

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

Reviewed Preprint

Revised by authors after peer review.

[About eLife's process](#)

Reviewed preprint version 2

February 7, 2024 (this version)

Reviewed preprint version 1

December 5, 2023

Posted to preprint server

September 14, 2023

Sent for peer review

September 13, 2023

Bo Lyu , Qisheng Song 

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

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

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.

eLife assessment

This **fundamental** study explores the relationship between guanine-quadruplex structures and pathogenicity islands in 89 bacterial strains representing a range of pathogens. Guanine-quadruplex structures were found to be non-randomly distributed within pathogenicity islands and conserved within the same strains. These **compelling** findings shed light on the molecular mechanisms of Guanine-quadruplex structure-pathogenicity island interactions and will be of interest to all microbiologists.

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

(20, 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*, SaPI_{bov} 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

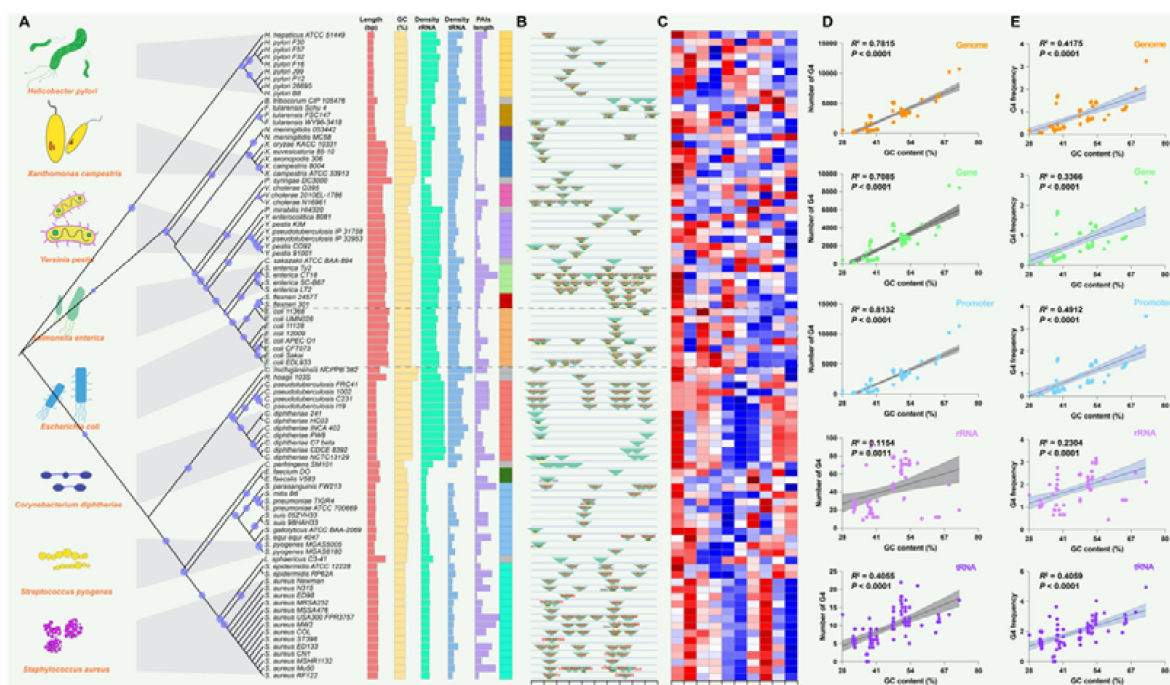


Figure 2.

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.

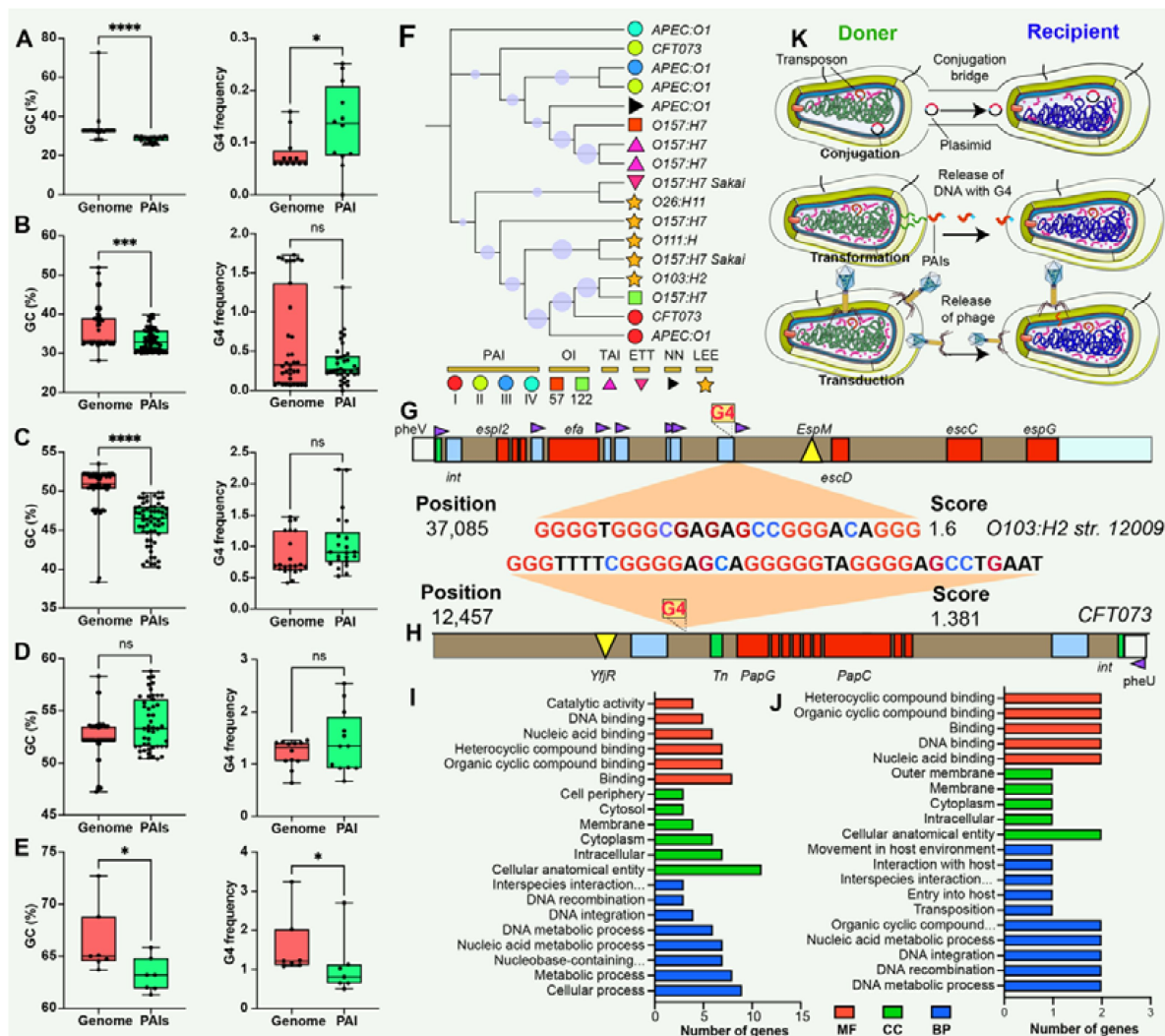


Figure 3.

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).

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.

Acknowledgements

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.

References

1. Watson J. D., Crick F. (1953) **A structure for deoxyribose nucleic acid** *Nature* **171**:737–738
2. Rhodes D., Lipps H. J. (2015) **G-quadruplexes and their regulatory roles in biology** *Nucleic acids research* **43**:8627–8637
3. Spiegel J., Adhikari S., Balasubramanian S. (2020) **The structure and function of DNA G-quadruplexes** *Trends in Chemistry* **2**:123–136
4. Stegle O., Payet L., Mergny J.-L., MacKay D. J., Huppert J. L. (2009) **Predicting and understanding the stability of G-quadruplexes** *Bioinformatics* **25**:i374–i1382
5. Huppert J. L. (2010) **Structure, location and interactions of GLquadruplexes** *The FEBS journal* **277**:3452–3458
6. Zybilov B., Sherpa M., Glazko G., Raney K., Glazko V. (2013) **G4-quadruplexes and genome instability** *Molecular Biology* **47**:197–204
7. Schmidt H., Hensel M. (2004) **Pathogenicity islands in bacterial pathogenesis** *Clinical microbiology reviews* **17**:14–56
8. Groisman E. A., Ochman H. (1996) **Pathogenicity islands: bacterial evolution in quantum leaps** *Cell* **87**:791–794
9. Hacker J., Kaper J. B. (2000) **Pathogenicity islands and the evolution of microbes** *Annual Reviews in Microbiology* **54**:641–679
10. Kers J. A., et al. (2005) **A large, mobile pathogenicity island confers plant pathogenicity on Streptomyces species** *Molecular Microbiology* **55**:1025–1033
11. Kombade S., Kaur N. (2021) **Pathogenicity island in Salmonella spp** *A Global Challenge*
12. Carniel E. (1999) **The Yersinia high-pathogenicity island** *International Microbiology* **2**:161–167
13. Syvanen M. (2012) **Evolutionary implications of horizontal gene transfer** *Annual Review of Genetics* **46**:341–358
14. Chen J., et al. (2015) **Intra- and inter-generic transfer of pathogenicity island-encoded virulence genes by cos phages** *The ISME Journal* **9**:1260–1263
15. Frees S., Menendez C., Crum M., Bagga P. S. (2014) **QGRS-Conserve: a computational method for discovering evolutionarily conserved G-quadruplex motifs** *Human Genomics* **8**
16. König S. L., Evans A. C., Huppert J. L. (2010) **Seven essential questions on G-quadruplexes** *Biomolecular Concepts* **1**:197–213
17. Yadav P., et al. (2021) **G-quadruplex structures in bacteria: Biological relevance and potential as an antimicrobial target** *Journal of Bacteriology* **203** <https://doi.org/10.1128/jb.00577-00520>

18. Harris L. M., Merrick C. J. (2015) **G-quadruplexes in pathogens: a common route to virulence control?** *PLoS Pathogens* **11**
19. Varshney D., Spiegel J., Zyner K., Tannahill D., Balasubramanian S. (2020) **The regulation and functions of DNA and RNA G-quadruplexes** *Nature Reviews Molecular Cell Biology* **21**:459–474
20. Cahoon L. A., Seifert H. S. (2009) **An alternative DNA structure is necessary for pilin antigenic variation in *Neisseria gonorrhoeae*** *Science* **325**:764–767
21. Franzin F. M., Sircili M. P. (2015) **Locus of enterocyte effacement: a pathogenicity island involved in the virulence of enteropathogenic and enterohemorrhagic *Escherichia coli* subjected to a complex network of gene regulation** *BioMed Research International* **2015**
22. Jores J., Rumer L., Wieler L. H. (2004) **Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*** *International Journal of Medical Microbiology* **294**:103–113
23. Bartas M., et al. (2019) **The presence and localization of G-quadruplex forming sequences in the domain of bacteria** *Molecules* **24**
24. Du Z., Zhao Y., Li N. (2009) **Genome-wide colonization of gene regulatory elements by G4 DNA motifs** *Nucleic Acids Research* **37**:6784–6798
25. Lombardi E. P., Londoño-Vallejo A., Nicolas A. (2019) **Relationship between G-quadruplex sequence composition in viruses and their hosts** *Molecules* **24**
26. Ivanov P., et al. (2014) **G-quadruplex structures contribute to the neuroprotective effects of angiogenin-induced tRNA fragments** *Proceedings of the National Academy of Sciences* **111**:18201–18206
27. Mestre-Fos S., et al. (2019) **G-quadruplexes in human ribosomal RNA** *Journal of Molecular Biology* **431**:1940–1955
28. Wu F., et al. (2021) **Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution** *Communications Biology* **4**
29. Deaton A. M., Bird A. (2011) **CpG islands and the regulation of transcription** *Genes Dev* **25**:1010–1022
30. Sushmita N., et al. (2020) **I-motif DNA: significance and future prospective** *Exploratory Animal and Medical Research* **10**:18–23
31. Niu K., et al. (2018) **BmILF and i-motif structure are involved in transcriptional regulation of BmPOUM2 in *Bombyx mori*** *Nucleic Acids Research* **46**:1710–1723
32. Xiang L., et al. (2022) **DNA G-quadruplex structure participates in regulation of lipid metabolism through acyl-CoA binding protein** *Nucleic Acids Research* **50**:6953–6967
33. Arkhipova I. R., Rice P. A. (2016) **Mobile genetic elements: in silico, in vitro, in vivo** *Mol Ecol* **25**:1027–1031
34. Wozniak R. A., Waldor M. K. (2010) **Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow** *Nature Reviews Microbiology* **8**:552–563

35. Cahoon L. A., Seifert H. S. (2013) **Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in *Neisseria gonorrhoeae*** *PLoS Pathogens* **9**
36. JaraLEspejo M., Line S. R. (2020) **DNA GLquadruplex stability, position and chromatin accessibility are associated with CpG island methylation** *The FEBS journal* **287**:483–495
37. Chen C., et al. (2020) **TBtools: An integrative toolkit developed for interactive analyses of big biological data** *Molecular Plant* **13**:1194–1202
38. Chen C., et al. (2023) **TBtools-II: A “one for all, all for one” bioinformatics platform for biological big-data mining** *Molecular Plant* **16**:1733–1742
39. Brázda V., et al. (2019) **G4Hunter web application: a web server for G-quadruplex prediction** *Bioinformatics* **35**:3493–3495
40. Brázda V., et al. (2020) **G-Quadruplexes in the archaea domain** *Biomolecules* **10**

Article and author information

Bo Lyu

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

For correspondence: bl3pt@missouri.edu

ORCID iD: [0000-0002-2744-730X](https://orcid.org/0000-0002-2744-730X)

Qisheng Song

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

For correspondence: songq@missouri.edu

Copyright

© 2023, Bo Lyu & Qisheng Song

This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Editors

Reviewing Editor

Bavesh Kana

University of the Witwatersrand, Johannesburg, South Africa

Senior Editor

Bavesh Kana

University of the Witwatersrand, Johannesburg, South Africa

Reviewer #1 (Public Review):

Summary:

This study explores the relationship between guanine-quadruplex (G4) structures and pathogenicity islands (PAIs) in 89 pathogenic strains.

Strengths:

The findings of this study hold significant implications for our understanding of bacterial pathogenicity and the role of guanine-quadruplex (G4) structures:

Molecular Mechanisms of Pathogenicity: The study highlights that G4 structures are not randomly distributed within pathogenicity islands (PAIs), suggesting a potential role in regulating pathogenicity. This insight into the uneven distribution of G4s within PAIs provides a basis for further research into the molecular mechanisms underlying bacterial pathogenicity.

Conservation of G4 Structures: The consistent conservation of G4 structures within the same pathogenic strains suggests that these structures might play a vital and possibly conserved role in the pathogenicity of these bacteria. This finding opens doors for exploring how G4s influence virulence across different pathogens.

Unique Nature of PAIs: The differences in GC content between PAIs and the core genome underscore the unique nature of PAIs. This distinction suggests that factors such as DNA topology and G4 structures might contribute to the specialized functions and characteristics of PAIs, which are often associated with virulence genes.

Regulatory Role of G4s: The identification of high-confidence G4 structures within regulatory regions of *Escherichia coli* implies that these structures could influence the efficiency or specificity of DNA integration events within PAIs. This finding provides a potential mechanism by which G4s can impact the pathogenicity of bacteria.

Weaknesses:

None

Overall, the study provides fundamental insights into the pathogenicity island and conservation of G4 motifs.

<https://doi.org/10.7554/eLife.91985.2.sa1>

Reviewer #2 (Public Review):

Summary: In the manuscript entitled "The Intricate Relationship of G-Quadruplexes and Pathogenicity Islands: A Window into Bacterial Pathogenicity" Bo Lyu explored the interactions between guanine-quadruplex (G4) structures and pathogenicity islands (PAIs) in 89 bacterial genomes through rigorous computational approach. This paper handles an intriguing and complex topic in the field pathogenomics, it has the potential to contribute significantly to the understanding of G4-PAI interactions and bacterial pathogenicity.

Strengths: Chosen research area and summarizing the results through neat illustrations

Weaknesses: I did not find any significant ones.

<https://doi.org/10.7554/eLife.91985.2.sa0>

Author Response

The following is the authors' response to the original reviews.

Reviewer #1 (Public Review):

Summary:

This study explores the relationship between guanine-quadruplex (G4) structures and pathogenicity islands (PAIs) in 89 pathogenic strains. G4 structures were found to be non-randomly distributed within PAIs and conserved within the same strains. Positive correlations were observed between G4s and GC content across various genomic features, suggesting a link between G4 structures and GC-rich regions. Differences in GC content between PAIs and the core genome underscored the unique nature of PAIs. High-confidence G4 structures in Escherichia coli's regulatory regions were identified, influencing DNA integration within PAIs. These findings shed light on the molecular mechanisms of G4-PAI interactions, enhancing our understanding of bacterial pathogenicity and G4 structures in infectious diseases.

Strengths:

The findings of this study hold significant implications for our understanding of bacterial pathogenicity and the role of guanine-quadruplex (G4) structures. Molecular Mechanisms of Pathogenicity: The study highlights that G4 structures are not randomly distributed within pathogenicity islands (PAIs), suggesting a potential role in regulating pathogenicity. This insight into the uneven distribution of G4s within PAIs provides a basis for further research into the molecular mechanisms underlying bacterial pathogenicity.

Conservation of G4 Structures: The consistent conservation of G4 structures within the same pathogenic strains suggests that these structures might play a vital and possibly conserved role in the pathogenicity of these bacteria. This finding opens doors for exploring how G4s influence virulence across different pathogens. Unique Nature of PAIs: The differences in GC content between PAIs and the core genome underscore the unique nature of PAIs. This distinction suggests that factors such as DNA topology and G4 structures might contribute to the specialized functions and characteristics of PAIs, which are often associated with virulence genes. Regulatory Role of G4s: The identification of high-confidence G4 structures within regulatory regions of Escherichia coli implies that these structures could influence the efficiency or specificity of DNA integration events within PAIs. This finding provides a potential mechanism by which G4s can impact the pathogenicity of bacteria.

Weaknesses:

No weaknesses were identified by this reviewer.

Overall, the study provides fundamental insights into the pathogenicity island and conservation of G4 motifs.

Thank you for your thorough review of our manuscript exploring the relationship between G4 structures and PAIs in 89 pathogenic strains. We appreciate your recognition of the strengths of our study and its potential implications for understanding bacterial pathogenicity. We are pleased that you highlighted the significance of our findings in revealing the non-random distribution and conservation of G4 structures within PAIs across various pathogenic strains.

Your insightful comments about the molecular mechanisms of pathogenicity, the conservation of G4 structures, the unique nature of PAIs, and the regulatory role of G4s within Escherichia coli are invaluable. We are encouraged by your positive evaluation of these aspects, which underscores the potential impact of our work on advancing the understanding of bacterial pathogenicity.

Reviewer #2 (Public Review):*Summary:*

In the manuscript entitled "The Intricate Relationship of G-Quadruplexes and Pathogenicity Islands: A Window into Bacterial Pathogenicity" Bo Lyu explored the interactions between guanine-quadruplex (G4) structures and pathogenicity islands (PAIs) in 89 bacterial genomes through a rigorous computational approach. This paper handles an intriguing and complex topic in the field of pathogenomics. It has the potential to contribute significantly to the understanding of G4-PAI interactions and bacterial pathogenicity.

Strengths:

- *The chosen research area.*
- *The summarizing of the results through neat illustrations.*

Weaknesses:

This reviewer did not find any significant weaknesses.

Thank you for your positive and encouraging feedback on our manuscript. We appreciate your specific mention of the strengths, particularly highlighting the chosen research area and the effectiveness of our illustrations in summarizing the results. Your acknowledgment of these aspects is motivating, and we are pleased that the content and presentation resonated well with you.

Reviewer #3 (Public Review):

The main problem with the work is that the results are only descriptive and do not allow any inferences or conclusions about the importance of the function of G4 structures. The discussion and conclusions are poor. The results are preliminary and in order to try to make the analysis more interesting, it should be further extended and the data must be explored in a much greater depth.

Thank you for your constructive feedback on our manuscript, and appreciate the time and effort you dedicated to evaluating our work. We acknowledge your concern regarding the descriptive nature of the results and the limitations in making inferences about the importance of G4 structures. To address this, we plan to enhance the depth of our analysis and provide more insightful interpretations in the discussion and conclusion sections. It's important to note that this study is intentionally a short report, emphasizing data mining findings rather than laboratory results. We understand the value of in-depth investigations and concur that our work lays the groundwork for more extensive studies in this area, aiming to provide a real-world scenario. We are committed to addressing your comments and refining our manuscript to contribute meaningfully to this field. Your insights are invaluable, and we look forward to presenting an improved version of our study.

Reviewer #2 (Recommendations For The Authors):

The authors could try a higher G-quadruplex score of 1.4 or higher values to substantiate their findings or pick up the bacterial genomes that relied on G4s for their pathogenicity.

We acknowledge your recommendation to explore a higher G-quadruplex score, and we would like to assure you that we have already conducted analyses using thresholds of 1.4 and

1.6. The findings consistently support the observations presented in the manuscript. We have updated the text to reflect this additional analysis, and the results are included in the revised version of the manuscript (Figure S1).

Reviewer #3 (Recommendations For The Authors):

Minor points

Introduction

Q1. The introduction is shallow. The concept and the importance of PAIs is vague. Why should these genes be different from other genes?

A1: Thank you for your valuable feedback and we have incorporated additional content to provide a more comprehensive understanding of PAIs and their distinctiveness from other genes in the Introduction section.

Changes: Lines 44-49 “G4 structures are ...innovative technologies.” were added.

Lines 51-55 “PAIs are distinct...such as plasmids.” were added.

Lines 60-66 “PAIs typically contain...recipient genome” were added.

Lines 77-80 “Growing evidence has...CpG islands, and PAIs” were added.

Material and Methods

Q2. It is not clear if the author used the TBTools or the G4Hunter software G4 structures. It would be interesting to include references to published articles that used this software.

A2: Thank you! Corrected and added more references that used TBTools to extract sequences and G4Hunter to identify G4 structures.

Q3. The statistical significance must not be based only on p-values. P-values are influenced by sample sizes. I strongly recommend the use of other parameters such as confidence interval and ROC analysis.

A3: Thank you! We have incorporated confidence intervals and ROC analysis to complement p-values, enhancing the robustness of our statistical analysis.

Changes: Lines 265-267 “The correlation's significance... sensitivity and specificity.” were added.

Results and discussion

Q4. The stability of G4 structures seems to be important for its function (doi:10.1111/febs.15065). Therefore it would be interesting if the analysis were carried out separating the G4 according to stability.

A4: Thank you for highlighting the importance of G4 structure stability for its function and suggesting an analysis based on stability. We have carefully reviewed the referenced paper (doi:10.1111/febs.15065) and note that their study focused on the stability analysis of individual G4s. In our current study, we identified a large number of G4s, and while stability analysis for each G4 is indeed an interesting avenue, it goes beyond the scope of this particular investigation. However, we agree that exploring the relationship between G4 stability and function is a valuable topic. We plan to delve deeper into this aspect in future work, as discussed in our response to your previous comment.

Changes: Lines 217-221 “Lastly, the stability of G4...molecular engineering.” were added.

| *Q5. The quality of the figures is poor. Is not possible to read the correlation and p-values from Figure 2.*

A5: The revised figure is now submitted with enhanced clarity to ensure that correlation and p-values can be easily discerned.

| *Q6. The analysis of promoter regions should be performed taking into account the distance between the G4 and the beginning of the gene.*

A6: Thank you and we have elaborated more in the revision.

Changes: Lines 198-106 “Additionally, considering the distance...of G4 structures in promoters.” were added.

| *Q7. The topic "Putative origin, transfer mechanisms, and functions of G4s in PAIs". The comments made on this topic are purely speculative and not backed up by data or any type of experimental analysis.*

A7: We appreciate the feedback and have revised the title to emphasize the focus on the functions of G4s in PAIs. We acknowledge that the content related to the putative origin and transfer mechanisms of G4s in PAIs is purely descriptive and speculative, we have made the adjustment to relocate this information to the discussion section for a more appropriate treatment.

| *Q8. The supplemental material is hard to follow. The meaning of each column should be better explained. Why was the data divided into 10 parts?*

A8: Following your suggestion, we have revised the tables for better clarity. To address concerns about the division into 10 parts, we have decided to remove this data from the tables as it was deemed unnecessary for presentation.

| *Q9. Why was the data of E. Coli strains 1 and 2 shown in Tables S3 and S4 and the other bacterial strains were not?*

A9: We appreciate your inquiry. The data of E. Coli strains 1 and 2 were specifically highlighted in Tables S3 and S4 as illustrative examples to demonstrate the putative functions of G4s in PAIs within the scope of our study. Given the extensive nature of function annotation analyses across various pathogenic strains, presenting additional tables for each strain would have resulted in an impractical volume of supplementary material.

| *Q10. The Results and Discussion should be separated.*

A10: Thank you! Corrected as suggested.