PLOS ONE

RESEARCH ARTICLE

Investigation of a Quadruplex-Forming Repeat Sequence Highly Enriched in *Xanthomonas* and *Nostoc* sp.

Charlotte Rehm[1], Lena A. Wurmthaler[1], Yuanhao Li[2], Tancred Frickey[2], Jörg S. Hartig[1]*

1 Department of Chemistry and Konstanz Research School Chemical Biology (KoRS-CB), University of Konstanz, Universitätsstr. 10, 78457 Konstanz, Germany, **2** Department of Biology, University of Konstanz, Universitätsstr. 10, 78457 Konstanz, Germany

* joerg.hartig@uni-konstanz.de

Citation: Rehm C, Wurmthaler LA, Li Y, Frickey T, Hartig JS (2015) Investigation of a Quadruplex-Forming Repeat Sequence Highly Enriched in *Xanthomonas* and *Nostoc* sp.. PLoS ONE 10(12): e0144275. doi:10.1371/journal.pone.0144275

Editor: Paul Jaak Janssen, Belgian Nuclear Research Centre SCK•CEN, BELGIUM

Received: August 10, 2015

Accepted: November 16, 2015

Published: December 22, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

In prokaryotes simple sequence repeats (SSRs) with unit sizes of 1–5 nucleotides (nt) are causative for phase and antigenic variation. Although an increased abundance of heptameric repeats was noticed in bacteria, reports about SSRs of 6–9 nt are rare. In particular G-rich repeat sequences with the propensity to fold into G-quadruplex (G4) structures have received little attention. In silico analysis of prokaryotic genomes show putative G4 forming sequences to be abundant. This report focuses on a surprisingly enriched G-rich repeat of the type GGGNATC in *Xanthomonas and* cyanobacteria such as *Nostoc*. We studied in detail the genomes of *Xanthomonas campestris pv*. *campestris* ATCC 33913 (*Xcc*), *Xanthomonas axonopodis pv*. *citri* str. 306 (*Xac*), and *Nostoc sp*. strain PCC7120 (*Ana*). In all three organisms repeats are spread all over the genome with an over-representation in non-coding regions. Extensive variation of the number of repetitive units was observed with repeat numbers ranging from two up to 26 units. However a clear preference for four units was detected. The strong bias for four units coincides with the requirement of four consecutive G-tracts for G4 formation. Evidence for G4 formation of the consensus repeat sequences was found in biophysical studies utilizing CD spectroscopy. The G-rich repeats are preferably located between aligned open reading frames (ORFs) and are under-represented in coding regions or between divergent ORFs. The G-rich repeats are preferentially located within a distance of 50 bp upstream of an ORF on the anti-sense strand or within 50 bp from the stop codon on the sense strand. Analysis of whole transcriptome sequence data showed that the majority of repeat sequences are transcribed. The genetic loci in the vicinity of repeat regions show increased genomic stability. In conclusion, we introduce and characterize a special class of highly abundant and wide-spread quadruplex-forming repeat sequences in bacteria.

Introduction

Non-B DNA structures have been identified in eukaryotes as well as prokaryotes [1, 2]. Z-DNA is formed by alternating purine/pyrimidine patterns [3, 4] and A- or H-DNA by oligo-purine or—pyrimidine runs [5, 6]. Other examples of sequences that can give rise to non-canonical DNA structures include palindromes and close inverted repeats [7], simple sequence repeats (SSRs) [8, 9] as well as G-quadruplex (G4) forming sequences [10, 11]. Among these different structural elements mutagenic effects on DNA have been associated especially to SSRs [12]. These perfect (or near-perfect) direct iterations of short DNA tracts in a head-to-tail manner with a motif size of 1–9 nt are also termed 'tandem repeats' [9]. In bacteria next to SSRs a number of other small repeat classes have been identified primarily in intergenic regions, e.g. Miniature Inverted-repeat Transposable Elements (MITEs) [13, 14], Repetitive Extragenic Palindromic sequences (REPs) [15] and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) [16, 17]. All three belong to the general class of inverted repeats. In addition to genomic instability there is increasing evidence for non-canonical nucleic acid structures to directly or indirectly influence replication, recombination, transcription and translation on the DNA or RNA level [1, 10, 18–23].

So far research on tandem repeats has primarily been focused on short 1–4 nt repeats of which every possible combination has been found to be vastly over-represented in the human genome [8]. In particular trinucleotide expansions in open reading frames (ORFs), introns or untranslated regions (UTRs) have been identified to give rise to human neurodegenerative disorders such as Huntington disease [24], spinobulbar muscular atrophy [25] and Fragile X syndrome [26]. Although microsatellites have been found in prokaryotes as well, they are present at lower numbers [27]. Especially longer repeat sequences are less abundant than in eukaryotes [8]. The distribution of SSRs across bacterial species has been shown to vary greatly even among close relatives [28–30]. In general, SSRs with smaller unit sizes of 1–4 nt are found more abundantly in smaller genomes, especially those of host-adapted pathogens and of low G+C content [31–33]. In contrast, longer repeat runs were more frequently found in non-pathogens and bacteria with large genomes (> 4Mb) and high G+C content (> 60%) [33]. Major differences were detected in the distribution of SSRs in coding and non-coding regions. In *Escherichia coli* (*E. coli*) most repeat sequences were found to concentrate in intergenic regions up to 200 nt upstream of the start codon, the region containing proximal regulators of gene expression. Investigation of tandem repeats in *E. coli* by Gur-Arie et al. also showed them to be under-represented in ORFs when exceeding a unit size of 3 nt [34]. SSRs play a role in bacterial evolution, where they allow for local sequence variation and thereby enable accelerated adaption to changing environmental conditions [35, 36]. By inducing local genetic instability SSRs have been shown to act as cis-regulatory motifs enabling the modulation of gene expression in a reversible manner, especially in phase and antigenic variation [22, 23, 37]. Both processes allow the switching of phenotypes in a bacterial population and thereby are thought to increase their fitness.

Research conducted on SSRs with longer repeat units of 5–9 nt is rare. In 1999 van Belkum et al. presented a study on the occurrence of pentameric tandem repeats in bacterial genomes [38]. Although heptameric repeats were found to be over-represented among SSRs in many bacterial genomes in 2007 [33] no detailed literature focusing on heptameric repeats is available to date. Van Belkum et al. report one example of a heptanucleotide `5′ -GTGATTA-3′` in *Helicobacter pylori* [38]. The presence of three different tandemly repetitive heptanucleotides has also been reported for the cyanobacterium *Calothrix* sp. strain PCC7601 [39]. However, no further characterization of these repeats has been carried out. Recently, Mrázek and Huang presented an extensive assessment of local sequence patterns with the potential to form non-canonical DNA conformations from 1424 bacterial chromosomes [20]. A different representation of short

versus long SSRs was reported with longer tandem repeats showing normal or slight over-representation. When analyzing Mrázek and Huang's data for γ-proteobacteria and cyanobacteria only, we noted a strong over-representation of heptameric SSRs in intergenic regions in *Xanthomonas* and *Nostoc* species, while other long SSRs in the range of 4–11 nt were normally represented. Furthermore, a slight over-representation of intergenic G4 forming sequences is present in xanthomonads, strong over-representation is evident for *Nostoc* species. G4s are four-stranded helical complexes that are assembled from multiple stacked guanine tetrads. These specialized secondary structures can be formed either by DNA or RNA consisting of consecutive runs of guanines. G-rich repeats are of special interest as in addition to being SSRs they also represent potential G4 forming sequences. G4 structures have been shown to be able to carry out a variety of cellular functions in eukaryotes, e.g. in replication and recombination [10] or as transcriptional regulators [40–42]. However, much less is known about their function in the eubacterial kingdom of life. In an earlier study Chowdhury and co-workers identified potential G4 forming sequences in 18 bacterial strains and report them to be over-represented in regulatory regions [43]. We have previously shown that G4s can be used as translational repressors in an artificial system in bacteria [44]. Recently, we have studied the multifaceted effects of G4s as potent transcriptional and translational regulators in *E. coli*. The influence of G4 sequences proved to depend strongly on strand orientation and the exact location within the promoter region, 5′-UTR or 3′-UTR [21]. In this report we focus on G-rich heptameric repeats of the type `GGGAATC` in the plant pathogens *Xanthomonas campestris pv. campestris* ATCC 33913 (*Xcc*) [45] and *Xanthomonas axonopodis pv. citri* str. 306 (*Xac*) [46]. In addition we studied similar `GGGGA(T/C)T` repeats in the cyanobacterium *Nostoc sp.* strain PCC7120 (*Ana*) [47].

Materials and Methods

Identification and characterization of repeat patterns

Potential G4 forming sequences were initially obtained from ProQuad Database (http://quadbase.igib.res.in/) [48]. Using the following query parameters for *Xcc*: pattern G (or C for minus strand), stem size G3 (or C3) and loop size L1-5, genomic location: all. For further studies the chromosomal sequences of *Xcc* (NC_003902), *Xac* (AE008923), plasmids pXAC33 (NC_003921) and pXAC64 (NC_003922) and Ana (NC_003272) were downloaded from the NCBI website. *Xcc* and *Xac* genomes were manually searched for repeats comprising at least two units and containing at least once the heptamer "GGGAATC" using the software Clone Manager 9 (Scientific & Educational Software). For *Ana* stem size G3-5 (or C3-5) and loop size L1-7 was used in the ProQuad search. From this set all patterns of the type G_{4}-$L_{1\text{-}4}$ containing at least twice the units `GGGGA(C/T)T` were selected. Frequency plots showing the consensus nucleotide sequence of a heptameric unit were created with WebLogo (http://weblogo.berkeley.edu/logo.cgi) [49]. Distances from the respective start or end point of the repeat to the start or stop codon to the next neighboring ORF were calculated. Subsequently repeats were grouped into three categories of increasing distance between the repeat motif and the start or stop codon of the neighboring ORF of 0–50 bp, 50–100 bp and $>$ 100 bp. Functions of annotated genes and their positions on the genome were collected from KEGG (http://www.kegg.jp/), NCBI as well as Cyanobase (http://genome.microbedb.jp/cyanobase) [50]. Repeat associated genes were sorted into functional categories using the KEGG pathway mapper (http://www.genome.jp/kegg/mapper.html) [51].

Circular Dichroism (CD) Measurements

Oligonucleotides (Table A in S3 File) for CD measurements and melting assays were synthesized by Sigma Aldrich (Steinheim, Germany) at the 1 μmol scale with HPLC purification. CD

spectra were recorded on a JASCO-J815 spectropolarimeter equipped with a MPTC-490S/15 multicell temperature unit using quartz cells with 1 cm optical path. Oligonucleotides were prepared in a reaction volume of 600 μL as a 5 μM solution in 10 mM Tris-HCl or 10 mM sodium acetate for C-rich oligonucleotides and adjusted to the indicated pH 4.5–7.5 with HCl. If noted, the solution was supplemented with either KCl, NaCl or LiCl to the indicated concentration. Oligonucleotides were denatured by heating to 98°C for 5 min, followed by slow cooling to 20°C over night. Scans were performed at 20°C over a wavelength range of 220–320 nm (5 accumulations) with a scanning speed of 500 nm/min, 0.5 s response time, 0.5 nm data pitch and 1 nm bandwidth. The buffer spectrum was subtracted and all spectra zero-corrected at 320 nm. For thermal denaturation oligonucleotides were prepared as previously described. Due to the temperature dependent pH change of tris buffer, melting experiments of C-rich oligonucleotides were carried out in sodium acetate buffer only. Samples were heated from 20°C to 100°C at a rate of 0.5°C/min. The CD signal was recorded every 0.5°C at the indicated wavelength. The temperature of the half-maximal decay of ellipticity $T_{1/2}$ was obtained from the normalized ellipticity decrease using the Boltzmann sigmoidal fit.

Analysis of sequence homology between *Xcc* and *Xac* in repeat containing regions

Nucleotide BLAST [52] (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to compare sequence similarity between *Xcc* and *Xac* applying the following parameters: algorithm: blastn (somewhat similar sequences), database: NCBI genomes (chromosome), organism: *Xanthomonas axonopodis pv. citri str*. 306 (taxid: 190486). The entire repeat containing intergenic region and the next up- and downstream neighboring ORFs or the entire ORF containing an intragenic repeat of *Xcc* were used as query sequence. Presence of the repeat was assessed. Sites where the alignments showed less homology or gaps were then checked directly in Clone Manager for repeat presence and compared for general changes in the intergenic regions and neighboring genes. 260 intergenic regions that did not contain `GGGAATC` repeats including the next neighboring ORFs were randomly chosen from the *Xcc* genome and subjected to the same blast analysis. Control sets were randomly assembled from this pool of controls to contain 117 queries each. From the same pool sequences for control 4 were chosen to show the same distribution along the *Xcc* genome and sequences for control 5 were chosen to show the same orientation of neighboring ORFs as the intergenic repeat containing sequences. One-sample t-tests were carried out using R (version 3.0.2) for each category. Distribution of the orientation of the neighboring genes relative to the repeats was analyzed for all controls.

Analysis of whole transcriptome sequencing data of *Xac*

Paired-end reads of *Xac* (referred to as XccA306 by Jalan et al. [53]) of NB sample 2 were downloaded from Gene Expression Omnibus database of NCBI (accession number GSE41519). Read quality was first checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (version 0.11.2) and then trimmed with Trimmomatic [54] (version 0.32). Trimmed reads were then mapped to the *Xac* genome using bowtie2 [55] (version 2.2.3). Uniquely mapped reads were assembled by Trinity [56, 57] (version r20140717). In total, 4266 transcripts were assembled, and their expression levels were calculated by aligning reads to each assembled transcript and normalizing them by Fragments Per Kilobase of exon per Million fragments mapped (FPKM). Assembled transcripts were then mapped to the *Xac* genome using blat [58] to obtain their respective coordinates on the genome. Number and orientation of repeat-containing transcripts was determined. Repeats were further classified as potential G4 forming repeats (at least 4 G-tracts without mutations) or short repeats unable to

form G4s (controls). Assembled transcripts were then sorted into the three categories according to the location of the repeat within the assembled transcript: start, middle or end. Reads mapping to repeats located in coding regions were excluded from the final analysis.

Results

GGGAATC repeat sequences in xanthomonads

The findings of Mrázek and Huang motivated us to investigate potential G4 forming sequences in the plant pathogen *Xanthomonas* in more detail. First we used the ProQuad Pattern Search [48] to gain an overview of potential G4 folding sequences in the *Xcc* genome (total of 270 potential G4s with G_3-$L_{1\text{-}5}$, Table A and B in S1 File). We hereby noticed an intriguing over-representation of GGGAATC repeat patterns among the putative quadruplex patterns, which led us to screen the genomes of *Xcc* and the related species *Xac* for GGGAATC-containing tandem repeats. The following parameters were used to define these G-rich SSRs: the total length must be ≥14 bp (at least 2 units) and contain at least once the GGGAATC heptamer. Repeats can be either perfect repeats $(GGGAATC)_n$ or heterogeneous $(GGGANTN)_n$. In total we identified 186 G-rich repeat patterns in *Xcc* and 183 in *Xac* (Table A and B in S2 File). The frequency plot in Fig 1A shows the consensus motif of a heptamer unit, in both organisms position 1–4 and 6 show high sequence conservation. Although extensive length variation was noted with repeats ranging from 2 to 26 units in *Xcc* and 2 to 18 units in *Xac*, the majority of the sequence motifs comprise four repeat units, as shown in the histogram in Fig 1B. 56% of all repeats in *Xcc* and 42% in *Xac* are made up of ≥4 units and have no point mutations in the G-tract, which would prevent G4 formation. An example for the longest perfect repeat with 14 GGGAATC units from *Xcc* is given in Fig 1C (top). Remarkably, in 70 cases in *Xcc* and 75 cases in *Xac* we found two repeat sites with convergent orientation in close proximity to each other, always located once on the plus and once on the minus strand of the genome. An example for such an inverted repeat is shown in Fig 1C (bottom). This rearrangement is of particular interest as inverted repeats have the potential to give rise to stem-loops or cruciform structures.

We found that GGGAATC repeat sequences are dispersed all over the genome in both species and do not show preference for a defined region on the chromosome, such as the origin or terminus of replication. Repeats are about equally distributed on the plus and minus strand of the chromosome and show no preference in regard to presence in the leading or lagging strand during replication (Fig 1D). In contrast to *Xcc*, *Xac* carries two plasmids, pXAC33 and pXAC64. No repeat sequences were identified on these plasmids. The repeats were most often found in intergenic regions (89% *Xcc*, 93% *Xac*) (Fig 1E) and are almost exclusively located at a shorter distance to the next 5' neighboring ORF (average distance 28 nt) than to the next downstream ORF (average distance 160 nt). Regarding the orientation of the neighboring ORFs to the intergenic repeats, we found that the majority of ORFs were oriented in the same direction, with the repeats localized in the intergenic region. In *Xcc* 30% of the G-rich patterns are present on the same strand as the aligned ORFs, and in 35% of cases are present on the opposite strand than the aligned ORFs. In *Xac* there are 35% of all repeats assigned to each of these categories. In both xanthomonads 16% of repeats were located between convergent ORFs, while only 8% in *Xcc* and 7% in *Xac* were located between divergent ORFs. Intragenic repeats are similarly rare, accounting to 11% in *Xcc* and 7% in *Xac* with at least partly overlap with the ORF (Fig 1E). Interestingly, although a high degree of sequence homology exists between *Xcc* and *Xac* [46, 59], repeats of similarly prominent length are not found in association with the same genes in the two species (Table B and C in S3 File).

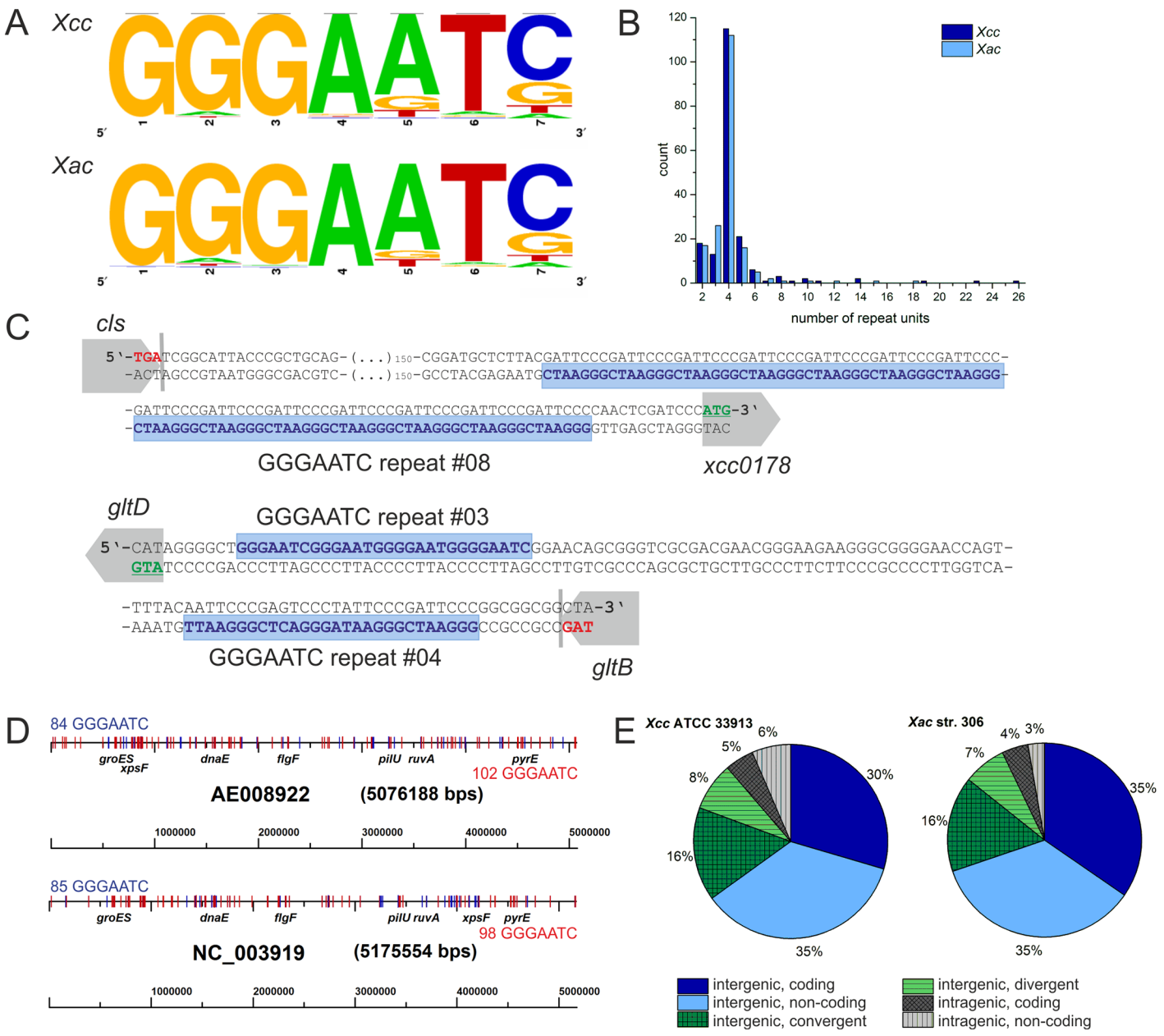

Fig 1. Overview of repeats in *Xanthomonas* species *Xcc* and *Xac*. (A) Frequency plot [49] shows the consensus nucleotide sequence of a heptameric repeat unit in *Xcc* (top) and *Xac* (bottom). (B) Histogram shows the count of repeat iterations per repeat sequence in *Xcc* (dark blue) and *Xac* (light blue). (C) Examples of GGGAATC patterns in *Xcc*. Repeat #08 located upstream of the hypothetical gene xcc0178 is the longest, perfect repeat present (top). Repeat #03 and #04 represent an inverted repeat with two short repeat sequences located in convergent orientation on the plus and on the minus strand of the genome (bottom). (D) Distribution of GGGAATC on the *Xcc* (AE008922, top) and *Xac* (NC_003919, bottom) genomes. Repeats located on the plus strand are marked in blue (84 *Xcc*, 85 *Xac*), repeats on the minus strand in red (102 *Xcc*, 98 *Xac*). Locations of repeat associated genes *groES*, *dnaE*, *flgF*, *pilU*, *ruvA*, *pyrE* and *xpsF* have been marked for orientation. (E) Orientation of neighboring genes relative to repeat sequences in *Xcc* (left) and *Xac* (right). Intergenic repeats can be located on the same strand that will serve as the coding strand of the aligned ORFs (dark-blue) or on the non-coding strand (light blue), between convergent (dark green) or divergent (light green) ORFs. Intragenic repeats can be located on the coding strand (dark gray) or non-coding strand (light gray).

doi:10.1371/journal.pone.0144275.g001

GGGGA(C/T)T repeat sequences in *Nostoc*

Generally, a very high repeat-coverage was found for cyanobacterial genomes [33]. Mrázek and Huang not only reported a strong over-representation of long SSRs, but in a later publication particularly of potential G4 forming sequences [20]. Even earlier, Swanson et al. noticed a long stretch of G-rich heptamer repeats in the *pec* (phycoerythrocyanin) locus of *Nostoc sp*. strain PCC7120 (*Ana*) (Fig 2A) [60], however to the best of our knowledge no further studies concerning this DNA pattern have been carried out to date. We therefore chose *Ana* for a more detailed examination. Despite its low G+C content of 41.3%, ProQuad pattern search ($G_{3\text{-}5}L_{1\text{-}5}$) yielded 471 hits (Table C and D in S2 File). Because of the high abundance of G-rich patterns we chose to focus our analysis on only repeat sequences containing at least twice the runs 5′ −GGGGATT−3′ or 5′ −GGGGACT−3′, similar to the patterns observed in xanthomonads. The analysis yielded 89 SSRs in total (Table C in S2 File). The frequency plot in Fig 2B shows the consensus nucleotide sequence GGGGA(T/C)T. The identified repeat patterns again varied strongly in length from 39 to 179 nt. The longest perfect GGGGATT pattern is a 26mer located within the *pec* operon (Fig 2A). Repeat patterns were again distributed all over the *Ana* chromosome, not restricted to specific genomic locations (Fig 2C) and almost equally distributed between the plus (43%) and minus strand (57%). Although the majority of repeats are located intergenically (68%), a significantly higher fraction of repeats is located intragenically than was the case for the xanthomonads. This is especially remarkable as the average codon usage in *Ana* shows a lower G+C content than codons used in xanthomonads (coding GC-content 42.34% in *Ana*, 65.58% *Xcc*, 65.06% *Xac*, http://www.kazusa.or.jp/codon). Regarding the orientation of the neighboring ORFs to the repeat only 18% of all repeats were oriented in the same direction with the neighboring ORFs and 34% are present on the opposite strand between aligned ORFs (Fig 2D). 11% of repeats were located between convergent ORFs and 9% between divergent ORFs. In contrast to *Xanthomonas* we found only a few paired repeats that could form inverted repeats. In addition to the chromosome *Ana* carries six plasmids, but repeat patterns were not found on the plasmids.

Oligonucleotides derived from repeat sequences form G4s in vitro

The majority of the sequence motifs comprise four (or more) repeat units. These consecutive runs of guanosines can give rise to G4s on the level of DNA as well as RNA. Hoogsteen base-pairing between the guanines arranges them in a square tetrameric formation, also called a tetrad. The quadruplex is then made up by several such tetrads stacking upon each other; stabilization of the compact structure is achieved by coordination of metal cations in the central cavity (Fig 3A) [61, 62]. We employed circular dichroism (CD) spectroscopy to study putative G4-formation of the repeat-derived DNA oligonucleotides *in vitro*. Stabilization of G4s by monovalent cations is dependent on the nature of the cation, in general the order of the degree of stabilization is $K^+ > Na^+ > Li^+$ [61]. We analyzed both the minimal motif needed to form a G4 consisting only of the four G-tracts and three loop regions, e.g. 5′ −$(GGGAATC)_3$GGG−3′, as well as the respective extended repeat motif 5′ −$(GGGAATC)_4$−3′. In case of the *Ana* sequences different G4 conformations are possible with the fourth guanine either being part of the loop sequence, e.g. 5′ −$(GGGACTG)_3$GGG−3′, or being located in the G-tract 5′ −$(GGGGACT)_3$GGGG−3′. Different G4 structures can be distinguished according to their signature in CD, a typical spectrum of an anti-parallel G4 shows a minimum at 260 nm and a maximum at 290 nm, while a G4 with parallel strand orientation shows a minimum at 240 nm and a maximum at 260 nm [63]. Different possible G4 topologies are shown in Fig 3B. For the G-rich motif from *Xcc* 5′ −$(GGGAATC)_3$GGG−3′ CD spectra in presence of K^+ showed a minimum in ellipticity at 240–250 nm, a shoulder at 270 nm and a maximum at 290 nm

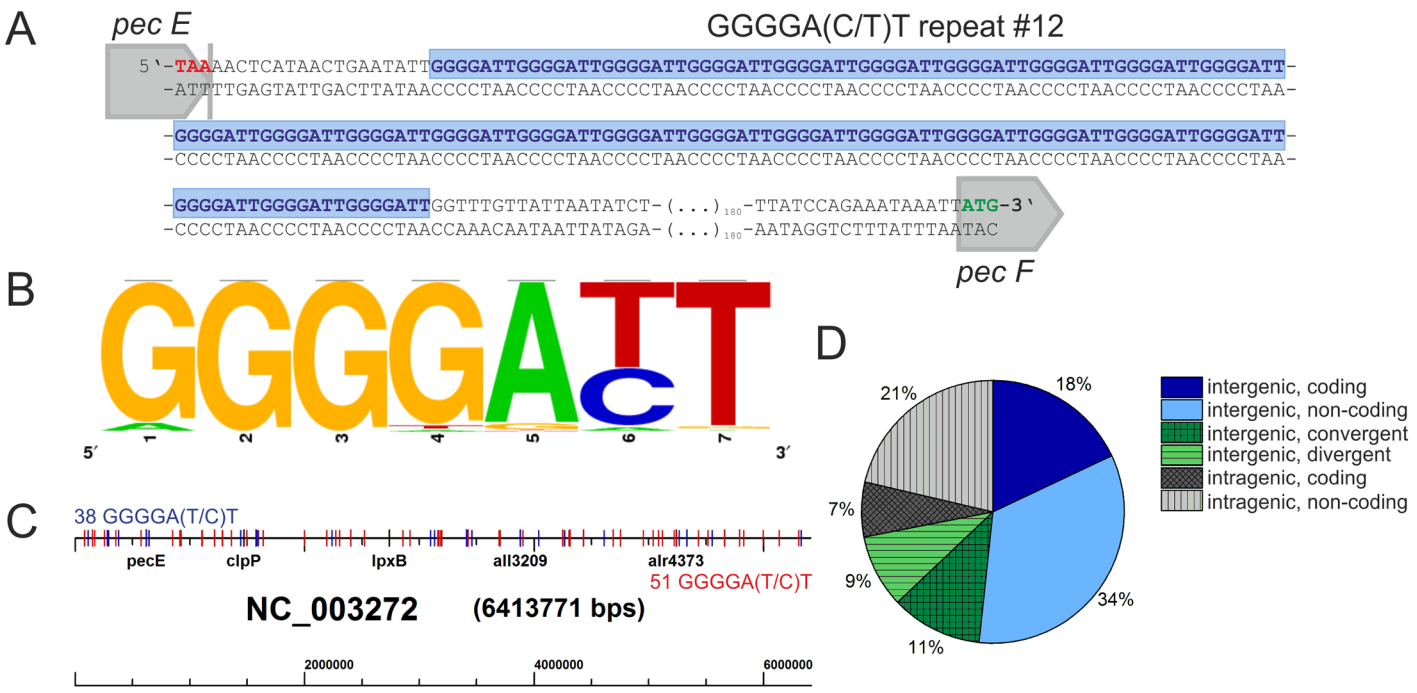

Fig 2. Overview of repeats in *Ana*. (A) Examples of a perfect 26mer GGGGATT repeat patterns in *Ana*. Repeat #12 is located in an intergenic region in the pec operon between *pecE* and *pecF*. (B) Frequency plot [49] shows the consensus nucleotide sequence of heptameric repeat units in Ana. (C) Distribution of GGGGA(T/C)T repeats on the *Ana* chromosome (NC_003272). Repeats located on the plus strand are marked in blue [38], repeats on the minus strand in red [51]. Locations of repeat associated genes *pecE*, *clpP*, *lpxB*, *all3209* and *alr4373* have been marked for orientation. (D) Orientation of neighboring genes relative to repeat sequences. Intergenic repeats can be located on the same strand that will serve as the coding strand of the aligned ORFs (dark-blue) or on the non-coding strand (light blue), between convergent (dark green) or divergent (light green) ORFs. Intragenic repeats can be located on the coding strand (dark gray) or non-coding strand (light gray).

doi:10.1371/journal.pone.0144275.g002

indicative for a (3+1) hybrid structure (Fig 3B **middle**, Fig 3C). The spectral change for the respective repeat motif is less pronounced (Fig 3D). As a control no structural changes could be observed in CD upon introduction of G to T mutations at the second position in the G-tract for the *Xcc* derived oligonucleotides (Fig A in S3 File). Possible quadruplex forming oligonucleotides from *Ana* showed clear formation of an antiparallel structure in the presence of KCl for 5′ – (GGGGACT)$_3$GGGG–3′ (Fig 3E), 5′ – (GGGGATT)$_3$GGGG–3′ (Fig 3F) and the repeat motifs 5′ – (GGGGACT)$_4$–3′ (Fig 3G) and 5′ – (GGGGATT)$_4$–3′ (Fig 3H). Peaks at 290 nm are also present in the spectra of 5′ – (GGGACTG)$_3$GGG–3′ (Fig 3I) and 5′ – (GGGATTG)$_3$GGG–3′ (Fig 3J) in solution with KCl. For these oligo types four guanines are present in the second and third G-tract which enables formation of a variety of G4 structures with three guanines in the G-tract. Spectra of these different structures formed may then overlap in CD. In all cases NaCl did not result to equally pronounced quadruplex formation as KCl and spectra in the presence of LiCl were similar to the unfolded state.

In order to assess thermodynamic stabilities of the structures formed in the presence of KCl and NaCl we performed thermal denaturation experiments. Melting temperatures $T_{1/2}$ are listed in Table D in S3 File. Melting profiles are shown in Fig B in S3 File. We determined moderate melting temperatures $T_{1/2}$ of 50.4°C for the *Xcc* quadruplex 5′ – (GGGAATC)$_3$GGG–3′ in the presence of 100 mM KCl. All sequences from *Ana* showed to be more stable than the *Xcc* quadruplex with $T_{1/2}$ higher than 74°C; in fact species with G-tracts comprising four guanines 5′ – (GGGGACT)$_3$GGGG–3′ , 5′ – (GGGGATT)$_3$GGGG–3′ and 5′ – (GGGGACT)$_4$–3′

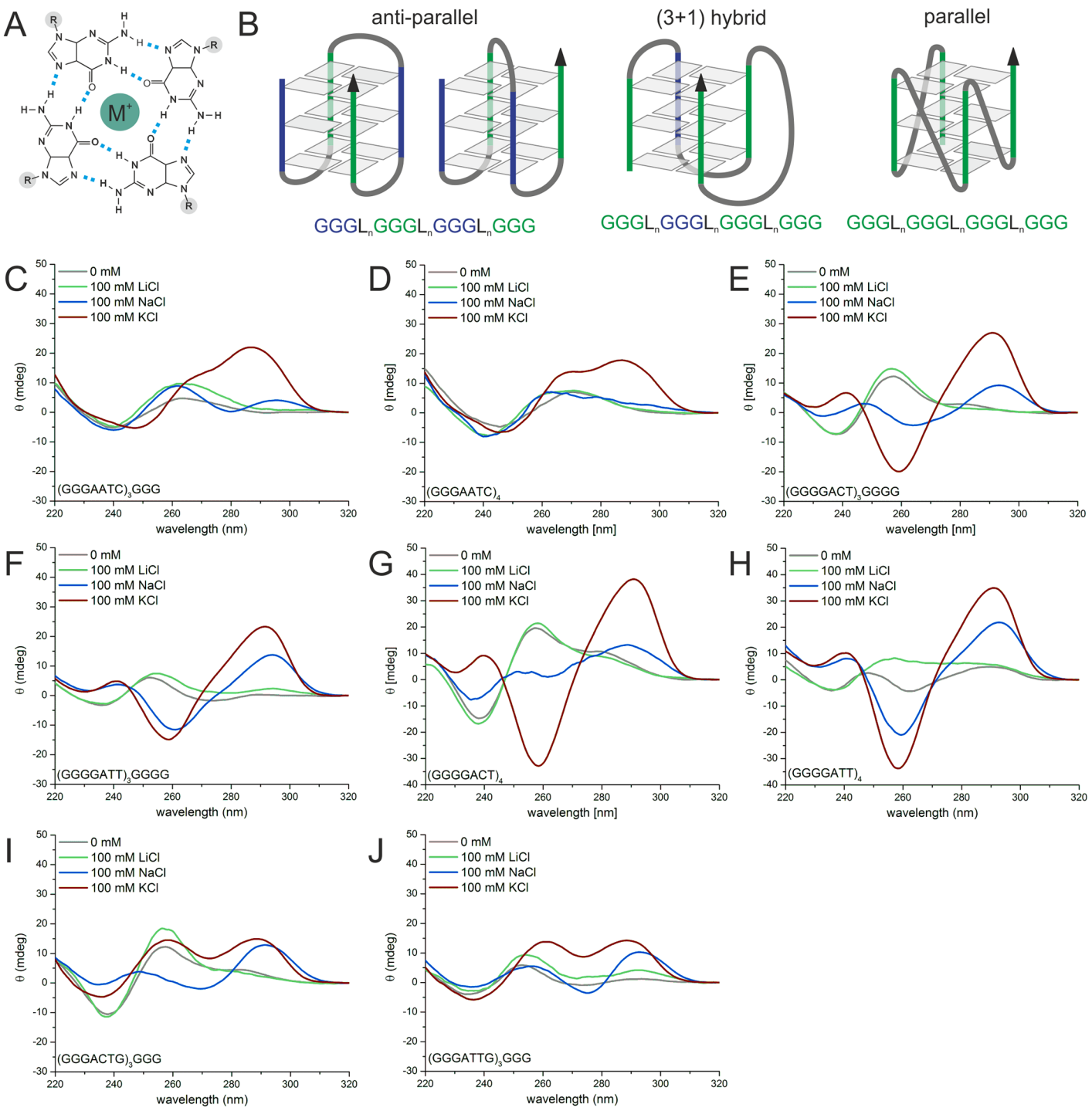

Fig 3. Circular dichroism analysis of G-rich repeat derived oligonucleotides. (A): Top view of a guanine tetrad formed by Hoogsteen base-pairing. Hydrogen bonds are depicted by light blue dashed lines. Monovalent cations M^+ (green) in the central cavity or between tetrads stabilize the structure. Sugar-phosphate backbone of the nucleic acid is depicted by R (highlighted in gray). (B): Schemes of different G4 topologies with three tetrads, from left to right: anti-parallel chair and anti-parallel basket structure, (3+1) hybrid structure and all-parallel propeller structure. Guanines forming a tetrad are represented by gray rectangles. General nucleic acids sequence is shown underneath, with Ln representing the nucleotides in the loop. Different strand orientations are indicated by blue (top to bottom) and green (bottom to top), the arrow indicates the 3' end. (C-J): CD spectra recorded from 220 to 320 nm of 5 μM oligonucleotide in 10 mM Tris-HCl (pH 7.5) in the presence of 100 mM LiCl (green), 100 mM NaCl (blue), 100 mM KCl (red) or tris buffer only (gray), $(GGGAATC)_3GGG$ (C), $(GGGAATC)_4$ (D), $(GGGGACT)_3GGGG$ (E), $(GGGGAAT)_3GGGG$ (F), $(GGGACTG)_3GGG$ (G), $(GGGATTG)_3GGG$ (H), $(GGGGACT)_4$ (I) and $(GGGGATT)_4$ (J).

doi:10.1371/journal.pone.0144275.g003

could not be fully denatured in presence of KCl with $T_{1/2}$ >95°C. In all cases structures folded in the presence of 100 mM NaCl were less stable than their K^+ stabilized counterparts.

Since the presence of a G-rich genomic repeat pattern is accompanied by the presence of a C-rich pattern on the complementary strand, we investigated the formation of a four-stranded structure of the C-rich motif. The so-called i-motif structure is formed from C-rich oligonucleotides at mild acidic conditions, which enables the formation of hemiprotonated cytosine-cytosine$^+$ base pairs (Fig 4A) [64]. Formation of the i-motif is favored at lower pH, although some sequences are able to stably fold i-motif structures even at neutral pH [65]. CD spectra show a characteristic minimum at about 260 nm and a maximum at around 290 nm [66]. We determined the folding behavior of the complementary C-rich repeat strands while decreasing pH from pH 7.5 to 4.5. CD spectra of the C-rich oligonucleotides derived from *Xcc* already showed a minimum at 240 nm and a maximum at about 270 nm suggesting a folded structure of unknown nature at neutral pH. As the pH of the buffer is decreased the spectrum shifts showing a minimum at 240 nm, shoulder at 260–270 nm and maximum at 280 nm at pH 4.5 suggesting overlapping spectra of different conformations, possibly including an i-motif at 290 nm (Fig 4B and 4C). I-motif signatures were readily detectable in all C-rich oligonucleotides derived from *Ana* (Fig 4D–4I). Remarkably all observed structures persisted even at the elevated pH of 6.5.

We also assessed the thermodynamic stability of the structures formed under acidic conditions (**Table E, Fig C** CD spectra and **Fig D** melting profiles in S3 File). At pH 4.5 all structures are fairly stable with $T_{1/2}$ ranging between 60–72°C. I-motifs have been reported to be destabilized by increased ion concentrations [67, 68], however we found that addition of 100 mM NaCl or KCl did not disturb i-motif formation at pH 4.5. Raising pH to 6.5 lead to a destabilization of the formed structures with $T_{1/2}$ dropping by 15–29°C in comparison to the $T_{1/2}$ determined at pH 4.5, except for $(AGTCCCC)_4$, which showed a weaker decrease of only 4°C.

In summary, characteristic changes in ellipticity and enhanced thermodynamic stability were indeed observed under conditions favoring either G4 or i-motif formation. K^+ has been reported to be the major cation in the bacterial cell, cytosolic concentrations of about 200 mM were determined for *E. coli* [69]. A concentration of 100 mM K^+ therefore represents a concentration likely to be achieved in a cellular environment to stabilize potential G4s.

Repeats in intergenic regions

During mapping of the repeat sequences we noticed that intergenic repeats are almost exclusively located at a shorter distance to the next 5' neighboring ORF than to the next downstream ORF irrespective of the ORFs orientation on the genome. We therefore decided to analyze the distance distribution of intergenic repeats in relation to the next neighboring ORF in more detail. We distinguished between a repeat's position upstream on the coding or non-coding strand of an ORF as well as downstream on the coding or non-coding strand. Repeats were grouped according to increasing distance from the ORF. In all three species intergenic repeats patterns showed a similar distribution (Fig 5A–5C): Upstream of the ORF the greatest fraction is localized within 0–50 bp from the ORF on the non-coding strand (Fig 5D). If the pattern is located on the coding strand the distance to the start codon increases. Downstream of the ORF the situation is reversed: most repeats are located within a distance of 0–50 bp from the stop codon on the coding strand. This includes all repeats overlapping with the stop codon (Fig 5E). When localized on the non-coding strand, the distance to the end of the ORF again increases. When considering only repeats able to form G4s for *Xcc*, we found the same distribution as when also taking into account shorter and mutated repeats (Fig E in S3 File). A preference for the non-coding strand can be observed for *Ana*.

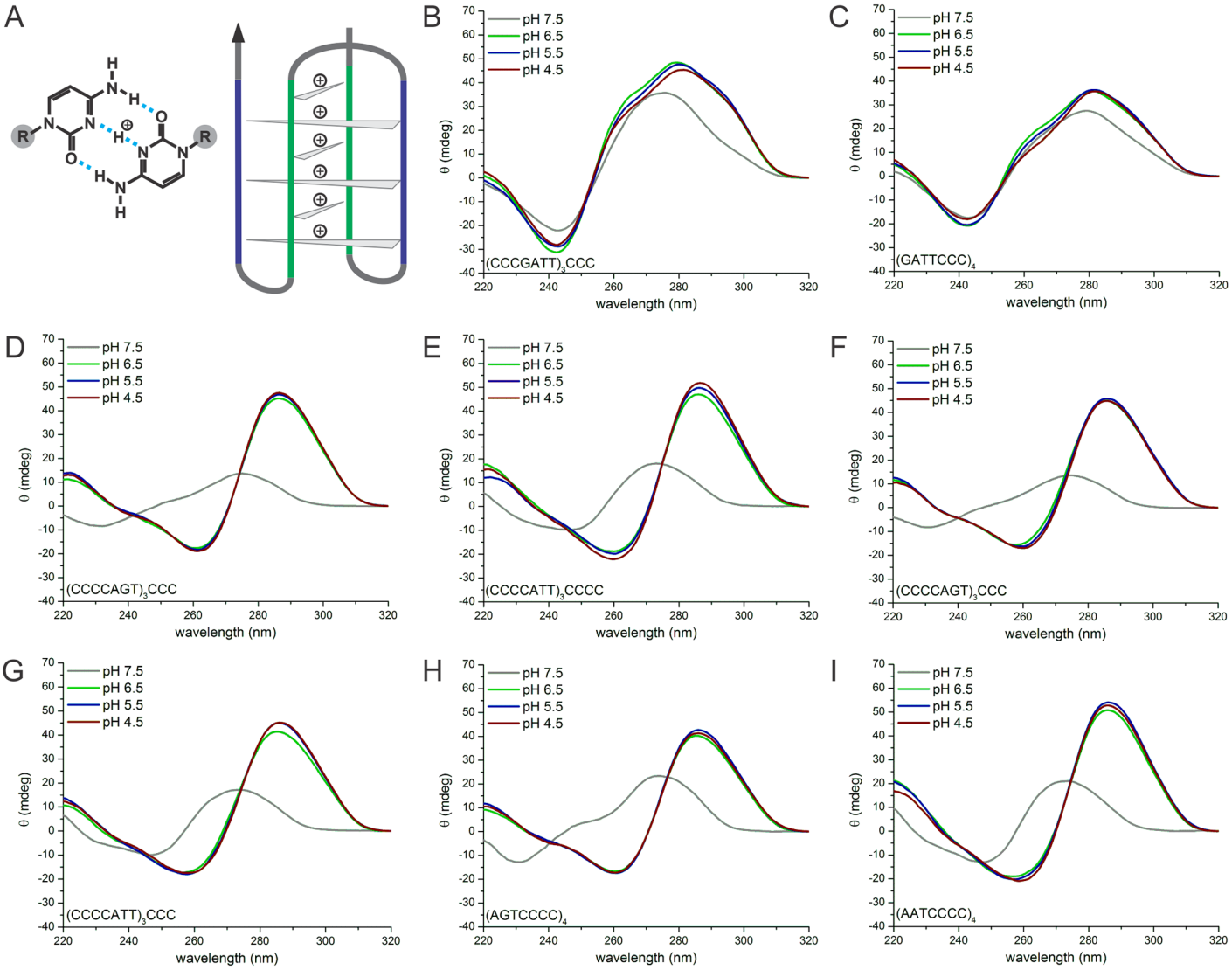

Fig 4. Circular dichroism analysis of C-rich complementary repeat oligonucleotides. (A) Left: Hemiprotonated cytosine-cytosine+ base pair. Hydrogen bonds are depicted by light blue dashed lines, sugar-phosphate backbone of the nucleic acid is depicted by R (highlighted in gray). Right: Scheme of an i-motif formed by a duplex between parallel oriented strands intercalated with anti-parallel duplex. Gray triangles represent cytosine-cytosine+ base pair. Different strand orientations are indicated by blue (top to bottom) and green (bottom to top), the arrow indicates the 3' end. (B-I) CD spectra recorded from 220 to 320 nm of 5 μM oligonucleotide in 10 mM Tris-HCl pH 7.5 (gray). pH 6.5 (green), pH 5.5 (blue) pH 4.5 (red) for $(CCCGATT)_3CCC$ (B), $(GATTCCC)_4$ (C), $(CCCCAGT)_3CCCC$ (D), $(CCCCAAT)_3CCCC$ (E), $(CCCCAGT)_3CCC$ (F), $(CCCCAAT)_3CCC$ (G), $(AGTCCCC)_4$ (H) and $(AATCCCC)_4$ (I).

doi:10.1371/journal.pone.0144275.g004

G4s have been shown to be potent modulators of gene expression in eukaryotes and bacteria [10, 21, 40–42, 44, 70] when they are located in close proximity to an ORF, e.g. in the promoter region or UTR. To gain further insight into a potential biological role of the repeat patterns we classified the neighboring genes according to functional classes using the KEGG database [51]. Many of the genes associated to the repeats sequences are hypothetical genes with no further functional description (55% in *Xcc* and *Xac*, 69% in *Ana*). The remaining genes belong mainly to general metabolism pathways. All three organisms show a similar distribution across the gene functional classes (overview and detailed lists in S4 File). Repeats are not exclusively associated to known cell surface structures or genes involved in adaption processes, making a possible function similar to SSRs in phase variation unlikely. In addition we generally did not find

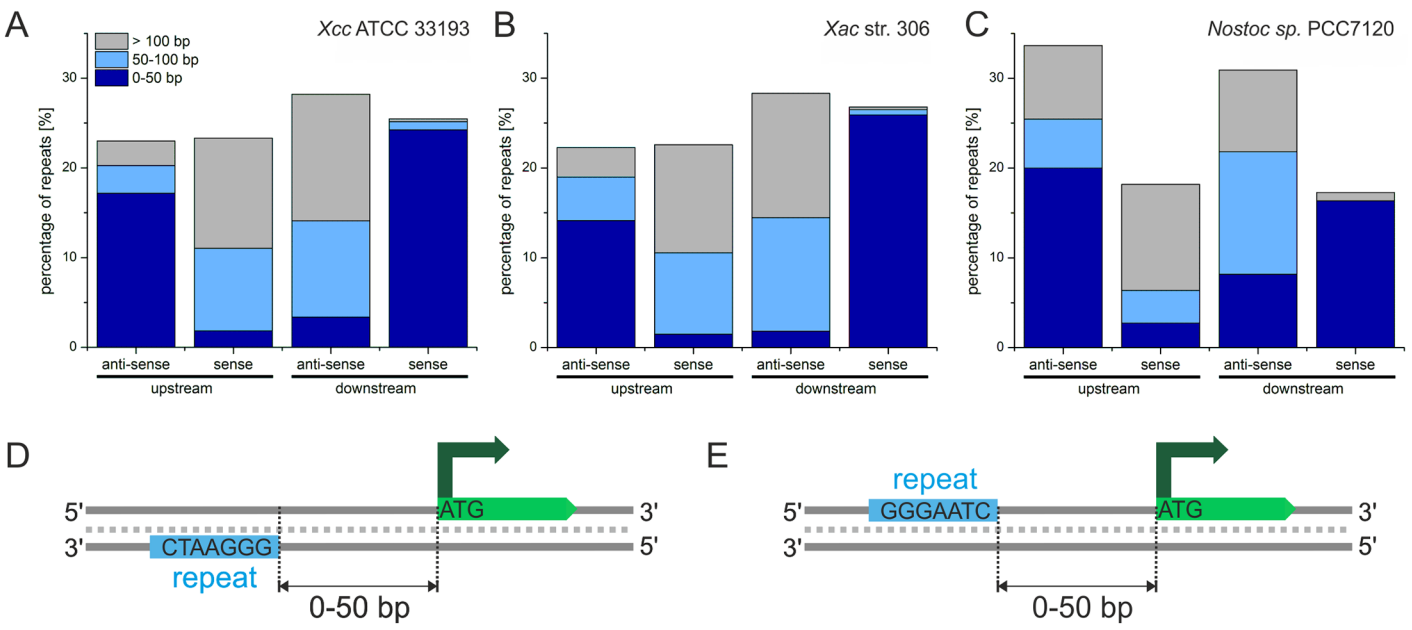

Fig 5. Distances of repeats to neighboring ORFs. (A-C) Analysis of the distance of repeat sequences relative to adjacent ORFs for *Xcc* (A), *Xac* (B) and *Ana* (C). Repeats can be either located upstream or downstream of the next neighboring ORF. Repeats were grouped into three categories according to increasing distance from the respective: distance of > 100 bp (gray), 50–100 bp (light blue) and repeats overlapping with ORFs or located in a distance of up to 50 bp from the respective start or stop codon are grouped together (dark blue). (D) Schematic of a repeat being located in close proximity, upstream of the neighboring ORF on the non-coding strand. (E) Schematic of a repeat being located in close proximity, downstream of the neighboring ORF on the non-coding strand.

doi:10.1371/journal.pone.0144275.g005

them associated with genetically mobile elements such as insertion sequences or transposable elements. However, in order to characterize whether the motif results in increased genetic instability we analyzed the genetic variability in repeat-containing regions.

Analysis of sequence homology in repeat-containing regions in xanthomonads

SSRs have been implicated as locations of genomic instability [1, 9, 37, 71, 72]. We used nucleotide blast (algorithm: blastn) to compare sequence similarity between the close relatives *Xcc* and *Xac* in repeat containing regions. Therefore all intergenic region containing a repeat and the complete neighboring ORFs, or complete ORFs containing an intragenic repeat of *Xcc* were aligned against the *Xac* genome (Table A in S5 File). We first assessed whether repeats from *Xcc* were also represented by G-rich repeat patterns at the same position in the *Xac* genome. 83% of the repeats were also present in the same gene context in *Xac*. For 16% we could not detect a G-rich pattern in the alignment or the G-rich stretch was strongly mutated. In two cases no alignment was possible between *Xcc* and *Xac* (Fig 6A). Furthermore we noticed differences in the length of the repeats between the two organisms, however the type of the repeat (singular repeat or inverted repeat pair) was usually preserved.

Next, we assessed changes of the identity of the neighboring genes for the repeats located in intergenic regions only (117 regions) (Fig 6B). Sequences were therefore grouped into the following categories according to their degree of sequence variability: "No homology" refers to all cases in which sequence alignment was impossible, "no homology of flanking region" refers to all cases in which one ORF was homologous, but the other neighboring ORF including the intergenic region was not homologous. We further distinguished between insertions of

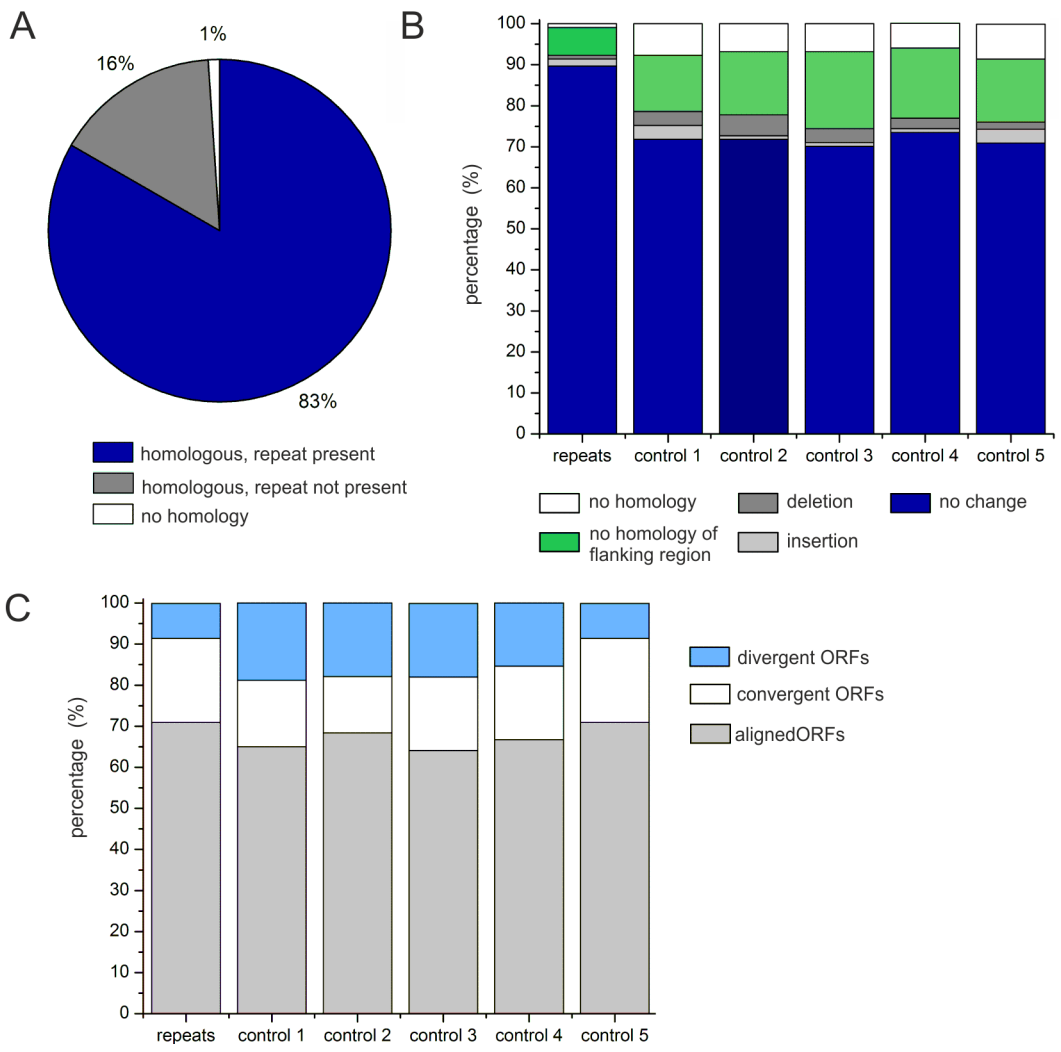

Fig 6. Sequence comparison between repeat containing regions in *Xcc* and *Xac*. (A): Presence of repeats patterns in *Xac* for repeat containing sequences from *Xcc*. Homologous repeats are depicted in blue, absent or mutated repeats depicted in gray, non-homologous alignments in white. (B) Analysis of changes of the identity of the neighboring genes for intergenic repeats from *Xcc* in comparison to *Xac*. Perfect alignments are grouped as "no change" (blue). Deletions (dark gray) or insertions (light gray) into intergenic regions were detected. Alignments showing only homology for one neighboring ORFs were grouped as "flanking region changes" (green). Non-homologous alignments are shown in white. (C) Orientation of neighboring genes relative to intergenic regions are shown for the repeat-containing intergenic regions from *Xcc* and the randomly chosen control sets 1–5. Sequences of control 5 were chosen to reflect the orientation of genes as found for the repeat containing intergenic regions in *Xcc*. ORFs can be either aligned (gray), convergent (white) or divergent (blue).

doi:10.1371/journal.pone.0144275.g006

fragments comprising one to several genes in the intergenic region and deletions of neighboring genes. Finally, alignments with high similarity throughout were grouped as "no change". For comparison we carried out the same analysis with 260 randomly chosen intergenic regions from *Xcc* that did not contain GGGAATC repeats (Table B in S5 File). From this pool of controls we randomly assembled three control sets with 117 sequences each (control 1 –control 3). In addition we assembled a fourth control set that mimics the overall distribution of the repeats along the *Xcc* genome (control). We found that 90% of the repeats were located between the same ORFs in *Xcc* and *Xac* (Fig 6B). Deletions or insertions in the intergenic regions, changes in flanking regions as well as no homology in the overall alignment were rare, altogether accounting to 10%. In contrast, these fractions of genomic changes were considerably higher in

the random control sets accounting to roughly 30%. We analysed the statistical relevance of the data presented in Fig 6B by carrying out one-sample t-tests for each category. Using the 5 frequency values for each category in the five controls as background, the probability of observing a value equal to, higher or lower than the repeat group was calculated. The t-test for the category "no change" between the repeat group and the five control groups shows significance with a p-value = 2.857e-06.

This indicates that overall the investigated repeats are located at more conserved genomic locations. This finding is in contrast to the genomic instability of many previously characterized SSRs. When analyzing the orientation of the neighboring ORFs of the repeat set and the control 1–4, we noticed a bias for the control groups containing more intergenic regions located between divergent ORFs. To rule out an effect of this arrangement on our analysis in Fig 6B, we assembled a fifth control in which the orientation of the neighboring ORFs with respect to the intergenic region is the same as for the repeat sample (control 5) (Fig 6C). Also for control 5 we found a higher fraction of deletions, insertions and changes in the flanking regions in comparison to the repeat set (Fig 6B).

Analysis of whole transcriptome sequencing data of *Xac*

Xanthomonads are plant pathogens. Since we identified the heptameric G-rich repeats in the genus *Xanthomonas* but not in other γ-proteobacteria, we considered a possible role of these putative G4-forming sequences in controlling a pathogenesis-related mechanism. Recently, whole transcriptome sequencing data became publicly available for *Xac* grown in full medium "NB" and hypersensitive response-elicitating medium "XVM2", the latter mimicking plant infection [53]. Jalan et al. identified 229 differentially expressed genes (≥3 fold up- or down-regulation) in XVM2 in comparison to NB. Reviewing this data we found that among the 173 up-regulated genes in XVM2 only 5 genes were associated with repeats (*aroG*, *kdpC*, *asnC*, *suc1*, *fecA*). Likewise of the 119 down-regulated genes 6 were connected with repeats /*cheA*, *flhB*, *cheV*, *flgA*, *cysJ*, *xac3999*). However, these genes did not show drastic changes in expression levels, nor do they exclusively feature very prominent members of repeats or show a trend regarding orientation of the differentially expressed gene to the respective repeat.

In addition to a clear preference for 4 units, a strong bias for repeats downstream of ORFs to be localized in very close proximity of the stop codon or even overlapping with the ORF had been noticed (see Fig 5). In order to gain insight into whether the repeats are transcribed and whether they play a role in transcription termination we assessed the location of the repeat sequences on transcripts by investigating the available RNA sequencing data of *Xac* grown in NB full medium (sample NB_2) [53] (S6 File).

First it was determined whether all repeat sequences are part of assembled transcripts. Of the 183 repeat sequences in *Xac* 24 repeats could not be assigned to a transcript in the analyzed sample. All transcripts mapping to repeats within coding regions were excluded from the following analysis and all repeats unable to fold putative G-quadruplexes with a G-tract of 3 guanines were allocated to a control set. In case of tandemly inverted repeats each repeat was analyzed individually. In the G4 group 49.3% showed the C-rich sequence on the transcript, 39.7% the G-rich sequence and for 11% of the repeats no transcript had been assembled. 37% of the control set showed the C-rich sequence on the transcript, 50% the G-rich sequence and for 14% of the repeats no transcript had been assembled (Fig 7A).

Next the assembled transcripts were sorted into the following groups according to the location of the repeat sequence within the transcript: 1) the transcript starts within the repeat, 2) the transcript ends within the repeat sequence or shortly thereafter (max. 30 nt) and 3) the repeat sequences is located anywhere in the middle of the transcript. Generally, putative G4

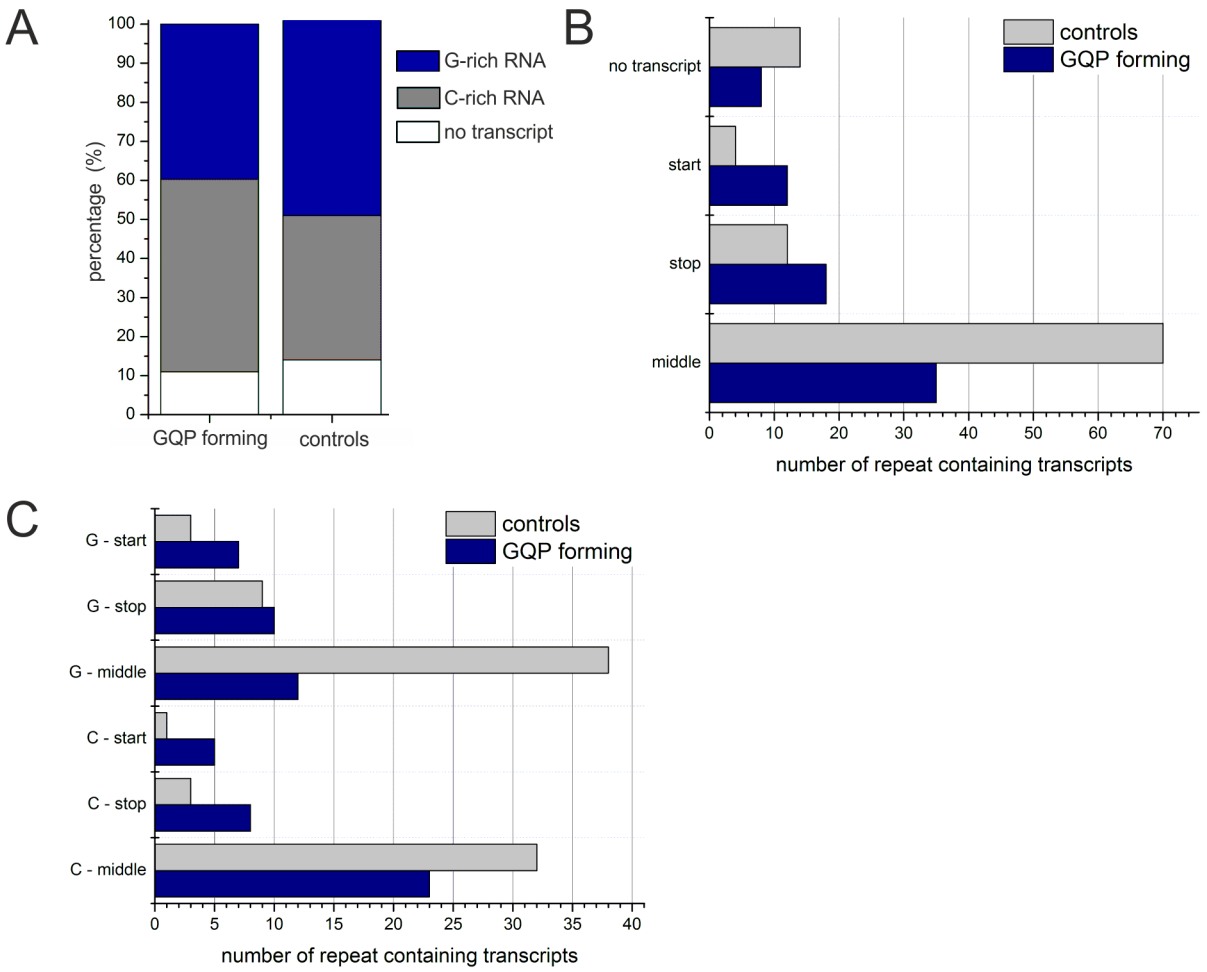

Fig 7. Analysis of repeat containing transcripts of *Xac* grown in NB medium. Analysis of assembled transcripts of *Xac* grown in NB medium that mapped to repeat containing regions. Control repeats are shown in gray, putative G4 forming repeats are shown in blue. The G4 forming set contains 65 transcripts, the control set contains 86 transcripts. In addition 8 G4 forming repeats and 14 control repeats are shown in A, for which no transcript could be assembled. (A): The overall distribution of a repeat's location on a transcript is shown. No transcript refers to repeats for which no transcript could be assembled. If a transcript was assembled, it may start within a repeat sequence (start), stop within a repeat sequence or maximum 30 nt after the repeat (stop) or the repeat may be located somewhere in the middle of transcript (middle). (B) The analysis of A is further split up to show whether the G- or C-rich strand was found in the respective transcript.

doi:10.1371/journal.pone.0144275.g007

forming sequences were under-represented in the middle of transcripts (Fig 7B). Interestingly, a transcript started or stopped more often in the G4 forming group than in the control group (Fig 7B). This effect was observed no matter if the G-rich or C-rich strand was found on the transcript (Fig 7C).

Discussion

GGGAATC / GGGGA(C/T)T repeat sequences are very abundant in xanthomonads and cyanobacteria. We investigated the occurrence of these motifs in the *Xcc*, *Xac* and *Ana* genomes. They represent a special type of SSR as in addition to being repetitive sequences they also have the capacity to form G4 structures. We found these repetitive patterns to be present all over the respective genomes with a strong bias for non-coding regions. Remarkably, a clear preference for a unit size of four was detected, which corresponds to the minimum number of G-tracts needed for G4 formation. Using CD spectroscopy we were able to show that repeat-comprising

DNA oligonucleotides readily formed secondary structures with moderate to very high thermodynamic stabilities and a clear preference for K^+, demonstrating that the adopted structures in presence of K^+ are G4s. In addition we observed characteristic spectral changes that suggest i-motif formation of the complementary C-rich oligonucleotide even at mildly acidic pH of 6.5. Increasing ionic strength did not disturb i-motif formation. In case of inverted repeats there is the possibility of formation of stem-loop structures as well as G4s, both secondary structures may also compete with each other. It is unclear whether such possible non-canonical nucleic acid structures are formed at the DNA or RNA level in the bacteria. However, analysis of RNA sequencing data published by Jalan et al. [53] showed that the majority of the repeat sequences in *Xac* are in fact transcribed. The G- as well as the C-rich strand was found to be part of transcripts. While DNA as well as RNA G4s exist, formation of an i-motif on RNA level is much less favored compared to G4s [73] as RNA i-motifs have been shown to be less stable than their DNA counterparts [74, 75].

A preference for these G-rich repeats to be located in close proximity to the ORF either upstream on the non-coding strand or downstream on the coding strand was detected in all three organisms. These locations are prone to allow for gene regulatory effects. A variety of possible cellular functions have been attributed to G4s as has been reviewed by Bochman et al. [10]. For instance putative regulative roles of G4 structures formed during transcription involve blocking of transcription via inhibition of the polymerase, facilitating transcription by keeping the DNA strands separated, or even promotion or repression of transcription by recruitment of G4 binding proteins that may in turn interact with the RNA polymerase. Recently, we showed hat in *E. coli* G4 sequences can have activating as well as inhibitory effects on gene expression that largely depend on the exact location of the quadruplex-forming sequence element within the promoter region or at the ribosomal binding site [21]. Gene regulatory effects have also been observed for SSRs involved in phase variation, e.g. by overlapping with binding sites of regulatory proteins or variation of spacing between promoter elements [76, 77]. However, we were not able to identify a role of the studied repeats in gene regulation.

Generally, we found repeats located between divergent ORFs to be under-represented. In this case G-rich repeats may overlap with promoter regions of several genes. Possible secondary structure formation or repeat expansion in this region may interfere with the promoter function of both genes. Under-representation of G-rich motifs at such a position may indicate that formation of non-canonical nucleic acid structures by the repeats might well be possible in vivo and therefore be avoided in this particular region. This goes hand in hand with repeats being underrepresented on the coding strand within ORFs in all three organisms. Apart from restrictions due to the coding function of the ORF, G4 formation may cause ribosome stalling or induce frame-shifts.[78, 79] Generally, Lin and Kussell found SSRs to be suppressed in the middle of coding regions in prokaryotes, but enriched near the termini. SSRs were especially over-representated close to the N-terminus indicating involvement in phase variation by frame-shifting [80].

Analysis of the repeat-associated genes in all three organisms showed them to be randomly distributed across the different functional gene classes. Repeats involved in phase variation have been shown to be associated with cell surface structures such as antigens [22, 23, 77, 81]. In addition a G4 sequence in *Neisseria gonorrhoeae* has been shown to promote antigenic variation [82–84]. While genes encoding cell wall and pili components were among the repeat-associated genes, the great number of genes belonging to general metabolism pathways makes a role of GGGAATC and GGGGA(C/T)T repeats in phase variation unlikely. The genus *Xanthomonas* shows a high degree of host plant specificity and may even show tissue specificity. In addition to infecting different dicotyledonous hosts, *Xcc* invades the vascular system of the plant while *Xac* infects the mesophyll tissue [45]. However repeats were often found

associated to similar genes in *Xac* and *Xcc* and not exclusive to pathogenicity-related genes. This makes a role of the repeats in pathogenicity or pathogen-host interactions unlikely.

While the majority of repeats are found between the same genes in *Xcc* and *Xac*, we found extensive length and sequence variation of the intergenic patterns even between these closely related organisms. It was hypothesized that the increased abundance of heptameric repeats in bacteria might be related to the size of the DNA segment that interacts with the active site of the DNA polymerase, which may lead to increased occurrence of polymerase slippage for this pattern type [33]. Joukhadar and Jighly hypothesized that microsatellites may even grant more stable flanking genes. SSRs may be able to discard weak DNA polymerases, thereby increasing the opportunity of the flanking genes to be replicated by more stable DNA polymerases [85]. In contrast to other SSRs, the sequences investigated here seem to be associated with genomic regions with increased genomic stability. While the over-representation of GGGAATC and GGGGA(C/T)T repeats in *Xcc*, *Xac* and *Ana*, respectively, is a remarkable feature of these prokaryotes, a potential functional role of these peculiar repeat motifs still remains to be elucidated.

Supporting Information

S1 File. G4 sequences in *Xanthomonas campestris pv*. *campestris* ATCC 33913 on plus strand (**Table A**). G4 sequences in *Xanthomonas campestris pv*. *campestris* ATCC 33913 on minus strand (**Table B**). G4 sequences in *Nostoc* sp. PCC7120 on plus strand (**Table C**). G4 sequences in *Nostoc* sp. PCC7120 on minus strand (**Table D**).
(DOCX)

S2 File. GGGAATC Repeats in *Xanthomonas campestris pv*. *campestris* ATCC 33913 (**Table A**). GGGAATC Repeats in *Xanthomonas axonopodis pv*. *ctri* str. 306 (**Table B**). GGGGA(C/T)T Repeats in *Nostoc* sp. PCC7120 (**Table C**).
(DOCX)

S3 File. DNA Oligonucleotides (**Table A**). Longest Repeats in *Xcc* (**Table B***)*. Longest Repeats in *Xac* (**Table C***)*. CD spectra of $(GGGAATC)_3GGG$ variants with G to T mutations in G-tract (**Fig A**). Melting temperatures of G-quadruplex structures (**Table D**). Melting profiles of G-rich repeat oligos at pH 7.5 (**Fig B**) Melting temperatures of structures formed by C-rich repeat oligonucleotides (**Table E**). CD spectra of C-rich repeat oligos in Na-acetate buffer (**Fig C**). Melting profiles of C-rich repeat oligos in Na-acetate buffer (**Fig D**). Distance of repeats to neighboring ORFs for potential quadruplex forming sequences in *Xcc* (**Fig E**).
(DOCX)

S4 File. Classification of repeat associated genes according to KEGG Pathways.
(DOCX)

S5 File. Sequence comparison between repeat containing regions in *Xcc* and *Xac* (**Table A**). Control sequences used for sequence comparison between *Xcc* and *Xac* (**Table B**).
(DOCX)

S6 File. Analysis of RNA sequencing data for repeat-containing transcripts.
(DOCX)

Acknowledgments

We thank the members of the Hartig group for helpful discussions. Yuanhao Li thanks the China Scholarship Council for financial support. Jörg S. Hartig thanks the Konstanz Research School Chemical Biology and the SFB 969 for financial support.

Author Contributions

Conceived and designed the experiments: CR TF JSH. Performed the experiments: CR LW YL. Analyzed the data: CR LW YL TF JSH. Wrote the paper: CR JSH.

References

1. Wang G, Vasquez KM. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. DNA repair. 2014 Jul; 19:143–51. doi: 10.1016/j.dnarep.2014.03.017 PMID: 24767258
2. Wang G, Vasquez KM. Non-B DNA structure-induced genetic instability. Mutation research. 2006 Jun 25; 598(1–2):103–19. PMID: 16516932
3. Malfoy B, Rousseau N, Vogt N, Viegas-Pequignot E, Dutrillaux B, Leng M. Nucleotide sequence of an heterochromatic segment recognized by the antibodies to Z-DNA in fixed metaphase chromosomes. Nucleic acids research. 1986 Apr 25; 14(8):3197–214. PMID: 3010230
4. Johnston BH. Generation and detection of Z-DNA. Methods in enzymology. 1992; 211:127–58. PMID: 1406305
5. Mirkin SM, Frank-Kamenetskii MD. H-DNA and related structures. Annual review of biophysics and biomolecular structure. 1994; 23:541–76. PMID: 7919793
6. Htun H, Dahlberg JE. Topology and formation of triple-stranded H-DNA. Science. 1989 Mar 24; 243 (4898):1571–6. PMID: 2648571
7. Brazda V, Laister RC, Jagelska EB, Arrowsmith C. Cruciform structures are a common DNA feature important for regulating biological processes. BMC molecular biology. 2011; 12:33. doi: 10.1186/1471-2199-12-33 PMID: 21816114
8. Ellegren H. Microsatellites: simple sequences with complex evolution. Nature reviews Genetics. 2004 Jun; 5(6):435–45. PMID: 15153996
9. Zhou K, Aertsen A, Michiels CW. The role of variable DNA tandem repeats in bacterial adaptation. FEMS microbiology reviews. 2014 Jan; 38(1):119–41. doi: 10.1111/1574-6976.12036 PMID: 23927439
10. Bochman ML, Paeschke K, Zakian VA. DNA secondary structures: stability and function of G-quadruplex structures. Nature reviews Genetics. 2012 Nov; 13(11):770–80. doi: 10.1038/nrg3296 PMID: 23032257
11. Murat P, Balasubramanian S. Existence and consequences of G-quadruplex structures in DNA. Current opinion in genetics & development. 2014 Apr; 25:22–9.
12. Rando OJ, Verstrepen KJ. Timescales of genetic and epigenetic inheritance. Cell. 2007 Feb 23; 128 (4):655–68. PMID: 17320504
13. Correia FF, Inouye S, Inouye M. A family of small repeated elements with some transposon-like properties in the genome of Neisseria gonorrhoeae. The Journal of biological chemistry. 1988 Sep 5; 263 (25):12194–8. PMID: 2842323
14. Delihas N. Impact of small repeat sequences on bacterial genome evolution. Genome biology and evolution. 2011; 3:959–73. doi: 10.1093/gbe/evr077 PMID: 21803768
15. Tobes R, Pareja E. Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. BMC genomics. 2006; 7:62. PMID: 16563168
16. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nature reviews Genetics. 2010 Mar; 11(3):181–90. doi: 10.1038/nrg2749 PMID: 20125085
17. Westra ER, Buckling A, Fineran PC. CRISPR-Cas systems: beyond adaptive immunity. Nature reviews Microbiology. 2014 May; 12(5):317–26. doi: 10.1038/nrmicro3241 PMID: 24704746
18. Choi J, Majima T. Conformational changes of non-B DNA. Chem Soc Rev. 2011; 40(12):5893–909. doi: 10.1039/C1CS15153C PMID: 21901191
19. Du X, Wojtowicz D, Bowers AA, Levens D, Benham CJ, Przytycka TM. The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in Escherichia coli. Nucleic acids research. 2013 Jul; 41(12):5965–77. doi: 10.1093/nar/gkt308 PMID: 23620297
20. Huang Y, Mrazek J. Assessing diversity of DNA structure-related sequence features in prokaryotic genomes. DNA research: an international journal for rapid publication of reports on genes and genomes. 2014 Jun; 21(3):285–97.
21. Holder Isabelle T, Hartig Jörg S. A Matter of Location: Influence of G-Quadruplexes on Escherichia coli Gene Expression. Chem Biol. 2014; 21(0):1511.

22. Henderson IR, Owen P, Nataro JP. Molecular switches—the ON and OFF of bacterial phase variation. Molecular microbiology. 1999 Sep; 33(5):919–32. PMID: 10476027

23. van der Woude MW, Baumler AJ. Phase and antigenic variation in bacteria. Clinical microbiology reviews. 2004 Jul; 17(3):581–611, table of contents. PMID: 15258095

24. MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993; 72(6):971–83. PMID: 8458085

25. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature. 1991 Jul 4; 352(6330):77–9. PMID: 2062380

26. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell. 1991 May 31; 65(5):905–14. PMID: 1710175

27. Schlotterer C, Imhof M, Wang H, Nolte V, Harr B. Low abundance of Escherichia coli microsatellites is associated with an extremely low mutation rate. Journal of evolutionary biology. 2006 Sep; 19(5):1671–6. PMID: 16910996

28. Yang J, Wang J, Chen L, Yu J, Dong J, Yao ZJ, et al. Identification and characterization of simple sequence repeats in the genomes of Shigella species. Gene. 2003 Dec 11; 322:85–92. PMID: 14644500

29. Mrazek J. Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. Molecular biology and evolution. 2006 Jul; 23(7):1370–85. PMID: 16618962

30. Kassai-Jager E, Ortutay C, Toth G, Vellai T, Gaspari Z. Distribution and evolution of short tandem repeats in closely related bacterial genomes. Gene. 2008 Feb 29; 410(1):18–25. doi: 10.1016/j.gene.2007.11.006 PMID: 18191346

31. Moxon R, Bayliss C, Hood D. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. Annual review of genetics. 2006; 40:307–33. PMID: 17094739

32. Treangen TJ, Abraham AL, Touchon M, Rocha EP. Genesis, effects and fates of repeats in prokaryotic genomes. FEMS microbiology reviews. 2009 May; 33(3):539–71. PMID: 19396957

33. Mrazek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. Proceedings of the National Academy of Sciences of the United States of America. 2007 May 15; 104(20):8472–7. PMID: 17485665

34. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y. Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome research. 2000 Jan; 10 (1):62–71. PMID: 10645951

35. Moxon ER, Rainey PB, Nowak MA, Lenski RE. Adaptive evolution of highly mutable loci in pathogenic bacteria. Current biology: CB. 1994 Jan 1; 4(1):24–33. PMID: 7922307

36. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual review of genetics. 2010; 44:445–77. doi: 10.1146/annurev-genet-072610-155046 PMID: 20809801

37. Bichara M, Wagner J, Lambert IB. Mechanisms of tandem repeat instability in bacteria. Mutation research. 2006 Jun 25; 598(1–2):144–63. PMID: 16519906

38. van Belkum A, van Leeuwen W, Scherer S, Verbrugh H. Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes. Research in microbiology. 1999 Nov-Dec; 150(9–10):617–26. PMID: 10673001

39. Mazel D, Houmard J, Castets AM, Tandeau de Marsac N. Highly repetitive DNA sequences in cyanobacterial genomes. Journal of bacteriology. 1990 May; 172(5):2755–61. PMID: 2110150

40. Siddiqui-Jain A, Grand CL, Bearss DJ, Hurley LH. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proceedings of the National Academy of Sciences of the United States of America. 2002 Sep 3; 99(18):11593–8. PMID: 12195017

41. McLuckie KI, Waller ZA, Sanders DA, Alves D, Rodriguez R, Dash J, et al. G-quadruplex-binding benzo[a]phenoxazines down-regulate c-KIT expression in human gastric carcinoma cells. Journal of the American Chemical Society. 2011 Mar 2; 133(8):2658–63. doi: 10.1021/ja109474c PMID: 21294544

42. Cogoi S, Xodo LE. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic acids research. 2006; 34(9):2536–49. PMID: 16687659

43. Beaume N, Pathak R, Yadav VK, Kota S, Misra HS, Gautam HK, et al. Genome-wide study predicts promoter-G4 DNA motifs regulate selective functions in bacteria: radioresistance of D. radiodurans involves G4 DNA-mediated regulation. Nucleic acids research. 2013 Jan 7; 41(1):76–89. doi: 10.1093/nar/gks1071 PMID: 23161683

44. Wieland M, Hartig JS. RNA quadruplex-based modulation of gene expression. Chemistry & biology. 2007 Jul; 14(7):757–63.
45. Ryan RP, Vorholter FJ, Potnis N, Jones JB, Van Sluys MA, Bogdanove AJ, et al. Pathogenomics of Xanthomonas: understanding bacterium-plant interactions. Nature reviews Microbiology. 2011 May; 9 (5):344–55. doi: 10.1038/nrmicro2558 PMID: 21478901
46. da Silva AC, Ferro JA, Reinach FC, Farah CS, Furlan LR, Quaggio RB, et al. Comparison of the genomes of two Xanthomonas pathogens with differing host specificities. Nature. 2002 May 23; 417 (6887):459–63. PMID: 12024217
47. Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, et al. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120. DNA research: an international journal for rapid publication of reports on genes and genomes. 2001 Oct 31; 8(5):205–13; 27–53.
48. Yadav VK, Abraham JK, 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 research. 2008 Jan; 36(Database issue):D381–5. PMID: 17962308
49. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004 Jun; 14(6):1188–90. PMID: 15173120
50. Nakao M, Okamoto S, Kohara M, Fujishiro T, Fujisawa T, Sato S, et al. CyanoBase: the cyanobacteria genome database update 2010. Nucleic acids research. 2010 10/30 09/15/received 10/07/accepted; 38(Database issue):D379–D81. doi: 10.1093/nar/gkp915 PMID: 19880388
51. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000 Jan 1; 28(1):27–30. PMID: 10592173
52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997 Sep 1; 25(17):3389–402. PMID: 9254694
53. Jalan N, Kumar D, Andrade MO, Yu F, Jones JB, Graham JH, et al. Comparative genomic and transcriptome analyses of pathotypes of Xanthomonas citri subsp. citri provide insights into mechanisms of bacterial virulence and host range. BMC genomics. 2013; 14:551. doi: 10.1186/1471-2164-14-551 PMID: 23941402
54. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics. 2014 April 1, 2014.
55. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. [Brief Communication]. 2012 04//print; 9(4):357–9.
56. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech. 2011 07//print; 29(7):644–52. doi: 10.1038/nbt.1883
57. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protocols. [Protocol]. 2013 08//print; 8(8):1494–512. doi: 10.1038/nprot.2013.084 PMID: 23845962
58. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome research. 2002 April 1, 2002; 12(4):656–64. PMID: 11932250
59. Qian W, Jia Y, Ren SX, He YQ, Feng JX, Lu LF, et al. Comparative and functional genomic analyses of the pathogenicity of phytopathogen Xanthomonas campestris pv. campestris. Genome research. 2005 Jun; 15(6):757–67. PMID: 15899963
60. Swanson RV, de Lorimier R, Glazer AN. Genes encoding the phycobilisome rod substructure are clustered on the Anabaena chromosome: characterization of the phycoerythrocyanin operon. Journal of bacteriology. 1992 Apr; 174(8):2640–7. PMID: 1556083
61. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S. Quadruplex DNA: sequence, topology and structure. Nucleic acids research. 2006; 34(19):5402–15. PMID: 17012276
62. Huppert JL. Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes. Chemical Society reviews. 2008 Jul; 37(7):1375–84. doi: 10.1039/b702491f PMID: 18568163
63. Kypr J, Kejnovska I, Renciuk D, Vorlickova M. Circular dichroism and conformational polymorphism of DNA. Nucleic acids research. 2009 Apr; 37(6):1713–25. doi: 10.1093/nar/gkp026 PMID: 19190094
64. Day HA, Pavlou P, Waller ZAE. i-Motif DNA: Structure, stability and targeting with ligands. Bioorg Med Chem. 2014 8/15/; 22(16):4407–18. doi: 10.1016/j.bmc.2014.05.047 PMID: 24957878
65. Han X, Leroy JL, Gueron M. An intramolecular i-motif: the solution structure and base-pair opening kinetics of d(5mCCT3CCT3ACCT3CC). Journal of molecular biology. 1998 May 22; 278(5):949–65. PMID: 9600855

66. Xu Y, Sugiyama H. Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb). Nucleic acids research. 2006; 34(3):949–54. PMID: 16464825

67. Mergny J-L, Lacroix L, Han X, Leroy J-L, Helene C. Intramolecular Folding of Pyrimidine Oligodeoxynucleotides into an i-DNA Motif. Journal of the American Chemical Society. 1995 1995/09/01; 117 (35):8887–98.

68. Benabou S, Ferreira R, Aviñó A, González C, Lyonnais S, Solà M, et al. Solution equilibria of cytosine- and guanine-rich sequences near the promoter region of the n-myc gene that contain stable hairpins within lateral loops. Biochimica et Biophysica Acta (BBA)—General Subjects. 2014 1//; 1840(1):41–52.

69. Shabala L, Bowman J, Brown J, Ross T, McMeekin T, Shabala S. Ion transport and osmotic adjustment in Escherichia coli in response to ionic and non-ionic osmotica. Environmental microbiology. 2009 Jan; 11(1):137–48. doi: 10.1111/j.1462-2920.2008.01748.x PMID: 18793315

70. Holder Isabelle T, Hartig Jörg S. A Matter of Location: Influence of G-Quadruplexes on Escherichia coli Gene Expression. Chemistry & biology.

71. Hannan AJ. TRPing up the genome: Tandem repeat polymorphisms as dynamic sources of genetic variability in health and disease. Discovery medicine. 2010 Oct; 10(53):314–21. PMID: 21034672

72. Wang G, Vasquez KM. Models for chromosomal replication-independent non-B DNA structure-induced genetic instability. Molecular carcinogenesis. 2009 Apr; 48(4):286–98. doi: 10.1002/mc.20508 PMID: 19123200

73. Joachimi A, Benz A, Hartig JS. A comparison of DNA and RNA quadruplex structures and stabilities. Bioorganic & medicinal chemistry. 2009 Oct 1; 17(19):6811–5.

74. Lacroix L, Mergny JL, Leroy JL, Helene C. Inability of RNA to form the i-motif: implications for triplex formation. Biochemistry. 1996 Jul 2; 35(26):8715–22. PMID: 8679634

75. Collin D, Gehring K. Stability of Chimeric DNA/RNA Cytosine Tetrads: Implications for i-Motif Formation by RNA. Journal of the American Chemical Society. 1998 1998/05/01; 120(17):4069–72.

76. Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER. Microsatellite instability regulates transcription factor binding and gene expression. Proceedings of the National Academy of Sciences of the United States of America. 2005 Mar 8; 102(10):3800–4. PMID: 15728391

77. van Ham SM, van Alphen L, Mooi FR, van Putten JP. Phase variation of H. influenzae fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. Cell. 1993 Jun 18; 73(6):1187–96. PMID: 8513502

78. Endoh T, Kawasaki Y, Sugimoto N. Suppression of gene expression by G-quadruplexes in open reading frames depends on G-quadruplex stability. Angewandte Chemie. 2013 May 17; 52(21):5522–6. doi: 10.1002/anie.201300058 PMID: 23589400

79. Endoh T, Sugimoto N. Unusual -1 ribosomal frameshift caused by stable RNA G-quadruplex in open reading frame. Analytical chemistry. 2013 Dec 3; 85(23):11435–9. doi: 10.1021/ac402497x PMID: 24191683

80. Lin W-H, Kussell E. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. Nucleic acids research. 2012 11/26 09/07/received 10/26/revised 10/28/accepted; 40(6):2399–413. doi: 10.1093/nar/gkr1078 PMID: 22123746

81. Stern A, Brown M, Nickel P, Meyer TF. Opacity genes in Neisseria gonorrhoeae: control of phase and antigenic variation. Cell. 1986 Oct 10; 47(1):61–71. PMID: 3093085

82. Cahoon LA, Seifert HS. An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. Science. 2009 Aug 7; 325(5941):764–7. doi: 10.1126/science.1175653 PMID: 19661435

83. Cahoon LA, Seifert HS. Focusing homologous recombination: pilin antigenic variation in the pathogenic Neisseria. Molecular microbiology. 2011 Sep; 81(5):1136–43. doi: 10.1111/j.1365-2958.2011.07773.x PMID: 21812841

84. Cahoon LA, Seifert HS. Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in Neisseria gonorrhoeae. PLoS pathogens. 2013 Jan; 9(1):e1003074. doi: 10.1371/journal.ppat.1003074 PMID: 23349628

85. Joukhadar R, Jighly A. Microsatellites grant more stable flanking genes. BMC research notes. 2012; 5:556. doi: 10.1186/1756-0500-5-556 PMID: 23035963