molecules

MDPI

Article

The Presence and Localization of G-Quadruplex Forming Sequences in the Domain of Bacteria

Martin Bartas [1,†], Michaela Čutová [2,†], Václav Brázda [2,3], Patrik Kaura [4], Jiří Šťastný [4,5], Jan Kolomazník [5], Jan Coufal [3], Pratik Goswami [3], Jiří Červeň [1] and Petr Pečinka [1,*]

1 Department of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, University of Ostrava, 710 00 Ostrava, Czech Republic; dutartas@gmail.com (M.B.); jiri.cerven@osu.cz (J.Č.)
2 Faculty of Chemistry, Brno University of Technology, Purkyňova 118, 612 00 Brno, Czech Republic; xcfricova@fch.vut.cz (M.Č.); vaclav@ibp.cz (V.B.)
3 Institute of Biophysics, Academy of Sciences of the Czech Republic v.v.i., Královopolská 135, 612 65 Brno, Czech Republic; jac@ibp.cz (J.C.); pratikgoswami@ibp.cz (P.G.)
4 Faculty of Mechanical Engineering, Brno University of Technology, Technicka 2896/2, 616 69 Brno, Czech Republic; 160702@vutbr.cz (P.K.); stastny@fme.vutbr.cz (J.Š.)
5 Department of Informatics, Mendel University in Brno, Zemedelska 1665/1, 61300 Brno, Czech Republic; jan.kolomaznik@gmail.com
* Correspondence: petr.pecinka@osu.cz; Tel.: +420-553-46-2318
† These authors contributed equally to this work.

Received: 16 April 2019; Accepted: 1 May 2019; Published: 2 May 2019

Abstract: The role of local DNA structures in the regulation of basic cellular processes is an emerging field of research. Amongst local non-B DNA structures, the significance of G-quadruplexes was demonstrated in the last decade, and their presence and functional relevance has been demonstrated in many genomes, including humans. In this study, we analyzed the presence and locations of G-quadruplex-forming sequences by G4Hunter in all complete bacterial genomes available in the NCBI database. G-quadruplex-forming sequences were identified in all species, however the frequency differed significantly across evolutionary groups. The highest frequency of G-quadruplex forming sequences was detected in the subgroup Deinococcus-Thermus, and the lowest frequency in Thermotogae. G-quadruplex forming sequences are non-randomly distributed and are favored in various evolutionary groups. G-quadruplex-forming sequences are enriched in ncRNA segments followed by mRNAs. Analyses of surrounding sequences showed G-quadruplex-forming sequences around tRNA and regulatory sequences. These data point to the unique and non-random localization of G-quadruplex-forming sequences in bacterial genomes.

Keywords: G-quadruplex; bacteria; bioinformatics; deinococcus; G4Hunter

1. Introduction

The discovery of the B-DNA structure by Crick and Watson started a rapid growth in genetic and molecular biology research [1]. However, it is now clear that, apart from this well-known double helical DNA structure, other forms of secondary structure participate in various basic processes [2]. The presence of various local DNA structures including cruciforms [3], quadruplexes [4] and triplexes [5] has been demonstrated by various methodological approaches. For example, G-quadruplexes (G4) were studied by crystallography as far back as 1962 [6]. G4s are secondary structures formed by guanine rich sequences which are widespread in DNA and RNA [4]. The building block for a G4 is a guanine quartet formed by G:G Hoogsteen base pairs (Figure 1). G4 formation requires the presence of monovalent cations such as Na^+ and K^+ [7]. Formation of this structure regulates various processes including gene expression [8], protein translation [9] and proteolysis pathways [10] in both prokaryotes

and eukaryotes. In human, G4s are formed in various regions including sub-telomeres, gene bodies and gene regulatory regions [11] and in telomere regions to suppress degradation and maintain genomic stability [12]. Formation of G4s in this region decreases telomerase activity and decreases the chances of cancer development [13,14]. In addition, the proto-oncogene *MYC* is bound by nucleolin in its hypersensitive region III and enhances G4 folding and suppresses *MYC* transcription [15]. Therefore, it has been suggested that anticancer therapy will be possible by targeting G4s [16–18]. Moreover, it has been demonstrated that G4-stabilizing ligands modulate gene transcription [11]. It is already known that clusters of G4-forming sequences induce gene expression and that they are distributed near promoters and 5′UTRs. Replication-dependent DNA damage evidenced by G4 ligands have also been discovered in tumor suppressor genes and oncogenes [19]. Another potential therapeutic option is to target G4-binding proteins. Many proteins are known to bind to G4s, including some proteins important in cancer [20,21]. Moreover, novel G4 binding proteins have been suggested, sharing the NIQI amino acid motif (RGRGR GRGGG SGGSG GRGRG) [22].

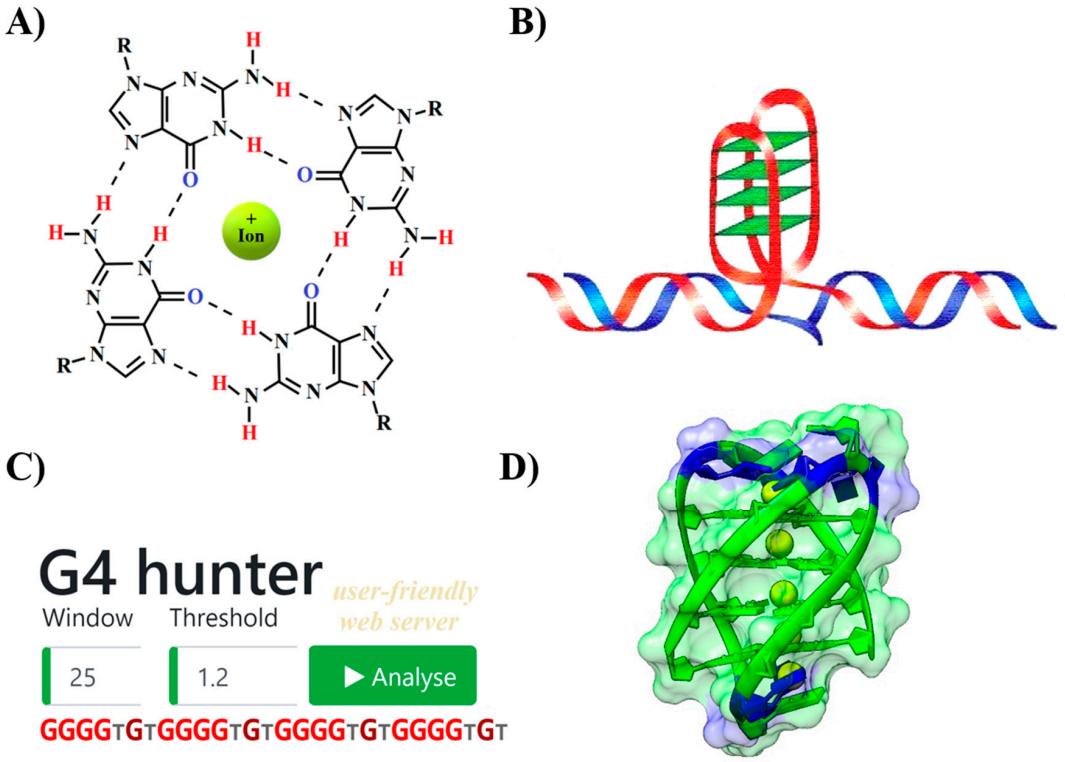

Figure 1. G-quadruplexes: (**A**) guanine tetrad stabilized by Hoogsten base pairing and positively charged central ion; (**B**) schematic drawing of intramolecular G4 structure arising from double stranded DNA; (**C**) G4Hunter, a new user-friendly web server for high throughput analyses of G4-forming sequences in DNA; and (**D**) 3D model of intramolecular antiparallel G4 formed from the sequence (5′-GGGGTGTGGGGTGT GGGGTGTGGGGTGT-3′) found in *Microcystis aeruginosa* built using 3D-NuS webserver [23].

Due to the roles of G4s in regulating basic cellular processes, it is essential to identify the location of G4s in genomes. Several algorithms for detecting expected matching patterns for G4 formation are already described. The first algorithm [$G_nN_mG_nN_oG_nN_pG_n$] was created by Balasubramanian and colleagues [24] and the second algorithm considering occurrence of repeating unit G_n ($n \geq 2$) was created by the group of Maizels [25]. Nevertheless, these algorithms only produce binary (yes/no; match/no match) results, rather than the quantitative analyses that are mandatory for correlation with quadruplex strength metrics. G4Hunter was developed to overcome this limitation, in which G4 propensity is calculated depending on G richness and G skewness [26].

Bacterial genetic material is stored mostly in circular chromosomes and plasmids [27]. It was demonstrated that secondary structures in bacterial genomes are responsible for genomic stability [28]. In addition, G4 structures are more stable than double stranded DNA due to slower unfolding kinetics [29]. Nevertheless, fewer studies on role of G4 in bacterial survival and virulence have been carried out [30]. A comparative functional analysis by Pooja et al. revealed that open reading frame (ORF) formulated amino acids biosynthesis and signal transduction are restrained by/controlled by G4 DNA in prokaryotes [31]. There are have been many reports on the role of G4s in eukaryotes over many years [32], although advances in prokaryotic G4s are not fully elucidated [33].

The formation of an intramolecular G4 requires the presence of a loop sequence between the G-tracts [34] and the density of G4 therefore broadly correlates with GC content. The GC content in bacterial genomes varies remarkably, from 17% to 75% [35]. It was demonstrated that G4 forming sequences are enriched and biased around transcription start sites of genes in the order Deinococcales [36]. Another function of G-tracts is in sustaining and maintaining duplex stability at higher temperatures in thermophiles; for example, *Thermus aquaticus* has a GC content of 68% [37]. Interestingly, the soil bacterium *Paracoccus denitrificans* contains 494 G4-forming sequences, which play roles in digestion of NO_3- through G4 formation upstream of *NasT* [38]. The presence of G4-forming motifs in genes *hsdS*, *recD*, and *pmrA* of *Streptococcus pneumoniae* participate in host–pathogen interactions [30]. Such observations show the significance role of G4 in bacteria and also in eukaryotic cell organelles such as chloroplasts and mitochondria with circular DNAs that originated from prokaryotic organisms. Several papers show the importance of local DNA structures in mitochondrial DNA including G4 using G4Hunter [26] and inverted repeats [39] using palindrome analyzer [40]. Similarly, cruciforms exist in various regulatory regions in chloroplast DNA [41].

The presence of G4 in bacteria remains poorly understood. In our study, we comprehensively analyzed the presence and locations of G4 in 1627 bacterial genomes using G4Hunter. These data bring more information about evolutionary changes of G4 frequency between phyla and provide evidence for the importance of G4 in prokaryotes.

2. Results

2.1. Variation in Frequency for G4-forming Sequences in Bacteria

We analyzed the occurrence of putative G4 sequences (PQS) by G4Hunter in all 1627 known bacterial genomes. The length of bacterial genomes in the dataset varies from 298 kbp to 20.20 Mbp. The GC content average is 50.44%, with minimum 20.2% for *Buchnera aphidicola* (Gammaproteobacteria) and maximum 74.7% for *Corynebacterium sphenisci* (Actinobacteria). Using standard values for G4Hunter algorithm—window size 25 and G4Hunter score above 1.2—we found 9,202,364 PQS in all 1547 bacteria with 1627 genomes (some bacteria have two genomes). The most abundant PQS are those with G4Hunter scores of 1.2–1.4 (97.9% of all PQS), much less abundant are PQS with G4Hunter scores 1.4–1.6 (1.96% of all PQS), followed by 1.6–1.8 (0.128% of all PQS) and 1.8–2.0 (0.0056% of all PQS) and the lowest number of PQS is above G4Hunter score 2 (0.0009% of all PQS). In general, a higher G4Hunter score means a higher probability of G4s forming inside the PQS [26]. A summary of all PQS found in ranges of G4Hunter score intervals and precomputed PQS frequencies per 1000 bp is shown in Table 1.

According to NCBI taxonomy classification, the fully sequenced organisms of Bacteria domain are divided into 18 groups (6 with 10 or more sequenced genomes) and 39 subgroups (14 with 10 or more sequenced genomes), as shown in the phylogenetic tree (Figure 2). For statistical analyses, we used only groups with 10 or more sequenced genomes (highlighted by colors).

The number of all analyzed sequences in individual phylogenetic categories, together with median genome length, shortest genome, longest genome, mean, minimal and maximal observed frequency of PQS per 1000 bp and total PQS counts are shown in Table 2. Five subgroups (Actinobacteria, Chloroflexi, Deinococcus-Thermus, Alphaproteobacteria and Betaproteobacteria) show >60% GC

content. On the other side, three subgroups (Spirochaetia, Thermotogae and Tenericutes) show < 40% GC content.

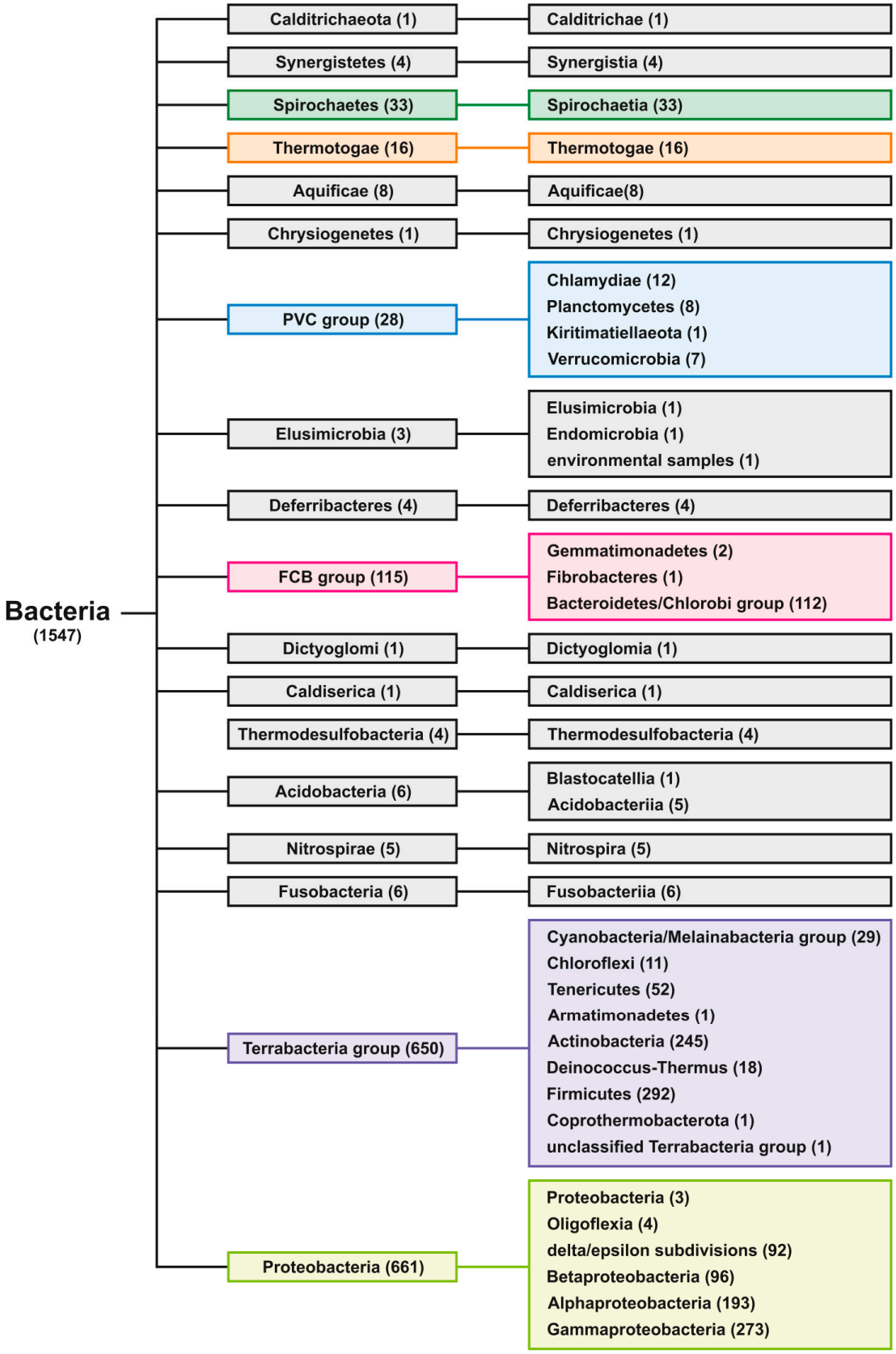

Figure 2. Phylogenetic tree of inspected Bacterial Groups and Subgroups.

Table 1. Total number of PQS and their resulting frequencies per 1000 bp in all 1547 representative bacteria, grouped by G4Hunter score. Frequency was computed by using total number of PQS in each category divided by total length of all analyzed sequences and multiplied by 1000.

Interval of G4Hunter Score	Number of PQS in Dataset	PQS Frequency per 1000 bp
1.2–1.4	9,009,593	1.315033
1.4–1.6	180,395	0.025058
1.6–1.8	11,779	0.00155
1.8–2.0	511	0.000055
2.0–more	86	0.000009

Table 2. Genomic sequences sizes, PQS frequencies and total counts. Seq (total number of sequences), Median (median length of sequences), Short (shortest sequence), Long (longest sequence), GC % (average GC content), PQS (total number of predicted PQS), Mean f (mean frequency of predicted PQS per 1000 bp), Min f (lowest frequency of predicted PQS per 1000 bp), Max f (highest frequency of predicted PQS per 1000 bp). Colors correspond to phylogenetic tree depiction.

Domain	Seq	Median	Short	Long	GC%	PQS	Mean f	Min f	Max f
Bacteria	1627	3,307,820	83,026	13,033,779	50.6	9,202,364	1.342	0.013	14.213
Group	**Seq**	**Median**	**Short**	**Long**	**GC%**	**PQS**	**Mean f**	**Min f**	**Max f**
Spirochaetes	38	2,646,038	277,655	4,653,970	39.7	87,109	0.809	0.079	6.668
Thermotogae	16	2,150,379	1,884,562	2,974,229	39.1	13,617	0.395	0.149	0.812
PVC group	28	2,917,407	1,041,170	9,629,675	50.7	198,358	1.646	0.388	4.802
FCB group	117	3,914,632	605,745	9,127,347	42.3	302,949	0.608	0.013	2.746
Terrabacteria	659	3,018,755	91,776	11,936,683	50.4	4,766,517	1.601	0.016	14.213
Proteobacteria	724	3,551,512	83,026	13,033,779	53.4	3,688,101	1.276	0.025	5.507
Other	45	2,157,835	1,012,010	6,237,577	44.3	145,713	1.103	0.062	5.855
Subgroup	**Seq**	**Median**	**Short**	**Long**	**GC%**	**PQS**	**Mean f**	**Min f**	**Max f**
Spirochaetia	38	2,646,038	277,655	4,653,970	39.7	87,109	0.809	0.079	6.668
Thermotogae	16	2,150,379	1,884,562	2,974,229	39.1	13617	0.395	0.149	0.812
Chlamydiae	12	1,168,953	1,041,170	3,072,383	40.3	12453	0.646	0.388	0.957
Bacteroidetes/Chlorobi	114	3,878,527	605,745	9,127,347	41.9	282,516	0.585	0.013	2.746
Cyanobacteria/Melainа	29	5,315,554	1,657,990	9,673,108	42.6	193,894	1.247	0.201	6.004
Chloroflexi	12	2,333,610	125,2731	5,723,298	60	62,688	1.89	1.223	3.222
Tenericutes	52	981,001	564,395	1,877,792	28	6460	0.136	0.016	0.834
Actinobacteria	246	3,960,961	775,354	11,936,683	66.2	3,590,884	2.821	0.143	8.556
Deinococcus-Thermus	18	2,895,913	2,035,182	3,881,839	66.8	311,949	6.626	1.885	14.213
Firmicutes	298	2,835,823	91,776	8,739,048	40.8	579,740	0.56	0.064	6.587
delta/epsilon subdiv.	92	3,136,746	1,457,619	13,033,779	50	807,281	1.681	0.034	5.282
Betaproteobacteria	110	3,763,620	820,037	6,987,670	60.6	585,984	1.306	0.195	4.007
Alphaproteobacteria	213	3,424,964	83,026	9,207,384	61.5	126,134	1.764	0.051	5.507
Gammaproteobacteria	302	3,777,066	298,471	7,783,862	48.8	31,686	0.799	0.025	4.264
other	75	2,406,157	1,012,010	9,629,675	48.4	432,683	1.406	0.0616	5.855

Mean frequency for all bacterial genomes was 1.342 PQS per 1000 bp. The lowest mean frequency is for Thermotogae (0.395) and the highest for the PVC group (1.646), followed by Terrabacteria (1.601). On the subgroup level, the lowest mean frequency was found in Tenericutes (0.136) and the highest in Deinococcus-Thermus (6.626), followed by Actinobacteria (2.821). The very highest PQS frequency of 14.213 PQS/kbp was found in *Thermus oshimai JL-2* (subgroup Deinococcus-Thermus) and the lowest frequency (0.013 PQS/kbp) in *Lacinutrix venerupis* (subgroup Bacteroidetes/Chlorobi) containing only 40 PQS in its 31,923,99 bp genome (0.0125 PQS/kbp). Detailed statistical inter group and inter subgroup comparisons are depicted in Supplementary Material S5 (SM_05).

Detailed statistical characteristics for PQS frequencies (including mean, variance, and outliers) are depicted in boxplots for all inspected subgroups (Figure 3).

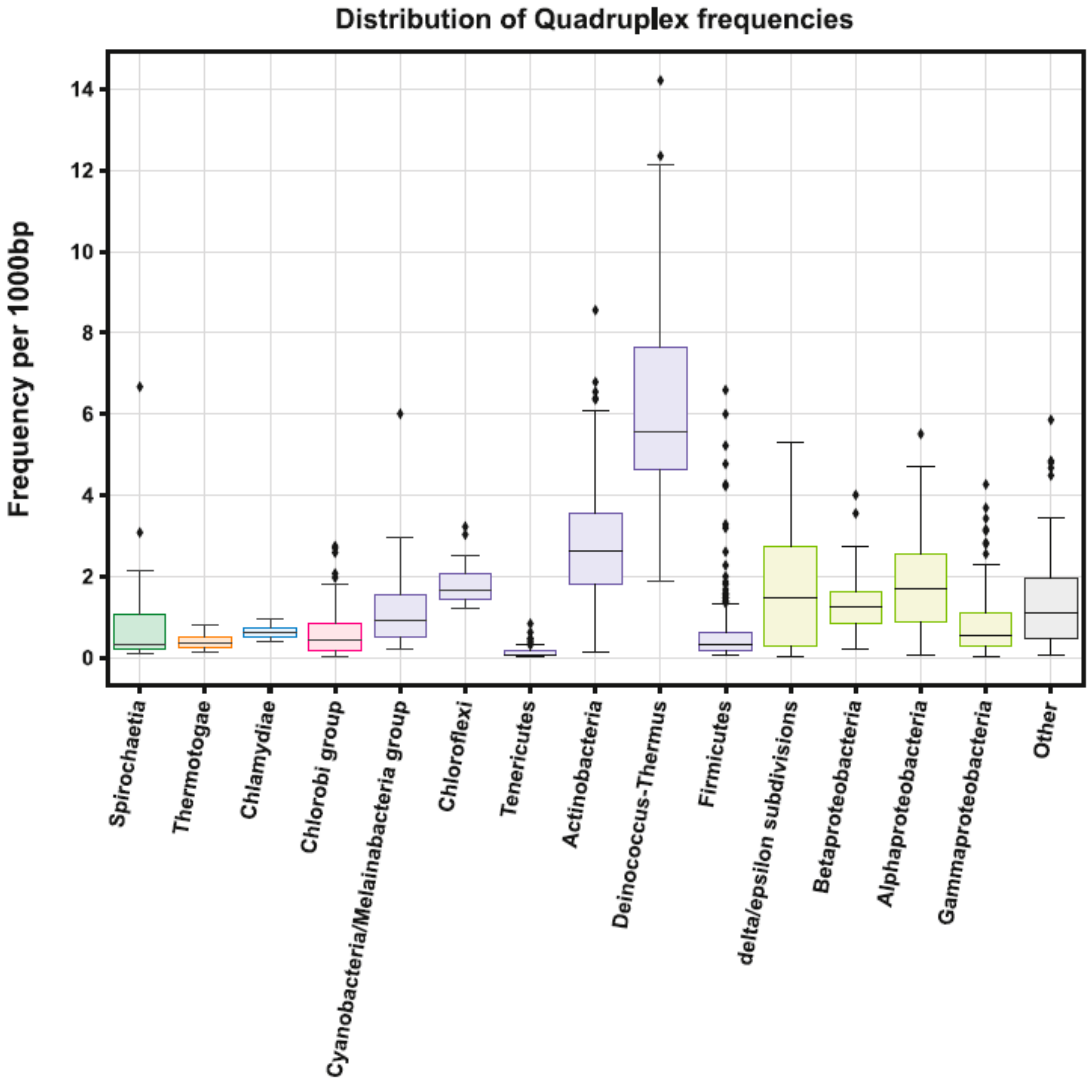

Figure 3. Frequencies of PQS in subgroups of the analyzed bacterial genomes. Data within boxes span the interquartile range and whiskers show the lowest and highest values within 1.5 interquartile range. Black diamonds denote outliers.

We visualized the relationship between %GC content in genomes with the frequency of PQS (Figure 4). In general, PQS frequencies usually correlate with GC content, however there are many exceptions to this rule. Organisms with high PQS frequencies relative to their GC content (over 50% of the maximal observed PQS frequency, Figure 4) are highlighted in color; the whole figure is separated into smaller segments according to inspected G4Hunter score intervals. Nearly all of the 10 outliers belong to the group Terrabacteria, except *Spirochaeta thermophila* DSM 6578 (group Spirochaetes). From the Terrabacteria group, six outliers belong to the small subgroup Deinococcus-Thermus (*Thermus oshimai* JL-2, *Thermus brockianus*, *Thermus aquaticus* Y51MC23, *Thermus scotoductus* SA-01, *Marinithermus hydrothermalis* DSM 14884, and *Deinococcus puniceus*), two outliers belong to the subgroup Actinobacteria (*Verrucosispora maris* AB-18-032 and *Rubrobacter xylanophilus* DSM 9941) and one outlier comes from the subgroup Cynobacteria/Melainabacteria (*Microcystis aeruginosa* NIES-843).

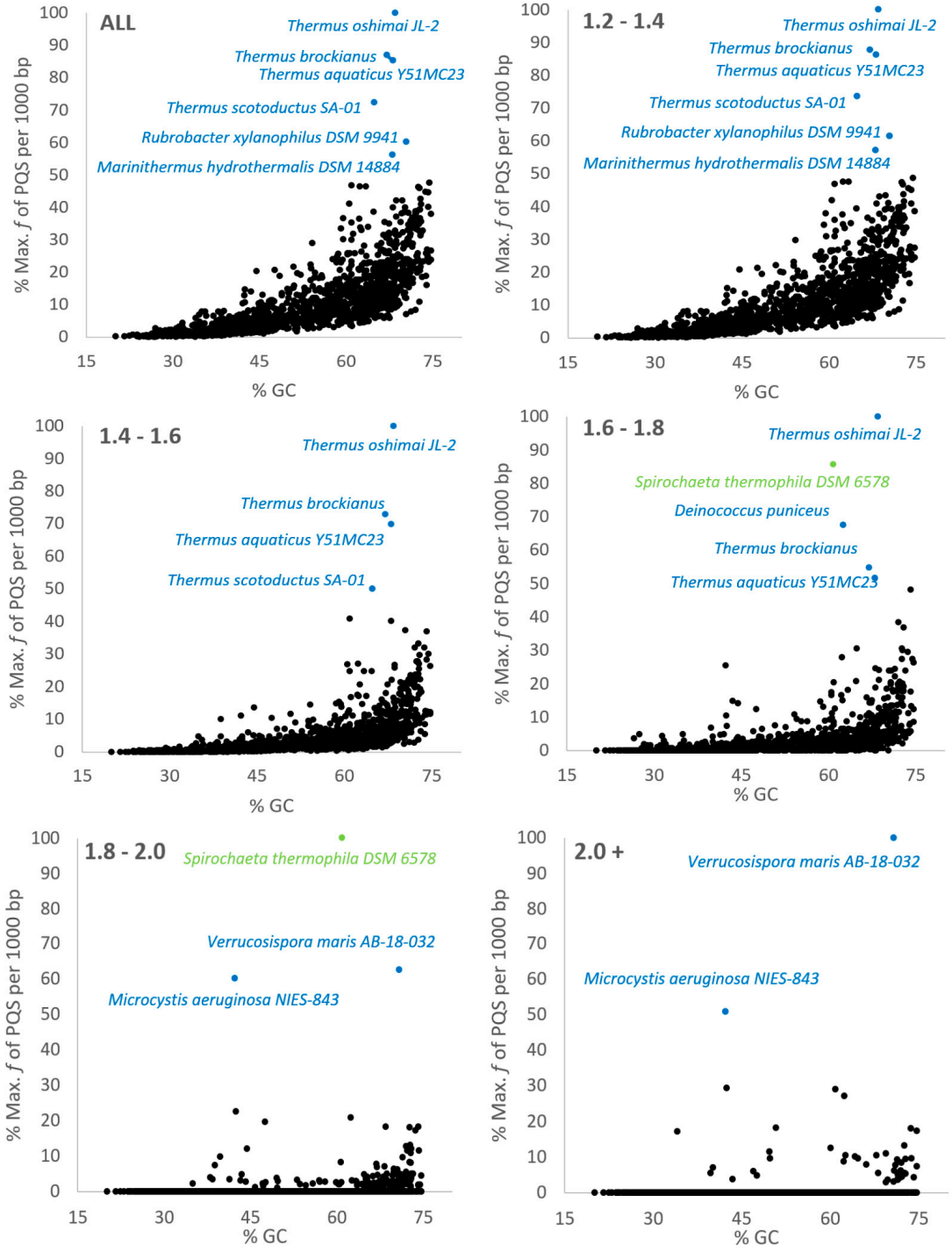

Figure 4. Relationship between observed frequency of PQS per 1000 bp and GC content in all analyzed prokaryotic sequences in various G4 Hunter score intervals. In each G4Hunter score interval miniplot, frequencies were normalized according to the highest observed frequency of PQS. Organisms with max. frequency per 1000 bp greater than 50% are described and highlighted in color.

2.2. Localization of PQS in Genomes

To evaluate the position of PQS in bacterial genomes, we downloaded the described "features" of all bacterial genomes and analyzed the presence of all PQS in each annotated sequences and its close proximity (100 bp before and after feature annotation). PQS frequencies around annotated genome sites are shown in Figure 5. The highest PQS frequencies are before and after transfer RNA

(tRNA), then inside transfer-messenger RNA (tmRNA) and inside ribosomal (rRNA). The lowest PQS frequencies were noticed before and after sequence-tagged sites (STS), then after and before rRNA and after miscellaneous features. If we consider only "inside" regions of inspected features, the differences between features are much smaller than within "before" and "after" regions.

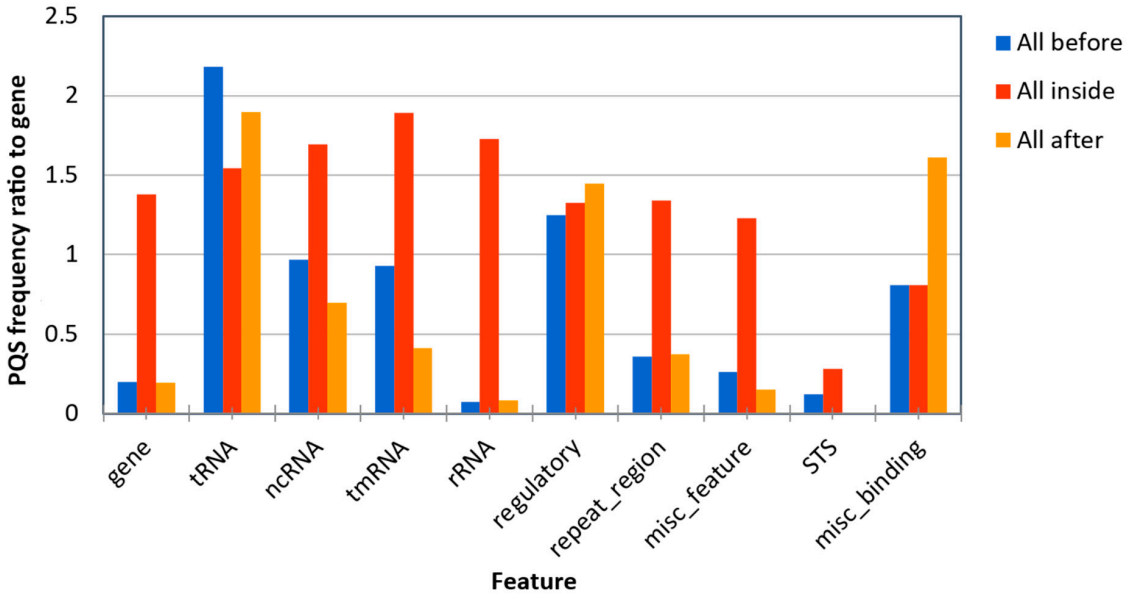

Figure 5. Differences in PQS frequency by DNA locus. The chart shows PQS frequencies according to "gene" annotation and other annotated locations from the NCBI database. We analyzed the frequencies of all PQS within (inside), before (100 bp) and after (100 bp) annotated locations.

As shown in Figure 5, there is no straight pattern in PQS occurrence in all annotated sequences but, in some groups, there are certain PQS distributions. For example, inside rRNA, tmRNA, ncRNA, misc_features, genes and repeat regions, there is higher amount of PQS in annotated sequences, but these PQS are not frequently present in DNA situated before and after these annotated sequences. In contrast, there is almost the same distribution of PQS before, inside and after annotated sequences in tRNA and regulatory groups.

3. Discussion

It has been demonstrated that G4s could be used as targets for therapy [42]. G4 ligands are suggested as a target in cancer [43] and show antiparasitic activity for *Trypanosoma brucei* binding to a G4 structure [44]. Therefore, it has been proposed that G4 sequences in bacterial genomes represent novel and promising targets for antimicrobial therapy [33], and dinuclear polypyridylruthenium(II) complexes are active against drug-resistant bacteria including Methicillin-resistant *Staphylococcus aureus* and Vancomycin-resistant *Enterococcus* [45,46]. Dinuclear ruthenium complexes are relatively well-characterized G4 DNA binding agents [47–49]. Interestingly, we found large numbers of PQS with G4Hunter scores greater than 1.8 in the cyanobacterium *Microcystis aeruginosa*. *Microcystis aeruginosa* is a ubiquitous cyanobacterium living in eutrophic fresh water, which produces harmful hepatotoxins and neurotoxins, and can cause economic loss and damage to the ecosystem [50]. Our analysis indicates that this organism contains unusual and perfectly repeated PQSs (for example, DNA repeat of $(GGGGTGT)_{58}$). Therefore, we hypothesize that this organism could be very sensitive to treatment with specific G4 binding compounds to inhibit its growth, as a possible alternative to commonly used algicides (the human genome does not contain these GGGGTGT repetitions). On the other hand, the lowest mean frequency of PQS was found in Terrabacteria subgroup Tenericutes (0.136 PQS/kbp) with the lowest average GC content (28%). The subgroup Tenericutes includes the genus *Mycoplasma* with many pathogens of clinical importance. On the other hand, a G4 was found in the promoter region

of *Mycobacterium tuberculosis* and G4 ligands inhibited *M. tuberculosis* growth in the low micromolar range [51]. Therefore, the presence of a G4 could be important not only in antiviral [42,52] but also in antibacterial therapy. According to a recent study by Ding et al., eukaryotic organisms have similar PQS frequencies of 0.3 PQS per 1000 bp, whereas prokaryote frequencies are more diverse [36]. Based on our analysis, prokaryote PQS frequencies span a range of 0.013 (*Lacinutrix venerupis*) to 14.213 (*Thermus oshimai JL-2*) PQS per 1000 bp. A similar observation was shown by Quadparser algorithm and leads to the hypothesis that thermophilic organisms are enriched with PSQs due to their living at high temperatures [36]. However, similar enrichment has been demonstrated also for organisms with resistance to other stress factors such as radioresistance [53,54], thus the direct correlation between temperature and G4 presence is not supported by these finding. Validation of the G4Hunter score was made based on biophysical measurements at room temperature [26], therefore the number of G4 sequences in thermophiles could be overestimated, especially for those sequences with G4Hunter scores close to 1.2. Moreover, the mostly thermophilic and hyperthermophilic bacteria in phylum Thermotogae strains has one of the lowest PSQ frequencies. Thus, it seems that Gram-negative thermophilic bacteria evolved according to G4 structures in a completely different way than Gram-positive thermophilic bacteria, and that correlation among thermophiles and G4s depends on the phylum. Contrary to the enrichment of PQS near transcriptional start sites (TSS), 5′-3′UTR sequences and coding regions in eukaryotes [36], our analyses showed the highest PQS frequencies inside tmRNA, ncRNA and rRNA regions in prokaryotes (Figure 5). tmRNAs play a key role in the so-called ribosome rescue process, if ribosomes cannot finish translation, e.g. due to lost stop codon in translated mRNA. The physiological role of ncRNAs in prokaryotes is not fully elucidated, although they are considered to be important regulators of pathogenic processes by controlling virulence gene expression in *Staphylococcus aureus* and *Vibrio cholerae* [55]. The comparison of PQS frequencies between different studies could be complicated due to various PQS thresholds and algorithms. In our study, we used the state-of-the-art algorithm, G4Hunter, developed by Mergny and colleagues. This algorithm takes into account G-richness and G-skewness and has been experimentally validated [26]. Moreover, the current G4Hunter web version allows easy analyses of multiple genomes [56] and our comprehensive analysis showed the broad variations of PQS frequencies and their locations in bacterial genomes.

4. Methods

4.1. Selection of DNA Sequences

The set of all complete bacterial genomic DNA sequences was downloaded from the Genome database of the National Center for Biotechnology Information [57]. We used for our analyses only completely assembly level and we have selected one genome (representative) for each species (Supplementary Material S1 (SM_01)) to avoid non-complete sequences and duplications. In total, we analyzed the presence of G4 sequences in 1627 genomes from the domain Bacteria, representing 5886 Mbp.

4.2. Process of Analysis

We used the computational core of our DNA analyzer software written in Java [40]. For these analyses, we used the G4Hunter algorithm implementation [56]. Parameters for G4Hunter was set to "25" for window size and G4 score above 1.2. An example of a putative G4 sequence found using such search criteria is provided in Supplementary Material S2 (SM_02). The overall results for each species group contained a list of species with size of genomic DNA and number of putative G4 sequences found (Supplementary Material S3 (SM_03)). These data were processed by Python jupyter using Pandas (contains statistical tools). Graphs were generated from the Pandas tables using "seaborn" graphical library.

4.3. Analysis of Putative G4 Sequences Around Annotated NCBI Features

We downloaded the feature tables from the NCBI database along with the genomic DNA sequences. Feature tables contain annotations of known features found in the DNA sequence. We analyzed the occurrence of G4-forming sequences inside and around (before and after) recorded features. Features were grouped by the name stated in the feature table file. From this analysis, we obtained a file with feature names and numbers of putative G4 forming sequences found inside and around features for each group of species analyzed. Search for putative G4 forming sequences took place in a predefined feature neighborhood (we used ±100 bp—this figure is important for calculation of putative G4-forming sequence frequencies in feature neighborhoods) and inside feature boundaries. We calculated the amount of all predicted putative G4-forming sequences in regions before, inside and after features. An example of categorizing a putative G4-forming sequence according to its overlap with a feature or feature's neighborhood is shown in Supplementary Material S2 (SM_02). Further processing was performed in Microsoft Excel and the data are available as Supplementary Material S4 (SM_04).

4.4. Phylogenetic Tree Construction

Exact taxid IDs of all analyzed groups were obtained from Taxonomy Browser via NCBI Taxonomy Database [58], downloaded to phyloT: a tree generator (http://phylot.biobyte.de) and a phylogenetic tree was constructed using function "Visualize in iTOL" in Interactive Tree of Life environment [59]. The resulting tree is shown in Figure 2.

4.5. Statistical Analysis

Statistical evaluations of differences in G4-forming sequences in phylogenetic groups were made by Kruskal–Wallis test in STATISTICA, with p-value cut-off 0.05; data are available in Supplementary Material S5 (SM_05).

5. Conclusions

In this research, we analyzed the presence of PQS in bacterial genomes. PQS were identified in all species, but the number of PQS differ remarkably among individual subgroups, showing evolutionary adaptations connected with G4. While the highest frequency of PQS was detected in Gram-positive extremophiles Deinococcus-Thermus subgroup, the lowest PQS frequency was found in Gram-negative thermophilic bacteria in Bacteroidetes/Chlorobi subgroup. Thus, it seems that evolution of these subgroups was driven by different strategies. PQS are enriched in ncRNA segments followed by mRNAs; analyses of surrounding sequences showed PQS enrichment also around tRNA and regulatory sequences. These data point to the unique and non-random localization of PQS in bacterial genomes.

Supplementary Materials: The supplementary material are available online. Supplementary Material S1 (SM_01): The accession codes and phylogenetic classification of all 1627 representative prokaryotic complete genomic DNA sequences, Supplementary Material S2 (SM_02): Example putative G4 sequence and predefined feature neighborhood, Supplementary Material S3 (SM_03): Overall results of PQS frequencies found in each analyzed genomic sequence (group or subgroup) together with GC content, sequence length and other parameters, Supplementary Material S4 (SM_04): Detailed results of PQS occurrence around defined genomic features, Supplementary Material S5 (SM_05): Statistical evaluations of differences in G4-forming sequences in phylogenetic groups

Author Contributions: Conceptualization, M.B. and P.P.; Data curation, P.K.; Formal analysis, M.B. and P.K.; Funding acquisition, P.P.; Investigation, M.Č. and J.C.; Methodology, V.B.; Project administration, P.P.; Resources, M.Č., J.Š.; and J.K.; Software, P.K., J.Š.; and J.K.; Supervision, J.Č. and P.P.; Validation, V.B.; Visualization, M.B., P.K. and P.G.; Writing—original draft, M.B., M.Č., V.B., J.C. and P.G.; and Writing—review and editing, J.Č.

Funding: This work was supported by the Grant Agency of the Czech Republic (18-15548S); the Ministry of Education, Youth and Sports of the Czech Republic in the "National Feasibility Program I" (LO1208 TEWEP); EU structural funding Operational Programme Research and Development for innovation, project No. CZ.1.05/2.1.00/19.0388; and project SGS/09/PrF/2019 financed by University of Ostrava.

Acknowledgments: We thank Philip J. Coates for proofreading and editing the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Watson, J.D.; Crick, F.H. Molecular structure of nucleic acids. *Nature* **1953**, *171*, 737–738. [CrossRef]
2. Szlachta, K.; Thys, R.G.; Atkin, N.D.; Pierce, L.C.T.; Bekiranov, S.; Wang, Y.-H. Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human. *Genome Biol.* **2018**, *19*, 89. [CrossRef] [PubMed]
3. Brázda, V.; Laister, R.C.; Jagelská, E.B.; Arrowsmith, C. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.* **2011**, *12*, 33. [CrossRef]
4. Sun, Z.-Y.; Wang, X.-N.; Cheng, S.-Q.; Su, X.-X.; Ou, T.-M. Developing Novel G-Quadruplex Ligands: From Interaction with Nucleic Acids to Interfering with Nucleic Acid–Protein Interaction. *Molecules* **2019**, *24*, 396. [CrossRef]
5. Nelson, L.D.; Bender, C.; Mannsperger, H.; Buergy, D.; Kambakamba, P.; Mudduluru, G.; Korf, U.; Hughes, D.; Van Dyke, M.W.; Allgayer, H. Triplex DNA-binding proteins are associated with clinical outcomes revealed by proteomic measurements in patients with colorectal cancer. *Mol. Cancer* **2012**, *11*, 38. [CrossRef] [PubMed]
6. Gellert, M.; Lipsett, M.N.; Davies, D.R. Helix Formation by Guanylic acid. *Proc. Natl. Acad. Sci. USA* **1962**, *48*, 2013–2018. [CrossRef]
7. Harkness, R.W.; Mittermaier, A.K. G-quadruplex dynamics. *Biochim. Biophys. Acta Proteins Proteom.* **2017**, *1865*, 1544–1554. [CrossRef]
8. Siddiqui-Jain, A.; Grand, C.L.; Bearss, D.J.; Hurley, L.H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 11593–11598. [CrossRef]
9. Lee, S.C.; Zhang, J.; Strom, J.; Yang, D.; Dinh, T.N.; Kappeler, K.; Chen, Q.M. G-Quadruplex in the NRF2 mRNA 5′ Untranslated Region Regulates De Novo NRF2 Protein Translation under Oxidative Stress. *Mol. Cell. Biol.* **2016**, *37*, e00122-16. [CrossRef] [PubMed]
10. Endoh, T.; Kawasaki, Y.; Sugimoto, N. Stability of RNA quadruplex in open reading frame determines proteolysis of human estrogen receptor α. *Nucleic Acids Res.* **2013**, *41*, 6222–6231. [CrossRef] [PubMed]
11. Lam, E.Y.N.; Beraldi, D.; Tannahill, D.; Balasubramanian, S. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.* **2013**, *4*, 1796. [CrossRef]
12. Long, X.; Stone, M.D. Kinetic Partitioning Modulates Human Telomere DNA G-Quadruplex StructuralPolymorphism. *PLoS ONE* **2013**, *8*, e83420. [CrossRef] [PubMed]
13. Sun, D.; Thompson, B.; Cathers, B.E.; Salazar, M.; Kerwin, S.M.; Trent, J.O.; Jenkins, T.C.; Neidle, S.; Hurley, L.H. Inhibition of human telomerase by a G-Quadruplex-Interactive compound. *J. Med. Chem.* **1997**, *40*, 2113–2116. [CrossRef] [PubMed]
14. Lee, H.-S.; Carmena, M.; Liskovykh, M.; Peat, E.; Kim, J.-H.; Oshimura, M.; Masumoto, H.; Teulade-Fichou, M.-P.; Pommier, Y.; Earnshaw, W.C.; et al. Systematic Analysis of Compounds Specifically Targeting Telomeres and Telomerase for Clinical Implications in Cancer Therapy. *Cancer Res.* **2018**, *78*, 6282–6296. [CrossRef]
15. Dickerhoff, J.; Onel, B.; Chen, L.; Chen, Y.; Yang, D. Solution Structure of a MYC Promoter G-Quadruplex with 1:6:1 Loop Length. *ACS Omega* **2019**, *4*, 2533–2539. [CrossRef]
16. Balasubramanian, S.; Hurley, L.H.; Neidle, S. Targeting G-quadruplexes in gene promoters: A novel anticancer strategy? *Nat. Rev. Drug Discov.* **2011**, *10*, 261–275. [CrossRef]
17. Cimino-Reale, G.; Zaffaroni, N.; Folini, M. Emerging Role of G-quadruplex DNA as Target in Anticancer Therapy. *Curr. Pharm. Design* **2016**, *22*, 6612–6624. [CrossRef]
18. Asamitsu, S.; Obata, S.; Yu, Z.; Bando, T.; Sugiyama, H. Recent Progress of Targeted G-Quadruplex-Preferred Ligands Toward Cancer Therapy. *Molecules* **2019**, *24*, 429. [CrossRef]
19. Yoshida, W.; Saikyo, H.; Nakabayashi, K.; Yoshioka, H.; Bay, D.H.; Iida, K.; Kawai, T.; Hata, K.; Ikebukuro, K.; Nagasawa, K.; et al. Identification of G-quadruplex clusters by high-throughput sequencing of whole-genome amplified products with a G-quadruplex ligand. *Sci. Rep.* **2018**, *8*, 1–8. [CrossRef] [PubMed]
20. Brázda, V.; Hároníková, L.; Liao, J.C.C.; Fojta, M. DNA and RNA Quadruplex-Binding Proteins. *Int. J. Mol. Sci.* **2014**, *15*, 17493–17517. [CrossRef]

21. Mishra, S.K.; Tawani, A.; Mishra, A.; Kumar, A. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci. Rep.* **2016**, *6*, 38144. [CrossRef] [PubMed]
22. Brázda, V.; Cerveň, J.; Bartas, M.; Mikysková, N.; Coufal, J.; Pečinka, P. The amino acid composition of quadruplex binding proteins reveals a shared motif and predicts new potential quadruplex interactors. *Molecules* **2018**, *23*, 2341. [CrossRef]
23. Patro, L.P.P.; Kumar, A.; Kolimi, N.; Rathinavelan, T. 3D-NuS: A web server for automated modeling and visualization of non-canonical 3-dimensional nucleic acid structures. *J. Mol. Biol.* **2017**, *429*, 2438–2448. [CrossRef]
24. Huppert, J.L.; Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **2005**, *33*, 2908–2916. [CrossRef] [PubMed]
25. Eddy, J.; Maizels, N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.* **2006**, *34*, 3887–3896. [CrossRef]
26. Bedrat, A.; Lacroix, L.; Mergny, J.L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.* **2016**, *44*, 1746–1759. [CrossRef]
27. diCenzo, G.C.; Finan, T.M. The Divided Bacterial Genome: Structure, Function, and Evolution. *Microbiol. Mol. Biol. Rev.* **2017**, *81*, e00019-17. [CrossRef]
28. Yadav, V.K.; Abraham, J.K.; Mani, P.; Kulshrestha, R.; Chowdhury, S. QuadBase: Genome-wide database of G4 DNA—Occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.* **2008**, *36*, D381–D385. [CrossRef]
29. König, S.L.B.; Huppert, J.L.; Sigel, R.K.O.; Evans, A.C. Distance-dependent duplex DNA destabilization proximal to G-quadruplex/i-motif sequences. *Nucleic Acids Res.* **2013**, *41*, 7453–7461. [CrossRef] [PubMed]
30. Mishra, S.K.; Jain, N.; Shankar, U.; Tawani, A.; Sharma, T.K.; Kumar, A. Characterization of highly conserved G-quadruplex motifs as potential drug targets in Streptococcus pneumoniae. *Sci. Rep.* **2019**, *9*, 1791. [CrossRef]
31. Rawal, P.; Kummarasetti, V.B.R.; Ravindran, J.; Kumar, N.; Halder, K.; Sharma, R.; Mukerji, M.; Das, S.K.; Chowdhury, S. Genome-wide prediction of G4 DNA as regulatory motifs: Role in Escherichia coli global regulation. *Genome Res.* **2006**, *16*, 644–655. [CrossRef]
32. Neidle, S. The structures of quadruplex nucleic acids and their drug complexes. *Curr. Opin. Struct. Biol.* **2009**, *19*, 239–250. [CrossRef] [PubMed]
33. Saranathan, N.; Vivekanandan, P. G-Quadruplexes: More than just a kink in microbial genomes. *Trends Microbiol.* **2018**, *27*, 148–163. [CrossRef]
34. Kaplan, O.I.; Berber, B.; Hekim, N.; Doluca, O. G-quadruplex prediction in E. coli genome reveals a conserved putative G-quadruplex-Hairpin-Duplex switch. *Nucleic Acids Res.* **2016**, *44*, 9083–9095. [PubMed]
35. Brocchieri, L. The GC Content of Bacterial Genomes. *J. Phylogenet. Evolut. Biol.* **2013**, *2*, 1–3. [CrossRef]
36. Ding, Y.; Fleming, A.M.; Burrows, C.J. Case studies on potential G-quadruplex-forming sequences from the bacterial orders Deinococcales and Thermales derived from a survey of published genomes. *Sci. Rep.* **2018**, *8*, 15679. [CrossRef]
37. Brumm, P.J.; Monsma, S.; Keough, B.; Jasinovica, S.; Ferguson, E.; Schoenfeld, T.; Lodes, M.; Mead, D.A. Complete Genome Sequence of Thermus aquaticus Y51MC23. *PLoS ONE* **2015**, *10*, e0138674. [CrossRef]
38. Waller, Z.A.E.; Pinchbeck, B.J.; Buguth, B.S.; Meadows, T.G.; Richardson, D.J.; Gates, A.J. Control of bacterial nitrate assimilation by stabilization of G-quadruplex DNA. *Chem. Commun.* **2016**, *52*, 13511–13514. [CrossRef] [PubMed]
39. Čechová, J.; Lýsek, J.; Bartas, M.; Brázda, V. Complex analyses of inverted repeats in mitochondrial genomes revealed their importance and variability. *Bioinformatics* **2018**, *34*, 1081–1085. [CrossRef]
40. Brázda, V.; Kolomazník, J.; Lýsek, J.; Hároníková, L.; Coufal, J.; Št'astný, J. Palindrome analyser—A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem. Biophys. Res. Commun.* **2016**, *478*, 1739–1745. [CrossRef]
41. Brázda, V.; Lýsek, J.; Bartas, M.; Fojta, M. Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs. *BioMed Res. Int.* **2018**, *2018*, 1097018. [CrossRef]
42. Ruggiero, E.; Richter, S.N. G-quadruplexes and G-quadruplex ligands: Targets and tools in antiviral therapy. *Nucleic Acids Res.* **2018**, *46*, 3270–3283. [CrossRef]
43. Chen, B.-J.; Wu, Y.-L.; Tanaka, Y.; Zhang, W. Small Molecules Targeting c-Myc Oncogene: Promising Anti-Cancer Therapeutics. *Int. J. Biol. Sci.* **2014**, *10*, 1084–1096. [CrossRef] [PubMed]

44. Belmonte-Reche, E.; Martínez-García, M.; Guédin, A.; Zuffo, M.; Arévalo-Ruiz, M.; Doria, F.; Campos-Salinas, J.; Maynadier, M.; López-Rubio, J.J.; Freccero, M.; et al. G-Quadruplex Identification in the Genome of Protozoan Parasites Points to Naphthalene Diimide Ligands as New Antiparasitic Agents. *J. Med. Chem.* **2018**, *61*, 1231–1240. [CrossRef]
45. Li, F.; Mulyana, Y.; Feterl, M.; Warner, J.M.; Collins, J.G.; Keene, F.R. The antimicrobial activity of inert oligonuclear polypyridylruthenium(II) complexes against pathogenic bacteria, including MRSA. *Dalton Trans.* **2011**, *40*, 5032–5038. [CrossRef]
46. Li, F.; Grant Collins, J.; Richard Keene, F. Ruthenium complexes as antimicrobial agents. *Chem. Soc. Rev.* **2015**, *44*, 2529–2542. [CrossRef] [PubMed]
47. Xu, L.; Chen, X.; Wu, J.; Wang, J.; Ji, L.; Chao, H. Dinuclear Ruthenium(II) Complexes That Induce and Stabilise G-Quadruplex DNA. *Chem. Eur. J.* **2015**, *21*, 4008–4020. [CrossRef]
48. Xu, L.; Zhang, D.; Huang, J.; Deng, M.; Zhang, M.; Zhou, X. High fluorescence selectivity and visual detection of G-quadruplex structures by a novel dinuclear ruthenium complex. *Chem. Commun.* **2010**, *46*, 743–745. [CrossRef] [PubMed]
49. Wilson, T.; Williamson, M.P.; Thomas, J.A. Differentiating quadruplexes: Binding preferences of a luminescent dinuclear ruthenium (II) complex with four-stranded DNA structures. *Org. Biomol. Chem.* **2010**, *8*, 2617–2621. [CrossRef] [PubMed]
50. Codd, G.A.; Lindsay, J.; Young, F.M.; Morrison, L.F.; Metcalf, J.S. Harmful cyanobacteria. In *Harmful Cyanobacteria*; Springer: Dordrecht, The Netherlands, 2005; pp. 1–23.
51. Perrone, R.; Lavezzo, E.; Riello, E.; Manganelli, R.; Palù, G.; Toppo, S.; Provvedi, R.; Richter, S.N. Mapping and characterization of G-quadruplexes in Mycobacterium tuberculosis gene promoter regions. *Sci. Rep.* **2017**, *7*, 5743. [CrossRef] [PubMed]
52. Lavezzo, E.; Berselli, M.; Frasson, I.; Perrone, R.; Palù, G.; Brazzale, A.R.; Richter, S.N.; Toppo, S. G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide. *PLOS Comput. Biol.* **2018**, *14*, e1006675. [CrossRef]
53. Beaume, N.; Pathak, R.; Yadav, V.K.; Kota, S.; Misra, H.S.; Gautam, H.K.; Chowdhury, S. Genome-wide study predicts promoter-G4 DNA motifs regulate selective functions in bacteria: Radioresistance of D. radiodurans involves G4 DNA-mediated regulation. *Nucleic Acids Res.* **2013**, *41*, 76–89. [CrossRef]
54. Kota, S.; Dhamodharan, V.; Pradeepkumar, P.I.; Misra, H.S. G-quadruplex forming structural motifs in the genome of Deinococcus radiodurans and their regulatory roles in promoter functions. *Appl. Microbiol. Biotechnol.* **2015**, *99*, 9761–9769. [CrossRef]
55. Repoila, F.; Darfeuille, F. Small regulatory non-coding RNAs in bacteria: Physiology and mechanistic aspects. *Biol. Cell* **2009**, *101*, 117–131. [CrossRef] [PubMed]
56. Brázda, V.; Kolomazník, J.; Lýsek, J.; Bartas, M.; Fojta, M.; Št'astný, J.; Mergny, J.-L. G4Hunter web application: A web server for G-quadruplex prediction. *Bioinformatics* **2019**, btz087. [CrossRef]
57. Sayers, E.W.; Agarwala, R.; Bolton, E.E.; Brister, J.R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2019**, *47*, D23–D28. [CrossRef]
58. Federhen, S. The NCBI taxonomy database. *Nucleic Acids Res.* **2011**, *40*, D136–D143. [CrossRef]
59. Letunic, I.; Bork, P. Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **2016**, *44*, W242–W245. [CrossRef]

Sample Availability: Not available.