17,010 genes, present in less than 15%) (Figures 1A and 1B). The core genome contained essential genes critical for survival, while the softcore genome likely encompassed genes with adaptive functions specific to environmental or host conditions. The shell genome contributed to the species' adaptability and specialization, while the cloud genome, comprising strain-specific genes, likely included unique pathogenic traits or adaptations to environmental factors.

Figure 1C illustrated the dynamic nature of genomes. Initially, the rapid increase in total gene counts indicated the wide genetic diversity across the strains. However, the slowing rate of gene discovery and the relatively slower growth of conserved genes suggested that genomes are open and evolving, continuously shaped by processes such as gene duplication, gene loss, and horizontal gene transfer. These processes drove the diversification and adaptation of organisms, leading to the development of unique genetic profiles for each strain.

Figure 1D depicted the relationship between the number of genomes analyzed and the number of unique genes identified. It demonstrated that while the addition of more genomes typically resulted in the discovery of more unique genes, there was substantial variability in the rate of gene discovery. This variability reflected the dynamic nature of genomes, influenced by factors such as genome size, sampling bias, and the phylogenetic relationships of the analyzed strains. The findings emphasized the complexity of genome composition and the continuous expansion of genetic knowledge as more genomes are sampled.

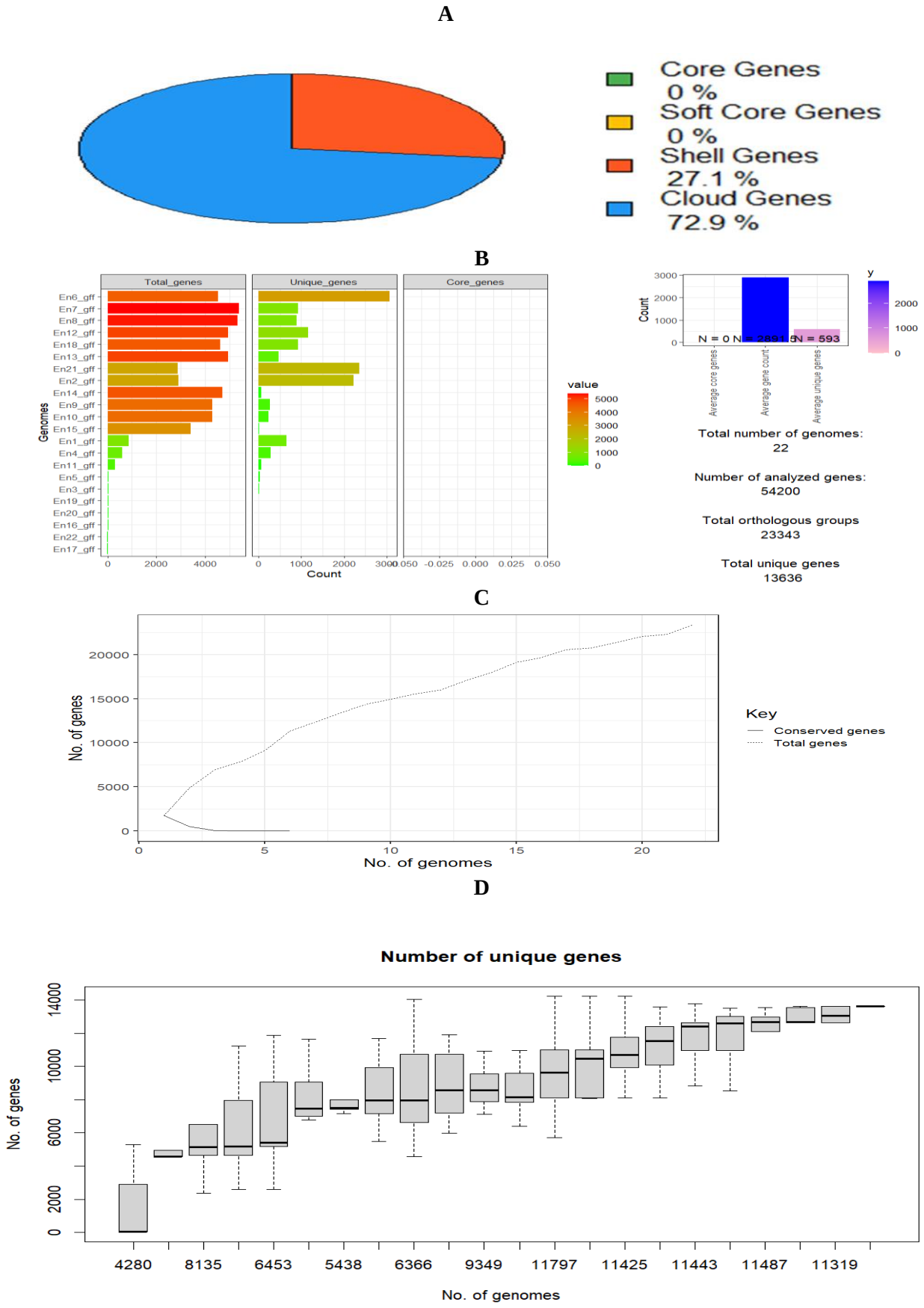


Figure 1. A. Pie chart depicting the distribution of gene categories within the *Enterobacteriaceae* pan-genome. **B.** Total gene counts, unique genes, core gene counts, and average gene counts across 22 members of *Enterobacteriaceae*. **C.** Ratio of conserved genes to total genes, illustrating the dynamic and open nature of the *Enterobacteriaceae* pan-genome. **D.** Ratio of unique genes to total genes, showcasing the degree of genetic diversity among the analyzed organisms.

The heatmap analysis revealed that no core genome genes were conserved across all 22 organisms of *Enterobacteriaceae*. Instead, the accessory genome, represented by a combination of green and purple, contained genes that conferred advantages such as virulence and antibiotic resistance, with variations across different strains. Strain-specific genes, marked in green in a few columns, highlighted the genetic diversity and suggested associations with traits like antibiotic resistance and pathogenicity. This emphasized the dynamic nature of the pan-genome, where essential functions are not universally shared, and adaptability is driven by the accessory genome (**Figure 2A**). The accessory gene tree revealed the evolutionary relationships among the strains, highlighting genetic divergence. Clusters in the tree indicated shared ancestry, while outliers suggested significant evolutionary differences, reflecting the diversity of accessory genes across strains (**Figure 2B**).

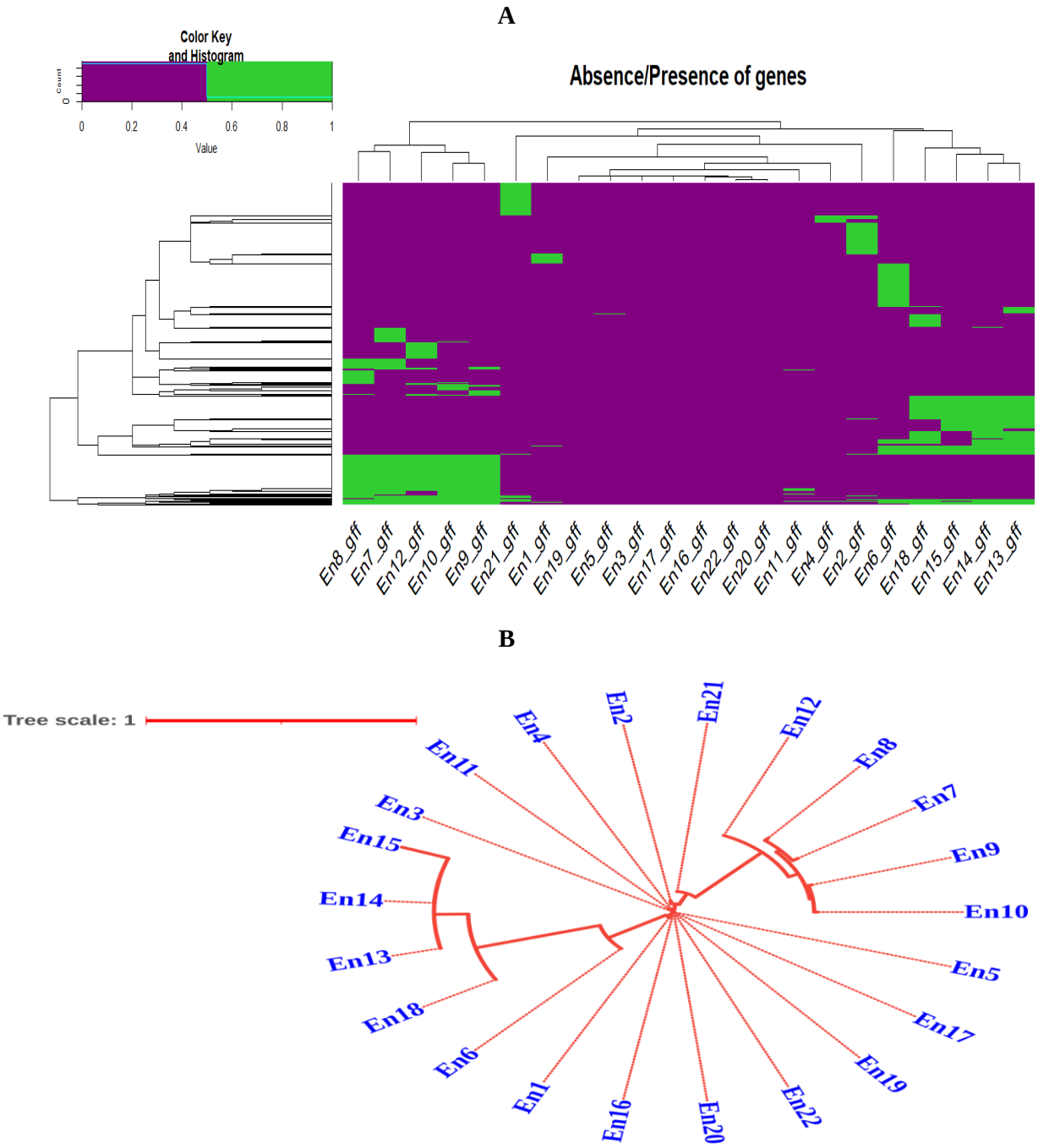


Figure 2. A) Accessory genes of *Enterobacteriaceae* members. The left tree illustrates strain clustering based on accessory genes, while the right matrix plot shows the presence (green) and absence (purple) of genes across all strains. **B)** Phylogenetic analysis of *Enterobacteriaceae* members based on accessory gene

3.2. Functional Analysis of Genes

To compare the functional distribution among the 22 strains, the categories of COG were analyzed. Generally, about 80% of the genes were annotated by the COG database, and the functional distribution within each strain was similar across most categories. However, there were variations in the proportions of sequences classified into different COG categories. The highest classification rate was observed in En22 with 100% (26/26) of sequences, while the lowest was found in En19 with 55.56% (20 / 36) of sequences (**Table 2**).

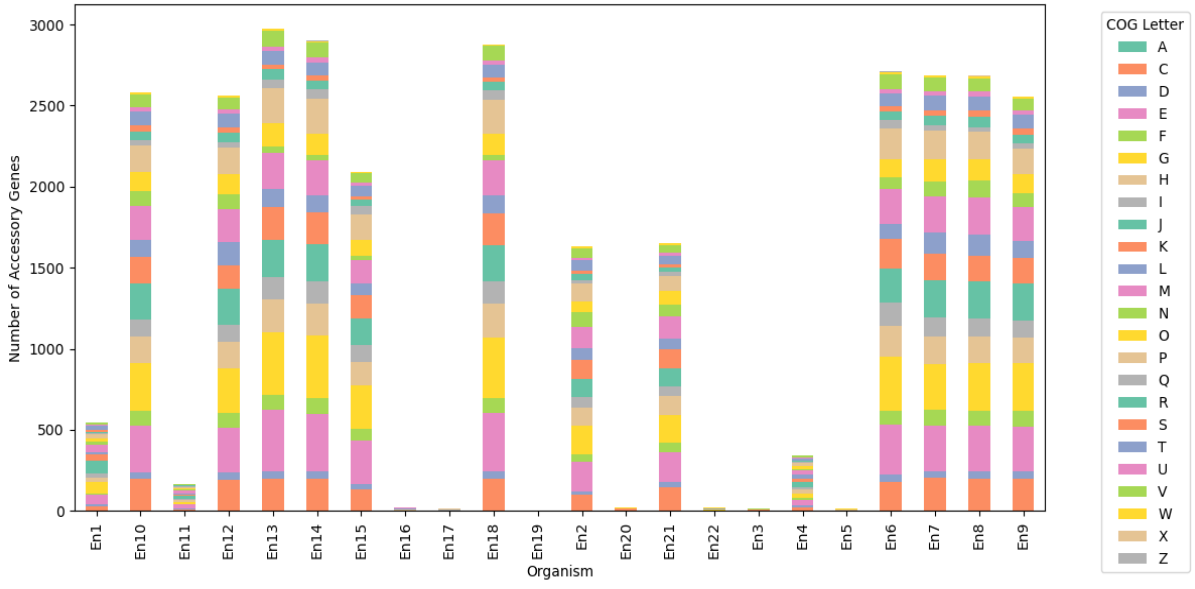
The overall distribution across the categories exhibited a relatively consistent trend, though some minor fluctuations were observed. Notably, strains En22, En20, En17, and En18 showed higher percentages of classified sequences, whereas strains En19 and En5 had slightly lower percentages. This suggested a generally consistent functional distribution of genes among the strains, but with slight deviations that could reflect the unique functional profiles and adaptations of individual strains (**Figure 5A**).

The largest COG group identified across the strains was related to Carbohydrate transport and metabolism (G), followed by Amino acid transport and metabolism (E), Translation, ribosomal structure and biogenesis (J), and Cell wall/membrane/envelope biogenesis (M). Conversely, some COG groups showed lower gene numbers across the 22 strains, such as Extracellular structures (W), Mobilome: prophages, transposons (X), RNA processing and modification (A), and Cytoskeleton (Z). This variation indicated that not all COG groups were present in all strains, with certain functional categories being conserved across the strains, while others exhibited considerable variability, highlighting functional diversity and strain-specific adaptations (**Figure 5B**).

Table 2. COG classification of the 22 Enterobacteriaceae Strains

Strain	Percentage (%)
En1	86.54
En2	85.32
En3	86.21
En4	81.58
En5	56.25
En6	86.21
En7	78.92
En8	78.63
En9	87.14
En10	87.10
En11	84.69
En12	84.91
En13	88.74
En14	89.77
En15	89.33
En16	89.66
En17	92.00
En18	90.20
En19	55.56
En20	96.55
En21	83.73
En22	100

A



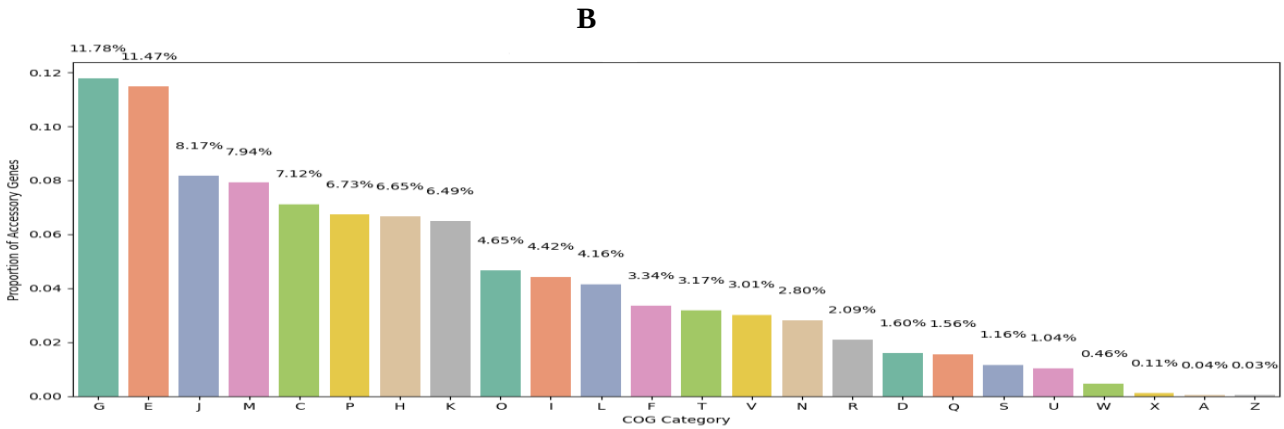


Figure 5. A. COG distribution across tested strains. **B.** COG distribution across accessory genes.

Discussion

The pan-genome analysis of 22 *Enterobacteriaceae* genomes highlighted significant genetic diversity, with no universally conserved core genes. Instead, the genome was dominated by accessory genes, including shell and cloud genes, suggesting an open pan-genome structure. The absence of core genes across all strains indicated extensive genomic variability, likely driven by horizontal gene transfer (HGT) and environmental adaptation [12]. This high level of genetic plasticity is characteristic of *Enterobacteriaceae*, as previous studies had demonstrated that frequent gene acquisition and loss shape their genomes, allowing them to adapt to various niches, including human hosts, animals, and environmental reservoirs [13].

The rapid initial increase in total gene counts observed in **Figure 1C** aligned with findings from other bacterial pan-genome studies, where the high genetic diversity among strains leaded to the continuous discovery of new genes [14]. The eventual slowing of gene accumulation suggested that while new genes continue to be identified, a substantial proportion of the genetic repertoire has been captured. This trend supported the concept of an open pan-genome, a common feature of bacterial species that frequently acquire genes via HGT [15]. The absence of core genes further underscored this point, as even essential housekeeping genes may be replaced by homologous genes from different sources, resulting in strain-specific variations.

Figure 1D illustrated that as more genomes were included in the analysis, the number of unique genes continued to rise, albeit with fluctuations. This trend suggested that *Enterobacteriaceae* exhibited substantial genomic fluidity, consistent with previous studies that have reported a high prevalence of strain-specific genes, including virulence factors and antimicrobial resistance genes [16]. The variability in unique gene discovery rates might be influenced by phylogenetic relationships among the strains, genome size, and sampling bias. These findings reinforced the notion that *Enterobacteriaceae* genomes are highly dynamic, with ongoing gene acquisition and loss shaping their evolution and adaptation [17]

Overall, the results highlighted the complexity of *Enterobacteriaceae* genome evolution, emphasizing the importance of accessory genes in shaping pathogenicity, antimicrobial resistance, and niche adaptation. The open nature of the pan-genome suggested that continuous surveillance of new strains was necessary to track emerging resistance genes and novel virulence determinants.

The absence of core genes across all 22 *Enterobacteriaceae* strains highlighted the open nature of their pan-genome, where adaptation was primarily driven by the accessory genome. This genomic flexibility, facilitated by horizontal gene transfer, enabled the acquisition of genes linked to virulence and antimicrobial resistance, allowing strains to thrive in diverse environments [13]. The presence of strain-specific genes, often associated with mobile genetic elements, underscores the role of genomic plasticity in bacterial evolution [18]. Phylogenetic analysis based on accessory genes revealed clusters of closely related strains, suggesting shared ancestry, while outliers reflected significant evolutionary divergence, likely influenced by niche specialization or

selective pressures [19]. These findings emphasized the importance of continuous genomic surveillance to monitor the emergence of novel resistance and virulence determinants in *Enterobacteriaceae*, which is crucial for public health interventions.

The functional annotation of genes using the Clusters of Orthologous Groups (COG) database revealed that approximately 80% of the genes across the 22 *Enterobacteriaceae* strains were successfully classified into functional categories. The distribution of COG categories was relatively consistent across strains, with minor variations in classification rates. Strains such as En22, En20, En17, and En18 exhibited higher percentages of classified sequences, whereas En19 and En5 displayed lower classification rates, suggesting variability in genome content and functional specialization. These variations might reflect strain-specific adaptations influenced by ecological niches, selective pressures, or horizontal gene transfer events [9, 10]. The relatively uniform functional distribution across strains suggested a conserved core of metabolic and cellular processes, yet the deviations in some strains indicated genetic plasticity that allowed adaptation to different environments [20]. The presence of highly classified strains with near-complete COG annotation, alongside others with lower classification rates, highlighted the influence of genome reduction, acquisition of novel genes, or differences in genome sequencing and annotation quality [21].

Among the COG categories, Carbohydrate transport and metabolism (G) emerged as the most abundant, followed by Amino acid transport and metabolism (E), Translation, ribosomal structure, and biogenesis (J), and Cell wall/membrane/envelope biogenesis (M). These categories were fundamental to bacterial survival and growth, playing key roles in nutrient uptake, protein synthesis, and cellular integrity [9]. In contrast, functional categories such as Extracellular structures (W), Mobilome: prophages, transposons (X), RNA processing and modification (A), and Cytoskeleton (Z) were less frequently detected, indicating their variable presence across strains. The reduced representation of these categories suggested that while some functions are conserved across the strains, others exhibited significant variability, likely driven by horizontal gene transfer and selective pressures in different environments [22, 23]. This functional diversity underscored the adaptability of *Enterobacteriaceae* and its capacity to evolve in response to environmental challenges, including antimicrobial resistance and niche specialization.

Conclusion

The pan-genome analysis of 22 *Enterobacteriaceae* strains revealed a highly dynamic and open genome structure, with no core genes conserved across all strains. Instead, the accessory genome played a critical role in genetic diversity, encompassing genes associated with virulence, antimicrobial resistance, and strain-specific adaptations. The absence of a core genome highlighted the extensive variability within *Enterobacteriaceae*, likely driven by horizontal gene transfer, gene loss, and environmental selection pressures. The functional classification of genes using the COG database demonstrated a generally conserved distribution of essential metabolic and cellular processes, with carbohydrate and amino acid metabolism being the most abundant categories. However, variability in certain functional groups suggested strain-specific adaptations, reflecting the genomic plasticity of *Enterobacteriaceae*. Overall, these findings underscored the complex evolutionary dynamics of *Enterobacteriaceae*, where genome flexibility enabled adaptation to diverse environments and selective pressures. Understanding this genomic variability was crucial for tracking antimicrobial resistance, designing effective therapeutic strategies, and developing targeted interventions to mitigate the spread of multidrug-resistant strains.

Conflicts of interest

The authors declare no competing interests.

References

1. Davin-Regli, A., J.-P. Lavigne, and J.-M. Pagès, (2019) Enterobacter spp.: update on taxonomy, clinical aspects, and emerging antimicrobial resistance. Clinical microbiology reviews,. **32**(4): p. 10.1128/cmr. 00002-19.

2. Lee, I.P.A. and C.P. Andam, (2019) Pan-genome diversification and recombination in Cronobacter sakazakii, an opportunistic pathogen in neonates, and insights to its xerotolerant lifestyle. BMC microbiology,. **19**: p. 1-14.
3. Koonin, E.V., (2002*)* The Clusters of Orthologous Groups (COGs) Database: phylogenetic classification of proteins from complete genomes. The NCBI handbook,.
4. Medicine, N.L.o., (1988*)* National Center for Biotechnology Information (NCBI). Bethesda (MD). Available from,.
5. Seemann, T., (2014*)* Prokka: rapid prokaryotic genome annotation. Bioinformatics,. **30**(14): p. 2068-2069.
6. Liu, N., et al., (2022) Pan-genome analysis of Staphylococcus aureus reveals key factors influencing genomic plasticity. Microbiology Spectrum,. **10**(6): p. e03117-22.
7. Pedersen, T.L., (2017) Hierarchical sets: analyzing pangenome structure through scalable set *visualizations.* Bioinformatics,. **33**(11): p. 1604-1612.
8. Kumar, S., et al., (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution,. **35**(6): p. 1547-1549.
9. Galperin, M.Y., et al., (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic acids research,. **49**(D1): p. D274-D281.
10. Tatusov, R.L., et al., (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic acids research,. **28**(1): p. 33-36.
11. Wickham, H., (2006) An introduction to ggplot: An implementation of the grammar of graphics in R. Statistics,. **1**.
12. Vernikos, G., et al., (2015) Ten years of pan-genome analyses. Current opinion in microbiology,. **23**: p. 148-154.
13. McInerney, J.O., A. McNally, and M.J. O'connell, (2017) Why prokaryotes have pangenomes. Nature microbiology,. **2**(4): p. 1-5.
14. Tettelin, H., et al., (2008) Comparative genomics: the bacterial pan-genome. Current opinion in microbiology,. **11**(5): p. 472-477.
15. Lapierre, P. and J.P. Gogarten, (2009) Estimating the size of the bacterial pan-genome. Trends in genetics,. **25**(3): p. 107-110.
16. Araújo, C.L., et al., (2019) Prediction of new vaccine targets in the core genome of Corynebacterium pseudotuberculosis through omics approaches and reverse vaccinology. Gene,. **702**: p. 36-45.
17. Touchon, M., et al., (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS genetics,. **5**(1): p. e1000344.
18. Pinzón-Latorre, D. and M.K. Deyholos, (2013) Characterization and transcript profiling of the pectin methylesterase (PME) and pectin methylesterase inhibitor (PMEI) gene families in flax (Linum usitatissimum). BMC genomics,. **14**: p. 1-25.
19. Mira, A., et al., (2010) The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol,. **13**(2): p. 45-57.
20. Caporale, A.L., C.M. Gonda, and L.F. Franchini, (2019) Transcriptional enhancers in the FOXP2 locus underwent accelerated evolution in the human lineage. Molecular biology and evolution,. **36**(11): p. 2432-2450.
21. Kristensen, D.M., et al., (2010) New dimensions of the virus world discovered through metagenomics. Trends in microbiology,. **18**(1): p. 11-19.
22. Huerta-Cepas, J., et al., (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular biology and evolution,. **34**(8): p. 2115-2122.
23. Rouli, L., et al., (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. New microbes and new infections,. **7**: p. 72-85.