The Intricate Relationship of G-Quadruplexes and Pathogenicity Islands: A Window into Bacterial Pathogenicity

Bo Lyu[1,*], Qisheng Song[1,*]

1 Division of Plant Science and Technology, University of Missouri, Columbia, MO 65211, USA

*Corresponding author:

Bo Lyu; E-mail: bl3pt@missouri.edu; Phone number: +1 573 289 3566;

Qisheng Song; E-mail: songq@missouri.edu; Phone number: +1 573 882 9798

Abstract

The dynamic interplay between guanine-quadruplex (G4) structures and pathogenicity islands (PAIs) represents a captivating area of research with implications for understanding the molecular mechanisms underlying pathogenicity. This study conducted a comprehensive analysis of a large-scale dataset from reported 89 pathogenic strains of bacteria to investigate the potential interactions between G4 structures and PAIs. G4 structures exhibited an uneven and non-random distribution within the PAIs and were consistently conserved within the same pathogenic strains. Additionally, this investigation identified positive correlations between the number and frequency of G4 structures and the GC content across different genomic features, including the genome, promoters, genes, tRNA, and rRNA regions, indicating a potential relationship between G4 structures and the GC-associated regions of the genome. The observed differences in GC content between PAIs and the core genome further highlight the unique nature of PAIs and underlying factors, such as DNA topology. High-confidence G4 structures within regulatory regions of *Escherichia coli* were identified, modulating the efficiency or specificity of DNA integration events within PAIs. Collectively, these findings pave the way for future research to unravel the intricate molecular mechanisms and functional implications of G4-PAI interactions, thereby advancing our understanding of bacterial pathogenicity and the role of G4 structures in pathogenic diseases.

Key Words: Guanine quadruplex; Pathogenicity island; Genome; *Escherichia coli*

Introduction

The discovery of the DNA double helix by Watson and Crick in 1953 revolutionized our understanding of genetics and laid the foundation for the modern field of molecular biology (*1*). Nonetheless, the intricate nature of DNA continues to surprise us even today. One such captivating feature is the DNA guanine (G)-quadruplex (G4) structure, a unique arrangement that defies the conventional double helix (*2, 3*). A G4 consists of four guanine bases and is stabilized by Hoogsteen hydrogen bonds. These stacked tetrads are interconnected by loop regions, which can vary in length and sequence, adding further complexity to the structure (Fig. 1). It is important to consider the inherent directionality of nucleic acids, with all four strands having the possibility to run in the same 5' to 3' direction, referred to as “parallel”, or alternatively, they can run in different directions, known as “antiparallel”. G4 regions can be very stable *in vitro*, particularly in the presence of K^+ (*4*). G4 structures are often found in regions of the genome with crucial regulatory functions, such as telomeres, promoters, and enhancers (*2, 5*). These structures play a role in various biological processes, including gene expression, DNA replication, and telomere maintenance (*2, 6*). Further research into G4 structures will undoubtedly uncover new insights into their functions and facilitate the development of innovative technologies.

Pathogenicity islands (PAIs) are genomic regions that contribute to the virulence and pathogenic potential of various microorganisms (*7, 8*). PAIs are distinct segments of the bacterial genome that exhibit unique characteristics compared to the rest of the DNA (*9*). They are often large in size, ranging from tens of kilobases to hundreds of

kilobases, and can be integrated into the chromosome or exist as extra-chromosomal elements, such as plasmids. PAIs often exhibit close proximity to tRNA genes, suggesting a putative mechanism where tRNA genes act as anchor points for the integration of foreign DNA acquired through horizontal gene transfer (Fig. 1E). One notable feature is their variable GC content, which tends to deviate from the average GC content of the genome in various organisms, such as *Streptomyces* (*10*), *Salmonella* (*11*), and *Yersinia* (*12*). PAIs typically contain clusters of genes involved in pathogenesis, including those encoding secretion systems (e.g., LEE (locus of enterocyte effacement) in *Escherichia coli*), superantigen (e.g., SaPI1 and SaPI2 in *Staphylococcus aureus*), and enterotoxin (e.g., *she* PAI in *Shigella flexneri*). PAIs can be acquired through the transfer of mobile genetic elements, such as plasmids, phages, or integrative and conjugative elements (ICEs), facilitating the incorporation of pathogenicity-associated genes into the recipient genome (*7, 13, 14*). One question raised in PAI is that PAIs often exhibit distinct base composition (G+C contents) compared to the core genome. The underlying reasons for this variation remain unknown, but the preservation of a genus- or species-specific base composition represents a noteworthy characteristic of bacteria (*7*). Herbert Schmidt et al. (2004) proposed a hypothetical mechanism to explain the observed variation, suggesting that factors such as DNA topology and codon message in the virulence regions present could contribute to the preservation of the distinct base composition (*7*). Hopefully, the availability of genome sequences from pathogenic bacteria and their non-pathogenic counterparts presents an exceptional opportunity to explore the intricate

structure variance and underlying mechanisms within PAIs.

Growing evidence has shown that G4 structures exhibit a striking colocalization with functional regions of the genome, and their high conservation across different species suggests a selective pressure to maintain these sequences at specific genomic regions (e.g., genome islands, resistance islands, CpG islands, and PAIs) (*2*, *15*, *16*). The possibility of interactions between G4 structures and pathogens has been suggested, although this field of study is still in its nascent phase. Some studies observed that bacterial genomes possess G4-forming sequences within their genome regions (*17*, *18*). G4 structures are formed by G-rich DNA sequences, and their stability is influenced by the G+C content and arrangement of G tetrads. Interestingly, PAIs often exhibit an altered GC content, putatively contributing to the propensity of G4 structure formation within these regions. The G4 structures in PAIs might modulate the accessibility of transcription factors, DNA-binding proteins, or RNA polymerase in pathogens, as documented in eukaryotes (*2*, *19*), thereby influencing the expression of virulence-associated genes (*20*). The formation of G4 structures within PAIs may serve as an additional layer of regulation that fine-tunes the expression of genes critical for pathogenesis. Hence, the investigation of G4 structures within PAIs may open new avenues for the development of therapeutic strategies aimed at disrupting the regulatory mechanisms of pathogenicity-associated genes.

Results

Genomic information, PAI patterns, and the presence of G4 structures in 89 reported pathogenic strains

A dataset of PAIs was compiled from 89 reported pathogenic strains of bacteria, encompassing 222 distinct types of PAIs. Pathogens exhibiting similar PAIs displayed closely clustered patterns on phylogenetic branches, such as LEE in *E. coli* strains (Fig. 2A). Additional information, including the genome length (bp), G+C content (%), rRNA density, tRNA density, and PAI length (bp), was present and showed conserved patterns in the same species (Fig. 2A & Table S1). PAIs commonly exhibit mosaic-like patterns, exemplified by the presence of distinct PAIs like FPI in *Francisella tularensis*, SaPIbov in *Staphylococcus aureus*, and Hrp PAI in *Xanthomonas campestris* (Fig. 2B). Many PAIs were present associated with tRNAs, such as the insertions of $tRNA^{Thr}$, $tRNA^{Phe}$, and $tRNA^{Gly}$ in *E. coli* strains (Fig. 2B & Table S2). The presence of PAIs distributes in similar genomic regions across different pathogens or strains, showing non-random patterns and functionally clustered. Employing the G4Hunter search algorithm, the study identified a total of 225,376 putative G4 sequences in these 89 pathogenic genomes (Table S1). The heatmap also showed that the number of G4 structures was diverse in the pathogen genomes (Fig. 2C).

Interaction between PAIs and G4 structures in different genomic features

The analysis of G4 structures across all pathogen species demonstrated a positive correlation between the number of G4 structures and the GC content in various genomic features, including the whole genome, gene, promoter, rRNA, and tRNA

regions (Fig. 2D). The frequency of G4 structures, measured as the frequency of predicted G4-forming sequences per 1000 base pairs (bp), also showed a positive correlation with the GC content across the analyzed genomic elements (Fig. 2E). A G4 score of 1.4 and 1.6 consistently supported a positive correlation between the number and frequency of G4 structures and the GC content across diverse genomic features (Fig. S1). Additionally, this study observed that the GC contents in the genome region were significantly higher compared to the corresponding PAIs region that was classified into five parts according to the genome datasets (Fig. 3A-E). Nonetheless, this study noted a unique pattern in the frequency of G4 structures within diverse regions of the PAIs, particularly in regions with GC contents less than 30% and greater than 60%.

Putative functions of G4 structures in PAIs

The study used *E. coli* as an example to investigate the potential regulatory role and function of genes covered by G4 structures in PAIs. *E. coli* contains at least ten types of PAIs in different strains, and one of the well-known PAIs is LEE (Fig. 3F), harboring genes responsible for causing attaching and effacing lesions (*21*, *22*). One stable G4 structure with a G4Hunter score of 1.6 was identified at position 37,085 in the LEE PAI of *E. coli* str. O103:H2 12009 (Fig. 3G), located between an IS element and a tRNA insertion site. The tRNA region generally contains a higher G4 frequency compared with transfer-messenger RNA (tmRNA) and rRNA regions in the bacterial genome (*23*). Interestingly, this G4 structure found in *E. coli* str. O103:H2 12009 was present in close proximity to a tRNA region, suggesting a potential regulatory role of

G4 structures in the tRNA gene, or upstream- and downstream-genes that are responsible for LEE virulence. Additionally, another stable G4 sequence with a score of 1.381 was discovered at position 12,457 in the *E. coli* str. CFT073 PAI II to provide more evidence of G4 in PAI regions (Fig. 3H). Functional enrichment analysis was conducted to explore the putative functions of G4-covered genes in the two *E. coli* strains (Table S3 & S4). The results revealed that the genes covered by G4 structures were predominantly involved in genetic information processes, including DNA binding, DNA integration, and nucleic acid metabolism processes (Fig. 3I & J).

Discussion

This study found that the non-random distribution of G4 structures within PAIs across different bacterial species, signifies a potential regulatory role in bacterial pathogenicity. The conservation of G4 structures within the same pathogenic strains suggests a crucial and possibly conserved function in regulating pathogenic traits. The findings are similar to previous reports that showed that the G4 structures display uneven distribution patterns in eukaryotic and prokaryotic genomes and are conserved evolutionary groups (*23-25*). To understand the origin of G4 structures within PAIs, we hypothesized that these G4 sequences could be acquired through three types of horizontal gene transfer mechanisms: conjugation, transformation, and transduction (Fig. 3K). These mechanisms serve as means for genetic material exchange between different organisms. Considering the presence of G4 sequences within the PAIs, it is plausible that these sequences are transferred along with the PAIs through these

horizontal gene transfer mechanisms. Additionally, the presence of G4 structures within the promoter, rRNA, and tRNA regions may have functional implications for the regulation of DNA replication, ribosome biogenesis, protein synthesis, and other RNA-related processes (*6, 26, 27*). Throughout evolution, there seems to be a greater frequency of G4 structures in regulatory genes, such as the tRNA region, compared to other genes, enabling intricate control of gene expression in signal transduction pathways (*28*).

The study found that the genomic regions surrounding the PAIs (i.e., core genome) tend to have a higher GC content than PAI regions, which was consistent with the fact that PAIs often exhibit distinct base composition compared with the core genome (*7*). The variation was explained by the presence of G4 sequences within the PAIs, whereas the results were surprising. This study observed a distinct pattern in the frequency of G4 structures within different regions of the PAIs. This differential distribution of G4 structures suggests that (i) specific genomic segments within the PAIs may be more prone to induce G4 formation discrepancy; (ii) the variation of base composition between core genome and PAIs is partially correlated with the presence of G4 structures; (iii) the frequency of G4 structures in PAIs present stable as the core genome in the most situation; (iv) an alternative hypothesis, other factors, such as i-motif (i.e., the anti-G4 structure) and CpG island, may work synergistically with G4 and potentially contribute the base composition variation (*29, 30*).

Enrichment analysis indicated a predominant involvement of these G4-covered genes in genetic information processes, encompassing DNA binding, DNA integration,

and nucleic acid metabolism. This suggests that G4 structures may play a regulatory role in these essential cellular processes, especially gene expression and DNA-related functions. For instance, G4 structures in the promoter regions of certain transcription factors may influence their binding affinity to DNA and subsequently affect downstream gene expression patterns (*31, 32*). These elements frequently utilize DNA integration mechanisms mediated by integrases, recombinases, or transposases to transfer or incorporate genetic material into the bacterial genome (*33, 34*). One compelling illustration is a study that identified a 16-base pair cis-acting G4 sequence near the pilin locus in *Neisseria gonorrhoeae*, demonstrating its pivotal role in antigenic variation and directing recombination to a specific chromosomal locus (*20, 35*). Disruption of the G4 structure in this context impeded pilin antigenic variation and recombination, highlighting its significance in immune evasion mechanisms. Additionally, considering the distance between G4 structures and the beginning site of gene (e.g., transcription start site (TSS)) in the analysis of promoter regions is pivotal for a comprehensive understanding of their regulatory impact on gene expression (*5*). The spatial proximity to the TSS influences interactions with regulatory elements, potentially modulating the binding of transcription factors and RNA polymerase. This spatial relationship affects accessibility, with G4 structures closer to the TSS potentially acting as direct impediments to transcription initiation. Acknowledging these spatial nuances would provide crucial insights into the functional implications of G4 structures in promoters.

Overall, the conserved evolutionary relatedness of PAIs, the detection of stable G4 structures in specific genomic positions, and the enrichment of G4-covered genes in genetic information processes collectively support the hypothesis that G4 structures may have regulatory functions in key biological processes in pathogens. However, it is important to acknowledge and address certain limitations that could potentially affect the interpretation of the results. One such limitation is the reliance on genome sequences obtained from external laboratories and datasets, which introduces a level of uncertainty regarding the accuracy and completeness. Furthermore, the dynamic nature of bacterial genomes, including genetic rearrangements and horizontal gene transfer events, can complicate the accurate assembly and annotation of genome sequences. Lastly, the stability of G4 structures seems to be important for their function according to recent evidence (*36*). Hence, exploring the relationship between G4 stability and function is a valuable and intriguing topic that could provide insights into the nuanced ways G4 structures contribute to cellular processes and potentially offer new avenues for therapeutic interventions or molecular engineering. To overcome these constraints, fostering collaboration among research teams and participating in data-sharing endeavors becomes imperative to guarantee access to high-quality genome data for exhaustive analyses. Moreover, it is crucial to interpret the results with caution and continue refining this understanding through validation experiments and collaborative efforts.

Methods

Selection and extraction of DNA Sequences

A total of 89 genomes corresponding to the identified pathogens from the Pathogenicity Island Database (PAIDB) were included in the study. The complete bacterial genomic DNA sequences and their corresponding annotation files in .gff and .fna formats were obtained from the Genome database of the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/genome). To ensure the reliability and completeness of the dataset, only completely assembled genomes were included in the analysis. To avoid redundancy and incomplete sequences, one representative genome was selected for each species or strain. The selection of representative genomes was based on a careful examination of the supplementary material (Table S1) accompanying the study. TBtools (https://cj-chen.github.io/tbtools), a versatile bioinformatics tool with extensive applications in both eukaryotes and prokaryotes (*37*, *38*), was employed for extracting genomic sequences. This tool facilitated the retrieval of gene regions, promoters (2 kb upstream of the genes), tRNA regions, and rRNA regions from the selected genomes. PAI regions were downloaded following previously documented information in PAIDB (Table S1&S2). Default thresholds and parameters were applied during extraction to maintain consistency across all genomes.

Data process and detection of G4 structures in genomic features

The G4Hunter algorithm, a widely used tool for G4 prediction, was employed for the identification of G4 motifs in the genomic sequences (*39*). The G4Hunter parameters were set to a window size of "25" and a G4 score threshold of 1.2, which ensured the

identification of potential G4 sequences (*23, 40*). The study additionally utilized G4 scores of 1.4 and 1.6 as a means of cross-verification for the results. The study quantified the predicted number of putative G4-forming sequences within different genomic features, including the whole genome, gene, promoter, tRNA, rRNA, and PAI regions. The density of G4 motifs was determined by dividing the number of G4 sequences by the total length of the genome, while the length ratio of G4 motifs was calculated by dividing the total length of the G4 sequences by the total length of the genome.

Relationship between G4 structures and PAIs

The heatmap was used to show the distribution of G4 motifs in the genome divided by ten parts as PAI regions using R package “pheatmap”. The correlation between the number of G4 structures and the GC content was analyzed across various genomic elements, including the whole genome, gene, promoter, rRNA, and tRNA regions. The analysis utilized the *R*-squared value (R^2) to determine the fit goodness of the correlation. The correlation's significance was evaluated through *p*-values along with a 95% confidence interval. Subsequently, a ROC analysis, yielding an area greater than 0.90, was employed to quantify sensitivity and specificity. The GC content in the genome regions and corresponding PAI regions was compared and classified into different ranges to explore the variation in base composition. GraphPad Prism (V.5.02, GraphPad Software, Inc) was employed to conduct Normality and Lognormality Tests. The K-S test and F-test were used to assess normal distribution and variances, and Student’s t-test was used to identify significant difference.

Phylogenetic Tree Construction

The exact Taxonomy ID (taxid) for each analyzed group was obtained from the NCBI Taxonomy Database using the Taxonomy Browser. The Neighbor-Joining (NJ) method was employed to construct the phylogenetic trees for the analyzed groups. The phylogenetic trees were generated using MEGA11 software (www.megasoftware.net), which offers robust algorithms and comprehensive tools for phylogenetic analysis. To assess the reliability and statistical support of the phylogenetic tree branches, bootstrap analysis was performed. One thousand bootstrap replicates were used to estimate the confidence levels of the branching patterns in the phylogenetic trees. The phylogenetic trees, along with the bootstrap support values, were displayed and visualized using the Interactive Tree of Life (ITOL) platform (https://itol.embl.de/).

Gene functional annotation

The gene sequences covered by G4 structures within PAIs were subjected to gene ontology (GO) annotation (https://geneontology.org/). The gene sequences were translated into protein sequences using the Expasy online toolkit (https://web.expasy.org/translate/). This tool performs the translation based on the standard genetic code, converting the DNA nucleotide sequence into its corresponding amino acid sequence. The GO annotation database assigned GO terms to the protein sequences based on their predicted functions and known biological process (BP), molecular function (MF), and cellular component (CC). Fisher's exact test was employed to determine the statistical significance of the enrichment results. The

obtained *p*-values indicated the overrepresentation of specific GO terms, with lower p-values suggesting higher significance.

Statistics and reproducibility

All genomic data utilized in this study, including the species-specific datasets, were obtained from publicly available sources. Statistical analyses, such as the Kruskal-Wallis test, F test, Student's t-test, Wilcoxon test, and chi-square test, were performed using GraphPad Prism software. The samples used in the statistical analyses corresponded to the genomic data, PAIs, or specific genes under investigation.

Acknowledgment

The sincere appreciation extends to Dr. Sung Ho Yoon and his colleagues for their dedicated efforts in identifying PAIs and establishing the Pathogenicity Island Database for public analysis. Their commitment to advancing the field of pathogen genomics has greatly facilitated this research. This study would like to thank Dr. Jingjing Li (Zhejiang University) and Dr. Mingyu Zhou (Sun Yat-Sen University) for their insightful suggestions and constructive comments regarding the exploration of G4 structures in genomes. Their expertise and guidance have significantly enriched the understanding of the potential roles and implications of G4 structures in the context of PAIs.

Competing Interests

The authors declare no competing interests.

Funding information

This research received no external funding.

Credit author statement

Bo Lyu: writing original draft, data analysis, and visualization.

Qisheng Song: writing original draft-reviewing and supervision.

Ethical approval

Ethical review and approval were not required for this study.

Availability of data and materials

The original reported PAIs datasets analyzed in this study are available from the publication with the DOI: 10.1093/nar/gku985. Additionally, supplementary materials accompanying this article provide further PAIs data analyzed in the study. These supplementary materials contain detailed information and datasets that support the findings and conclusions presented in this research. Researchers and interested individuals can access supplementary materials to explore and validate the results obtained in this study.

Consent for publication

Not applicable.

1. J. D. Watson, F. Crick, A structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

2. D. Rhodes, H. J. Lipps, G-quadruplexes and their regulatory roles in biology. *Nucleic acids research* **43**, 8627-8637 (2015).

3. J. Spiegel, S. Adhikari, S. Balasubramanian, The structure and function of DNA G-quadruplexes. *Trends in Chemistry* **2**, 123-136 (2020).

4. O. Stegle, L. Payet, J.-L. Mergny, D. J. MacKay, J. L. Huppert, Predicting and understanding the stability of G-quadruplexes. *Bioinformatics* **25**, i374-i1382 (2009).

5. J. L. Huppert, Structure, location and interactions of G□quadruplexes. *The FEBS journal* **277**, 3452-3458 (2010).

6. B. Zybailov, M. Sherpa, G. Glazko, K. Raney, V. Glazko, G4-quadruplexes and genome instability. *Molecular Biology* **47**, 197-204 (2013).

7. H. Schmidt, M. Hensel, Pathogenicity islands in bacterial pathogenesis. *Clinical microbiology reviews* **17**, 14-56 (2004).

8. E. A. Groisman, H. Ochman, Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* **87**, 791-794 (1996).

9. J. Hacker, J. B. Kaper, Pathogenicity islands and the evolution of microbes. *Annual Reviews in Microbiology* **54**, 641-679 (2000).

10. J. A. Kers *et al.*, A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. *Molecular Microbiology* **55**, 1025-1033 (2005).

11. S. Kombade, N. Kaur, Pathogenicity island in *Salmonella spp.*-A Global Challenge. (IntechOpen, 2021).

12. E. Carniel, The *Yersinia* high-pathogenicity island. *International Microbiology* **2**, 161-167 (1999).

13. M. Syvanen, Evolutionary implications of horizontal gene transfer. *Annual Review of Genetics* **46**, 341-358 (2012).

14. J. Chen *et al.*, Intra- and inter-generic transfer of pathogenicity island-encoded virulence genes by cos phages. *The ISME Journal* **9**, 1260-1263 (2015).

15. S. Frees, C. Menendez, M. Crum, P. S. Bagga, QGRS-Conserve: a computational method for discovering evolutionarily conserved G-quadruplex motifs. *Human Genomics* **8**, 8 (2014).

16. S. L. König, A. C. Evans, J. L. Huppert, Seven essential questions on G-quadruplexes. *Biomolecular Concepts* **1**, 197-213 (2010).

17. P. Yadav *et al.*, G-quadruplex structures in bacteria: Biological relevance and potential as an antimicrobial target. *Journal of Bacteriology* **203**, 10.1128/jb. 00577-00520 (2021).

18. L. M. Harris, C. J. Merrick, G-quadruplexes in pathogens: a common route to virulence control? *PLoS Pathogens* **11**, e1004562 (2015).

19. D. Varshney, J. Spiegel, K. Zyner, D. Tannahill, S. Balasubramanian, The regulation and functions of DNA and RNA G-quadruplexes. *Nature Reviews Molecular Cell Biology* **21**, 459-474 (2020).

20. L. A. Cahoon, H. S. Seifert, An alternative DNA structure is necessary for pilin antigenic variation in *Neisseria gonorrhoeae*. *Science* **325**, 764-767 (2009).

21. F. M. Franzin, M. P. Sircili, Locus of enterocyte effacement: a pathogenicity island involved in the virulence of enteropathogenic and enterohemorragic *Escherichia coli* subjected to a complex network of gene regulation. *BioMed Research International* **2015**, 534738 (2015).

22. J. Jores, L. Rumer, L. H. Wieler, Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*. *International Journal of Medical Microbiology* **294**, 103-113 (2004).

23. M. Bartas *et al.*, The presence and localization of G-quadruplex forming sequences in the domain of bacteria. *Molecules* **24**, 1711 (2019).

24. Z. Du, Y. Zhao, N. Li, Genome-wide colonization of gene regulatory elements by G4 DNA motifs. *Nucleic Acids Research* **37**, 6784-6798 (2009).

25. E. P. Lombardi, A. Londoño-Vallejo, A. Nicolas, Relationship between G-quadruplex sequence composition in viruses and their hosts. *Molecules* **24**, 1942 (2019).

26. P. Ivanov *et al.*, G-quadruplex structures contribute to the neuroprotective effects of angiogenin-induced tRNA fragments. *Proceedings of the National Academy of Sciences* **111**, 18201-18206 (2014).

27. S. Mestre-Fos *et al.*, G-quadruplexes in human ribosomal RNA. *Journal of Molecular Biology* **431**, 1940-1955 (2019).

28. F. Wu *et al.*, Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. *Communications Biology* **4**, 98 (2021).

29. A. M. Deaton, A. Bird, CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-1022 (2011).

30. N. Sushmita *et al.*, I-motif DNA: significance and future prospective. *Exploratory Animal and Medical Research* **10**, 18-23 (2020).

31. K. Niu *et al.*, BmILF and i-motif structure are involved in transcriptional regulation of BmPOUM2 in *Bombyx mori*. *Nucleic Acids Research* **46**, 1710-1723 (2018).

32. L. Xiang *et al.*, DNA G-quadruplex structure participates in regulation of lipid metabolism through acyl-CoA binding protein. *Nucleic Acids Research* **50**, 6953-6967 (2022).

33. I. R. Arkhipova, P. A. Rice, Mobile genetic elements: in silico, in vitro, in vivo. *Mol Ecol* **25**, 1027-1031 (2016).

34. R. A. Wozniak, M. K. Waldor, Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology* **8**, 552-563 (2010).

35. L. A. Cahoon, H. S. Seifert, Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in *Neisseria gonorrhoeae*. *PLoS Pathogens* **9**, e1003074 (2013).

36. M. Jara□Espejo, S. R. Line, DNA G□quadruplex stability, position and chromatin accessibility are associated with CpG island methylation. *The FEBS journal* **287**, 483-495 (2020).

37. C. Chen *et al.*, TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant* **13**, 1194-1202 (2020).

38. C. Chen *et al.*, TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Molecular Plant* **16**, 1733-1742 (2023).

39. V. Brázda *et al.*, G4Hunter web application: a web server for G-

quadruplex prediction. *Bioinformatics* **35**, 3493-3495 (2019).

40. V. Brázda *et al.*, G-Quadruplexes in the archaea domain. *Biomolecules* **10**, 1349 (2020).

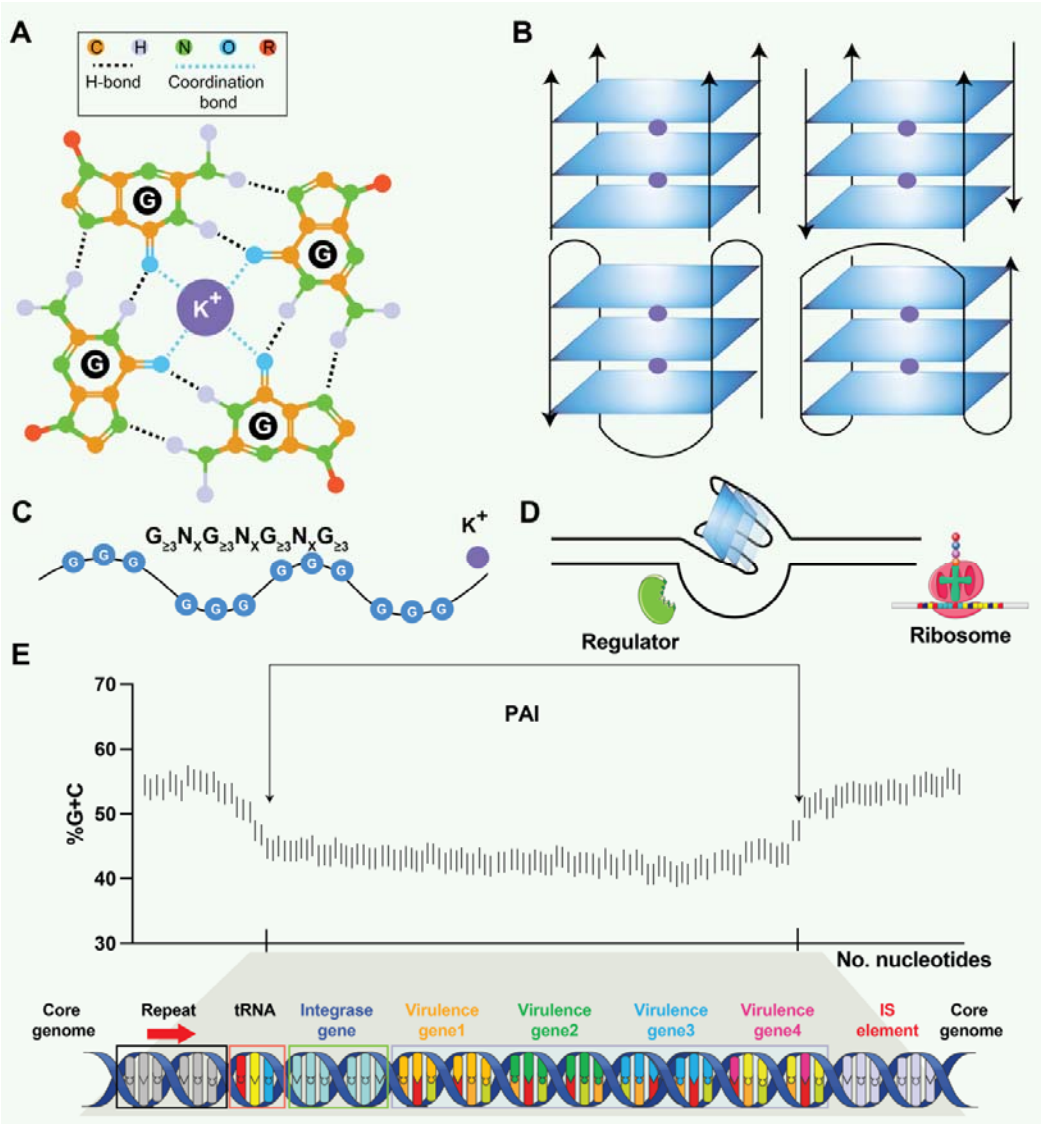


Figure 1. Structural and Functional Aspects of G-quadruplex (G4) Structures and Pathogenicity Islands (PAIs). (A) Schematic representation of a guanine tetrad stabilized by Hoogsten base pairing and a positively charged central ion, illustrating the key elements of G4 structures. (B) Structural heterogeneity of G4 structures. G4 structures exhibit polymorphism and can be categorized into different families, such as parallel or antiparallel, based on the orientation of the DNA strands. They can fold either intramolecularly or intermolecularly, leading to diverse structural

configurations. (C) General sequence formula for G4, highlighting the repeated occurrence of guanine-rich sequences that form G4 structures. (D) Regulatory roles of G4 in transcription. G4 can regulate transcription by blocking RNA polymerase from binding to promoter sequences or aiding in single-stranded DNA (ssDNA) formation, thereby enhancing transcription. (E) General structure of PAI. PAIs are characteristic regions of DNA found within the genomes of pathogenic bacteria, distinguishing them from nonpathogenic strains of the same or related species. Repeat sequences are DNA segments duplicated within the PAI and can serve as recognition sites for various enzymes involved in the integration and excision of the PAI from the bacterial chromosome. tRNA genes act as anchor points for the insertion of foreign DNA acquired through horizontal gene transfer. Virulence genes encode proteins or factors that play crucial roles in the virulence and pathogenicity of the bacterium, contributing to adhesion, invasion, immune evasion, toxin production, or other pathogenic mechanisms. Insertion elements include transposons, bacteriophages, or plasmids, enabling the PAI to be transferred between bacterial cells and potentially disseminated to different strains or species.

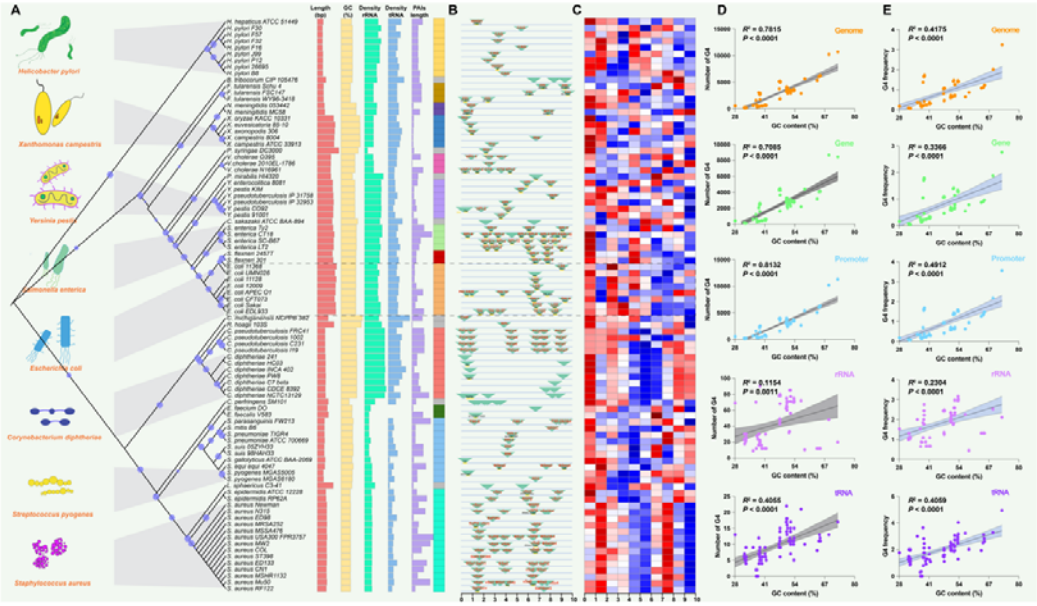


Figure 2. Analysis of Pathogenicity Islands (PAIs) and G-quadruplexes (G4) in Pathogen Genomes. (A) Phylogenetic analysis of pathogen genomes based on 89 bacterial strains, showing the evolutionary relationships among species. Additional genomic information, including genome size, GC content, rRNA density, tRNA density, and PAI length, is provided. The same color indicates the same species. (B) Genomic location of specific PAIs in bacterial genomes, divided into ten regions. PAIs are represented by green triangles, and their names are indicated. The tRNA insertion sites are also marked. (C) Heatmap illustrating the relative abundance of G4 structures in bacterial genomes, divided into ten regions. Red indicates a higher relative abundance, while blue indicates a lower relative abundance. (D & E) Correlation analysis between the number of G4 structures, the frequency of G4 structures and GC content in various genomic features, including the whole genome, genes, promoters, rRNA, and tRNA.

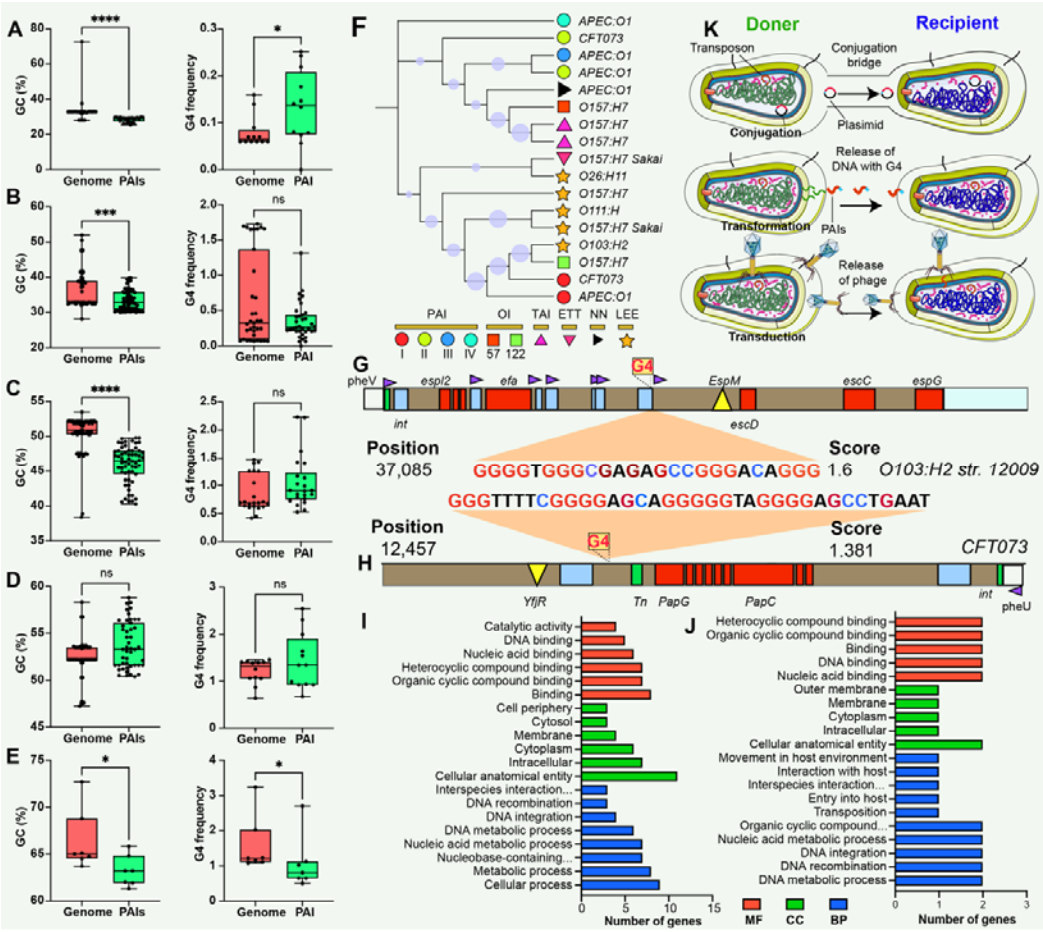
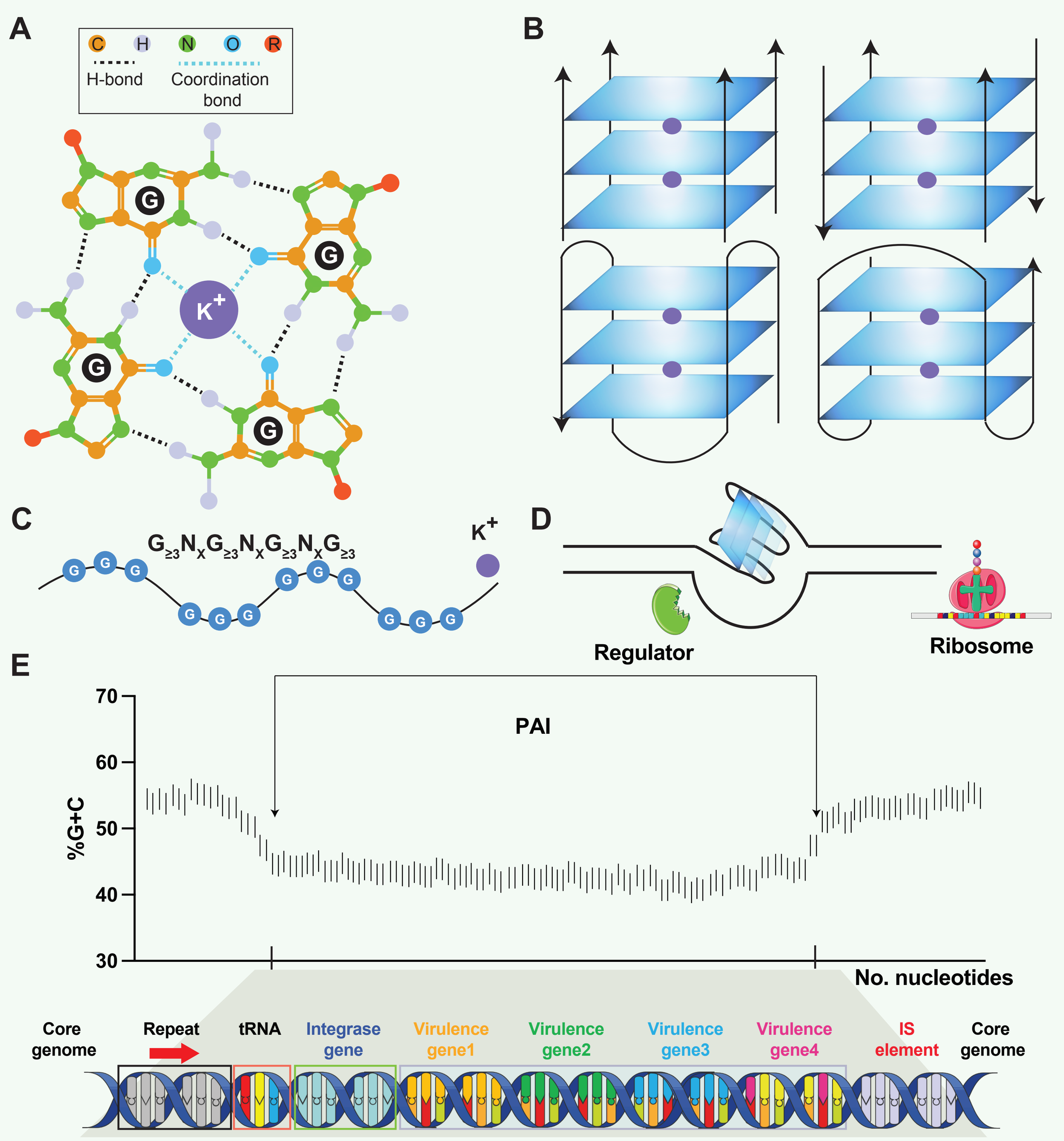
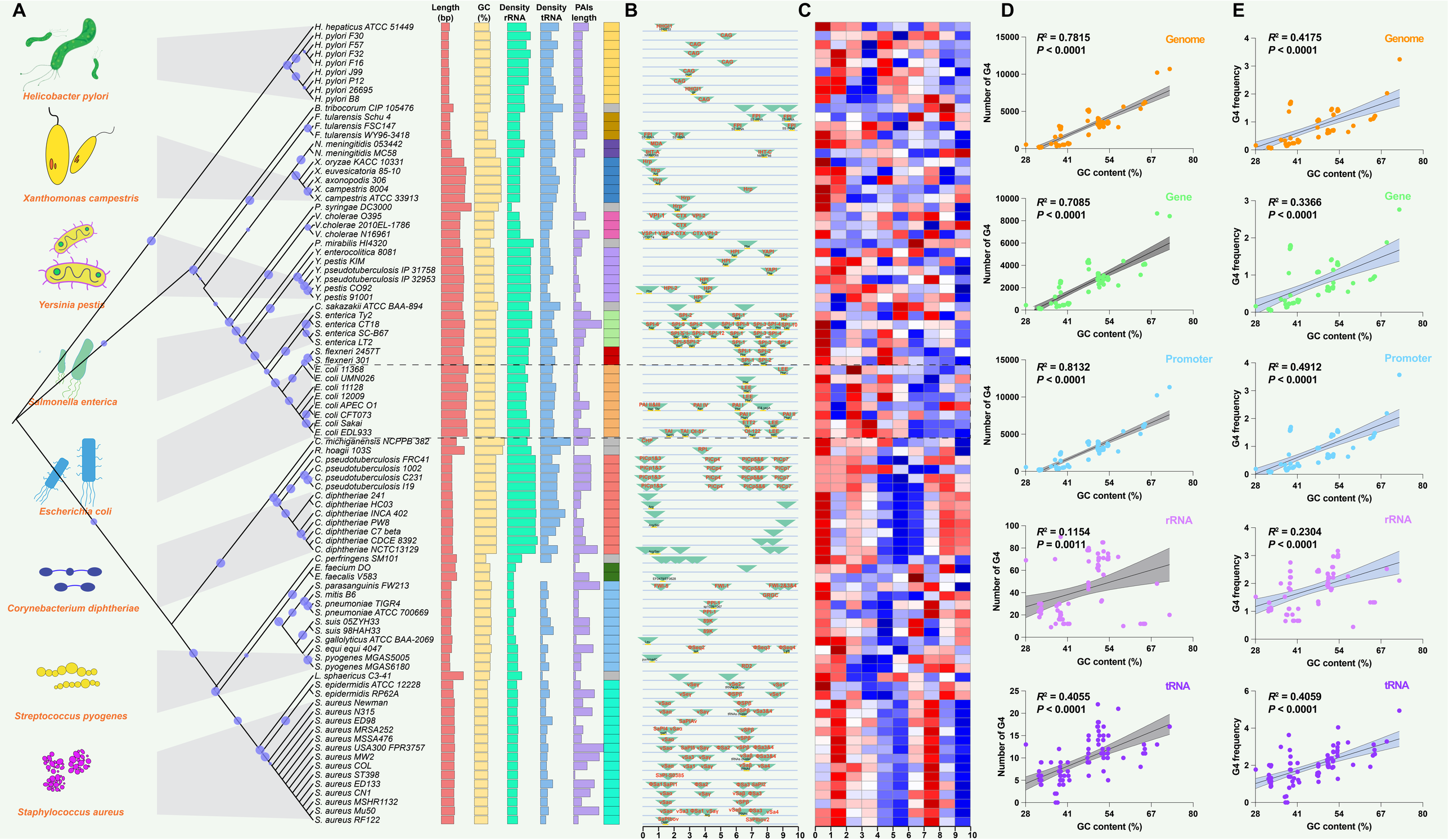
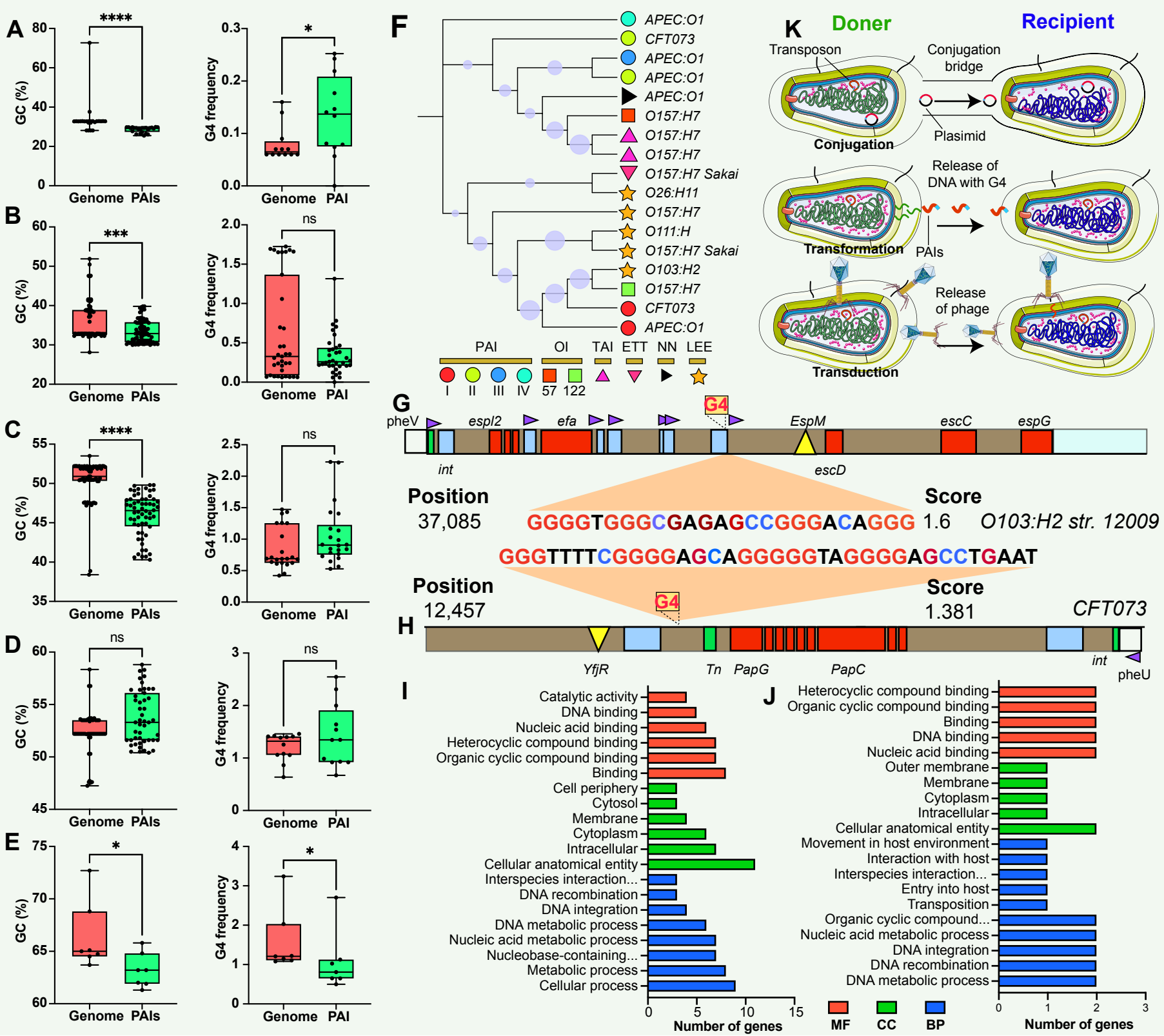


Figure 3. Comparison and Functional Annotation of G-quadruplexes (G4) within Pathogenicity Islands (PAIs). (A-E) Comparison of GC content (left panel) and GC frequency (right panel) between the genome and PAIs, categorized into five regions (20%-30%, 30%-40%, 40%-50%, 50%-60%, and 60%-70%). */**/***/**** indicates significant difference ($p < 0.05/0.01/0.001/0.0001$). (F) Evolutionary relatedness of 10 types of PAIs (categorized into six main categories) in *E. coli* strains. (G & H) Examples of G4 structures within PAIs in *E. coli* strains. The grey bar represents the virulence region, the red box indicates a virulence gene, the blue box represents an insertion site region or repeat, the green box denotes an integrase, the purple triangle indicates a tRNA insertion site, and the yellow triangle indicates an effector. (I &J)

Functional annotation analysis of G4-covered genes within PAIs in two *E. coli* strains, including biological process (BP), cellular component (CC), and molecular function (MF) categories. (K) Hypotheses on the origin of G4 structures within PAIs, involving gene horizontal transfer mechanisms (conjugation, transduction, and transformation).


A
C
H
N
O
R
H-bond
Coordination bond
G
K+
B
C
G≥3NXG≥3NXG≥3NXG≥3
K+
D
Regulator
Ribosome
E
PAI
%G+C
70
60
50
40
30
No. nucleotides
Core genome
Repeat
tRNA
Integrase gene
Virulence gene1
Virulence gene2
Virulence gene3
Virulence gene4
IS element
Core genome


A
Helicobacter pylori
Xanthomonas campestris
Yersinia pestis
Salmonella enterica
Escherichia coli
Corynebacterium diphtheriae
Streptococcus pyogenes
Staphylococcus aureus
Length (bp)
GC (%)
Density rRNA
Density tRNA
PAIs length
B
C
D
E
Genome
Gene
Promoter
rRNA
tRNA
Number of G4
G4 frequency
GC content (%)
R² = 0.7815
P < 0.0001
R² = 0.4175
P < 0.0001
R² = 0.7085
P < 0.0001
R² = 0.3366
P < 0.0001
R² = 0.8132
P < 0.0001
R² = 0.4912
P < 0.0001
R² = 0.1154
P = 0.0011
R² = 0.2304
P < 0.0001
R² = 0.4055
P < 0.0001
R² = 0.4059
P < 0.0001


A
B
C
D
E
F
G
H
I
J
K
GC (%)
G4 frequency
Genome
PAIs
PAI
****
***
*
ns
APEC:O1
CFT073
O157:H7
O157:H7 Sakai
O26:H11
O111:H
O103:H2
PAI
OI
TAI
ETT
NN
LEE
I
II
III
IV
57
122
Doner
Recipient
Transposon
Conjugation bridge
Conjugation
Plasmid
Release of DNA with G4
Transformation
PAIs
Release of phage
Transduction
pheV
espI2
efa
G4
EspM
escC
espG
int
escD
Position
37,085
Score
GGGGTGGGCGAGAGCCGGGACAGGG 1.6
O103:H2 str. 12009
GGGTTTTCGGGGAGCAGGGGGTAGGGGAGCCTGAAT
Position
12,457
Score
1.381
CFT073
YfjR
Tn
PapG
PapC
int
pheU
Catalytic activity
DNA binding
Nucleic acid binding
Heterocyclic compound binding
Organic cyclic compound binding
Binding
Cell periphery
Cytosol
Membrane
Cytoplasm
Intracellular
Cellular anatomical entity
Interspecies interaction...
DNA recombination
DNA integration
DNA metabolic process
Nucleic acid metabolic process
Nucleobase-containing...
Metabolic process
Cellular process
Number of genes
Outer membrane
Movement in host environment
Interaction with host
Entry into host
Transposition
Organic cyclic compound...
MF
CC
BP