

Prepared for *Biochem J.*

Running title: Simple DNA Repeats in Bacteria

**Structures and Stability of Simple DNA Repeats from Bacteria**

Vaclav Brazda<sup>1</sup>, Miroslav Fojta<sup>1</sup> & Richard P. Bowater<sup>2</sup>

<sup>1</sup> Institute of Biophysics of the Czech Academy of Sciences, Královopolská  
135, 612 65 Brno, Czech Republic

<sup>2</sup> School of Biological Sciences, University of East Anglia, Norwich Research  
Park, Norwich NR4 7TJ, United Kingdom

\*To whom correspondence should be addressed:

Tel.: 01603 592186

Fax: 01603 592250

e-mail: R.Bowater@uea.ac.uk

## **Abstract**

DNA is a fundamentally important molecule for all cellular organisms due to its biological role as the store of hereditary, genetic information. On the one hand genomic DNA is very stable, both in chemical and biological contexts, and this assists its genetic functions. On the other hand it is also a dynamic molecule, and constant changes in its structure and sequence drive many biological processes, including adaptation and evolution of organisms. DNA genomes contain significant amounts of repetitive sequences, which have divergent functions in the complex processes that involve DNA, including replication, recombination, repair and transcription. Through their involvement in these processes, repetitive DNA sequences influence the genetic instability and evolution of DNA molecules and they are located non-randomly in all genomes. Mechanisms that influence such genetic instability have been widely studied in many organisms, including within human genomes where they are linked to a variety of human diseases. Here, we review our understanding of short, simple DNA repeats across a diverse range of bacteria, comparing the prevalence of repetitive DNA sequences in different genomes. We describe the range of DNA structures that have been observed in such repeats, focusing on their propensity to form local, non-B-DNA structures. Finally, we discuss the biological significance of such unusual DNA structures and relate this to studies where the impacts of DNA metabolism on genetic stability are linked to human diseases. Overall, we show that simple DNA repeats in bacteria serve as excellent and tractable experimental models for biochemical studies of their cellular functions and influences.

## **Keywords**

DNA metabolism, DNA structure, microsatellites, nucleic acids, repetitive DNA sequences

## **Abbreviations**

bp: basepairs

CD: circular dichroism

DNA: deoxyribonucleic acid

MMR: mismatch repair

NCBI: National Center for Biotechnology Information

NER: nucleotide excision repair

NMR: nuclear magnetic resonance

ORF: open reading frame

## **Simple DNA repeats**

DNA molecules are the store of genetic information for all cellular organisms. The arrangements of individual bases in the DNA sequences of an organism, its genome, are specific to that organism, and elucidation of massive numbers of genome sequences have impacted on our understanding of the phylogenetic tree of life [1]. The organisation of sequences in any genome is critical for its function and, from the earliest days of genome sequence analysis, it was recognized that natural DNA molecules contain a wide array of repeating sequences [2]. In fact, this was particularly important in many genomic studies because such sequences are challenging to obtain accurate data [3]. Repeat sequences of approximately 1 to 6 basepairs (bp) in their unit structure are termed simple repeating sequences, due to their sequence being less complex (“simpler”) than random sequences [4, 5]. Such simple sequences are often called microsatellites and the term “short tandem repeats” is also used frequently in the literature. Although most base sequences will be found within double-stranded DNA molecules, within this review we generally refer to sequences via a single strand, given in the 5’-3’ direction.

Simple repeating sequences can be distinguished by their sequence motif and base composition [4-7]. The various sequence motifs consist of different lengths of the repeat unit, such as mono-, di-, tri- or tetra-nucleotide repeats, etc.. For example, mononucleotide repeats are tracts of a single nucleotide in the sequence. Within repeating units there is some redundancy within DNA sequences e.g. (CT)<sub>n</sub> also contains (TC)<sub>m</sub>, where “n” and “m” refer to numbers of repeats – see Figure 1. (Depending on the sequences that flank the repeat, “n” and “m” may be equal, or they may differ by 1.) Importantly, DNA molecules have a directionality associated with them, with the 5’- and 3’-ends usually containing terminal

phosphate and terminal hydroxyl groups, respectively [8]. Following the convention of writing sequences in a 5'-3' direction and anti-parallel arrangement of complementary chains in double-stranded DNA molecules, there are just two options for mononucleotide repeats (A/T or C/G basepairs) and four different types of dinucleotide repeats, (AT)<sub>n</sub>, (GT)<sub>n</sub>, (GA)<sub>n</sub> and (GC)<sub>n</sub>. Similar analysis of trinucleotide repeats identifies ten different repeat sequences [9]. Classical examples of microsatellites consist of uninterrupted sequence of tandem repeats of the same motif (Figure 1). When one or more bases interrupt the repeat array, the microsatellite is termed "interrupted" (also sometimes called "imperfect").

Juxtapositions of two types of repeat (called "compound" or sometimes "composite" microsatellites) also occur frequently in genomes (Figure 1).

Some repetitive elements are referred to as "inverted repeats" because the rules of complementary base pairing mean that their sequence is the same when the complementary strand is read in its 5'-3' direction (Figure 2A) [10]. Since inverted repeats will occur on both strands at the specific location, they can adopt a specific structure referred to as a cruciform (Figure 2B) – see below for more details. Such sequences are targets for many architectural and regulatory proteins and their importance has been demonstrated for several basic biological processes. As we discuss below, such processes may be regulated by the formation of specific types of localised DNA structures at these sequences.

### **Prevalence of DNA repeats in bacterial genomes**

Advances in DNA sequencing technologies have generated massive numbers of genome sequences for prokaryotes due to their relatively small size and ease of experimental manipulation [1]. Most genome sequences are deposited in databases

that make them publicly available. One such archive is the genome database at the National Center for Biotechnology Information (NCBI) and it contains DNA sequences from over two hundred thousand bacteria (206,445) as of 13/09/2019).

One of the first sequenced and best characterized bacterial genome is that of *Escherichia coli*, which contains a 4.6 million base pair genome with 4288 annotated protein-coding genes, seven ribosomal RNA operons, and 86 transfer RNA genes [11]. It is clear that there is massive variation in phenotypes of bacteria, which is reflected in the huge variety of sizes and types of sequences found within their genomes. The vast majority of bacterial genomes are circular, consisting usually of large chromosomes and small plasmids. However, this is not always the case and there are notable examples of bacteria that harbour linear genomes, including some that are industrially-important, such as *Streptomyces coelicolor* [12, 13]. Indeed, there is vastly more evolutionary divergence among bacteria than is found among all other organisms on earth [1]. Many of the examples discussed in this review refer to *E. coli* because that system allows good correlation between bioinformatics and laboratory-based biological studies, but representative details from other organisms are discussed as appropriate.

All DNA genomes contain amounts of repetitive sequences that are larger than expected for random distribution of bases, but the percentage of repetitive sequences varies greatly across different organisms. For example, while the genome of *E. coli* contains only 0.7% of repeats in non-coding regions [11], at least 50% of the human genome is repetitive or repeat-derived [3]. As discussed in more detail below, through their involvement in DNA metabolism, repetitive DNA sequences have a dramatic influence on the genetic instability and evolution of genomes and organisms. These factors are some of the major forces that drive the increased

prevalence of repeats within genomes compared to what would be expected if all bases were distributed randomly.

While simple DNA repeats are over-represented in the human genome and, generally, in eukaryotic genomes [14], in bacteria they are less common and are often subjected to negative selection [15]. However, significant differences in the amounts of simple DNA repeats exist, even among closely related species, as shown in mycoplasma [16]. An algorithm was developed to search specifically for tandem repeats [17]. Refinement of these approaches have developed computer-based analyses of microbial whole genome sequences that reveal overrepresentation of several simple DNA repeats. Such screening of the genome sequence of *E. coli* strain K12 identified thousands of tandem simple sequence repeat tracts, with motifs ranging from 1 to 6 nucleotides [18]. In addition to simple microsatellites, the repeats also consist of transposable genetic elements.

Comprehensive analyses of DNA sequence frequencies in various genomes has been published in the Genome Composition database (GCD) [19]. The genome-wide analysis of *E. coli* strain K12 already referred to showed a significant excess of mono- and trinucleotide repeats only [18]. The presence of the mononucleotide repeats is unequal for the two types and differs according to the GC contents of individual organisms [20]. For example, the GC content of *E. coli* K12 strain is 50.79%, but 93% of the mononucleotide repeats in its genome are formed by A (or T, its complement), both in open reading frames (ORFs) and in noncoding regions [18]. Similarly, the distribution of dinucleotide repeats in the genome of *E. coli* strain K12 is not random, with the (CG)<sub>n</sub> motif being very abundant in coding regions (49.1% of all

dinucleotide repeats, compared to 17.3% expected).<sup>1</sup> In non-coding regions, the (AT)<sub>n</sub> motif is over-represented relative to its expected value (24.4% compared to 17.9% expected), as is (CG)<sub>n</sub> (23.1% compared to 15.4% expected). Trinucleotide repeats are of particular interest to researchers because genetic instabilities within some of them are associated with a range of human diseases (see below). In *E. coli* strain K12 there is a significant excess of trinucleotide repeats, although their maximum observed number of repetitions is only 5 [18].

Similar analyses of repeats with larger unit lengths also showed that not all combinations are equally distributed in genomes. In *E. coli* strain K12 the maximum observed repeat length is 4 for tetranucleotide repeats, there are no pentanucleotide repeats and only 3 hexanucleotide repeats [18]. Furthermore, the frequencies of repeats with a specific motif of 3 and more bases was not distributed equally across all possible combinations. Most notably, of 52 examples of tetranucleotide repeats, (TGGC)<sub>n</sub> (and its complement (GCCA)<sub>n</sub>) occurred 21 times in coding sequences. The finding that the *E. coli* genome is rich in (TGGC)<sub>n</sub> has been attributed to the activity of very short patch repair, which corrects T:G mismatches to C:G, thus increasing GC dinucleotide content in the genome [21].

The length and type of simple repeat sequences also vary significantly in different locations of genomes. For example, simple repeats that are rich in G bases on one strand (and C bases on the other) are often located at the ends of chromosomes.

---

<sup>1</sup> The expected frequencies referred to here were determined by observing those in 10 computer-generated genomes constructed by random ordering of nucleotides according to their overall frequencies in the genome, with departures tested using parametric statistics.



Known as telomeres, these repeats have been best characterized in the genomes of eukaryotes [22], but they also occur in some bacteria [12, 13].

Analyses of short simple repeats among different strains of *E. coli* shows that the number of repeats is polymorphic [23]. Determination of the size of repeat tracts can be used to identify different strains as long as care is taken to be aware of the potential for variable sizes to be identified in short repeats [24]. This approach can quickly diagnose the presence of different strains of bacteria, allowing identification of those that may be pathogenic, as demonstrated with *E. coli* [25, 26], *Staphylococcus aureus* [27], *Mycobacterium leprae* [28] and many others [24].

### **DNA structures formed by DNA repeats**

DNA molecules, including those containing repetitive sequences, mostly form the two-stranded, right-handed helical B-form structure [8]. This structure maximises thermodynamic stability of the molecule and is crucial for fundamental biological processes that store, replicate and transcribe genetic information. Nevertheless, various alternative (non-B) structures can also occur in DNA. These structures are usually characterized by the occurrence of single-stranded regions (loops) and/or sites of disrupted base pair stacking (junctions between continuous B-form DNA and the alternative structure). Since disruption of hydrogen bonds and stacking interactions represents a loss of enthalpic contribution to the free energy of the molecule, any transition from B-form DNA to an alternative structure requires an input of energy. An alternative structure can be favoured if there are alterations to the sequence of one strand, for example when the complementary strand is absent or present in a sub-stoichiometric amount (as in the structure depicted in Figure 3B). However, some environmental (and cellular) conditions promote formation of

alternative structures due to their improved thermodynamic stability compared to B-form DNA under the given conditions. This type of situation occurs for some types of repetitive DNA sequences *in vitro*, with increasing evidence that such structures also exist within cells (see below). The types of structure adopted by repetitive DNA sequences – and their thermodynamic stabilities – are influenced by the length and type of bases within the repeat. Further, topological stress, which is inherent to the majority of DNA molecules inside cells, is another important factor that influences local DNA structures. Typically, DNAs in bacterial cells exist as negatively supercoiled molecules, which can lead to destabilisation of right-handed, double-helical DNA [29, 30]. In the presence of suitable nucleotide sequences, certain levels of negative superhelical stress can be locally absorbed via transition from the B-form DNA to an open local structure. This can assist formation of non B-DNA structures, as shown *in vitro* for various types of repeats [31-34]. Evidence is particularly strong to show that higher levels of negative supercoiling increase the extent of cruciform formation in dinucleotide repeats. This has been confirmed for  $(AT)_n$  sequences *in vitro* and in *E. coli* [29, 35]. Variations in levels of DNA superhelicity naturally occurs *in vivo* in “active” regions of the genome, where processes that involve unravelling of the DNA double helix take place, such as transcription, replication and recombination.

Due to complementary base pairing in double-stranded DNA, mononucleotide repeats are inherently homopurine on one strand and homopyrimidine on the other. While A tracts are prone to DNA bending [36], homopurine/homopyrimidine tracts in general are able to form triplex structures (Figure 3A). Mononucleotide repeats naturally possess mirror symmetry, which is a feature favouring triplex structures via formation of Hoogsteen triads, as shown in Figure 4. Hoogsteen hydrogen bonding

occurs between the purine-rich strand of the duplex DNA and either a pyrimidine-rich or a purine-rich third strand. Pyrimidine-rich third strand interactions are stabilized by Hoogsteen hydrogen bonds that are favoured at low pH, which facilitates the requirement for cytosine protonation required for its Hoogsteen pairing. By contrast, purine-rich third strand interactions form reverse-Hoogsteen hydrogen bonds, which do not require acidic pH and are stabilized by bivalent cations.

Mononucleotide repeats can also undergo strand slipping transitions, resulting in extrusion of a hairpin (Figure 3B) or a pair of hairpins that are separated from each other (Figure 3C). Proclivity to strand slipping is a common feature of simple repeats, playing a crucial role in their change in size during replication [31, 37]. Conditions for good thermodynamic stability of hairpins have been well characterized *in vitro* for trinucleotide repeats such as (CGG)<sub>n</sub>, (CAG)<sub>n</sub> and (CTG)<sub>n</sub>, even though these contain base mismatches or wobble pairs, such as T•T, A•A or G•G [38, 39].

For dinucleotide repeats the length observed in typical microsatellites varies from 5 to 50 repeats. Importantly, while all dinucleotide sequences are direct repeats, some are also inverted repeats (e.g. (AT)<sub>n</sub> and (CG)<sub>n</sub>), whereas others are not (e.g. (AG)<sub>n</sub> and (AC)<sub>n</sub>). This is significant because those that are inverted repeats are able to form cruciform structures (Figures 2 and 3D). At the same time these sequences are composed of (purine-pyrimidine)<sub>n</sub> motifs that are capable of forming a segment of left-handed, Z-form, double helix under certain conditions [40].

Tandem repeats involving G<sub>n</sub> blocks and mononucleotide repeats consisting of G-tracts are able to form quadruplex structures (Figure 3E). Such structures are typically formed when four G nucleotides can be brought together in a planar arrangement to form guanine quartets involving Hoogsteen G-G pairing (see Figure 4) and are usually stabilised by the presence of monovalent cations in the middle of

each G-quartet. Note that the presence of G-tracts on one strand means that C-tracts must be present on the complementary strand, and such sequences can adopt other non-B-DNA structures, such as the i-motif, which we discuss in more detail below.

A strikingly wide range of sequences have been demonstrated to form stable G-quadruplexes under different environmental conditions [37, 41]. All of these sequences are not classically considered as simple DNA repeats, but G-quadruplexes can be formed by various types of short repeats of G bases within longer sequences. Some of the sequences that can form G-quadruplexes are simple microsatellite sequences, such as trinucleotide and hexanucleotide repeats [42, 43]. Other sequences that are more complex in base composition can also form G-quadruplexes, but they all contain G tracts that are repeated with specific periodicities. Within any particular sequence that can form G-quadruplexes the bases that separate the G-tracts may be different in type and number and, thus, they represent a complicated type of interrupted repeat tract (see Figure 1). A wide array of sequences have been shown to form quadruplexes, but longer G-tracts and shorter interruptions form more stable G-quadruplexes, although the size of the loop also impacts on the type of folding seen in stable quadruplexes [44]. Importantly, the likelihood of G-quadruplexes forming in genomes varies dramatically in different locations of DNA molecules [45]. For example, simple repeats that are rich in G bases are often found at telomeric ends of chromosomes and there is significant evidence that such sequences form complexes of proteins specifically bound to four-stranded structures [46]. Telomeres have been best characterized in the genomes of eukaryotes, including humans, but they also occur in some bacteria [12, 13, 22].

Non B-DNA structures are also able to form within sequences that would not typically be able to form significant levels of base pairing. For example, mononucleotide C<sub>n</sub>

sequences and repeats with  $C_n$  blocks are able to form hairpins (Figure 3B) and i-motif structures (Figure 3F) under conditions allowing formation of hemi-protonated  $C^+/C$  base pairs [47, 48]. Following similar arguments presented above for G-quadruplexes, sequences that can form i-motifs are not all classically considered as simple DNA repeats. However, all of these sequences do contain C tracts that are repeated with specific periodicities and, thus, are relevant to topics discussed in this review. The i-motif structures require four C-rich strands containing bases, which can be formed from four distinct strands, two hairpins each carrying two cytosine stretches, or from a single strand with four cytosine stretches [49, 50]. Recent observations have indicated that it is possible to achieve stable i-motifs at physiological pH without the use of crowding agents, if there are at least five cytosine bases per tract [48, 51].

Trinucleotide repeat sequences also adopt many of the structures described above that are dependent on environmental conditions and type of sequences. For example, they can form slipped-stranded DNA and hairpins, but  $(CGG)_n$  have been shown to form G-quadruplexes under specific conditions [52, 53]. R-loops (Figure 3G) are another altered structure, which can be thermodynamically stable in  $(CAG)_n$  and  $(GAA)_n$  [54, 55]. Major structures formed by  $(GAA)_n$  are triplexes in which the third strand can be derived from the either pyrimidine strand or the purine strand [56, 57]. One related structure that has particularly high thermodynamic stability in these sequences has been referred to as “sticky DNA” because of the way it brings together multiples triplexes [58].

Thus, many molecular and biochemical studies demonstrate that simple repeating DNA sequences form a wide array of non-B-DNA structures *in vitro*. Whether such

structures influence biological processes and consequences are questions that have been addressed in different cell types, including several bacteria, as we now discuss.

### **Biochemical and cellular impacts of simple repeat sequences in bacteria**

Within the highly complex environment in cells, various local structures in long, genomic DNA molecules appear to serve as markers of the location of specific activities or functions. Examples of the types of cellular functions that they are involved in are highlighted in Figure 5. The biological relevance of these types of non-B-DNA motifs in recombination, replication and the regulation of gene expression has long been proposed [59]. Furthermore, several studies have demonstrated the important role of non-B-DNA structures in the context of gene regulation in bacteria [30, 60, 61]. For example, cruciforms have been shown to be important for dynamic genome organization [62], and for replication of the circular molecules of genomes, plasmids, mitochondrial DNAs [63] and chloroplast DNAs [64]. Cruciforms are targets for many architectural and regulatory proteins [10] and their importance has been demonstrated for regulation of transcription of some genes [65].

Three-stranded triplex structures can be formed in a range of simple repeats, and structures of many different types have been characterized [66]. Genomic loci containing motifs that can form triplexes are significantly more likely to undergo genome rearrangement compared to control sites, as demonstrated in certain Enterobacteria and Cyanobacteria species [67]. A systematic search of 5246 different bacterial plasmids and genomes for intrastrand triplex motifs was conducted and the results summarized in the ITxF database [66]. This database points to the

importance of these types of sequences (and their potential to form non-B-DNA structures) in influencing the genetic stability of bacterial genomes.

Several bioinformatics tools have been developed to identify potential quadruplex sequences in genomes, such as QGRS Mapper [68] and G4Hunter [69]. In another example, the ProQuad database developed simple rules for G quadruplex forming patterns and used them to assess the occurrence of repeating G-tracts and their association with different genomic regions. This initially identified potential quadruplex sequences within the genomes of 146 bacterial species [70], and an updated database, QuadBase2, mined motifs across genes and their promoter sequences in 1719 prokaryotes [71]. This database can be used to identify the number and location of repeats within large genome sequences. As an example, we use this to identify potential quadruplex forming sequences in the genome of *E. coli* K12 strain, highlighting 69 sequences, 37 in the plus strand and 28 in the minus strand (Figure 6). A separate genome-wide analysis of 18 microbes indicated enrichment of G-quadruplex DNA motifs in putative promoters, with detailed analysis in *E. coli* suggesting a global role for them in “turning-on” transcription during certain growth phases [72]. Along with *in vitro* data that demonstrates quadruplexes are bound by some proteins [46, 73], these findings point towards physiological functions for G-quadruplexes. In this respect, it is significant that genomes with high G+C content are more able to form 4-stranded structures with relatively high thermodynamic stability [37, 74]. There is increasing evidence that these types of structures provide opportunities to regulate DNA metabolism in bacteria [51, 75-77]. The genome of the bacterium *Paracoccus denitrificans* PD1222 has a relatively high G+C content (around 67%) and a range of biophysical, molecular and microbiological

studies show that targeting of 4-stranded structures can be controlled under cellular conditions, allowing regulation of expression of some genes [48, 78-81].

Scientific interest in the genetic stability of simple DNA repeats took on much wider significance when it was recognized that length changes within them is linked to human diseases and disorders. In the 1990s, genetic instability of microsatellites was identified as a useful diagnostic tool for some types of cancer and is associated with some hereditary neurological disorders in humans [54, 82-84]. Much effort has been put into analysing cellular mechanisms that lead to genetic instabilities of trinucleotide repeats, aiming to understand why some are more prevalent in human disorders, the most common of which are CAG, CTG, CGG, and GAA. Recent molecular studies have confirmed that other simple repeats are also important for human diseases [54, 58]. These links have driven many studies that focus on DNA repeats in bacteria where it is often more tractable to conduct genetic analyses.

Different models have been proposed to explain genetic instabilities observed in simple repeats. Many of them involve DNA synthesis, including DNA replication, and various types of DNA repair and recombination [7, 31, 33, 82, 84-86]. Extensive experiments using *E. coli* confirmed that length changes in plasmid-based DNA trinucleotide repeats are affected by replication. The observations are consistent with known biochemical properties of replication forks and lead to suggestions that the sequence within the repeat influences the thermodynamic stability of unusual structures in the DNA [31, 33, 84, 86, 87]. Other processes acting on DNA can impact on mechanisms by which DNA synthesis influences the genetic stability of simple repeats in *E. coli*. For example, transcription of DNA mononucleotide repeats blocked their subsequent replication [88], and transcription into trinucleotide repeats in plasmids influenced the frequency of deletions to the repeat [89-91]. These



experiments highlight that interactions between different processes acting on DNA combine to influence their genetic instability. Interactions may be particularly relevant for processes that use similar proteins, such as DNA polymerases in DNA replication and repair.

The link to DNA repair systems have intriguing roles in relation to genetic instabilities of simple DNA repeats because some of them recognize any aspect of genome structure that is different to the standard base pairs and double helix, including non-B-DNA structures [33, 74, 92]. All cells contain proteins that recognize and repair such genome alterations, protecting genomic integrity by different pathways, which include mismatch repair (MMR), nucleotide and base excision repair, and the repair of double- and single-strand breaks [83, 93, 94]. Generally, the DNA repair pathways and their proteins are well conserved, which means that there is much to be gained from studies of these systems in simpler experimental models, such as bacteria [95-97]. As described below, experiments using bacteria, particularly different strains of mycobacteria, have been very useful for understanding how DNA repair systems influence the genetic stability of simple DNA repeats.

An important physiological role for some DNA repair pathways is to prevent significant changes to the type and number of bases within the genome. However, the genetic instabilities observed within DNA repeats indicate that modifications to the size of the genome are not always repaired. Possibly, cells may not be able to repair some types of length changes to repeats due to non-recognition of certain structures or inaccessibility of DNA processed by some events. Alternatively, mutations in repair proteins may induce length alterations to repeats. Numerous studies show that the impact of DNA repair pathways on repeat tract stability is

complex [84, 85]. Importantly, some non-B DNA structures are identified as modifications to be removed, at least in some contexts or under certain conditions.

MMR and nucleotide excision repair (NER) are fundamental cellular systems involved in maintaining genomic integrity [82, 83, 85, 93, 94]. MMR is able to detect and replace mismatched base pairs that are introduced during inaccurate DNA synthesis. Without such repair, these mismatched base-pairs are a source of mutations within genomes. Upon inactivation of MMR, increased heterogeneities have been observed at simple repetitive DNA (e.g. mono- and dinucleotides) in bacteria [82, 87], suggesting that the genetic stability of simple repeats indicates the increased rate of mutation throughout the whole genome. Due to this phenomenon, such deficiencies within DNA repair systems have been termed the "mutator phenotype" [98].

Generally, NER systems recognize a wide range of lesions and damage due to distortion of the DNA double helix, and unusual DNA structures that could form in repeat tracts are likely to be activators of NER [83, 93, 94]. Studies in *E. coli* observed that their constituent NER proteins influenced the genetic stability of long plasmid-based DNA trinucleotide repeats in a complex fashion [33, 82, 87].

Associations between defective MMR and NER and elevated microsatellite instability are linked to some human diseases, and are particularly strong for hereditary nonpolyposis cancer.

In contrast to their usual cellular functions, the excision repair systems can enhance the genetic instabilities of DNA repeats since they provide opportunities for non-B DNA structures to form on single-stranded regions that are presented as the damage is excised from the DNA helix. Therefore, the repair processes themselves can lead to further consequences, such as addition or deletion of bases, which would be observed as genetic instability [82, 85, 93, 94]. Furthermore, abundant evidence

demonstrates that unusual DNA structures may be recognized as “damaged DNA” by DNA repair systems, sometimes leading to the deletion of the sequence [92, 99, 100]. To reduce such potential problems, cells also take advantage of enzymatic processes to dissolve unusual DNA structures, such as DNA helicases [101]. For example, the RecQ helicases are capable of unwinding G-quadruplex DNA and this family of enzymes is conserved and is essential for genomic stability in organisms from *E. coli* to humans [102, 103].

Genetic instabilities within mono- and dinucleotide repeats increase for longer runs of consecutive repeats and, therefore, are decreased by interruptions in the repeat sequence [87]. These observations are consistent with the hypothesis that slipped-strand mispairing during DNA synthesis generates misaligned intermediates. Such parameters are intrinsic to the DNA repeat, but flanking sequences also influence the genetic stability of simple repeat sequences. These observations suggest that many aspects of DNA metabolism affect the genetic stability of all microsatellite sequences.

Through their effects on DNA metabolism, repetitive DNA sequences have a dramatic influence on the genetic instability and evolution of genomes and organisms. The high levels of genetic instability of repetitive DNA sequences may act to promote evolution of genomic sequences [84, 104]. It has been suggested that length changes to simple repeats can normally be tolerated because they do not have dramatic consequences for the organism in question and that deleterious consequences occur only at extreme length changes [104, 105], as described for the trinucleotide repeat diseases. However, it is clear that simple DNA repeats in bacteria represent hypermutable loci associated with reversible changes in the number of repeats [2, 106]. Variability of the single DNA repeats can lead to increased antigenic variance of the pathogen population [107]. Such length changes

have been clearly demonstrated in bacteria, where this property means that simple DNA repeats can act as prerequisites for bacterial phase variation and adaptation, providing clear evidence that length variations to repeat tracts are used as a means of modulating gene expression. For example, in some bacteria, such as *Haemophilus influenzae*, the susceptibility of microsatellites to reversible length changes is used to control specific genes that allow environmental adaptation [104, 108]. Thus, the hypermutable repeat sequence allows the bacterium to respond swiftly to changes in environmental conditions and adapt to different situations [104, 109]. Such variability in repeat tracts can even impact on the virulence of some bacteria, as seen in *H. influenzae* and *Neisseria meningitides* [110, 111]. Variation in the overall size of the repetitive domains was detected even among bacteria sub-cultured from a single colony, highlighting that the altered size of the repeat was intrinsic to the sequence.

## **Conclusions**

From the earliest studies of natural DNAs it became clear that repetitive DNA sequences are common, leading to expectations that there must be biological reasons to explain this. The advent of large numbers of genome sequences has reinforced these observations, but biologists continue to assess the full biological significance of repetitive regions of genomes. Different aspects of DNA metabolism influence genetic instabilities within these sequences, and many of the studies that have improved knowledge have originated in bacteria, where the experiments are most tractable. An important corollary of the results from such studies is that many of

the biochemical pathways are found in all organisms, meaning that many of the conclusions are relevant to all organisms.

Genetic instabilities of simple repeats may be mediated by many biochemical processes, including DNA replication-based slipped-strand mispairing, small slipped-register DNA synthesis, tandem duplications, and gene conversion-recombination processes. These processes may occur independently or in concert with each other and/or other DNA metabolic processes such as MMR, NER, DNA polymerase proofreading, SOS repair, transcription, etc. It is also clear that structural properties of the simple repeats (hairpin loop formation, slipped structures, triplexes, etc) play a consequential role in their genetic instabilities. The involvement of unusual DNA structures may occur because they are inherent within simple repeats inside cells, or because enzymes manipulating DNA may promote their formation. Either way, the presence of unusual structures within simple repeats is likely to influence the interaction of the DNA with proteins, which, in turn, facilitates the genetic instability of simple repeats.

Rapid progress in obtaining and interpreting genome information will continue to extend knowledge about the genetic variations that exist for simple repeating DNA sequences across all organisms. In this review we have summarized current understanding obtained from biochemical and cellular studies of such repeat sequences in bacteria. Combination of these different experiments in bacteria will shed further insight into the biological impacts of simple DNA repeats, including enhancement in understanding their roles in bacterial metabolism (with possible impact in treatment of bacterial pathogens) as well as in a range of human diseases.

## **Acknowledgements and Declarations of Interests**

We thank colleagues at the University of East Anglia, Institute of Biophysics and Central European Institute of Technology, Masaryk University in Brno, Czech Republic for discussions that assisted development of this manuscript. The authors have no conflicts of interest to declare.

## **Funding Information**

This work was supported by the Czech Science Foundation (18-15548S) and the European Union Horizon 2020 research and innovation programme under grant agreement No 692068, for the BISON project and by the SYMBIT project reg. no. CZ.02.1.01/0.0/0.0/15\_003/0000477 financed from the ERDF. VB and MF acknowledge institutional support from the Czech Academy of Sciences (68081707).

## **Author Contribution Statement**

All authors were involved in the planning and writing of the manuscript. Figures were prepared by VB and RPB.

## References

1. Castelle, C.J. and Banfield, J.F. (2018) Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181-97
2. Kashi, Y. and King, D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* **22**, 253-9
3. Treangen, T.J. and Salzberg, S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36-46
4. Schlötterer, C. (2016) Simple Repeats. *eLS*,
5. Mrazek, J., Guo, X. and Shah, A. (2007) Simple sequence repeats in prokaryotic genomes. *Proc Natl Acad Sci U S A* **104**, 8472-7
6. van Belkum, A., Scherer, S., van Alphen, L. and Verbrugh, H. (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* **62**, 275-93
7. Neil, A.J., Kim, J.C. and Mirkin, S.M. (2017) Precarious maintenance of simple DNA repeats in eukaryotes. *Bioessays*,
8. Bowater, R.P. and Waller, Z.A.E. (2014) DNA Structure. *eLS*,
9. Ohshima, K., Kang, S., Larson, J.E. and Wells, R.D. (1996) Cloning, characterisation, and properties of seven triplet repeat DNA sequences. *J Biol Chem* **271**, 16773-83
10. Brazda, V., Laister, R.C., Jagelska, E.B. and Arrowsmith, C. (2011) Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol Biol* **12**, 33
11. Blattner, F.R., Plunkett, I., G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-62
12. Chen, C.W., Huang, C.H., Lee, H.H., Tsai, H.H. and Kirby, R. (2002) Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends Genet* **18**, 522-9
13. Hopwood, D.A. (2006) Soil To Genomics: The *Streptomyces* Chromosome. *Annual Review of Genetics* **40**, 1-23
14. Toth, G., Gaspari, Z. and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* **10**, 967-81
15. Field, D. and Wills, C. (1998) Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *Proc Natl Acad Sci U S A* **95**, 1647-52
16. Mrazek, J. (2006) Analysis of distribution indicates diverse functions of simple sequence repeats in *Mycoplasma* genomes. *Molecular biology and evolution* **23**, 1370-85
17. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80
18. Gur-Arie, R., Cohen, C.J., Eitan, Y., Shelef, L., Hallerman, E.M. and Kashi, Y. (2000) Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. *Genome Res* **10**, 62-71
19. Kryukov, K., Sumiyama, K., Ikeo, K., Gojobori, T. and Saitou, N. (2012) A new database (GCD) on genome composition for eukaryote and prokaryote genome sequences and their initial analyses. *Genome biology and evolution* **4**, 501-12
20. Wu, H., Zhang, Z., Hu, S. and Yu, J. (2012) On the molecular mechanism of GC content variation among eubacterial genomes. *Biology direct* **7**, 2
21. Bhagwat, A.S. and Lieb, M. (2002) Cooperation and competition in mismatch repair: very short-patch repair and methyl-directed mismatch repair in *Escherichia coli*. *Mol Microbiol* **44**, 1421-8
22. Aksenova, A.Y. and Mirkin, S.M. (2019) At the Beginning of the End and in the Middle of the Beginning: Structure and Maintenance of Telomeric DNA Repeats and Interstitial Telomeric Sequences. *Genes* **10**,
23. Metzgar, D., Thomas, E., Davis, C., Field, D. and Wills, C. (2001) The microsatellites of *Escherichia coli*: rapidly evolving repetitive DNAs in a non-pathogenic prokaryote. *Mol Microbiol* **39**, 183-90
24. van Belkum, A. (2007) Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). *FEMS immunology and medical microbiology* **49**, 22-7
25. Dyet, K.H., Robertson, I., Turbitt, E. and Carter, P.E. (2011) Characterization of *Escherichia coli* O157:H7 in New Zealand using multiple-locus variable-number tandem-repeat analysis. *Epidemiol Infect* **139**, 464-71
26. Byrne, L., Elson, R., Dallman, T.J., Perry, N., Ashton, P., Wain, J., Adak, G.K., Grant, K.A. and Jenkins, C. (2014) Evaluating the use of multilocus variable number tandem repeat analysis of Shiga toxin-producing *Escherichia coli* O157 as a routine public health tool in England. *PLoS One* **9**, e85901
27. Melles, D.C., Schouls, L., Francois, P., Herzig, S., Verbrugh, H.A., van Belkum, A. and Schrenzel, J. (2009) High-throughput typing of *Staphylococcus aureus* by amplified fragment length polymorphism (AFLP)

- or multi-locus variable number of tandem repeat analysis (MLVA) reveals consistent strain relatedness. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* **28**, 39-45
28. Mohanty, P.S., Bansal, A.K., Naaz, F., Arora, M., Gupta, U.D., Gupta, P., Sharma, S. and Singh, H. (2019) Multiple strain infection of *Mycobacterium leprae* in a family having 4 patients: A study employing short tandem repeats. *PLoS One* **14**, e0214051
  29. Bowater, R.P., Chen, D. and Lilley, D.M.J. (1994) Elevated unconstrained supercoiling of plasmid DNA generated by transcription and translation of the tetracycline resistance gene in eubacteria. *Biochemistry* **33**, 9266-75
  30. Hatfield, G.W. and Benham, C.J. (2002) DNA topology-mediated control of global gene expression in *Escherichia coli*. *Annu Rev Genet* **36**, 175-203
  31. Pearson, C.E., Edamura, K.N. and Cleary, J.D. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* **6**, 729-42
  32. Mirkin, S.M. (2006) DNA structures, repeat expansions and human hereditary disorders. *Curr Opin Struct Biol* **16**, 351-8
  33. Wells, R.D. (2007) Non-B DNA conformations, mutagenesis and disease. *Trends Biochem Sci* **32**, 271-8
  34. Sjakste, T., Paramonova, N. and Sjakste, N. (2016) Structural and functional significance of microsatellites. *Biopolymers and Cell* **32**, 334-46
  35. McClellan, J.A., Boublikova, P., Palecek, E. and Lilley, D.M.J. (1990) Superhelical torsion in cellular DNA responds directly to environmental and genetic factors. *Proc Natl Acad Sci USA* **87**, 8373-7
  36. Gimenes, F., Takeda, K.I., Fiorini, A., Gouveia, F.S. and Fernandez, M.A. (2008) Intrinsically bent DNA in replication origins and gene promoters. *Genetics and molecular research : GMR* **7**, 549-58
  37. Murat, P. and Balasubramanian, S. (2014) Existence and consequences of G-quadruplex structures in DNA. *Curr Opin Genet Dev* **25**, 22-9
  38. Mitas, M. (1997) Trinucleotide repeats associated with human disease. *Nucleic Acids Res* **25**, 2245-54
  39. Kiliszek, A. and Rypniewski, W. (2014) Structural studies of CNG repeats. *Nucleic Acids Res* **42**, 8189-99
  40. Wang, G. and Vasquez, K.M. (2007) Z-DNA, an active element in the genome. *Front Biosci* **12**, 4424-38
  41. Bochman, M.L., Paeschke, K. and Zakian, V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**, 770-80
  42. Malgowska, M., Gudanis, D., Kierzek, R., Wyszko, E., Gabelica, V. and Gdaniec, Z. (2014) Distinctive structural motifs of RNA G-quadruplexes composed of AGG, CGG and UGG trinucleotide repeats. *Nucleic Acids Res* **42**, 10196-101207
  43. Zamiri, B., Mirceta, M., Bomsztyk, K., Macgregor, R.B., Jr. and Pearson, C.E. (2015) Quadruplex formation by both G-rich and C-rich DNA strands of the C9orf72 (GGGGCC)<sup>8</sup>(GGCCCC)<sup>8</sup> repeat: effect of CpG methylation. *Nucleic Acids Res* **43**, 10055-64
  44. Cheng, M., Cheng, Y., Hao, J., Jia, G., Zhou, J., Mergny, J.L. and Li, C. (2018) Loop permutation affects the topology and stability of G-quadruplexes. *Nucleic Acids Res* **46**, 9264-75
  45. Bartas, M., Cutova, M., Brazda, V., Kaura, P., Stastny, J., Kolomaznik, J., Coufal, J., Goswami, P., Cerven, J., et al. (2019) The Presence and Localization of G-Quadruplex Forming Sequences in the Domain of Bacteria. *Molecules* **24**,
  46. McRae, E.K.S., Booy, E.P., Padilla-Meier, G.P. and McKenna, S.A. (2017) On Characterizing the Interactions between Proteins and Guanine Quadruplex Structures of Nucleic Acids. *Journal of nucleic acids* **2017**, Article ID 9675348
  47. Qiu, J., Liu, J., Chen, S., Ou, T.M., Tan, J.H., Gu, L.Q., Huang, Z.S. and Li, D. (2015) Role of Hairpin-Quadruplex DNA Secondary Structural Conversion in the Promoter of hnRNP K in Gene Transcriptional Regulation. *Organic letters* **17**, 4584-7
  48. Abdelhamid, M.A., Fabian, L., MacDonald, C.J., Cheesman, M.R., Gates, A.J. and Waller, Z.A. (2018) Redox-dependent control of i-Motif DNA structure using copper cations. *Nucleic Acids Res* **46**, 5886-93
  49. Dembska, A., Bielecka, P. and Juskowiak, B. (2017) pH-Sensing fluorescence oligonucleotide probes based on an i-motif scaffold: a review. *Analytical Methods* **9**, 6092-106
  50. Abou Assi, H., Garavis, M., Gonzalez, C. and Damha, M.J. (2018) i-Motif DNA: structural features and significance to cell biology. *Nucleic Acids Res*, in press
  51. Wright, E.P., Huppert, J.L. and Waller, Z.A.E. (2017) Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH. *Nucleic Acids Res* **45**, 2951-9
  52. Fojtik, P., Kejnovska, I. and Vorlickova, M. (2004) The guanine-rich fragile X chromosome repeats are reluctant to form tetraplexes. *Nucleic Acids Res* **32**, 298-306



53. Renciuik, D., Zemanek, M., Kejnovska, I. and Vorlickova, M. (2009) Quadruplex-forming properties of FRAXA (CGG) repeats interrupted by (AGG) triplets. *Biochimie* **91**, 416-22
54. Schmidt, M.H. and Pearson, C.E. (2016) Disease-associated repeat instability and mismatch repair. *DNA Repair (Amst)* **38**, 117-26
55. Freudenreich, C.H. (2018) R-loops: targets for nuclease cleavage and repeat instability. *Current genetics* **64**, 789-94
56. Kumari, D., Lokanga, R., Yudkin, D., Zhao, X.N. and Usdin, K. (2012) Chromatin changes in the development and pathology of the Fragile X-associated disorders and Friedreich ataxia. *Biochim Biophys Acta* **1819**, 802-10
57. Wang, G. and Vasquez, K.M. (2014) Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair (Amst)* **19**, 143-51
58. Bacolla, A. and Wells, R.D. (2009) Non-B DNA conformations as determinants of mutagenesis and human disease. *Molecular carcinogenesis* **48**, 273-85
59. Rocha, E.P. (2008) The organization of the bacterial genome. *Annu Rev Genet* **42**, 211-33
60. West, B.J., Allegrini, P., Buiatti, M. and Grigolini, P. (2000) Non-normal Statistics of DNA Sequences of Prokaryotes. *Journal of biological physics* **26**, 17-25
61. Jansen, R., Embden, J.D., Gaastra, W. and Schouls, L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565-75
62. Kolsto, A.B. (1997) Dynamic bacterial genome organization. *Mol Microbiol* **24**, 241-8
63. Cechova, J., Lysek, J., Bartas, M. and Brazda, V. (2017) Complex analyses of inverted repeats in mitochondrial genomes revealed their importance and variability. *Bioinformatics*,
64. Brazda, V., Lysek, J., Bartas, M. and Fojta, M. (2018) Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs. *Biomed Res Int* **2018**, 1097018
65. Horwitz, M.S. and Loeb, L.A. (1988) An E. coli promoter that regulates transcription by DNA superhelix-induced cruciform extrusion. *Science* **241**, 703-5
66. Holder, I.T., Wagner, S., Xiong, P., Sinn, M., Frickey, T., Meyer, A. and Hartig, J.S. (2015) Intrastrand triplex DNA repeats in bacteria: a source of genomic instability. *Nucleic Acids Res* **43**, 10126-42
67. Bacolla, A., Wang, G. and Vasquez, K.M. (2015) New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity. *PLoS Genet* **11**, e1005696
68. Kikin, O., D'Antonio, L. and Bagga, P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* **34**, W676-82
69. Brazda, V., Kolomaznik, J., Lysek, J., Bartas, M., Fojta, M., Stastny, J. and Mergny, J.L. (2019) G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics*, in press
70. Yadav, V.K., Abraham, J.K., Mani, P., Kulshrestha, R. and Chowdhury, S. (2008) QuadBase: genome-wide database of G4 DNA--occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res* **36**, D381-5
71. Dhapola, P. and Chowdhury, S. (2016) QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res* **44**, W277-83
72. Rawal, P., Kummarasetti, V.B., Ravindran, J., Kumar, N., Halder, K., Sharma, R., Mukerji, M., Das, S.K. and Chowdhury, S. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome Res* **16**, 644-55
73. Brazda, V., Haronikova, L., Liao, J.C. and Fojta, M. (2014) DNA and RNA quadruplex-binding proteins. *International journal of molecular sciences* **15**, 17493-517
74. Day, H.A., Pavlou, P. and Waller, Z.A. (2014) i-Motif DNA: structure, stability and targeting with ligands. *Bioorg Med Chem* **22**, 4407-18
75. Harris, L.M. and Merrick, C.J. (2015) G-quadruplexes in pathogens: a common route to virulence control? *PLoS pathogens* **11**, e1004562
76. Rhodes, D. and Lipps, H.J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**, 8627-37
77. Saranathan, N. and Vivekanandan, P. (2019) G-Quadruplexes: More Than Just a Kink in Microbial Genomes. *Trends Microbiol* **27**, 148-63
78. Day, H.A., Wright, E.P., MacDonald, C.J., Gates, A.J. and Waller, Z.A. (2015) Reversible DNA i-motif to hairpin switching induced by copper(II) cations. *Chem Commun (Camb)* **51**, 14099-102
79. Waller, Z.A., Pinchbeck, B.J., Buguth, B.S., Meadows, T.G., Richardson, D.J. and Gates, A.J. (2016) Control of bacterial nitrate assimilation by stabilization of G-quadruplex DNA. *Chem Commun (Camb)* **52**, 13511-4
80. Abdelhamid, M.A.S., Gates, A.J. and Waller, Z.A.E. (2019) Destabilization of i-Motif DNA at Neutral pH by G-Quadruplex Ligands. *Biochemistry* **58**, 245-9

81. Pinchbeck, B.J., Soriano-Laguna, M.J., Sullivan, M.J., Luque-Almagro, V.M., Rowley, G., Ferguson, S.J., Roldan, M.D., Richardson, D.J. and Gates, A.J. (2019) A dual functional redox enzyme maturation protein for respiratory and assimilatory nitrate reductases in bacteria. *Mol Microbiol*, in press
82. Iyer, R.R., Pluciennik, A., Napierala, M. and Wells, R.D. (2015) DNA triplet repeat expansion and mismatch repair. *Annu Rev Biochem* **84**, 199-226
83. Chatterjee, N. and Walker, G.C. (2017) Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen* **58**, 235-63
84. Shah, K.A. and Mirkin, S.M. (2015) The hidden side of unstable DNA repeats: Mutagenesis at a distance. *DNA Repair (Amst)* **32**, 106-12
85. Lahue, R.S. and Slater, D.L. (2003) DNA repair and trinucleotide repeat instability. *Front Biosci* **8**, s653-65
86. Zhao, X.N. and Usdin, K. (2015) The Repeat Expansion Diseases: The dark side of DNA repair. *DNA Repair (Amst)* **32**, 96-105
87. Bowater, R.P. and Wells, R.D. (2001) The intrinsically unstable life of DNA triplet repeats associated with human hereditary disorders. *Prog Nucl Acids Res Mol Biol* **66**, 159-202
88. Krasilnikova, M., Samadashwily, G.M., Krasilnikov, A.S. and Mirkin, S.M. (1998) Transcription through a simple DNA repeat blocks replication elongation. *EMBO J* **17**, 5095-102
89. Bowater, R.P., Jaworski, A., Larson, J.E., Parniewski, P. and Wells, R.D. (1997) Transcription increases the deletion frequency of long CTG•CAG triplet repeats from plasmids in *Escherichia coli*. *Nucl Acids Res* **25**, 2861-8
90. Lin, Y., Dent, S.Y., Wilson, J.H., Wells, R.D. and Napierala, M. (2010) R loops stimulate genetic instability of CTG.CAG repeats. *Proc Natl Acad Sci U S A* **107**, 692-7
91. Reddy, K., Tam, M., Bowater, R.P., Barber, M., Tomlinson, M., Nichol Edamura, K., Wang, Y.H. and Pearson, C.E. (2011) Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. *Nucleic Acids Res* **39**, 1749-62
92. Zhao, J., Bacolla, A., Wang, G. and Vasquez, K.M. (2010) Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci* **67**, 43-62
93. Hoeijmakers, J.H. (2001) Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366-74
94. Friedberg, E.C. (2003) DNA damage and repair. *Nature* **421**, 436-40
95. Gorna, A.E., Bowater, R.P. and Dziadek, J. (2010) DNA repair systems and the pathogenesis of *Mycobacterium tuberculosis*: varying activities at different stages of infection. *Clin Sci (Lond)* **119**, 187-202
96. van der Veen, S. and Tang, C.M. (2015) The BER necessities: the repair of DNA damage in human-adapted bacterial pathogens. *Nat Rev Microbiol* **13**, 83-94
97. Uphoff, S. and Sherratt, D.J. (2017) Single-Molecule Analysis of Bacterial DNA Repair and Mutagenesis. *Annu Rev Biophys* **46**, 411-32
98. Loeb, K.R. and Loeb, L.A. (1999) Genetic instability and the mutator phenotype. *Amer J Pathol* **154**, 1621-6
99. Bacolla, A., Wojciechowska, M., Kosmider, B., Larson, J.E. and Wells, R.D. (2006) The involvement of non-B DNA structures in gross chromosomal rearrangements. *DNA Repair (Amst)* **5**, 1161-70
100. Wojcik, E.A., Brzostek, A., Bacolla, A., Mackiewicz, P., Vasquez, K.M., Korycka-Machala, M., Jaworski, A. and Dziadek, J. (2012) Direct and inverted repeats elicit genetic instability by both exploiting and eluding DNA double-strand break repair systems in mycobacteria. *PLoS One* **7**, e51064
101. Mendoza, O., Bourdoncle, A., Boule, J.B., Brosh, R.M., Jr. and Mergny, J.L. (2016) G-quadruplexes and helicases. *Nucleic Acids Res* **44**, 1989-2006
102. Shen, J.C. and Loeb, L.A. (2000) The Werner syndrome gene: the molecular basis of RecQ helicase-deficiency diseases. *Trends Genet* **16**, 213-20
103. Wu, X. and Maizels, N. (2001) Substrate-specific inhibition of RecQ helicase. *Nucleic Acids Res* **29**, 1765-71
104. Moxon, R., Bayliss, C. and Hood, D. (2006) Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* **40**, 307-33
105. Kashi, Y., King, D. and Soller, M. (1997) Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet* **13**, 74-8
106. Karlin, S., Campbell, A.M. and Mrazek, J. (1998) Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32**, 185-225
107. Groisman, E.A. and Casadesus, J. (2005) The origin and evolution of human pathogens. *Mol Microbiol* **56**, 1-7
108. Power, P.M., Sweetman, W.A., Gallacher, N.J., Woodhall, M.R., Kumar, G.A., Moxon, E.R. and Hood, D.W. (2009) Simple sequence repeats in *Haemophilus influenzae*. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **9**, 216-28

109. Zhou, K., Aertsen, A. and Michiels, C.W. (2014) The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol Rev* **38**, 119-41
110. Peak, I.R., Jennings, M.P., Hood, D.W. and Moxon, E.R. (1999) Tetranucleotide repeats identify novel virulence determinant homologues in *Neisseria meningitidis*. *Microbial pathogenesis* **26**, 13-23
111. Hood, D.W., Deadman, M.E., Jennings, M.P., Bisercic, M., Fleischmann, R.D., Venter, J.C. and Moxon, E.R. (1996) DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proc Natl Acad Sci USA* **93**, 11121-215

## Figure Legends

**Figure 1:** Nomenclature to illustrate variation of microsatellites repeats. Microsatellite sequences consist of up to 6 bases per repeat and examples are shown for microsatellite repeats consisting of one base (mononucleotide), two different bases (dinucleotide) and three different bases (trinucleotide). Note that shifting of the frame of the sequence highlights redundancy within each repeat, meaning that it covers multiple types of sequences – the green box highlights (TC) repeats within (CT) repeats. Classical examples of microsatellites consist of uninterrupted repeats of the same sequence. When one or more bases interrupt the repeat array (shown by the bases in red), the microsatellite is termed “interrupted” (sometimes referred to as “imperfect”). Two types of repeat that neighbour each other are also found frequently in genomes, and are called “compound” (or sometimes “composite”) microsatellites. Adapted from [4].

**Figure 2:** Inverted repeat DNA sequences can adopt different types of three-dimensional structure. “Inverted repeats” are repetitive DNA elements where the 5’-3’ sequence of one strand is the same when the complementary strand is read in the 5’-3’ direction. The sequence shown is the inverted repeat from *E.coli K12* genome 3144772-3144797. (A) Such DNA sequences can exist in a regular double-stranded, anti-parallel form. (B) Intra-strand base pairing within the inverted repeat allows the formation of a cruciform.

**Figure 3:** Ribbon scheme of localised non-B DNA structures: (A) triplex; (B) hairpin; (C) slip-stranded DNA; (D) cruciform; (E) G-quadruplex; (F) i-motif. Black and red represents individual DNA strands, and G-quartets are highlighted by rhomboids.

**Figure 4:** Watson-Crick and Hoogsteen hydrogen bonds in triplex DNA molecules. A variety of triplex structures are shown involving three separate bases. The most common sequences, both in intramolecular and intermolecular triplexes, include A•A–T, G•G–C, and T•A–T for R•R–Y type triplexes (bottom left), and C+•G–C and T•A–T for Y•R–Y type triplexes. Each triplex includes two bases that form hydrogen bonds following the standard (Watson-Crick) pattern (red), plus one additional base form base-pair where the interactions are stabilised by Hoogsteen pairing (green) [66]. Note that in some cases the additional hydrogen bonds are stabilised by positive charges on a cytosine base and, thus, are favoured at low pH.

**Figure 5:** Suggested biological roles of simple DNA repeats. Central part: single DNA repeats (red) can form various local DNA structures (e.g. see Figure 3), which can participate in: A) protein recognition; B) genetic instability; C) genome evolution; D) regulation of transcription; E) genome organization; F) DNA replication. Colours highlight proteins with specificity for transcription (green), replication (yellow) or simply to the DNA structure or single stranded DNA (blue).

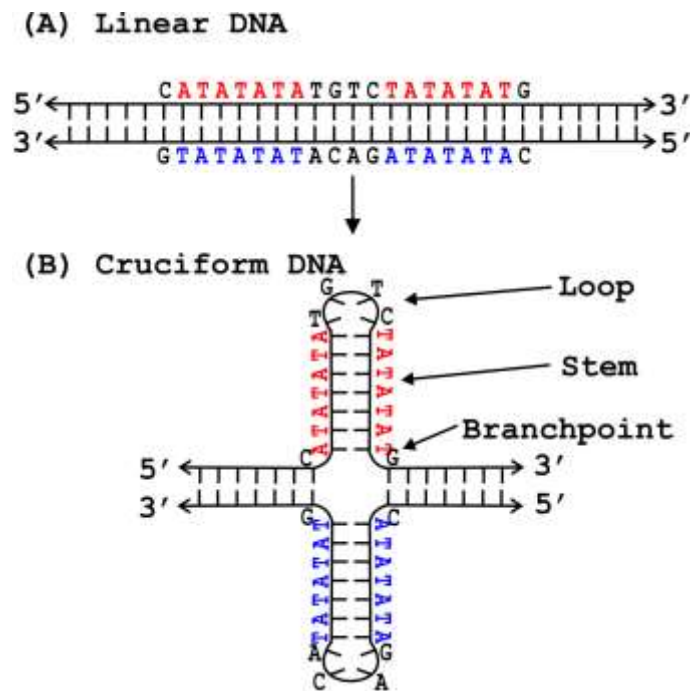
**Figure 6:** Potential quadruplex forming sequences are dispersed throughout the *Escherichia coli* genome. The presence of potential quadruplex forming sequences in *Escherichia coli* 55989 was visualized by Quadbase [70, 71]. The complete length of

the genome is 5,154,862 bp, with a chromosomal CG content of 50.10% (outer black circle). The presence of quadruplex motifs is highlighted by the lines emanating from the inner circles: the “+ strand” has 37 (middle blue circle), the “- strand” has 28 (inner red circle).

## Figures

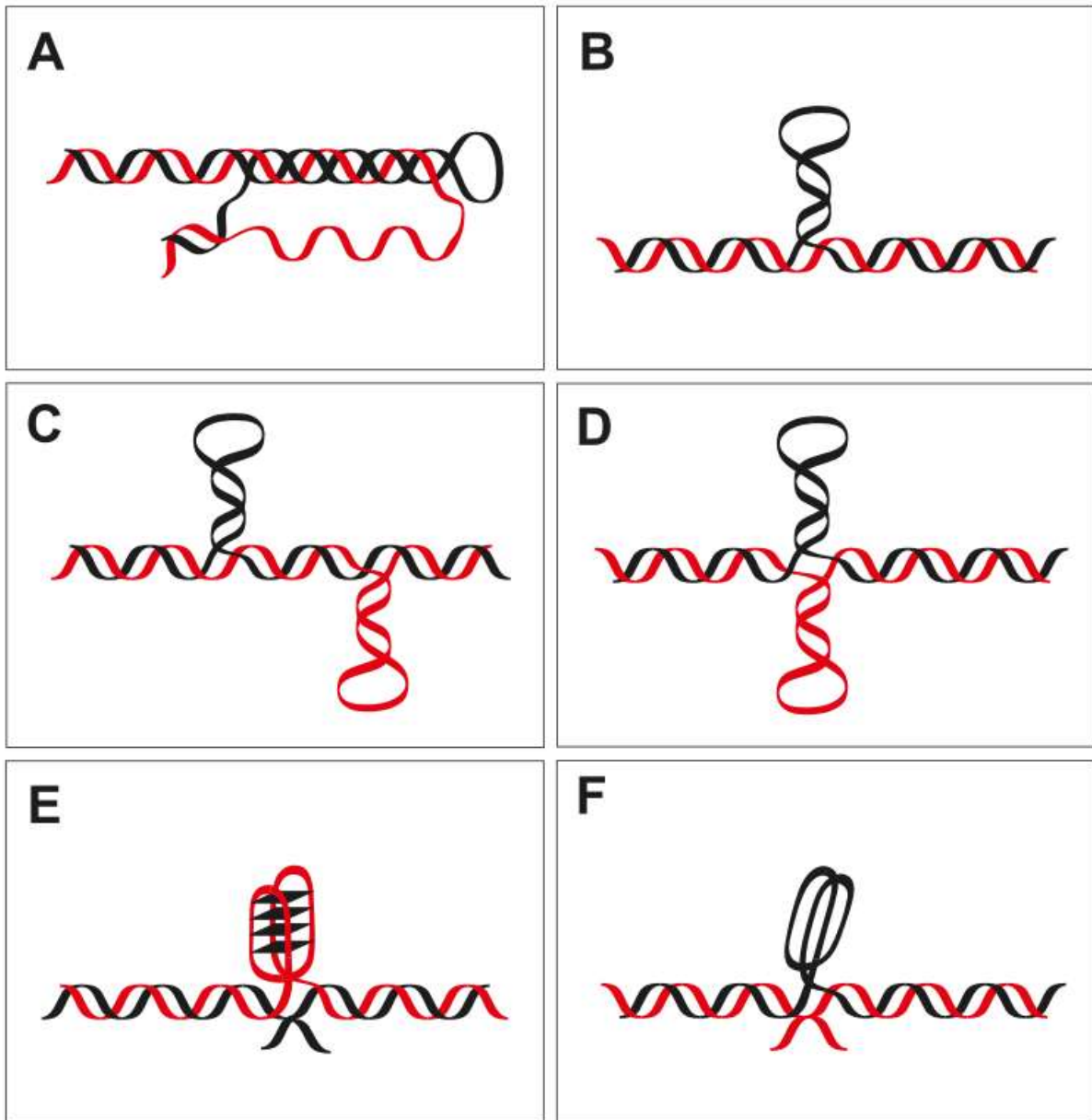
Type of microsatellite	Example sequence	Shorthand nomenclature
Uninterrupted	CCCCCC CTCTCTCTCTCT CTGCTGCTGCTGCTGCTG	$(C)_6$ $(CT)_6$ or $(TC)_5$ $(CTG)_6$
Interrupted	CC <b>A</b> CCC CTCT <b>A</b> ACTCTCT CTGCTGCTG <b>AA</b> ACTGCTG	$(C)_2A(C)_3$ $(CT)_2AA(CT)_3$ $(CTG)_3AAA(CTG)_2$
Compound	CCC <b>AAA</b> CTCTCT <b>CACACA</b> CTGCTGCTG <b>CAGCAGCAG</b>	$(C)_3(A)_3$ $(CT)_3(CA)_3$ $(CTG)_3(CAG)_3$

**Figure 1:** Nomenclature to illustrate variation of microsatellites repeats. Microsatellite sequences consist of up to 6 bases per repeat and examples are shown for microsatellite repeats consisting of one base (mononucleotide), two different bases (dinucleotide) and three different bases (trinucleotide). Note that shifting of the frame of the sequence highlights redundancy within each repeat, meaning that it covers multiple types of sequences – the green box highlights (TC) repeats within (CT) repeats. Classical examples of microsatellites consist of uninterrupted repeats of the same sequence. When one or more bases interrupt the repeat array (shown by the bases in red), the microsatellite is termed “interrupted” (sometimes referred to as “imperfect”). Two types of repeat that neighbour each other are also found frequently in genomes, and are called “compound” (or sometimes “composite”) microsatellites. Adapted from [4].

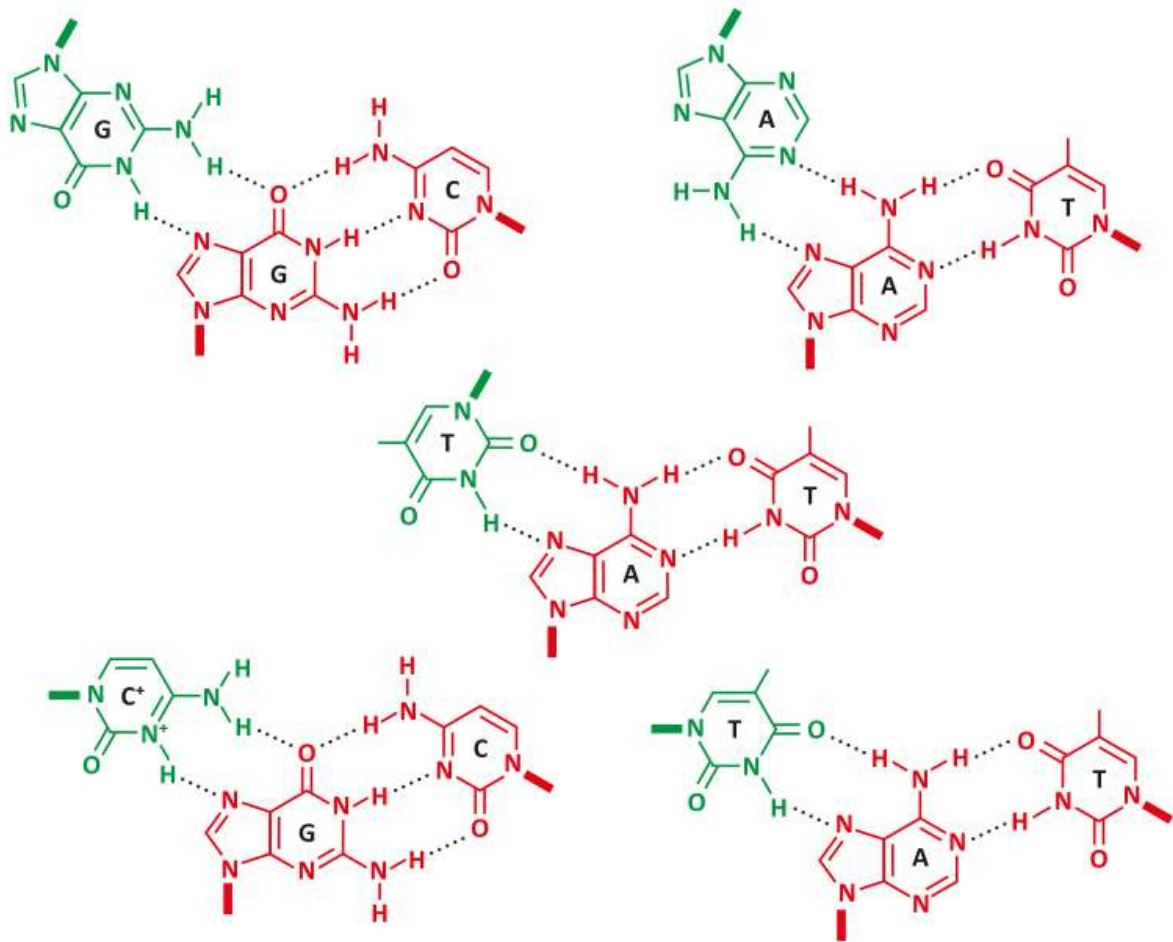


**Figure 2:** Inverted repeat DNA sequences can adopt different types of three-dimensional structure. “Inverted repeats” are repetitive DNA elements where the 5’-3’ sequence of one strand is the same when the complementary strand is read in the 5’-3’ direction. The sequence shown is the inverted repeat from *E.coli K12* genome 3144772-3144797. (A) Such DNA sequences can exist in a regular double-stranded, anti-parallel form. (B) Intra-strand base pairing within the inverted repeat allows the formation of a cruciform.

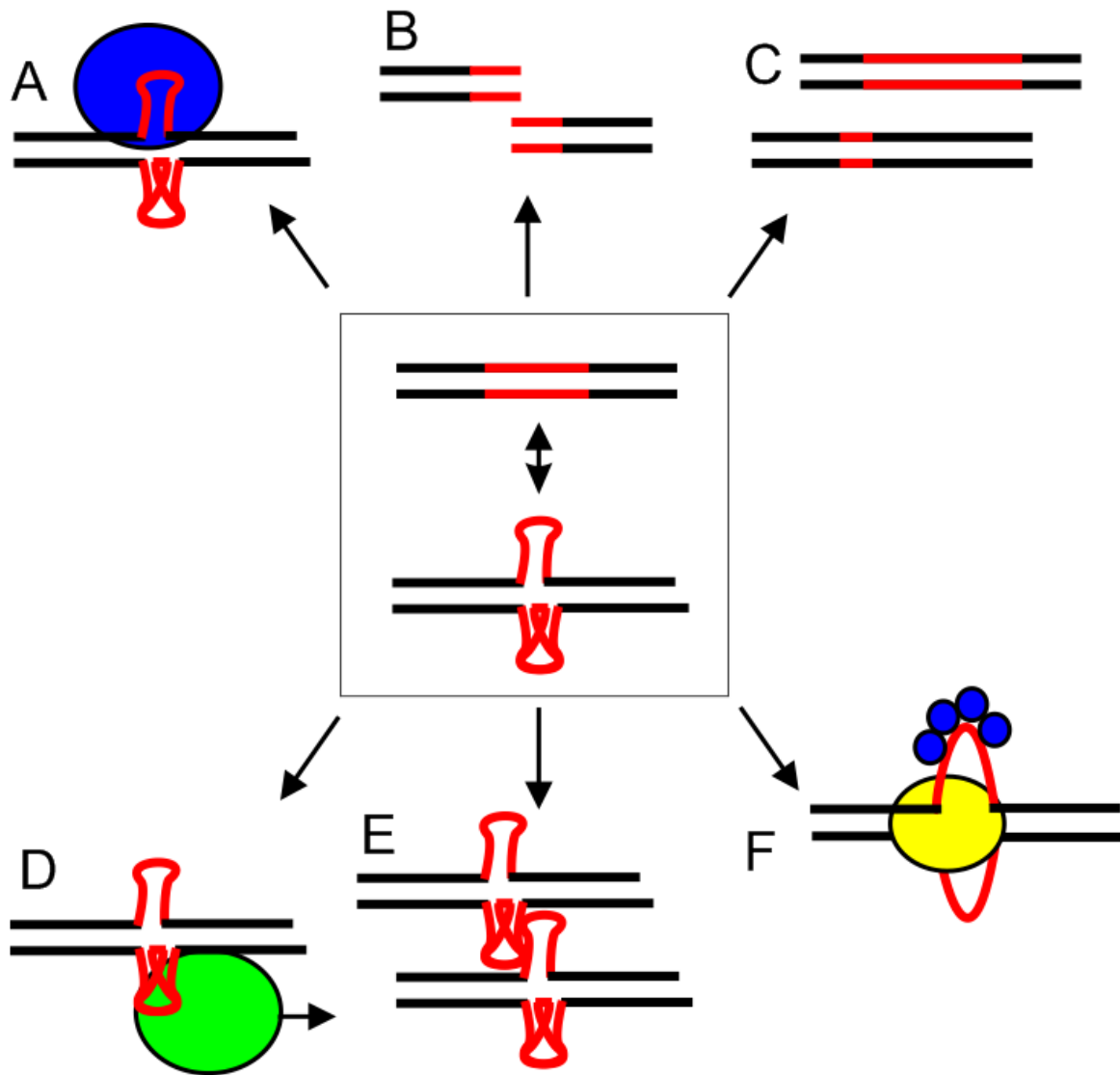




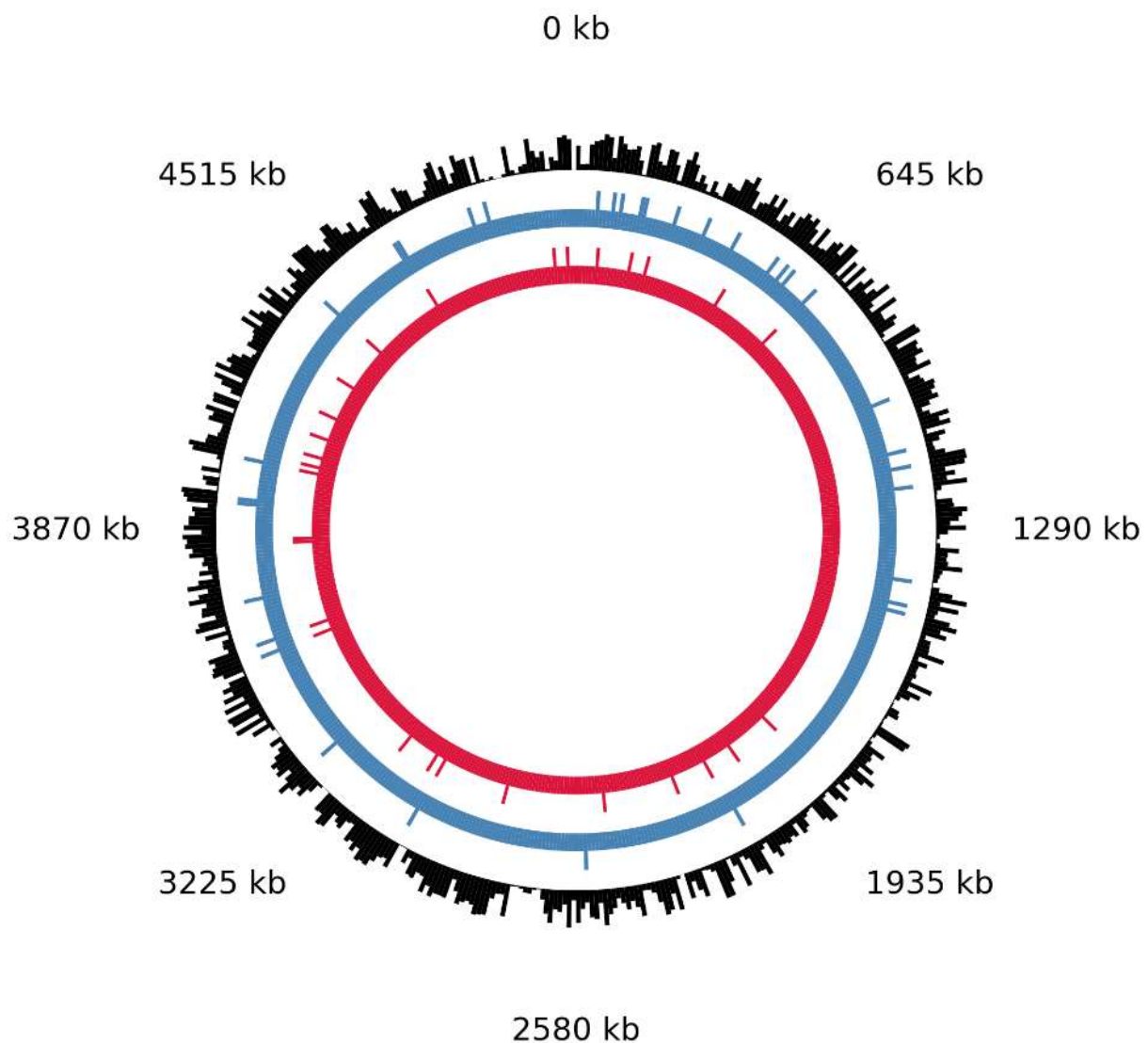
**Figure 3:** Ribbon scheme of localised non-B DNA structures: (A) triplex; (B) hairpin; (C) slip-stranded DNA; (D) cruciform; (E) G-quadruplex; (F) i-motif. Black and red represents individual DNA strands, and G-quartets are highlighted by rhomboids.



**Figure 4:** Watson-Crick and Hoogsteen hydrogen bonds in triplex DNA molecules. A variety of triplex structures are shown involving three separate bases. The most common sequences, both in intramolecular and intermolecular triplexes, include A•A–T, G•G–C, and T•A–T for R•R–Y type triplexes (bottom left), and C••G–C and T•A–T for Y•R–Y type triplexes. Each triplex includes two bases that form hydrogen bonds following the standard (Watson-Crick) pattern (red), plus one additional base form base-pair where the interactions are stabilised by Hoogsteen pairing (green) [66]. Note that in some cases the additional hydrogen bonds are stabilised by positive charges on a cytosine base and, thus, are favoured at low pH.



**Figure 5:** Suggested biological roles of simple DNA repeats. Central part: Single DNA repeats (red) can form various local DNA structures (e.g. see Figure 3), which can participate in: A) protein recognition; B) genetic instability; C) genome evolution; D) regulation of transcription; E) genome organization; F) DNA replication. Colours highlight proteins with specificity to transcription (green), replication (yellow) or simply to the DNA structure or single stranded DNA (blue).



**Figure 6:** Potential quadruplex forming sequences are dispersed throughout the *Escherichia coli* genome. The presence of potential quadruplex forming sequences in *Escherichia coli* 55989 was visualized by Quadbase [70, 71]. The complete length of the genome is 5,154,862 bp, with a chromosomal CG content of 50.10% (outer black circle). The presence of quadruplex motifs is highlighted by the lines emanating from the inner circles: the “+ strand” has 37 (middle blue circle), the “- strand” has 28 (inner red circle).