Exploring the Links between G-Quadruplex Structures and Pathogenicity Islands: Insights from Large-Scale Dataset Analysis

Bo Lyu*

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;

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 with 222 types of PAIs 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 G4s and the GC content across different genomic features, including the genome itself, 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. Some high-confidence G4 structures within regulatory regions of *Escherichia coli* were identified, modulating the efficiency or specificity of DNA integration events within PAIs. In conclusion, 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 infectious 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*). The DNA G4 structure arises from G-rich DNA sequences that possess the ability to form stacked guanine tetrads. A G4 consists of four guanine bases, arranged in a planar square, and stabilized by Hoogsteen hydrogen bonding. 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" (Fig. 1). 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 critical 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*). Genome sequencing in the past decade has unveiled the abundance of sequence motifs with G4-forming potential in genomes, their distribution patterns, and their correlation with functionally significant genomic regions (*7, 8*). The DNA G4 structure represents an attracting departure from the traditional double helix and has garnered

45 attention for its diverse biological roles and potential applications. Further research
46 into G4 structures will undoubtedly uncover new insights into their functions and
47 facilitate the development of innovative technologies.

48 Pathogenicity islands (PAIs) are genomic regions that contribute to the virulence
49 and pathogenic potential of various microorganisms (*9, 10*). PAIs are distinct
50 segments of the bacterial genome that exhibit unique characteristics compared to the
51 rest of the DNA (*11*). They are often large in size, ranging from tens of kilobases to
52 hundreds of kilobases, and can be integrated into the chromosome or exist as extra-
53 chromosomal elements, such as plasmids. PAIs often exhibit close proximity to tRNA
54 genes, suggesting a putative mechanism where tRNA genes act as anchor points for
55 the integration of foreign DNA acquired through horizontal gene transfer (Fig. 1E).
56 One notable feature is their variable GC content, which tends to deviate from the
57 average GC content of the genome in various organisms, such as *Streptomyces* (*12*),
58 *Salmonella* (*13*), and *Yersinia* (*14*). PAIs typically contain clusters of genes involved
59 in pathogenesis, including those encoding secretion systems (e.g., the LEE (locus of
60 enterocyte effacement) in *Escherichia coli*), superantigen (e.g., SaPI1 and SaPI2 in
61 *Staphylococcus aureus*), and enterotoxin (e.g., *she* PAI in *Shigella flexneri*). PAIs can
62 be acquired through the transfer of mobile genetic elements, such as plasmids, phages,
63 or integrative and conjugative elements (ICEs), facilitating the incorporation of
64 pathogenicity-associated genes into the recipient genome (*9, 15, 16*). One question
65 raised in PAI is that PAIs often exhibit distinct base composition (G+C contents)
66 compared to the core genome. The underlying reasons for this variation remain

67 unknown, but the preservation of a genus- or species-specific base composition
68 represents a noteworthy characteristic of bacteria (*9*). Herbert Schmidt et al. (2004)
69 proposed a hypothetical mechanism to explain the observed variation, suggesting that
70 factors such as DNA topology and codon message in the virulence regions present
71 could contribute to the preservation of the distinct base composition (*9*). Hopefully,
72 the availability of genome sequences from pathogenic bacteria and their non-
73 pathogenic counterparts currently presents an exceptional opportunity to explore the
74 intricate structure variance and underlying mechanisms within PAIs.

75 Growing evidence has shown that G4s exhibit a striking colocalization with
76 functional regions of the genome, and their high conservation across different species
77 suggests a selective pressure to maintain these sequences at specific genomic regions
78 (e.g., genome islands, resistance islands, CpG islands, and PAIs) (*2, 8, 17*). The
79 possibility of interactions between G4 structures and pathogens has been suggested,
80 although this field of study is still in its nascent phase. Some studies observed that
81 bacterial genomes possess G4-forming sequences within their genome regions (*18,*
82 *19*). As mentioned, G4 structures are formed by G-rich DNA sequences, and their
83 stability is influenced by the G+C content and arrangement of G tetrads. Interestingly,
84 PAIs often exhibit an altered GC content, putatively contributing to the propensity of
85 G4 structure formation within these regions. The G4 structures in PAIs might
86 modulate the accessibility of transcription factors, DNA-binding proteins, or RNA
87 polymerase in pathogens, as documented in eukaryotes (*2, 20*), thereby influencing
88 the expression of virulence-associated genes. The formation of G4 structures within

89 PAIs may serve as an additional layer of regulation that fine-tunes the expression of
90 genes critical for pathogenesis. Hence, the investigation of G4 structures within PAIs
91 may open new avenues for the development of therapeutic strategies aimed at
92 disrupting the regulatory mechanisms of pathogenicity-associated genes or targeting
93 the stability of PAIs in order to attenuate pathogenicity.
94

95 # Results and Discussion

96 **Genomic information and PAI patterns in 89 reported pathogenic strains**

97 A dataset of PAIs was compiled from 89 reported pathogenic strains, encompassing
98 222 distinct types of PAIs. These datasets were sourced from the Pathogenicity Island
99 Database (PAIDB, http://www.paidb.re.kr), the primary repository offering
100 comprehensive information on reported PAIs and candidate PAIs in prokaryotic
101 genomes (*21*). Those pathogens that share closely related PAIs on phylogenetic
102 branches, such as LEE in *E. coli* strains (i.e., str. 11368, 11128, 12009, Sakai, and
103 EDL933) (Fig. 2A), may have acquired these elements through recent common
104 ancestors or horizontal gene transfer events, indicating potential inter-species
105 transmission or shared evolutionary history (*9, 22, 23*). The tree provides insights into
106 the evolution of pathogenicity and the emergence of specific virulence traits
107 associated with different branches or clades on the tree. Additional information,
108 including the genome length (bp), G+C content (%), rRNA density (the proportion of
109 rRNA length in its whole genome), tRNA density (the proportion of tRNA length in
110 its whole genome length), and PAI length (bp), was present and showed conserved

111 patterns in the same species (Fig. 2A & Table S1). PAIs commonly exhibit mosaic-
112 like patterns, exemplified by the presence of distinct PAIs like FPI in *Francisella*
113 *tularensis*, SaPIbov in *Staphylococcus aureus*, and Hrp PAI in *Xanthomonas*
114 *campestris* (Fig. 2B). Many PAIs were present associated with tRNAs, such as the
115 insertions of $tRNA^{Thr}$, $tRNA^{Phe}$, $tRNA^{Asp}$, $tRNA^{Asn}$, $tRNA^{Ser}$, and $tRNA^{Gly}$ in *E. coli*
116 strains (Fig. 2B & Table S2). The presence of PAIs distributes in similar genomic
117 regions across different pathogens or strains, showing non-random patterns and
118 functionally clustered.

119 **The presence of G4 structure in 89 reported pathogenic strains**

120 The study utilized G4Hunter, a recently developed user-friendly web server designed
121 for high-throughput analysis of G4-forming sequences in DNA
122 (https://bioinformatics.ibp.cz/) (*24*), to predict the presence and quantity of G4
123 structures in the complete genomes of pathogenic organisms. The bacterial genomes
124 analyzed in the dataset exhibited a range of lengths, spanning from 1.58 Mbps to 6.54
125 Mbps (Table S1 & Table S2). The average GC content was 44.58%, with *Clostridium*
126 *perfringens* SM101 having the lowest value of 22.81% and *Clavibacter michiganensis*
127 subsp. michiganensis NCPPB 382 having the highest value of 72.70%. Employing the
128 G4Hunter search algorithm with a window size of 25 and using the default threshold
129 of 1.2, the study identified a total of 225,376 putative G4 sequences in these 89
130 pathogenic genomes (Table S1). The heatpmap showed that the number of G4
131 structures was diverse in the pathogen genomes (Fig. 2C), similar to previous reports
132 that showed that the G4 structures display uneven distribution patterns in eukaryotic

133 and prokaryotic genomes and are conserved in evolutionary groups (*25-27*). The
134 presence of G4 structures exhibited a remarkable level of conservation within the
135 same pathogenic strains that displayed conserved and non-random distribution
136 patterns. This observation suggests that the arrangement and distribution of G4
137 structures are not solely determined by genomic position but are also influenced by
138 strain-specific factors or evolutionary constraints. Hence, the conserved presence
139 patterns of G4 structures within specific pathogenic strains provide evidence of their
140 functional significance or potential regulatory role in the genome.

141 **Interaction between PAIs and G4s in different genomic features**

142 The analysis of G4 structures across all pathogen species demonstrated a positive
143 correlation between the number of G4s and the GC content in various genomic
144 features, including the whole genome ($R^2 = 0.7815$, $p < 0.0001$), gene ($R^2 = 0.7085$, p
145 < 0.0001), promoter ($R^2 = 0.8132$, $p < 0.0001$), rRNA ($R^2 = 0.1154$, $p < 0.01$), and
146 tRNA ($R^2 = 0.4055$, $p < 0.0001$) regions (Fig. 2D). Furthermore, the density of G4s,
147 measured as the frequency of predicted G4-forming sequences per 1000 base pairs
148 (bp), also showed a positive correlation with the GC content across the analyzed
149 genomic elements (Fig. 2E). The presence of a loop sequence between the G-tracts is
150 required for the formation of an intramolecular G4, and the density of G4 structures is
151 generally associated with the GC content (*28*). Additionally, the whole genome
152 characteristics analysis showed that the number and density of G4s were elevated
153 during evolution based on 37 evolutionarily representatives (*29*). The presence of G4
154 structures within the promoter, rRNA, and tRNA regions may have functional

155 implications for the regulation of DNA replication, ribosome biogenesis, protein
156 synthesis, and other RNA-related processes (*6, 30, 31*). Throughout evolution, there
157 seems to be a greater enrichment of G4 structures in regulatory genes compared to
158 other genes, potentially enabling intricate control of gene expression in cascade signal
159 transduction pathways (*29*). The positive correlation observed underscores the
160 potential role of GC content in influencing the propensity for G4 formation and
161 suggests that the GC-associated regions of pathogen genomes may be particularly
162 relevant for G4-related functions and regulatory mechanisms.

163 This study observed that the GC contents in the genome region were significantly
164 higher compared to the corresponding PAIs region that was classified into five parts
165 (20%-30%, 30%-40%, 40%-50%, 50%-60%, and 60%-70%, respectively) according
166 to the genome datasets (Fig. 3A-E & Table S1). This suggests that the genomic
167 regions surrounding the PAIs (i.e., core genome) tend to have a higher GC content
168 than PAI regions, which was consistent with the fact that PAIs often exhibit distinct
169 base composition (G+C contents) compared with the core genome (*9*). The variation
170 was explained by the presence of G4 sequences within the PAIs, whereas the results
171 were surprising. This study observed a distinct pattern in the frequency of G4
172 structures within different regions of the PAIs. Specifically, the frequency of G4
173 structures was found to be increased in the 20-30% GC content regions (Fig. 3A),
174 decreased in the 60-70% GC content regions (Fig. 3E), and remained unchanged in
175 30%-60% GC content regions (Fig. 3B-D). This differential distribution of G4
176 structures suggests that (i) specific segments (e.g., <30% and >60% GC contents)

177 within the PAIs may be more prone to induce G4 formation discrepancy, as these
178 segments showed unstable gene structure; (ii) the variation of base composition
179 between core genome and PAIs is partially correlated with the presence of G4s; (iii)
180 the frequency of G4s in PAIs present stable as the core genome in the most situation
181 (the majority of pathogenic bacterial species exhibit G+C contents ranging from 40%
182 to 60%); (iv) an alternative hypothesis, other factors, such as i-motif (i.e., the anti-G4
183 structure) and CpG island, may work synergistically with G4 and potentially
184 contribute the base composition variation (*32, 33*).

185 **Putative origin and transfer mechanisms of G4s in PAIs**

186 In order to understand the origin of G4 structures within PAIs, the study hypothesized
187 that these G4 sequences could be acquired through three types of horizontal gene
188 transfer mechanisms: conjugation, transformation, and transduction (Fig. 3F). These
189 mechanisms serve as means for genetic material exchange between different
190 organisms. Considering the presence of G4 sequences within the PAIs, it is plausible
191 that these sequences are transferred along with the PAIs through these horizontal gene
192 transfer mechanisms. Conjugation involves the direct transfer of genetic material
193 through cell-to-cell contact, typically mediated by plasmids or conjugative
194 transposons (*34*). If G4 sequences are present within the donor cell's genome or
195 plasmids, they can be transferred to the recipient cell during conjugation, thereby
196 becoming incorporated into the PAIs of the recipient. Transformation, on the other
197 hand, involves the uptake and incorporation of free DNA from the environment (*35,*
198 *36*). If G4 sequences are present in the extracellular DNA, they can be taken up and

199 integrated into the PAIs of the recipient bacterium during the transformation process.
200 Transduction occurs when genetic material is transferred between bacteria through
201 bacteriophages, which are viruses that infect bacteria (*37*). If G4 sequences are
202 present within the genome of the infecting bacteriophage, they can be introduced into
203 the recipient bacterium during the phage-mediated transfer, potentially becoming
204 integrated into the PAIs.

205 **Evolutionary and presence of G4 within PAIs in *E. coli* strains**

206 Later, the study used *E. coli* as an example to investigate the potential regulatory role
207 and function of genes covered by G4 structures in PAIs. *E. coli* contains at least ten
208 types of PAI in different strains, and one of the well-known PAI is LEE (Fig. 3G),
209 harboring genes responsible for causing attaching and effacing lesions (*23, 38*). One
210 stable G4 structure with a G4hunter score of 1.6 was identified at position 37,085 in
211 the LEE PAI of *E. coli* str. O103:H2 12009 (Fig. 3H), located between an IS element
212 and a tRNA insertion site. The tRNA region generally contains a higher G4 frequency
213 compared with transfer-messenger RNA (tmRNA) and rRNA regions in the bacterial
214 genome (*26*). Interestingly, this G4 structure found in *E. coli* str. O103:H2 12009 was
215 present in close proximity to a tRNA region, suggesting a potential regulatory role of
216 G4s in the tRNA gene, or upstream- and downstream-genes that are responsible for
217 LEE virulence. Additionally, another stable G4 sequence with a score of 1.381 was
218 discovered at position 12,457 in the *E. coli* str. CFT073 PAI II to provide one more
219 evidence of G4 in PAI regions (Fig. 3I).

220 **Functional annotation analysis of G4s within PAIs in *E. coli* strains**

221 Functional enrichment analysis was conducted to explore the putative functions of
222 G4-covered genes in two *E. coli* strains (Table S3&S4). The results revealed that the
223 genes covered by G4 structures were predominantly involved in genetic information
224 processes, including DNA binding, DNA integration, and nucleic acid metabolism
225 processes (Fig. 3J & K). This suggests that G4 structures may play a regulatory role in
226 these essential cellular processes, especially gene expression and DNA-related
227 functions. For instance, G4 structures in the promoter regions of certain transcription
228 factors may influence their binding affinity to DNA and subsequently affect
229 downstream gene expression patterns (*39, 40*). Pathogens often possess mobile
230 genetic elements, including plasmids, bacteriophages, and genomic islands, which
231 play a significant role in their virulence and adaptation. These elements frequently
232 utilize DNA integration mechanisms mediated by integrases, recombinases, or
233 transposases to transfer or incorporate genetic material into the bacterial genome (*41,*
234 *42*). The presence of G4 structures within PAI-related genes suggests a potential role
235 in modulating the integration of these islands into the bacterial genome. Furthermore,
236 G4 structures may influence the binding of integrases or recombinases to the specific
237 target sites within the bacterial chromosome, potentially regulating the successful
238 acquisition and integration of PAIs.

239 This study focused on selecting pathogens with high-quality sequenced genomes
240 to ensure reliability and availability. However, it is important to acknowledge and
241 address certain limitations that could potentially affect the interpretation of the results.
242 One such limitation is the reliance on genome sequences obtained from external

243 laboratories and datasets, which introduces a level of uncertainty regarding the
244 accuracy and completeness. Additionally, the rapidly evolving field of genomics
245 poses challenges in terms of the availability and quality of genome data. The
246 sequencing technologies and methodologies used by different laboratories may vary,
247 leading to variations in data quality and potential biases. Furthermore, the dynamic
248 nature of bacterial genomes, including genetic rearrangements and horizontal gene
249 transfer events, can complicate the accurate assembly and annotation of genome
250 sequences. To address these limitations, future studies could benefit from the
251 advancement of sequencing technologies, improved genome assembly algorithms, and
252 standardized quality control measures. Collaboration between research groups and
253 data sharing initiatives also help ensure the availability of high-quality genome data
254 for comprehensive analyses. It is crucial to interpret the results with caution and
255 continue refining this understanding through validation experiments and
256 collaborations. By addressing these limitations, we can enhance the robustness and
257 reliability of genomic analyses, further advancing our knowledge in the field of PAIs
258 and G4 structures.
259

260 **Conclusion**

261 The present study conducted a comprehensive analysis to explore the links between
262 G4s and PAIs. Overall, the conserved evolutionary relatedness of PAIs, the detection
263 of stable G4 structures in specific genomic positions, and the enrichment of G4-
264 covered genes in genetic information processes collectively support the hypothesis

265 that G4 structures may have regulatory functions in key biological processes in
266 pathogens.
267

268 Methods

269 Selection and extraction of DNA Sequences

270 A total of 89 genomes corresponding to the identified pathogens from the
271 Pathogenicity Island Database (PAIDB) were included in the study. The complete
272 bacterial genomic DNA sequences and their corresponding annotation files in .gff
273 and .fna formats were obtained from the Genome database of the National Center for
274 Biotechnology Information (NCBI). To ensure the reliability and completeness of the
275 dataset, only completely assembled genomes were included in the analysis. To avoid
276 redundancy and incomplete sequences, one representative genome was selected for
277 each species or strain. The selection of representative genomes was based on a careful
278 examination of the supplementary material (Table S1) accompanying the study. The
279 study employed TBtools (https://cj-chen.github.io/tbtools), a versatile bioinformatics
280 tool, to investigate the presence of G-quadruplex (G4) sequences in specific genomic
281 features. The gene regions, promoters (2 kb upstream of the genes), tRNA regions,
282 and rRNA regions were extracted from the selected genomes using TBtools. PAI
283 regions were downloaded following previously documented information in PAIDB
284 (Table S1&S2). Default thresholds and parameters were applied during extraction to
285 maintain consistency across all genomes.

286 Data process and detection of G4s in genomic features

287 The G4Hunter algorithm, a widely used tool for G4 prediction, was employed for the
288 identification of G4 motifs in the genomic sequences. The G4Hunter parameters were
289 set to a window size of "25" and a G4 score threshold of 1.2, which ensured the
290 identification of potential G4 sequences. The analysis was performed for each species
291 group, where the genomic DNA size and the number of putative G4 sequences found
292 were determined. The study quantified the predicted number of putative G4-forming
293 sequences within different genomic features, including the whole genome, gene,
294 promoter, tRNA, rRNA, and PAI regions. The density of G4 motifs was determined
295 by dividing the number of G4 sequences by the total length of the genome, while the
296 length ratio of G4 motifs was calculated by dividing the total length of the G4
297 sequences by the total length of the genome.

298 **Relationship between G4s and PAIs**

299 The heatmap was used to show the distribution of G4 motifs in the genome divided by
300 ten parts as PAI regions using R package “pheatmap”. The correlation between the
301 number of G4 structures and the GC content was analyzed across various genomic
302 elements, including the whole genome, gene, promoter, rRNA, and tRNA regions.
303 The analysis utilized the *R*-squared value (R^2) to determine the strength of the
304 correlation. The significance of the correlation was assessed using *p*-values. The GC
305 content in the genome regions and corresponding PAI regions was compared and
306 classified into different ranges to explore the variation in base composition. GraphPad
307 Prism (V.5.02, GraphPad Software, Inc) was employed to conduct Normality and

308 Lognormality Tests. The K-S test and F-test were used to assess normal distribution
309 and variances, and Student's t-test was used to identify significant difference.

310 **Phylogenetic Tree Construction**

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

322 **Gene functional annotation**

323 The gene sequences covered by G4 structures within PAIs were subjected to gene
324 ontology annotation. The gene sequences were translated into protein sequences using
325 the Expasy online toolkit (https://web.expasy.org/translate/). This tool performs the
326 translation based on the standard genetic code, converting the DNA nucleotide
327 sequence into its corresponding amino acid sequence. The GO annotation database
328 assigned GO terms to the protein sequences based on their predicted functions and
329 known biological process (BP), molecular function (MF), and cellular component

330 (CC). Fisher's exact test was employed to determine the statistical significance of the
331 enrichment results. The obtained *p*-values indicated the overrepresentation of specific
332 GO terms, with lower p-values suggesting higher significance.

333 **Statistics and reproducibility**

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

339

340 ## Acknowledgment

341 The sincere appreciation extends to Dr. Sung Ho Yoon and his colleagues for their
342 dedicated efforts in identifying PAIs and establishing the Pathogenicity Island
343 Database for public analysis. Their commitment to advancing the field of pathogen
344 genomics has greatly facilitated this research. This study would like to thank Prof.
345 Qisheng Song and Dr. Jingjing Li for their insightful suggestions and constructive
346 comments regarding the exploration of G4 structures in genomes. Their expertise and
347 guidance have significantly enriched the understanding of the potential roles and
348 implications of G4 structures in the context of PAIs. The study is grateful to Dr.
349 Minyu Zhou for her valuable discussions and exchange of ideas regarding the
350 interplay between G4 structures and pathogenicity islands.

351 **Competing Interests**

The author declares no competing interests.

Funding information

This research received no external funding.

Credit author statement

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

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.

369 1. J. D. Watson, F. Crick, A structure for deoxyribose nucleic acid. (1953).

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

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

374 4. O. Stegle, L. Payet, J.-L. Mergny, D. J. MacKay, J. L. Huppert, Predicting
375 and understanding the stability of G-quadruplexes. *Bioinformatics* **25**, i374-i1382
376 (2009).

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

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

381 7. J. L. Huppert, S. Balasubramanian, Prevalence of quadruplexes in the human
382 genome. *Nucleic Acids Res* **33**, 2908-2916 (2005).

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

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

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

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

391 12. J. A. Kers *et al.*, A large, mobile pathogenicity island confers plant
392 pathogenicity on Streptomyces species. *Molecular Microbiology* **55**, 1025-1033
393 (2005).
394 13. S. Kombade, N. Kaur, in *Salmonella spp.-A Global Challenge*. (IntechOpen,
395 2021).
396 14. E. Carniel, The Yersinia high-pathogenicity island. *Int Microbiol* **2**, 161-167
397 (1999).
398 15. M. Syvanen, Evolutionary implications of horizontal gene transfer. *Annu Rev*
399 *Genet* **46**, 341-358 (2012).
400 16. J. Chen *et al.*, Intra- and inter-generic transfer of pathogenicity island-
401 encoded virulence genes by cos phages. *Isme j* **9**, 1260-1263 (2015).
402 17. S. Frees, C. Menendez, M. Crum, P. S. Bagga, QGRS-Conserve: a
403 computational method for discovering evolutionarily conserved G-quadruplex motifs.
404 *Hum Genomics* **8**, 8 (2014).
405 18. P. Yadav *et al.*, G-quadruplex structures in bacteria: Biological relevance and
406 potential as an antimicrobial target. *Journal of Bacteriology* **203**, 10.1128/jb. 00577-
407 00520 (2021).
408 19. L. M. Harris, C. J. Merrick, G-quadruplexes in pathogens: a common route to
409 virulence control? *PLoS pathogens* **11**, e1004562 (2015).
410 20. D. Varshney, J. Spiegel, K. Zyner, D. Tannahill, S. Balasubramanian, The
411 regulation and functions of DNA and RNA G-quadruplexes. *Nature reviews*
412 *Molecular cell biology* **21**, 459-474 (2020).

413 21. S. H. Yoon, Y.-K. Park, J. F. Kim, PAIDB v2. 0: exploration and analysis of
414 pathogenicity and resistance islands. *Nucleic acids research* **43**, D624-D630 (2015).
415 22. E. L. Hartland, J. M. Leong. (Frontiers Media SA, 2013), vol. 3, pp. 15.
416 23. J. Jores, L. Rumer, L. H. Wieler, Impact of the locus of enterocyte
417 effacement pathogenicity island on the evolution of pathogenic Escherichia coli.
418 *International Journal of Medical Microbiology* **294**, 103-113 (2004).
419 24. V. Brázda *et al.*, G4Hunter web application: a web server for G-quadruplex
420 prediction. *Bioinformatics* **35**, 3493-3495 (2019).
421 25. Z. Du, Y. Zhao, N. Li, Genome-wide colonization of gene regulatory
422 elements by G4 DNA motifs. *Nucleic acids research* **37**, 6784-6798 (2009).
423 26. M. Bartas *et al.*, The presence and localization of G-quadruplex forming
424 sequences in the domain of bacteria. *Molecules* **24**, 1711 (2019).
425 27. E. P. Lombardi, A. Londoño-Vallejo, A. Nicolas, Relationship between G-
426 quadruplex sequence composition in viruses and their hosts. *Molecules* **24**, 1942
427 (2019).
428 28. O. I. Kaplan, B. Berber, N. Hekim, O. Doluca, G-quadruplex prediction in E.
429 coli genome reveals a conserved putative G-quadruplex-Hairpin-Duplex switch.
430 *Nucleic acids research* **44**, 9083-9095 (2016).
431 29. F. Wu *et al.*, Genome-wide analysis of DNA G-quadruplex motifs across 37
432 species provides insights into G4 evolution. *Communications biology* **4**, 98 (2021).
433 30. P. Ivanov *et al.*, G-quadruplex structures contribute to the neuroprotective
434 effects of angiogenin-induced tRNA fragments. *Proceedings of the National Academy*

435 *of Sciences* **111**, 18201-18206 (2014).

436 31. S. Mestre-Fos *et al.*, G-quadruplexes in human ribosomal RNA. *Journal of*
437 *molecular biology* **431**, 1940-1955 (2019).

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

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

442 34. P. Trieu-Cuot, C. Carlier, P. Martin, P. Courvalin, Plasmid transfer by
443 conjugation from Escherichia coli to Gram-positive bacteria. *FEMS Microbiology*
444 *Letters* **48**, 289-294 (1987).

445 35. M. G. Lorenz, W. Wackernagel, Bacterial gene transfer by natural genetic
446 transformation in the environment. *Microbiological reviews* **58**, 563-602 (1994).

447 36. H. Ochman, J. G. Lawrence, E. A. Groisman, Lateral gene transfer and the
448 nature of bacterial innovation. *Nature* **405**, 299-304 (2000).

449 37. C. L. Schneider, Bacteriophage-mediated horizontal gene transfer:
450 transduction. *Bacteriophages: biology, technology, therapy*, 151-192 (2021).

451 38. F. M. Franzin, M. P. Sircili, Locus of enterocyte effacement: a pathogenicity
452 island involved in the virulence of enteropathogenic and enterohemorragic
453 Escherichia coli subjected to a complex network of gene regulation. *Biomed Res Int*
454 **2015**, 534738 (2015).

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

457 (2018).

458 40. L. Xiang *et al.*, DNA G-quadruplex structure participates in regulation of
459 lipid metabolism through acyl-CoA binding protein. *Nucleic Acids Research* **50**,
460 6953-6967 (2022).

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

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

466

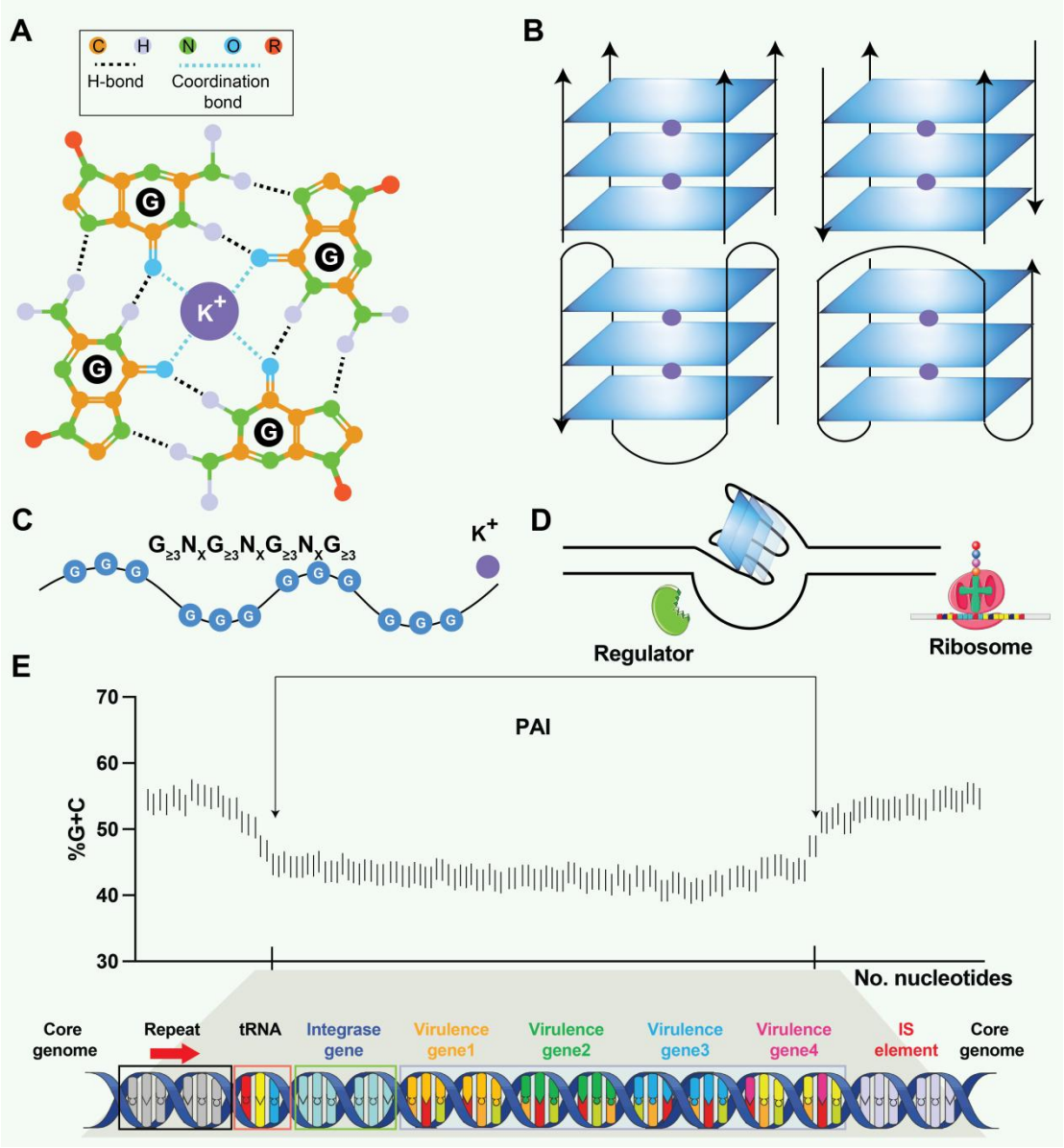

Figure 1. Structural and Functional Aspects of G-quadruplex (G4) Structures and Pathogenicity Islands (PAIs). (A) Schematic representation of a guanine tetrad stabilized by Hoogsten base pairing and a positively charged central ion, illustrating the key elements of G4 structures. (B) Structural heterogeneity of G4s. G-quadruplex 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 G4s, highlighting the repeated

occurrence of guanine-rich sequences that form G-quartets. (D) Regulatory roles of G-quadruplexes in transcription. G4 structures 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.

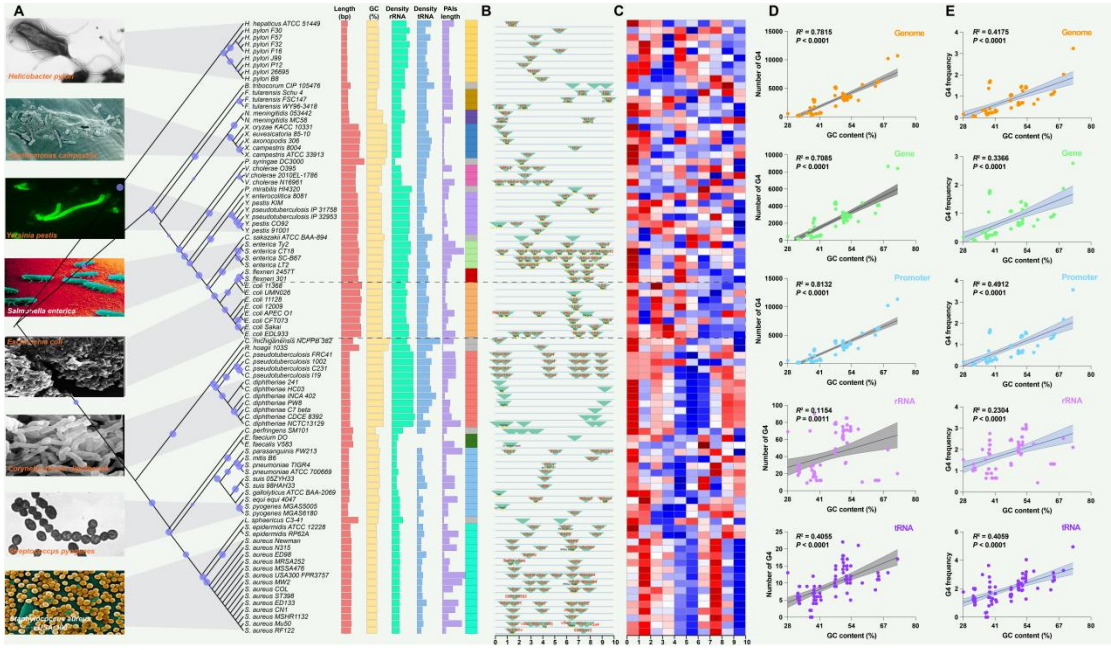

Figure 2. Analysis of Pathogenicity Islands (PAIs) and G-quadruplexes (G4s) in Pathogen Genomes. (A) Phylogenetic analysis of pathogen genomes based on 89 bacterial strains, showing the evolutionary relationships among species. Additional genomic information, including genome size, GC content, rRNA density, 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 G4s 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 G4s, the frequency of G4s and GC content in various genomic features, including the whole genome, genes, promoters, rRNA, and tRNA.

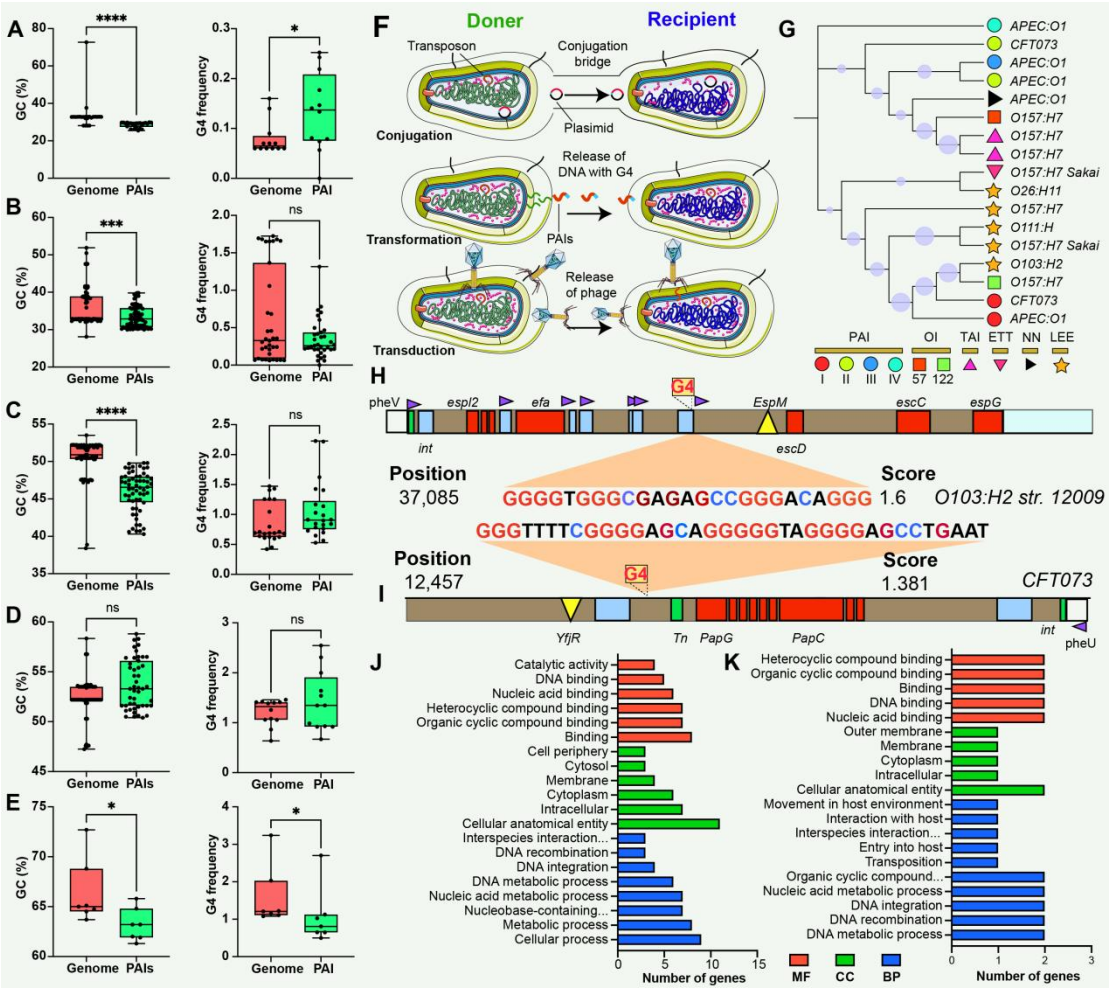

Figure 3. Putative Origin and Functional Annotation of G-quadruplexes (G4s) within Pathogenicity Islands (PAIs) in *Escherichia coli.* (A-E) Comparison of GC content (left panel) and GC frequency (right panel) between the genome and pathogenicity islands (PAIs), categorized into five regions (20%-30%, 30%-40%, 40%-50%, 50%-60%, and 60%-70%). */**/***/**** indicates significant difference (p < 0.05/0.01/0.001/0.0001). (F) Hypotheses on the origin of G4 structures within PAIs, involving gene horizontal transfer mechanisms (conjugation, transduction, and transformation). (G) Evolutionary relatedness of 10 types of PAIs (categorized into six main categories) in E. coli strains. (H & I) Examples of G4 structures within PAIs in E. coli strains. The grey bar represents the virulence region, the red box indicates a

50 virulence gene, the blue box represents an insertion site region or repeat, the green
51 box denotes an integrase, the purple triangle indicates a tRNA insertion site, and the
52 yellow triangle indicates an effector. (J &K) Functional annotation analysis of G4-
53 covered genes within PAIs in two E. coli strains, including biological process (BP),
54 cellular component (CC), and molecular function (MF) categories.