Acta Biologica Szegediensis

Volume 68(1):46-53, 2024
Acta Biologica Szegediensis
http://abs.bibl.u-szeged.hu/index.php/abs

DOI:10.14232/abs.2024.1.46-53

ARTICLE

Insights into the structure, function, and pathophysiology of *Shigella dysenteriae* through pangenome analysis

Asif Mir*, Danial Ahmed Hashmi, Muhammad Mustajeeb UL Haq Qureshi, Muhammad Saleem Faisal, Wajid Iqbal

Department of Biological Sciences, International Islamic University, Islamabad, Pakistan

ABSTRACT *Shigella dysenteriae*, the causative agent of shigellosis, poses a significant global health threat due to its role in causing millions of cases of bacillary diarrhea and escalating antibiotic resistance. This study utilized bioinformatics analysis with the Pan Explorer to delve into the pangenome of *S. dysenteriae*. The aim was to uncover key bacterial functions, elucidate its pathogenicity and virulence, and identify factors contributing to genetic variability among strains. Results revealed a larger dispensable genome compared to the core genome and strain-specific genes. Metabolism-related Cluster of Orthologous Groups were predominant, followed by cellular processing and signaling pathways, while poorly characterized Cluster of Orthologous Groups had modest representation and those associated with information and storage processing were least prevalent. Notably, genes linked to the pathogenicity and virulence of *S. dysenteriae* were found in both dispensable and core genome regions, indicating their significance. Overall conservation was observed among strain genomes, but the open pangenome nature suggests potential for genetic exchange with other sources. These findings provide valuable insights for future microbial genomics research on *S. dysenteriae*.
Acta Biol Szeged 68(1):46-53 (2024)

KEY WORDS
antibiotic resistance
bioinformatics
genetic variability
pangenome analysis
pathogenicity
Shigella dysenteriae

ARTICLE INFORMATION
Submitted
17 July 2024
Accepted
26 November 2024

***Corresponding author**
E-mail: asif.mir@iiu.edu.pk

Introduction

Shigella dysenteriae (*S. dysenteriae*) is a Gram-negative bacterium that lacks motility and possesses a rod-shaped structure (Kadhim et al. 2023). *S. dysenteriae* is responsible for shigellosis, a disease that takes the lives of over one million people annually, making it the leading cause of bacterial diarrhea worldwide, accounting for 13.2% of all deaths related to diarrheal diseases (Bengtsson et al. 2022). The common antimicrobial drugs used against *S. dysenteriae* include beta-lactam, tetracycline, aminoglycoside, and fluoroquinolone antibiotics (Kadhim et al. 2023). However, there is an increasing trend in *S. dysenteriae* strains to attain resistance against these antibiotics, therefore, the World Health Organization enlist *S. dysenteriae* as a potential biothreat agent in low- and middle-income countries (LMICs) (Baker et al. 2023). LMICs are most susceptible to shigellosis due to inadequate sewage systems and limited access to clean drinking water, thus shigellosis is emerging as a major public health threat in these countries (Baker et al. 2023).

To address the challenges of shigellosis management, there is a pressing need to develop effective alternative treatments to antibiotics (Khan et al. 2023). Computational drug discovery has emerged as a promising approach, outpacing traditional wet-lab methods that are often time-consuming and costly. The advent of big data and advanced informatics pipelines has facilitated the shift from conventional therapeutic discovery to computational approaches, enabling improved drug efficacy and faster development timelines.

A crucial strategy in optimizing drug discovery and identifying novel targets is leveraging genomic data from multiple strains of a species. Pan-genomics, a method that examines the collective genome of multiple strains, has proven instrumental in this regard. Pan-genomes are categorized into the core genome (shared by all strains), the accessory genome (present in some strains), and strain-specific genes.

In this study, we employed computational approaches to investigate the genomic diversity of *S. dysenteriae*. Our objectives were to delineate the core genome, accessory genome, and strain-specific genes, and to analyze the distribution of COG functional categories across *S. dysenteriae* genomes. This analysis offers critical insights into the evolutionary dynamics, selective pressures, and pathogenic mechanisms of *S. dysenteriae*. Additionally, it may shed light on the functional roles of gene clusters in disease progression, susceptibility, and resistance, thereby

Table 1: Collection of analyzed genomes of *S. dysenteriae*

No	Strain	Gene Bank ID's
1	*S. dysenteriae* ATCC 13313	GCA_002949675.1
2	*S. dysenteriae* SWHEFF_51	GCA_022354065.1
3	*S. dysenteriae* E670/74	GCA_002943855.1
4	*S. dysenteriae* ATCC 12039	GCA_002950055.1
5	*S. dysenteriae* 204/96	GCA_002949345.1
6	*S. dysenteriae* 93-119	GCA_002949555.1
7	*S. dysenteriae* 96-265	GCA_002949615.1
8	*S. dysenteriae* CFSAN029786	GCA_009662135.1
9	*S. dysenteriae* ATCC 9753	GCA_002950155.1
10	*S. dysenteriae* ATCC 49347	GCA_002950095.1
11	*S. dysenteriae* ATCC 12037	GCA_002950015.1
12	*S. dysenteriae* NCDC 599-52	GCA_032363495.1
13	*S. dysenteriae* ATCC 49346	GCA_002950075.1
14	*S. dysenteriae* 2017C-4522	GCA_002949415.1
15	*S. dysenteriae* ATCC 12021	GCA_002949935.1
16	*S. dysenteriae* 08- 3380	GCA_002949815.1
17	*S. dysenteriae* BU53M1	GCA_002741615.1
18	*S. dysenteriae* 80- 547	GCA_002949855.1
19	*S. dysenteriae* 69- 3818	GCA_002949715.1
20	*S. dysenteriae* 53- 3937	GCA_002949775.1
21	*S. dysenteriae* 07- 3308	GCA_002949755.1
22	*S. dysenteriae* NCTC 9718	GCA_002949835.1
23	*S. dysenteriae* SC595	GCA_032357325.1
24	*S. dysenteriae* Sd197	GCA_000012005.1
25	*S. dysenteriae* CFSAN010954	GCA_002949875.1

contributing to the development of future diagnostic and therapeutic strategies for managing *S. dysenteriae* infections.

Material and Methods

Sequence Retrieval and Filtering

Shigella dysenteriae genomes were retrieved from the NCBI database (https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=622&assembly_level=2:3). A total of 273 publicly available genomes (as of October 23, 2023) were obtained. These genomes underwent a two-step filtering process: first, genomes with an assembly status of "complete" and labeled as "chromosome" were selected. Additionally, genomes with "complete" annotation status were included, resulting in 35 strains meeting these criteria.

To ensure quality and eliminate incomplete or poorly annotated sequences that could bias the analysis, further filtering was conducted. Eight genomes were excluded due to frameshifted proteins, reducing the dataset to 27 strains. Additional quality control was performed using the Pan-Explorer program, which filters partially annotated or fragmented genome assemblies, leading to the removal of two more genomes. After these steps, the final dataset comprised 25 genomes of *S. dysenteriae* strains, as detailed in Table 1.

Software

The pangenome analysis of *S. dysenteriae* was conducted using the Pan-Explorer tool (Dereeper et al. 2022), which facilitates comparative genomics and evolutionary analyses in bacteria. Pan-Explorer implements the PGAP pipeline for pangenome analysis and provides a user-friendly interface for exploring gene clusters and interpreting data.

Data Processing

The 25 genomes of *S. dysenteriae* were indexed in GenBank using their assembly identifiers. To process the data efficiently, the latest pan-genomics software, PanACoTA (Perrin et al. 2021), was employed. The compatibility of the GenBank assemblies was subsequently verified using Pan-Explorer. No strains were rejected during this compatibility check. The final dataset included 25 *S. dysenteriae* genomes, with BLAST analysis set to a minimum identity threshold of 80%.

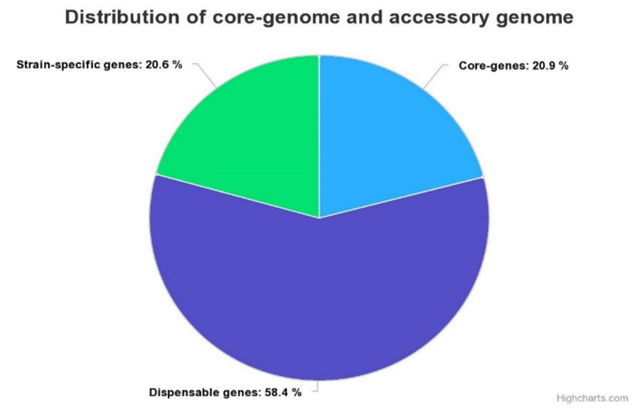


Fig. 1. Dispensable (accessory) and core genome distribution in *S. dysenteriae*. Gene percentages for core, dispensable, and strain-specific genes are shown in the pie-chart.

Examination of Core and Dispensable Genomes

Following data processing, the distribution of genes within the dispensable genome (genes in the accessory genome) and the core genome (genes shared by all strains) was analyzed. Additionally, strain-specific genes were identified for each genome. To explore gene clusters present in subsets of strains, an interactive presence/absence matrix was analyzed, revealing intersections within the dispensable genome. The magnitude and prevalence of these intersections were also assessed.

Examination of COG Functional Classifications

The functional classifications of genes were analyzed using Pan-Explorer's implementation of RPS-BLAST (Tatusov et al. 2000). This analysis determined the representation and distribution of COG functional categories across *S. dysenteriae* genomes.

Exploratory Analysis

A Circos model was employed to visualize similarities and differences between core and strain-specific genes in the 25 *S. dysenteriae* strains (Krzywinski et al. 2009). Gene clusters were color-coded according to their COG classifications for easier interpretation. Additionally, hive plots were generated to illustrate improvements and insights derived from the analysis.

Results

General Features

In this study, comparative genome analysis was conducted on 25 strains of *S. dysenteriae*. Fully annotated and completely assembled genomes were retrieved from the NCBI database. The genome sizes of these strains ranged from 4.4 to 5.2 MB. The total gene count per strain varied between 4715 and 5393, with protein-coding genes ranging from 3446 to 4798.

The Pan, Core, and Dispensable Genome

The pangenome of the 25 *S. dysenteriae* genomes comprised 7616 gene clusters, representing the collective genome pool of the species (Fig. 1). This pool was divided into 1595 core genes (20.9%), 4450 dispensable genes (58.4%), and 1571 strain-specific genes (20.6%) (Fig. 2).

The core genome, consisting of 1595 genes, did not include any hypothetical proteins. These core genes were associated with critical functions such as metabolism, survival, and virulence. Specifically, they encompassed genes involved in transcription, translation, cell division, signaling pathways, and pathogenesis, underscoring their essential roles in the biology of *S. dysenteriae*.

The core genome of *S. dysenteriae* lacks hypothetical

Table 2. Genetic intersections in the dispensable genome of *S. dysenteriae*

Intersection ID	Number of genes (Nb)	Number of strains	Strains involved	Meaning and importance
1	Largest	24	53_3937, 07_3308, BU53M1, SC595, NCTC_9718, ATCC_13313, 80_547, 08_3380, swheff_49, swHEFF_51, NCDC_599_52, ATCC_12021, 93_199, 96_265, ATCC_9753, ATCC_12037, 2017C_4522, ATCC_49346, CF-SAN029786, ATCC_49347	Represents the core genetic features shared across a significant number of *S. dysenteriae* strains, suggesting a cohesive genetic foundation important for the bacterium's pathogenicity and survival.
2	Small (occurring twice)	2	SWHEFF_49, SWHEFF_51	Indicates a localized or unique genetic overlap, possibly highlighting specialized functions or adaptations shared between these two strains.
3	Small (occurring twice)	2	ATCC_12021, NCDC_599_52	Suggests a strain-specific or niche adaptation within these two strains, which could be crucial for understanding variations in pathogenicity or ecological roles.

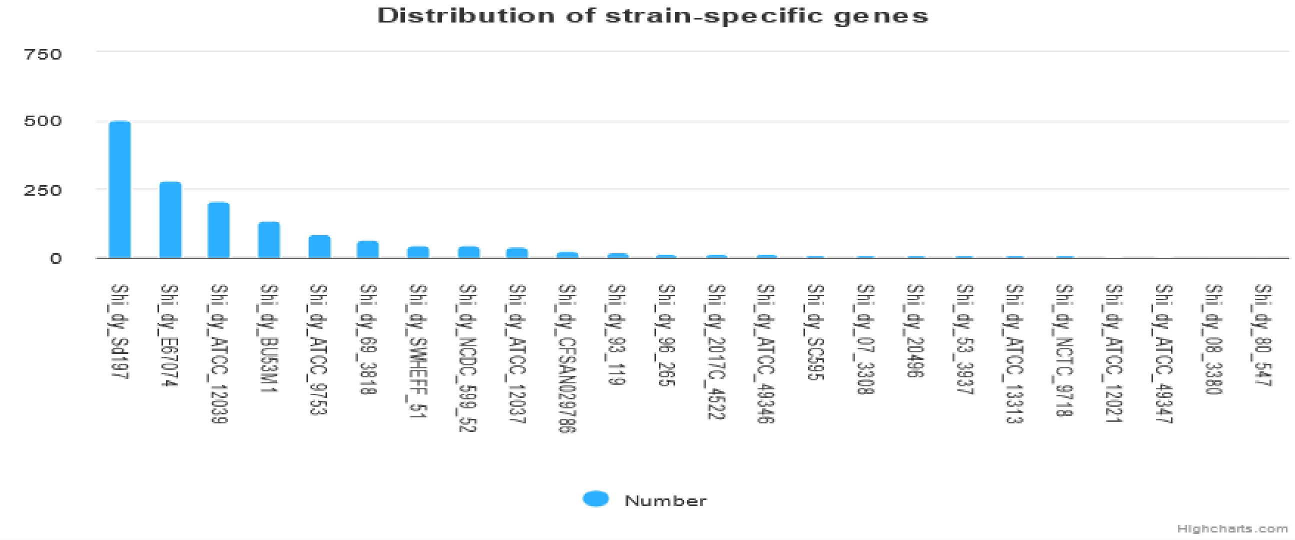


Fig. 2. Distribution of strain-specific genes in *S. dysenteria*e.

proteins and includes genes essential for critical cellular and metabolic processes, such as transcription, translation, cell division, and the metabolism of sugars, fats, and proteins. It also plays vital roles in DNA repair, motility, post-translational modifications, and pathogenesis. Key components of the core genome include ATP-binding proteins, ABC transporters, transcriptional regulators, transcription elongation factors, cell division proteins, endonucleases, type III secretion system proteins, type II toxin-antitoxin systems, NADPH nitroreductase, lipoprotein antitoxins, oxidoreductases, ribosomal proteins (50S and 30S), HlyD family secretion proteins, dihydrolipoyl dehydrogenase, DNA-cytosine methyltransferase, cytochrome c maturation proteins (CcmE), DNA-binding transcriptional regulator KdgR, RNA chaperone/antiterminator CspA, RNA polymerase sigma-54 factor, and TerC family proteins.

The dispensable genome contains 685 hypothetical protein genes out of 4450 genes. The remaining 3765 genes encode important features, including those related to pathogenesis and virulence. Fig 2. demonstrates the distribution of strain-specific genes among *S. dysenteriae* strains. Shigella_dysentriae_Sd197 contains the greatest number of strain specific genes (501) followed by Shi_dy_E67074 (282), ATCC_12039 (208), BU53M1 (136), ATCC_9753 (88), 69_3818 (66), respectively. Shi_dy_SWHEFF-51, NCDC_599_52, ATCC12037, CFSAN029786, 93_119, 96_265, 2017C_4522, ATCC_49346, SC595 has 47, 46, 42, 23, 20, 15, 14, 13, 12, respectively.

While 07_3308, 20496, 53_3937, ncdc_59952 have 8 genomes each. Whereas ATCC_12021, ATCC_49347, have 7 genomes, respectively. The Shi_dy_08_3380 and Shi_dy_80_547 have 2 strain specific genes, making them the strains with least number of strain specific genomes.

In Fig. 3, distribution of gene cluster orders in strains is shown based on hierarchical clustering. The results of pangenome analysis indicated that the pangenome of *S. dysenteriae* has a smaller core genome than the dispensable genome. Moreover, *S. dysenteriae* has an open pangenome, integrating more gene families as the new genomes are added during the pangenome analysis.

Moreover, the construction of an upset diagram illustrated the abundance of intersections between the dispensable genome of *S. dysenteriae*. It revealed that there are 14 intersections in total. The largest intersections incorporated genes shared between 24 strains of *S. dysenteriae*, while the smallest intersection consisted of genes shared among 2 strains. 21 strains of *S. dysenteriae*, 53_3937, 07_3308, BU53M1, SC595, NCTC_9718, ATCC_13313, 80_547, 08_3380, SWHEFF_49, SWHEFF_51, NCDC_599_52, ATCC_12021, 93_199, 96_265, ATCC_9753, ATCC_12037, 2017C_4522, ATCC_49346, CFSAN029786, ATCC_49347 were closely related due to their presence in the top 4 intersections.

Evolution of COG functional categories

Analysis of the functional distribution of COGs within the pangenome revealed significant patterns. The categories included: 1) Information and storage processing {sub categories A, J, K, L}, Cellular processing and signaling {Sub categories D, Y, V, T, N, M, Z, W, U, O}, Metabolism{sub categories C, E, F, G, H, I, P, Q} & Poorly {sub categories R, S}.

Metabolism-related COGs were the most abundant, followed by those associated with cellular processes and signaling. Poorly characterized COGs had a modest presence, while those linked to information storage and processing were the least represented. A significant proportion of genes (132 core and 1539 dispensable) were not assigned to any COG category.

In the core genome, 77.3% of metabolism-related genes were identified, compared to 29.1% in the dispensable genome. Key metabolic subcategories included energy production and conversion (C), carbohydrate transport and metabolism (G), lipid transport and metabolism (I), amino acid transport and metabolism (E), nucleotide transport and metabolism (F), coenzyme transport and metabolism (H), and secondary metabolite biosynthesis, transport, and metabolism (Q).

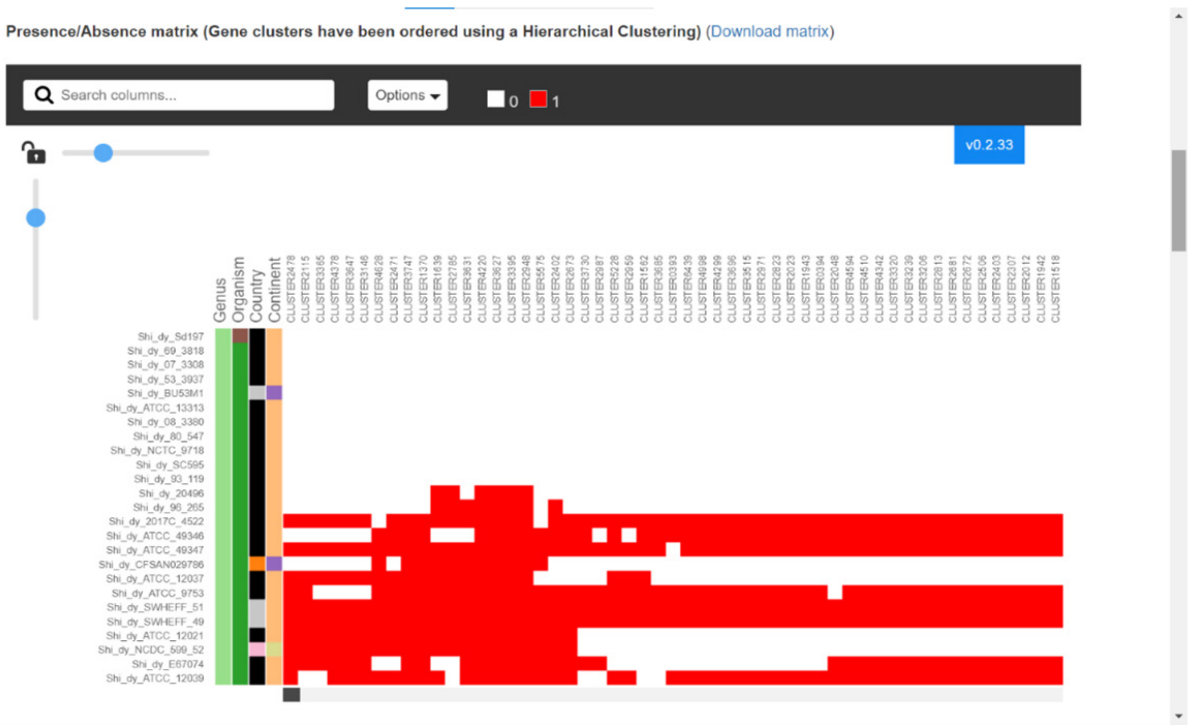
Among poorly characterized categories, subcategories such as general functional prediction (R) and function unknown (S) were prominent, with the latter accounting for 19.46% of dispensable genes. The core genome also contained genes involved in cellular processes, signaling, and information storage, with notable representation in subcategories such as defense mechanisms (V), cell wall/membrane/envelope biogenesis (M), signal transduction (T), cell motility (N), post-translational modifications (O), translation (J), transcription (K), and replication, recombination, and repair (L).

At the strain-specific level, all *S. dysenteriae* strains had a high abundance of genes in the general functional prediction (R) subcategory. In contrast, categories such as RNA processing and modification (A) and nucleotide transport and metabolism (F) were the least represented. Other modestly represented subcategories included transcription (K), replication, recombination, and repair (L), and metabolism-related categories such as energy production and conversion (C), carbohydrate transport and metabolism (G), and amino acid transport and metabolism (E).

Strain Sd197 had 24.2% of its strain-specific genes assigned to COGs, while strain E67074 had 12.78%. In other strains, the percentage ranged from 4.3% to 10.94%.

Physical maps of S. dysenteriae

The results section presents the physical mapping of 25 *S. dysenteriae* strains, illustrated in Fig. 4(a-y) using Circos plots. These plots provide a detailed depiction of genomic features, with tracks representing genes on the forward strand (blue), reverse strand (red), core genome (purple), strain-specific genome (green), and the GC skew. Notably, the Sd-197 strain contained the highest number of strain-specific genes. Additionally, the physical maps revealed a balanced distribution of forward and reverse genome sizes across all strains, with consistent GC content throughout the dataset.



Presence/Absence matrix (Gene clusters have been ordered using a Hierarchical Clustering) (Download matrix)

Fig. 3. A representation of the presence/absence matrix illustrating the hierarchical clustering-based ordering of gene clusters. The colors red and white stand for presence and absence, respectively.

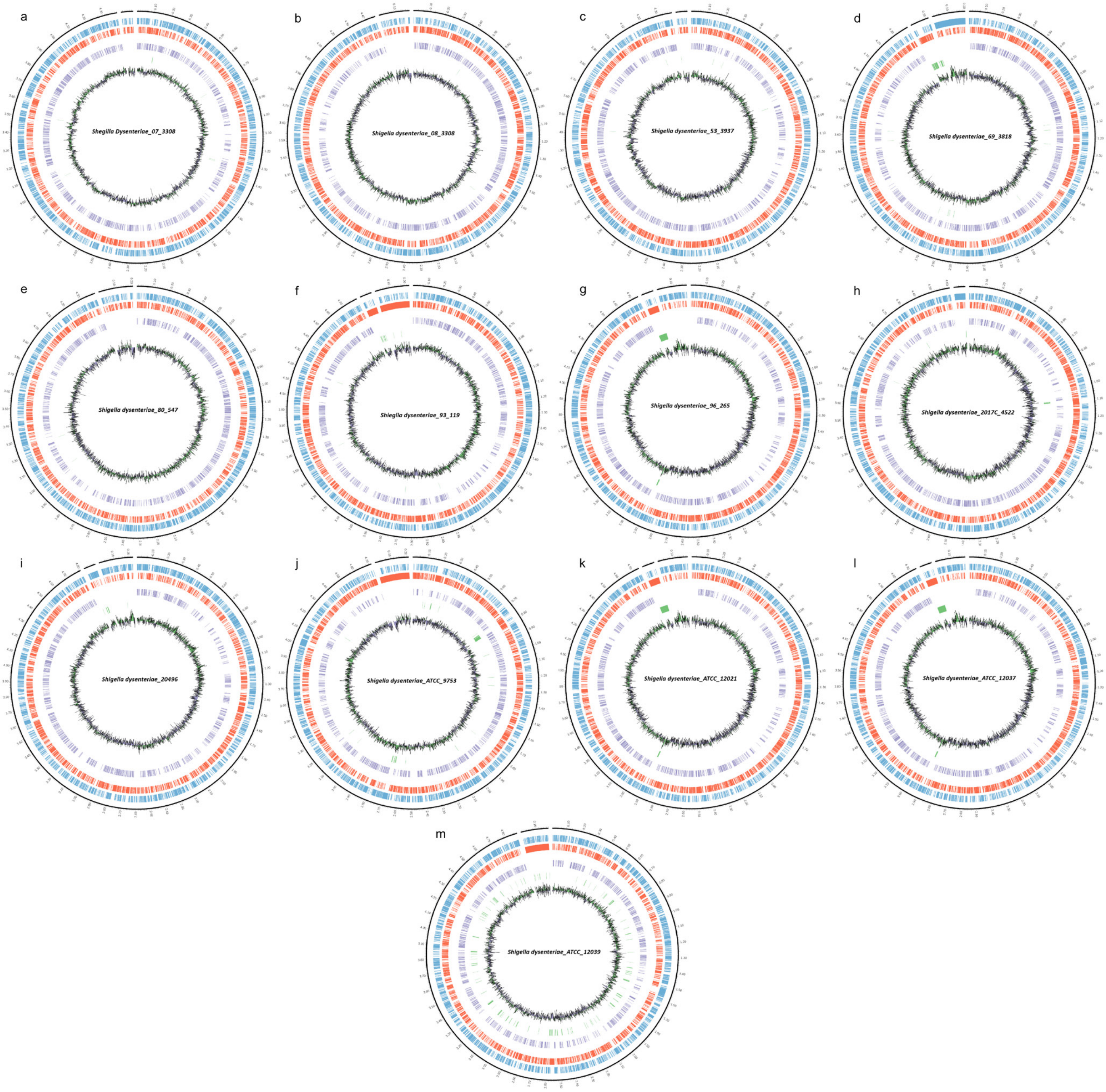


Fig. 4(a-y). Graphical representation depicting the arrangement of core and strain-specific genes in a physical map. The outer-to-inner tracks illustrate genes on the Forward strand (blue), Reverse strand (red), Core-genome (purple), Strain-specific genome (green), followed by the GC skew. Sd-197 strain was found to be the strain possessing the greatest count of genes in strain specific genome. Moreover, the physical maps showed that all the strains had almost equal size of forward and reverse genome. Similarly, the GC content was also almost constant in all 25 strains.

Discussion

Diseases caused by *S. dysenteriae* are classified by the WHO as a global health threat, particularly in underdeveloped countries. Populations in these regions are especially vulnerable due to poor sanitation and contaminated drinking water (Ahmed et al. 2003). Shigellosis, the primary health complication caused by *S. dysenteriae*, is associated with a high incidence of hospital admissions, significant morbidity, and mortality (Ahmed et al. 2003).

To address these challenges, this study conducted an up-to-date pangenomic analysis to investigate the genetic variability of *S. dysenteriae*, identify core and strain-specific genes involved in virulence and essential

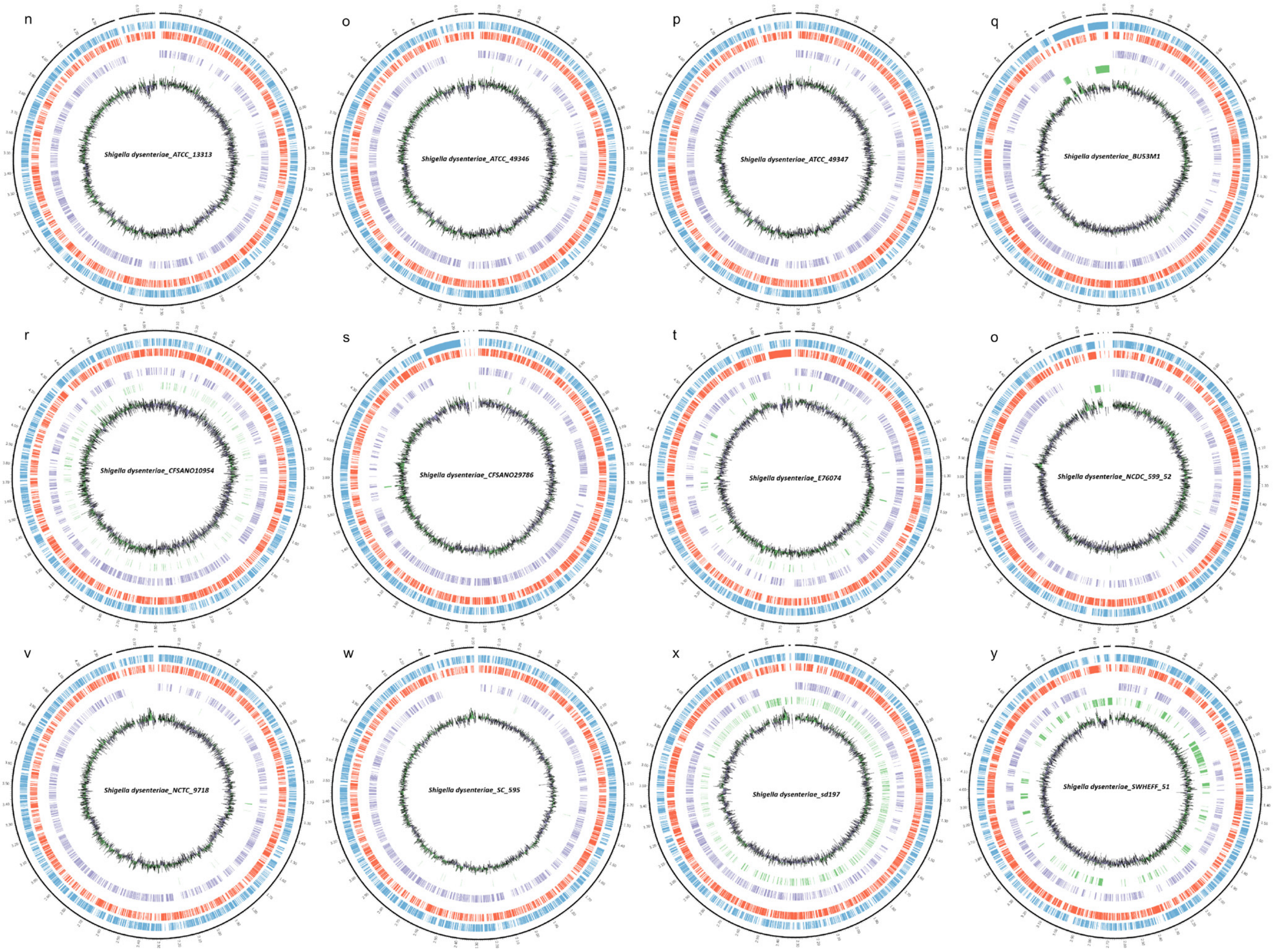


Fig. 4. Continued.

functions, and enhance our understanding of microbial evolution and biology. These insights aim to inform more effective treatment strategies for shigellosis, particularly considering increasing antibiotic resistance.

The genetic repertoire of 25 *S. dysenteriae* strains was analyzed to construct the pangenome and determine factors influencing the bacterial lifestyle. All strains displayed GC content between 49% and 51%, indicating close genetic relationships and uniform taxonomy.

These genomes demonstrated minor variations in size, gene count, and protein-coding genes, alongside relatively large variations in pseudogenes, potentially reflecting events of genome gain or loss. The dataset included 7616 gene clusters, comprising core, dispensable, and strain-specific genes. Notably, the core genome, containing 1595 genes, was much smaller than the dispensable genome, which contained 4450 genes. This supports the classification of *S. dysenteriae* as having an open pangenome, where sequencing new strains contributes to the discovery of additional genes and genetic material.

The core genome exhibited a low proportion of hypothetical proteins, whereas the dispensable genome contained a higher proportion. This highlights opportunities for experimental and computational research to elucidate the functions of these proteins and their roles in virulence, metabolism, and survival processes. Strain-specific genes varied across strains, suggesting genetic diversity influenced by geographical and environmental factors.

The core genome encodes essential functions, including nutrient uptake via ABC transporters and major facilitator superfamily (MFS) transporters. Genes associated with pathogenicity included those involved in invasion (e.g., SirB1, SirB2), secretion (e.g., HlyD family proteins), and toxicity (e.g., type II toxin-antitoxin systems).

Genes supporting homeostasis and energy balance during infection were also identified, including Fe-S cluster assembly genes (e.g., SufC, IscU) and ferrichrome porins (e.g., FhuA). Antibiotic resistance genes (e.g., YcbX, TehB) and stress response genes (e.g., GadE, YodD) were also present, underscoring the adaptability of *S. dysenteriae* to adverse conditions.

Numerous core genes facilitated metabolism, including those involved in electron transport (e.g., cytochrome oxidase subunits), fatty acid oxidation (e.g., FadB), and carbohydrate metabolism (e.g., glucose-6-phosphate dehydrogenase). These findings reveal the critical roles of core genes in sustaining bacterial survival and pathogenicity, while highlighting potential targets for therapeutic interventions.

Conclusion

This study employed PanExplorer to perform a comparative pangenome analysis of *S. dysenteriae*, providing insights into its genomic composition and lifestyle. The analysis of 25 *S. dysenteriae* strains revealed that the dispensable genome is significantly larger than the core genome, confirming the species' open pangenome.

The core and dispensable genomes were found to contain vital genes essential for bacterial survival and pathogenicity, offering valuable information for the development of novel therapeutic targets and reverse vaccines against *S. dysenteriae*. Additionally, the presence of antibiotic resistance genes and the ability to survive in harsh environments contribute to the pathogen's increased virulence.

Despite the close genetic relationships and shared evolutionary background among *S. dysenteriae* strains, their genomic variability is shaped by rearrangement events and geographic factors. The findings from this research can inform improved strategies for managing *S. dysenteriae* infections, addressing its role as a significant global health threat.

Acknowledgement

The authors extend sincere appreciation to the Human and Molecular Genetic Lab at International Islamic University Islamabad, Pakistan, for their invaluable support in facilitating the research endeavors, as well as to the university itself for providing the necessary infrastructure and resources for this study.

References

Ahmed K, Shakoori FR, Shakoori AR (2003) Aetiology of shigellosis in northern Pakistan. J Health Popul Nutr 1:32-39.

Baker S, Scott TA (2023) Antimicrobial-resistant *Shigella*: where do we go next? Nat Rev Microbiol 7:409-410.

Bengtsson RJ, Simpkin AJ, Pulford CV, Low R, Rasko DA, Rigden DJ, Hall N, Barry EM, Tennant SM, Baker KS (2022) Pathogenomic analyses of *Shigella* isolates inform factors limiting shigellosis prevention and control across LMICs. Nat Microbiol 2:251-261.

Dereeper A, Summo M, Meyer DF (2022) PanExplorer: a web-based tool for exploratory analysis and visualization of bacterial pan-genomes. Bioinformatics 18:4412-4414.

Kadhim BA, Alqaseer K, Al-Ganahi SA (2023) Identification and characterization of a novel lytic peptidoglycan transglycosylase (MltC) in *Shigella dysenteriae*. Braz J Microbiol 2:609-618.

Khan K, Jalal K, Uddin R (2023) Pangenome profiling of novel drug target against vancomycin-resistant *Enterococcus faecium*. J Biomol Struct Dyn 24:15647-15660.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 9:1639-1645.

Perrin A, Rocha EPC (2021) PanACoTA: a modular tool for massive microbial comparative genomics. NAR Genom Bioinform 1:lqaa106.

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 1:33-36.