Non-canonical Nucleic Acids in Bacteria

-Structural Characterization and Functional Properties

of Quadruplex and Triplex Conformations-

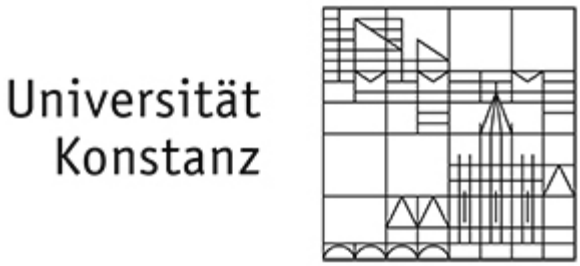
Dissertation submitted for the degree of

Doctor of Natural Sciences (Dr. rer. nat.)

Presented by

Isabelle T. Holder (née Seemann)

at the



Faculty of Natural Sciences

Department of Chemistry

Date of the oral examination: 18.12.2014

First referee: Prof. Dr. J. S. Hartig

Second referee: Prof. Dr. A. Marx

„The truth is rarely pure and never simple. “

Oscar Wilde

Parts of this work were published in:

I.T. Holder, S. Wagner, P. Xiong, T. Frickey, M. Sinn, K. Halder, A. Meyer, J.S. Hartig "DNA triplex repeats in *Escherichia coli*: A source of genetic instability", *Nucleic Acids Research*, **2014,** in preparation

I.T. Holder, J.S. Hartig "A matter of location: Influence of G-quadruplexes on *Escherichia coli* gene expression", *Chemistry & Biology*, **2014**, 21, 1511

Additional publications:

C. Rehm, I.T. Holder, A. Gross, F. Wojciechowski, M. Urban, M. Sinn, M. Drescher, J.S. Hartig "A bacterial DNA quadruplex with exceptional K^+ selectivity and unique structural polymorphism", *Chemical Science*, **2014**, 5, 2809

I.T. Holder, M. Drescher, J.S. Hartig "Structural Characterization of Quadruplex DNA with in-cell EPR approaches", *Bioorganic & Medicinal Chemistry*, **2013**, 21, 6156

M. Azarkh, V. Singh, O. Okle, I.T. Seemann, D.R. Dietrich, J.S. Hartig, M. Drescher "Site directed spin labeling of nucleotides and the use of in-cell DEER: An EPR method for determination of long-range distances in cellulo", *Nature Protocols*, **2013**, 8, 131

I.T. Seemann, V. Singh, M. Azarkh, M. Drescher, J.S. Hartig "Small molecule-triggered manipulation of DNA three-way-junctions", *Journal of the American Chemical Society*, **2011**, 133, 4706

I.T. Seemann, J.S. Hartig "Artificial Ribozyme-Based Regulators of Gene Expression (Account)", *Synlett*, **2011**, 11, 1486

M. Azarkh, O. Okle, V. Singh, I.T. Seemann, J.S. Hartig, D.R. Dietrich, and M. Drescher "Long-Range Distance Determination in a DNA Model System inside Xenopus Laevis Oocytes by In-Cell Spin-Label EPR", *ChemBioChem*, **2011**, 12, 1992

CONTENTS

1 INTRODUCTION

Nucleic acids are the storage of the genetic information in every living organism. The structure of DNA was determined by Watson and Crick in 1953 (1). Since then, research investigating nucleic acid conformations and functions has continually increased. The ability of nucleic acids to serve as regulatory units influencing gene expression has been shown, e.g. with riboswitches. In 2001 the sequence of the human genome was determined by whole genome shotgun sequencing (2). This paved the way for several studies investigating gene functions, genetic disorders and human diseases. With the elucidation of eukaryotic and prokaryotic genomes, genomic repeat sequences were discovered that might be able to fold into secondary nucleic acid structures. Such non-canonical conformations have been assumed to play important roles in a variety of biological processes, but ultimate proof for their specific functions is scarce. Easy growing conditions, fast proliferation and well described genetic mechanisms make prokaryotic systems particularly useful to investigate general mechanisms induced by such motifs. Therefore it is necessary to examine the structural und functional properties and the *in vivo* occurrence of non-canonical nucleic acids in bacteria.

1.1 Non-canonical nucleic acid structures

Inside the nucleus, DNA usually occurs as a double-stranded, right-handed helix – the so called B-DNA (3). Apart from this well known, canonical structure, DNA can adopt several non-canonical (non-B) conformations (4). Helical three-way junctions are the simplest branched nucleic acid molecules that can arise. They are comprised of three double helical arms which are connected at a branch-point. Branched DNA molecules appear during DNA metabolism processes such as replication, repair or recombination (5-8). Holliday junctions (9) are well described cruciforms, containing four double-helical arms that branch out from a central junction. More complex DNA structures occur when more than two DNA strands interact with each other. In triple-stranded DNA, one strand binds via Hoogsteen or reverse Hoogsteen hydrogen bonds to the major groove of a B-form double-helix. Such structures can either be formed inter- or intra-molecularly in pyrimidine- or purine-rich regions (10-12). DNA-quartet structures, more often called G-quadruplexes, are made up of Hoogsteen hydrogen bonded G-tetrads that can stack on one another. Interestingly, most of these

structures can arise from particular sequence patterns, such as guanosine rich regions. Computational studies have shown that non-canonical motifs are widely distributed in human and bacterial genomes (13-15). It is unclear whether they form alternative DNA structures *in vivo*, although their potential to do so das been described *in vitro* by a variety of methods.

1.1.1 G-quadruplex structures

Repetitive guanine-rich nucleic acids are prone to adopt G-quadruplex structures. G-quadruplexes are made up of at least two tetrad layers that stack upon each other via π-π interactions. Each tetrad is composed of four guanines stabilized by Hoogsteen base pairs in a coplanar arrangement (16) (*see Figure 1.1 A*).

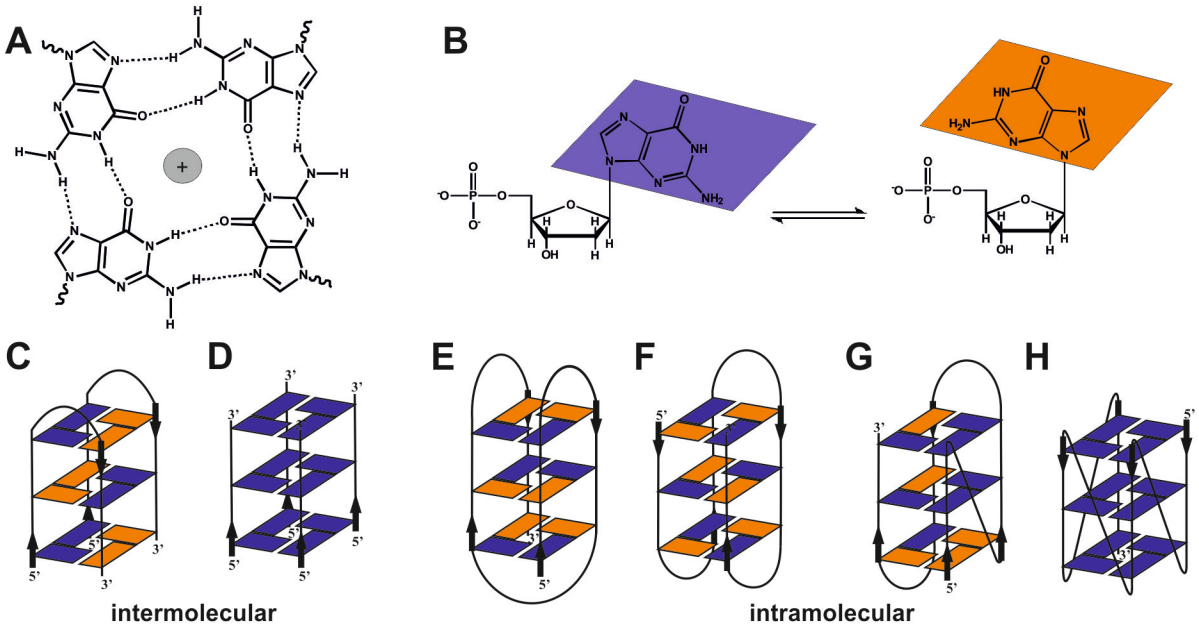


Figure 1.1: Quadruplex structure and topologies.

A Coplanar arrangement of guanines by non-canonical Hoogsteen base pairing. The tetrad can be stabilized by monovalent cations (especially potassium). **B** Guanosine glycosidic conformations. Syn-conformation is depicted in orange and anti is shown in blue. **C, D** Examples of intermolecular G-quadruplexes, with **C** depicting an antiparallel quadruplex made up of 2 strands and **D** showing a parallel quadruplex formed by 4 distinct strands. **E-H** Different quadruplex topologies, shown for intramolecular structures: **E** antiparallel basket type; **F** antiparallel chair type; **G** (3+1) hybrid; **H** parallel propeller type. The respective glycosidic conformations of the guanosines are shown in blue (anti) and orange (syn). Figure modified from (17).

In intramolecular quadruplexes, the guanosines are located on the same nucleic acid strand, where the interjacent sequences are usually bulged out as single-stranded loops of different

lengths (*see Figure 1.1 E-H*). The guanosines can also appear on different strands where the Hoogsteen bonds occur intermolecularly (*see Figure 1.1 C&D*). Quadruplexes that are able to fold within one strand usually can also form multimolecular structures, depending on the concentration of the respective molecule (18-20). G-quadruplex structures can fold within DNA and RNA. DNA quadruplexes can adopt a variety of topologies based on the orientation of the strands: parallel, antiparallel or hybrid structures are known (*see Figure 1.1 E-H*). The conformation of the glycosidic bond between nucleobase and sugar differs depending on the topology: it can adopt *syn*- or *anti*-conformation (*see Figure 1.1 B*). While in parallel G-quadruplexes the base orientation is always identical – leading to a highly symmetrical structure –, in G-quadruplexes with an antiparallel strand orientation the glycosidic bonds of two neighboring guanines in one tetrad differ (*examples in Figure 1.1*). In contrast to DNA quadruplexes, RNA-quadruplexes are exclusively found in the parallel conformation (21). It has been shown that quadruplex structures adopted by RNA sequences are more stable than their respective DNA counterparts (21-24). In both RNA and DNA quadruplexes, the adopted structure and the stability are also influenced by the loop length and the sequence composition of the total quadruplex motif (21,25): The shorter the loop, the more stable the G-quadruplex (26).

1.1.1.1 G-quadruplex stabilizing compounds

In recent years several G-quadruplex stabilizing small molecule compounds have been identified. In general, quadruplexes are stabilized by monovalent cations or small molecule compounds interacting with the G-tetrads and thereby neutralizing the negative, electrostatic repulsion of inwardly pointing guanine oxygens (27-30). Several chemical molecule classes possess quadruplex affinity, for example acridines (28,31), ethidium bromide derivatives (32), cationic porphyrins (33), perylenes (34), anthraquinones (35), steroids (36) or macrocyclic compounds (37). Most of these compounds include a planar aromatic system, which can interact with the terminal G-tetrads (end-stacking) or intercalate between two G-layers. Some compounds only interact with the grooves or loops of the G-quadruplex.

In this study, mainly quadruplex compounds belonging to the bisquinolinium family or to cationic porphyrin derivatives were used. 6,6'-disubstituted-2,2'-bipyridine (Phen-DC_3) and 2,9-disubstituted-1,10-phenanthroline (Phen-DC_6) (38) (*see Figure 1.2*) are examples for promising molecules that display strong quadruplex stabilizing abilities and a preference for binding quadruplex over duplex DNA. The 2,6-pyridine-dicarboxamide bisquinolinium derivative 360A (*see Figure 1.2*) was reported as being one of the most selective G-

quadruplex ligands able to actively induce the formation of G-quadruplexes (39,40). The internal hydrogen bonds and the electrostatic properties through the two quinolinium side arms result in perfect recognition of the quadruplex target by bisquinolinium derivatives (41). Another compound for quadruplex stabilization used in this study is TMPyP z, a derivative of 5,10,15,20-tetrakis(N-methyl-4-pyridyl)porphyrin (*see Figure 1.2*). TMPyP is a well studied planar cationic porphyrin that can bind different quadruplex structures and is also known to inhibit telomerase activity (42). Furthermore the compound N-methyl mesoporphyrin IX (NMM) was used, especially because it was reported to stabilize bacterial quadruplexes in *in vivo* experiments (43,44).The quadruplex interaction modes of these molecules are end-to-end stacking (45-47) and intercalation between two adjacent guanine quartets (47-49).

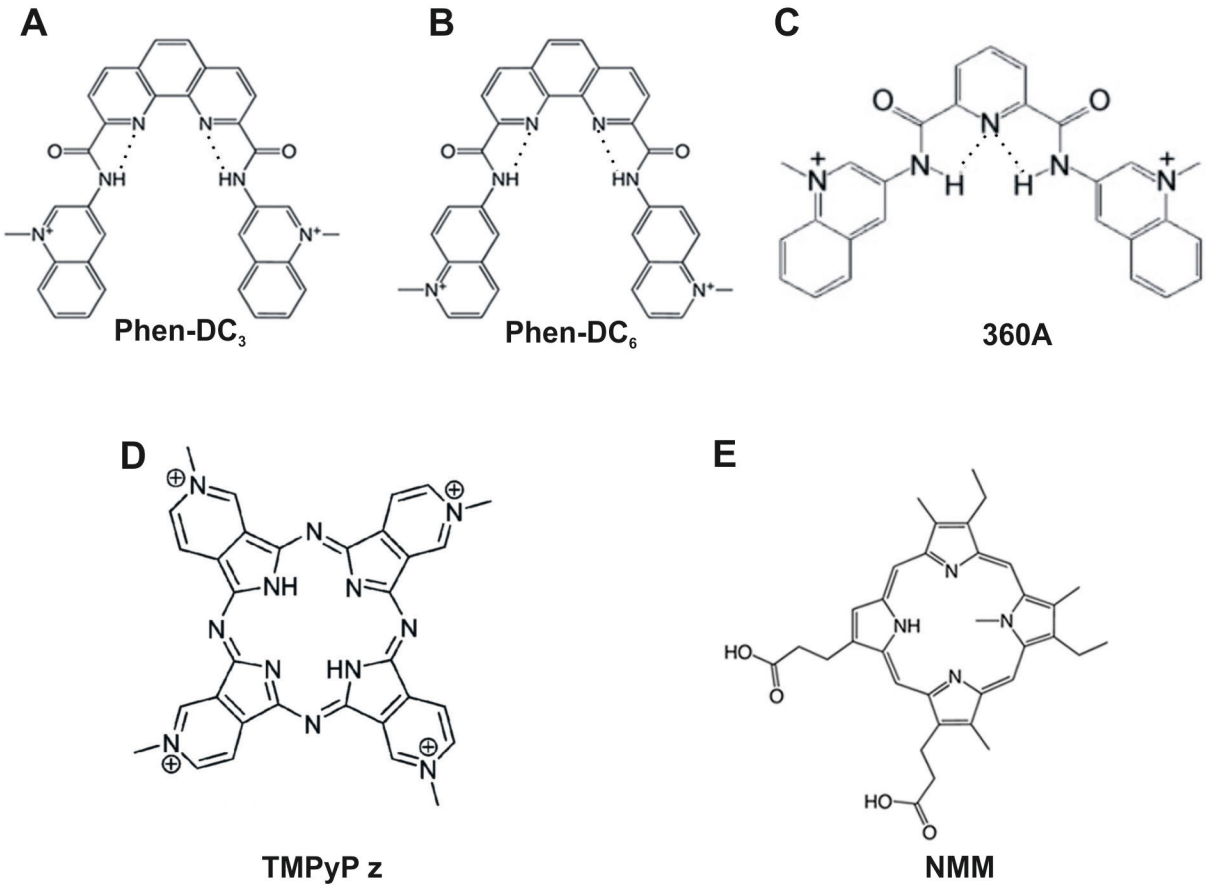


Figure 1.2: Quadruplex stabilizing compounds.

A-C Examples of quadruplex stabilizing compounds of the bisquinolinium family: Phen-DC_3, Phen-DC_6 (38) and 360A (50). Internal hydrogen bonds are represented by dotted black lines. **D, E** Examples of porphyrin compounds: TMPyP z (51) and NMM (38).

1.1.2 Nucleic acid triplex structures

Triple-helical nucleic acids were first described in 1957 (52). Triplex structures form between three nucleic acid strands. They occur in purine-rich DNA strands that form Hoogsteen hydrogen bonds. Two different triplex motifs have been described: 1. the purine motif and 2. the pyrimidine motif (10). Both require a purine rich Watson-Crick duplex binding the third strand in the major groove. In the purine (R) motif the third strand has an antiparallel orientation to the duplex purine strand and forms A(T)AT and GGC triplexes in reverse Hoogsteen configuration (*see Figure 1.3 A*). The pyrimidine (Y) motif contains TAT and CGC triplets in Hoogsteen configuration, thus having the third strand in parallel orientation (*see Figure 1.3. B*). The cytosine containing Y motif is stabilized under acidic conditions (cytosine in third strand is protonated C^+) (*see Figure 1.3. B*).

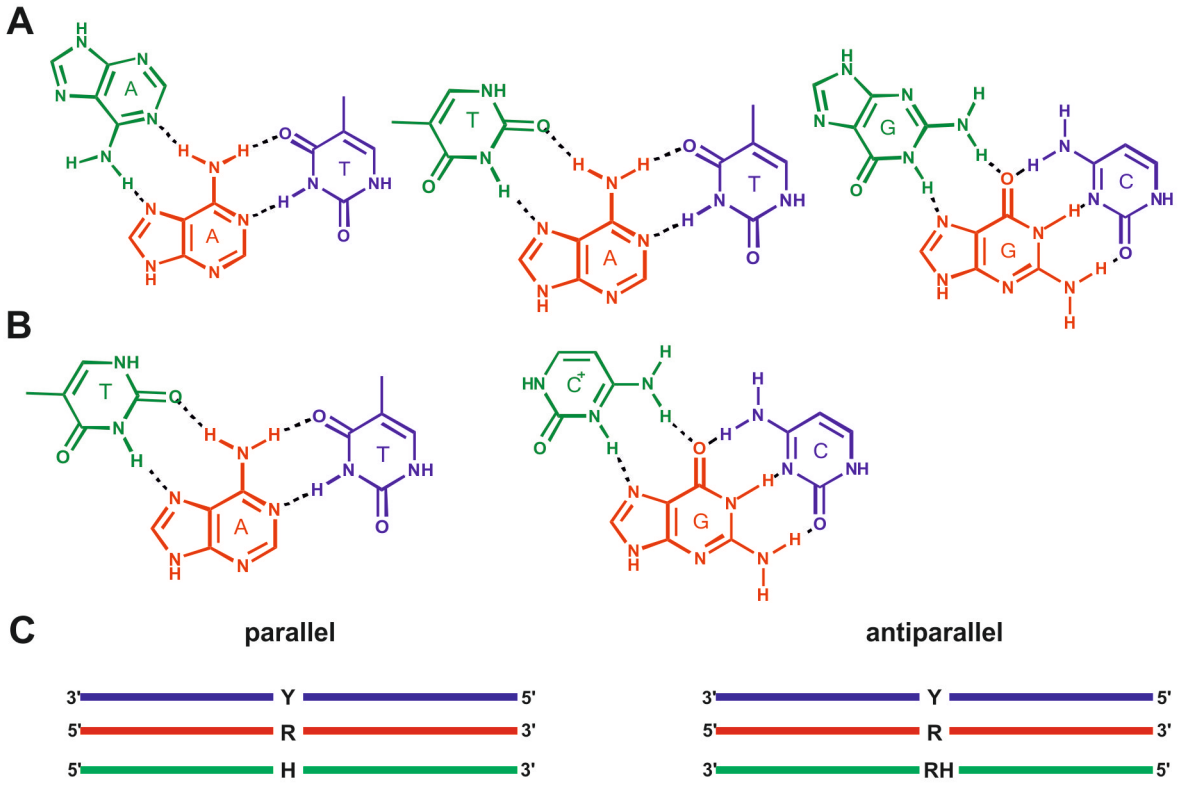


Figure 1.3: Purine and Pyrimidine type triplexes.

A Purine motif base triplets. Purine bases are colored red, pyrimidine bases are colored blue and reverse Hoogsteen bases are shown in green. **B** Pyrimidine motif base triplets. Purine bases are colored red, pyrimidine bases are colored blue and Hoogsteen pairing bases are shown in green. **C** Strand orientations for parallel and antiparallel triplex motifs. Color code as in **A**/**B**, Y: pyrimidine rich strand; R: purine rich strand; H: strand bound by Hoogsteen base pairing; RH: strand bound by reverse Hoogsteen base pairing. **A** and **B** modified from (10).

Steric properties make the triplex of the pyrimidine motif more stable compared to the purine motif triplex, especially for intermolecular formations (53). Triplex structures can be formed intra- and intermolecularly. Intermolecular structures are formed out of two or three distinct DNA strands – most often between a DNA duplex and a single stranded triplex-forming oligonucleotide (TFO) (54) (*see Figure 1.4. A*). In intramolecular triplexes the third strand is physically tethered to the DNA duplex. Most studies investigating intramolecular triplexes focus on H-DNA. For the formation of a H-DNA (*see Figure 1.4 B*) the homopurine-homopyrimidine sequence must be a mirror repeat (*see Chapter 1.3*). That way, half of the pyrimidine tract swivels its backbone towards the purine strand of the duplex or the purine strand binds to the purine part of the underlying duplex, forming a parallel or antiparallel H-DNA structure, respectively (55). Different H-DNA isoforms can occur, depending on whether the 3' half or the 5' half of the third strand is involved in triplex structure formation (*see Figure 1.4 B*).

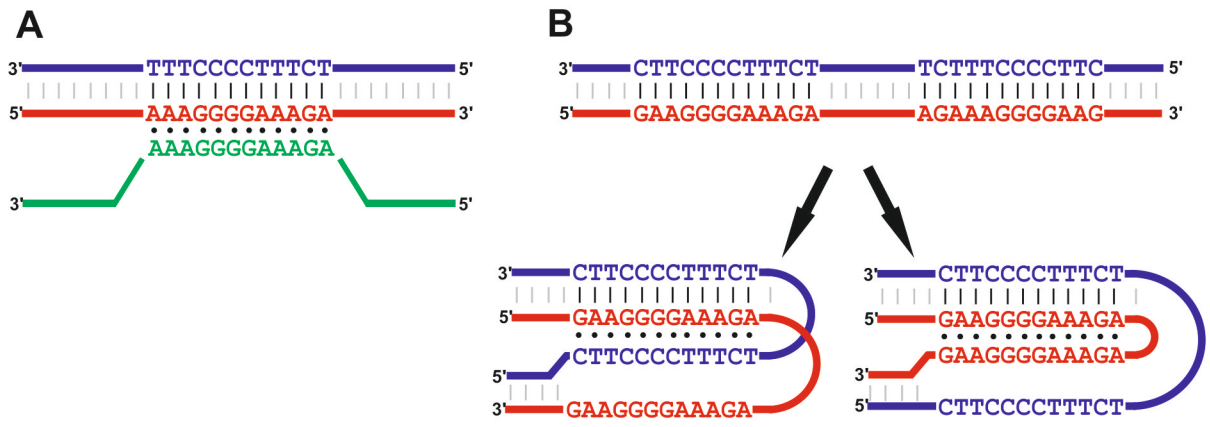


Figure 1.4: Schematics of inter- and intramolecular triplex structures.

A Schematics of an intermolecular purine motif triplex with antiparallel strand orientation formed by 3 distinct DNA strands. **B** Examples of intramolecular H-DNA structures that can form from a mirror repeat sequence within a DNA double strand. Pyrimidine motif H-DNA with parallel strand orientation is shown on the left side. Purine motif H-DNA with antiparallel orientation is shown on the right side. Pyrimidine-rich strands are shown in blue, purine-rich strands are depicted in red and the reverse Hoogsteen strand is colored green. Figure modified from (56).

Different groups have reported the existence of imperfect triplexes with mismatches between the strands (57-60) which lead to destabilizing effects. Such effects increase with the number of contiguous mismatches (61) and further depend on their position in the triplex: Mismatches in the center of a triplex are more disruptive than those at terminal sites (62). Furthermore, triplex stability is influenced by the presence of multivalent cations. They compensate the negative electrostatic repulsion of the three nucleic acid backbones and therefore stabilize triplex structures (63-65). G-rich triplex sequences can compete with

quadruplex formation, which is favored in the presence of potassium (66). Other factors that can influence triplex stability are pH, temperature, solvent and the presence of charge-neutralizing polyamines (67) or polypeptides (68). Apart from DNA, triplex structures also form in combination with RNA strands. RNA triplexes are found in different structured RNAs: They occur in pseudoknots (minor groove triplexes) (69,70), riboswitches (71) and other structured RNAs (72). Pyrimidine motif triplexes have been the main focus of most investigations, but other types may form as well (73,74).

1.1.2.1 Intrastrand triplexes

A different type of intramolecular triplex structure arises from the folding of polypurine/polypyrimidine units along one single strand of DNA or RNA. Although those intrastrand triplexes have been investigated *in vitro* (75-84), studies investigating their *in vivo* occurrence (10) and function (85) are sparse. Intrastrand triplex motifs have been assigned to four different conformational classes, depending on the strand orientation of their base triplets (10) (*see Figure 1.5*). Class I and II refer to purine motif triplexes, with class I having the reverse Hoogsteen domain at the 5' terminus, followed by the purine- and the pyrimidine-rich domain. Class II triplexes have the pyrimidine-rich domain at the 5' end, followed by the purine rich domain and the reverse Hoogsteen domain at the 3' end of the sequence.

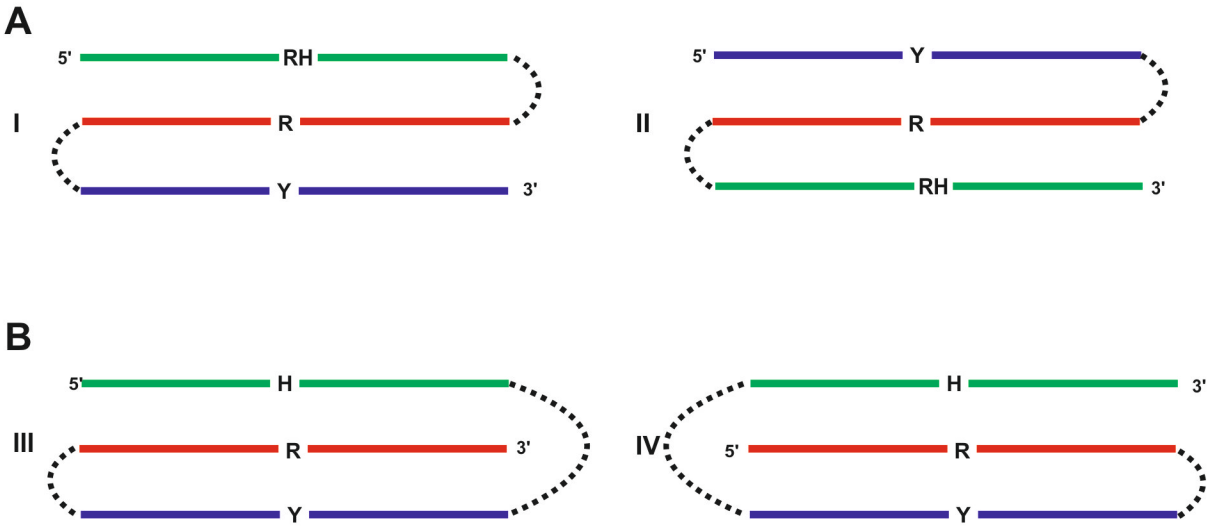


Figure 1.5: Intrastrand triplex classes.

A R motif triplexes: class I and class II. **B** Y motif triplexes: class III and class IV. Pyrimidine rich strands are shown in blue, purine rich strands are depicted in red and (reverse) Hoogsteen strands are colored green. Dashed lines represent arbitrary spacer sequences. Figure modified from (10).

Class III and IV correspond to the Y motif triplex structures: Class III triplexes progress from the 5' purine-rich domain through the pyrimidine-rich domain to the 3' Hoogsteen domain, and class IV triplexes start with the Hoogsteen domain at the 5' terminus, followed by the pyrimidine-rich domain and ending with the purine rich domain at the 3' terminus. It has been shown that both purine and pyrimidine type triplex DNA structures can form under physiological conditions. Pyrimidine-type intrastrand triplex structures even occur on the RNA level (11).

1.2 Functions of non-canonical nucleic acids

Evidence pointing towards the *in vivo* existence of non-canonical nucleic acid structures is increasing. So far, most of these structures have only been hypothesized to have functional roles *in vivo*. The increasing amount of genomic sequencing data available allowed for the detection of G-rich sequences in functional regions of many eukaryotic and prokaryotic genomes. G-rich sequences are prone to form quadruplex or triplex structures. As the functional properties of non-canonical nucleic acid sequences rely on their structure-forming ability, mechanisms of interference with biological functions are similar for different motifs.

When formed during transcription or translation, secondary nucleic acid structures might influence gene expression. Non-canonical nucleic acid structures formed during transcription (e.g. close to promoter sites) can have incremental or decremental effects on transcription efficiencies (*see Figure 1.6 A*): 1. The structure could facilitate the continuation of the RNA polymerase and the transcription machinery by stabilizing the single-stranded DNA conformation. 2. Transcription could be inhibited by blockade of the transcription machinery via the non-canonical structure (4). Similar mechanisms occur during translation (*see Figure 1.6 B*) where the secondary structure could form on RNA level. It might block ribosome binding or continuation when formed in proximity to the ribosome binding site (RBS). In some mRNAs the RBS is not accessible because of hairpin formation or other interactions; in those cases the formation of a stable non-B DNA structure (e.g. in front of the RBS) could counteract competing interactions and liberate the RBS for translation.

Sequences with the potential to form alternative nucleic acid structures can also affect translation when occurring in open reading frames (ORF) and formed on RNA level. Secondary structure formation in ORFs is known to play an important role in ribosomal pausing and frameshifting in viruses, eukaryotes and bacteria (86-89). The exact interaction mechanism of secondary structures with the ribosome is not known; however, a decrease in

the rate constants of both translocation and tRNA dissociation steps was postulated for pseudoknot structures (89). Ribosomal stalling can have the following effects: 1. The timing of the co-translational protein folding might be affected and could lead to an altered protein function (90). 2. Nonsense or non-stop mRNA could be forced to decay which protects the organism against the formation of truncated proteins. 3. Frameshifting can lead to the production of different proteins from one mRNA template.

Furthermore, secondary nucleic acid structures can interfere with replication (*see Figure 1.6 C*). During replication, the DNA double helix is separated and replication proceeds continuous on the leading strand and discontinuously on the lagging strand. Those transient single-stranded states (especially on the lagging strand) facilitate secondary structure formation. Replication fork progression can be slowed down or blocked when meeting obstacles like previously described non B-DNA structures. A disassembly of the replication fork may lead to double-strand breaks, polymerase stalling or replication slippage. Replication slippage proceeds as follows: 1. During replication the DNA polymerase pauses at the secondary structure. 2. The template and the nascent strand separate allowing for secondary structure formation in the single strand. 3. The nascent strand anneals back to the template, the polymerase reassembles and replication proceeds. However, during the process of reassembling the polymerase can backtrack at guanine repeat sites. Depending on whether the secondary structure is located in the template or the nascent strand, this can result in deletion or expansion of the G-rich sequence (91,92). Stalled replication can be reinitiated by primase, which creates a new primer that binds after the obstacle and leaves a gap in the DNA sequence (93). Thus, genomic instability can be induced by the formation of alternative DNA structures.

Non-B DNA structures themselves have been suggested to be identified by DNA repair proteins as they represent distortions of the DNA double helix (94,95). A consequence of DNA repair is the introduction of mutations or small deletions, leading to genomic instability. Non-B DNA structures formed during DNA repair could also alter the repair process and have been suggested to contribute to error-generating repair and genomic instability when analyzed in a plasmid system in mammalian cells (96).

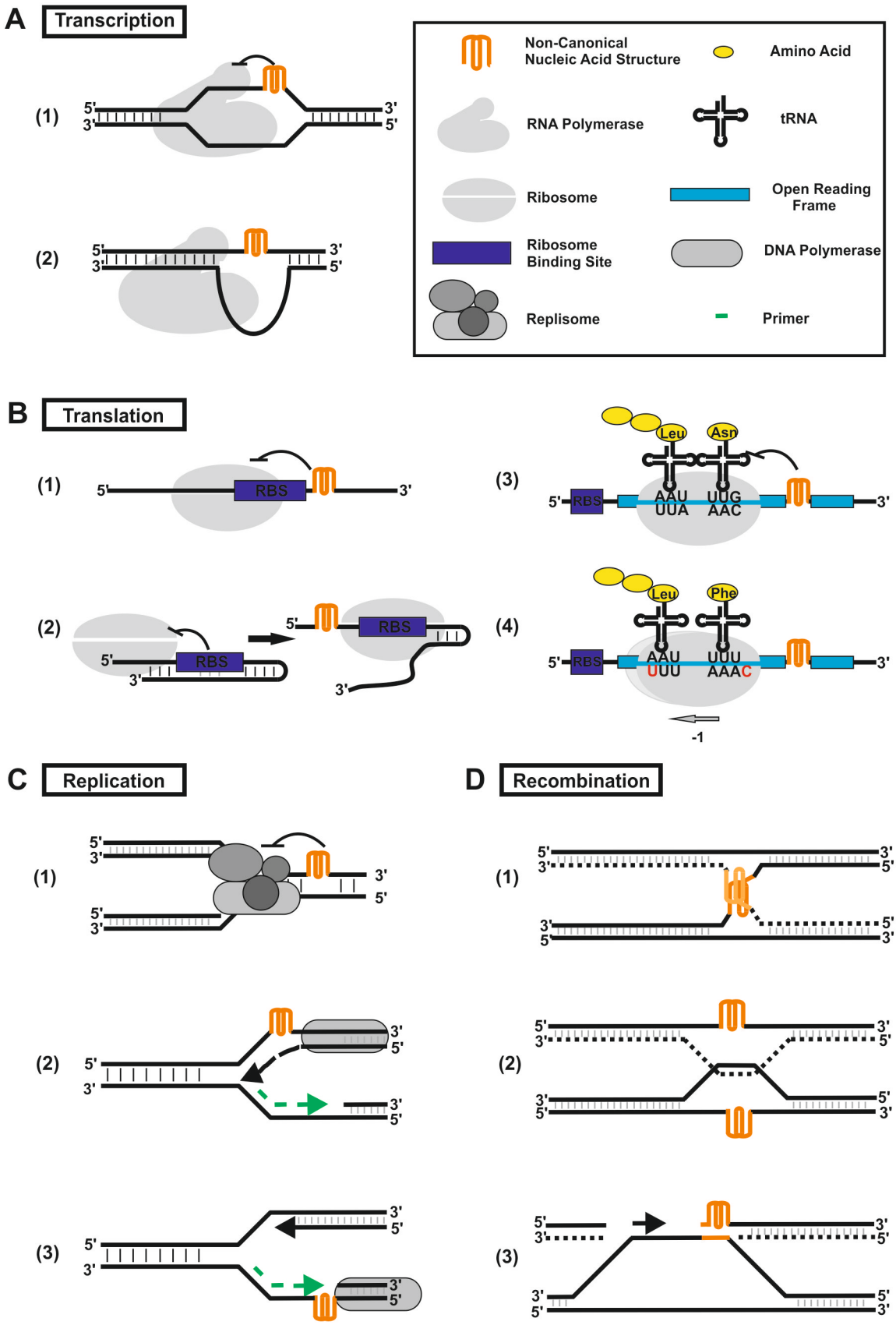


Figure 1.6: Potential *in vivo* functions of non-canonical nucleic acids.

A Interferences of non-B DNA with transcription: (1) Continuation of RNA polymerase is blocked due to the physical hindrance posed by secondary structure. (2) Binding of the RNA polymerase is facilitated, because non-canonical structure stabilizes the single-stranded conformation. **B** Interferences of non-canonical nucleic acids with translation: (1) Secondary structure formation adjacent to or within the RBS can block the binding of the ribosome and stall translation. (2) A blocked RBS can be liberated by the formation of a secondary structure, thus facilitating ribosome binding. (3) Alternative mRNA structure occurs in the ORF. Continuation of the ribosome is blocked, possibly leading to the production of a truncated protein or mRNA decay. (4) The downstream secondary mRNA structure causes the elongating ribosome to pause. Repositioning of the ribosome during opening of the secondary structure leads to a shifted reading frame and the production of a different protein (example for -1 frameshifting is shown, similar mechanism also possible in +1 direction). **C** Interferences of non-canonical DNA with replication: (1) Inhibition of replication by blockade of replication fork progression via secondary structure. (2) Polymerase stutters and reads irregularly over the non-canonical structure on the leading strand, possibly more than one time, thus creating sequence expansions. (3) Secondary structure on the lagging strand blocks polymerase. Replication is re-initiated at the next primer, resulting in a gap on the newly synthesized DNA strand. **D** Interferences of non-B-DNA structures with recombination: (1) If there is a sequence overlap between secondary motifs on different DNA double strands, an interaction could occur which initiates homologous recombination. (2) Secondary structure formation causes the complementary strand to be single-stranded and facilitates initiation of homologous recombination. (3) Alternative structure induces strand break or deletion. Illegitimate recombination occurs between short regions of homology (4-10 bp, which could be the G-rich sites).

Non-canonical DNA can also influence recombination (*see Figure 1.6 D*). Recombination events at non-homologous sites, such as illegitimate recombination, could be induced by DNA breakage or strand slippage near the secondary structure. Furthermore, the formation of the alternative motif would provide the complementary strand in a single-stranded state which could be used as a locus for homology searching and trigger recombination events. Alternatively, recombination could occur between two secondary structures forming at homologous regions. Naturally, protein binding can also be influenced by the formation of secondary nucleic acid structures.

Specific proteins that bind to the alternative structure could act as enhancers or repressors of transcription, translation, replication or recombination (97). On the other hand, the binding of certain proteins to the double stranded DNA could be blocked by the formation of the alternative structure. Generally, in double-stranded DNA the formation of non-canonical nucleic acids always competes with the annealing of the complementary strand. Single-stranded states occur transiently during replication, recombination, transcription, translation or can be caused by destabilization of the DNA via negative supercoiling. Interestingly, most of the stress-induced DNA supercoiling destabilization (SIDD) sites are found in regulatory regions (such as promoters) in eukaryotes and prokaryotes (98-100).

The following sub-chapters shall give an overview of studies investigating different *in vivo* functions of G-quadruplexes (*see Chapter 1.2.1*) and triplexes (*see Chapter 1.2.2*), respectively.

1.2.1 G-quadruplexes in vivo

Several computational studies have screened genomic sequences for potential quadruplex motifs with four runs of guanines, in which each G-tract is composed of at least 3 guanines (13-15). For the human genome more than 375,000 motifs have been found (101,102). The studies mentioned above showed them to be unevenly distributed in human (102,103), yeast (13,15) and bacterial genomes (14,104). Quadruplexes are over-represented in certain regulatory elements such as promoter-proximal regions (13,14,102,105), nuclease-hypersensitive sites (106), CpG islands, enhancers, insulators (103) and conserved elements like transcription factor binding sites (107,108). Furthermore, they occur within micro- and minisatellites (109,110) as well as in ribosomal (111) and telomeric DNA (15,112). Essential guanosines have been described as being more conserved than nucleotides (nt) which do not interfere with G-quadruplex formation (108). Strong evidence of the actual formation of G-quadruplex structures *in vivo* is increasing. Antibodies or fluorescent biomarkers for different G-quadruplex structures have been developed (113-115). In two successive studies, Balasubramanian and co-workers have described specific immunostaining of DNA (116,117) and RNA (118) quadruplexes. Their findings strongly suggest that G-quadruplexes have important functions in cellular and genetic processes (4). Furthermore, various studies investigating the regulatory roles of G-quadruplexes in eukaryotes have been published. One prominent example is the human telomeric G-quadruplex sequence (17). Telomere sequences are located at the ends of chromosomes. They contain a double-stranded region with a single-stranded 3'-overhang. The whole human telomeric DNA region consists of 5'-d(TTAGGG)-3' repeats and is typically 5-8 kb long with a 3´-end overhang of the G-rich strand which is up to 200 nt in length (119,120). After each somatic cell division the single-stranded overhang progressively decreases in length until the cell undergoes apoptosis (121,122). The enzyme telomerase, a reverse-transcriptase, can elongate the telomeric ends after cell division. It is usually inactivated in most somatic cells, but highly activated in 80-90% of cancer cells (123). Intramolecular and antiparallel quadruplex structures have been shown to block telomerase activity, whereas intermolecular and parallel structures did not (124-126). Therefore, targeting the human telomeric quadruplex (HTQ) is of great research interest for cancer therapeutics (127-129). An understanding of the folding of human telomeric G-quadruplexes *in vivo* and their formed structures under physiological conditions will be very beneficial for a rational-based drug design. In addition, the formation of telomeric quadruplexes has been implicated in processes related to aging (130,131) and genetic stability (132). In the absence of the specific quadruplex-resolving Pif1 helicase, a slowdown

of replication or an occurrence of double-strand breaks has been shown for G-quadruplex-comprising DNA in *Saccharomyces cerevisae* (133).

In addition to effects on replication and genetic stability, G-quadruplex structures have been shown to interfere with both transcription and translation. Potential quadruplex-forming sequences were identified in the promoter regions of many human proto-oncogenes, such as *C-MYK*(134), *C-KIT*(135), *KRAS* (136), *NRAS* (137) and *BCL-2* (138). The quadruplexes occurring in the promoter regions of *C-MYC* (139), *C-KIT*(140) and *KRAS* (136) have been proven to silence transcription when stabilized by small molecule compounds *in vivo*. Using a whole-transcriptome assay, Hartig and co-workers were able to show specific effects on genes containing G-quadruplexes in their promoter regions upon treatment of mammalian cells with quadruplex-selective bisquinolinium compounds (27). Recent studies investigated the influence of transcription-induced quadruplex formation in a double-stranded DNA template. G-quadruplex formation was observed in the upstream but not the downstream region of an *in vitro* transcribed sequence. G-quadruplexes can be induced thousands of base pairs away from a transcription start site (TSS), acting as silencer or enhancer of transcription (141,142). In an earlier study, G-quadruplexes have been shown to cause transcriptional arrest when located in front of the moving T7 RNA polymerase with the *C-MYB* quadruplex repeat (143). However, in addition to 5'-untranslated regions (UTRs) potential G-quadruplexes have also been identified in the 3'-UTRs near transcription termination, splicing and polyadenylation sites (144). In a recent study by Beaudoin et al., potential quadruplex-forming sequences were found enriched in the 3'-UTR of human mRNAs. Two quadruplex sequences were investigated in detail and identified as cis-regulatory elements which might increase the efficiency of alternative polyadenylation sites and could interfere with miRNA regulatory actions by mRNA shortening (145). Beyond transcriptional control, G-quadruplexes have also been described to interfere with translation. Different RNA G-quadruplex structures have been linked to the translational processing of human transcripts (146). One prominent example are the G-rich telomeric repeat RNAs (TERRA) which are suggested to be involved in chromatin remodeling and regulation of telomerase activity (147,148). Hartig and co-workers demonstrated by way of a luciferase reporter assay that artificial RNA G-quadruplex sequences inserted into the 5'-UTR can act as translational suppressors in mammalian cells (149,150). In 1996, Horsburgh et al. suggested G-rich sequences occurring in ORFs to be the reason for frameshifting in Herpes simplex virus thymidine kinase mRNA (151). In recent years, this topic was re-investigated by different groups: Yu et al. showed that G-quadruplexes are able to induce -1 and +1 ribosomal frameshifts in artificial constructs in eukaryotic cells (152); Sugimoto and co-workers reported increased -1 frameshift efficiency when eukaryotic cells were incubated

with a quadruplex-stabilizing compound (153); they further showed translational halt and a truncated protein product caused by a quadruplex found in the human estrogen receptor (154); also, the group of Balasubramanian observed translational inhibition in eukaryotic cells and suggested clusters of 13 G-quadruplexes within the EBNA1 mRNA to act as *cis*-regulatory elements in translation (155).

1.2.1.1 G-quadruplexes in prokaryotes

Most of the studies described in *Chapter 1.2.1* cover the influence of quadruplexes on gene regulation in eukaryotic cells. Clues to potential functions of quadruplex-forming sequences in bacteria are rare; nonetheless, some very specific roles of quadruplex formation have been described. In the pathogenic bacterium *Neisseria gonorrhoeae* pilin antigenic variation is necessary to evade the human immune system. Seifert and co-workers demonstrated pilin antigenic variation to be regulated by a quadruplex-based system. The quadruplex forming upstream of the *pilE* gene locus induces DNA nicks which are further processed by the recombination machinery. Non-homologous recombination takes place between the *pilE* locus and many silent *pilS* donor loci, thereby enabling antigenic variation. Mutation of the potential quadruplex-forming sequence inhibited recombinational switching (44,156). An involvement of quadruplex structures in antigenic variation was also suggested for the recombinational switching of the Lyme disease pathogen *Borrelia burgdorferi* (157). Furthermore, G-quadruplex motifs were found in the antigenic variation loci of *Treponema pallidum* (158). One of those motifs was recently characterized by Rehm et al. and found to be significantly enriched in bacteria (20).

Apart from intergenetically occurring quadruplexes, four-stranded motifs also occur in open reading frames (ORFs). In a series of publications, the group of Sugimoto described the influence of G-quadruplexes occurring in protein-coding sequences. They identified different stable quadruplexes occurring in *E. coli* ORFs and showed their ability to suppress translation elongation by *in vitro* translation studies. Subsequently, *in vivo* experiments carried out in mammalian cells also demonstrated an influence of quadruplex sequences on translation (153,154,159,160). It seems that quadruplex formation increases the potential for ribosomal stalling and frameshifting (152,155). Although not within a bacterial ORF but immediately preceding it, in an artificial setup Hartig and co-workers showed that translation is strongly influenced by masking the ribosome binding site of *E. coli* utilizing a G-quadruplex motif. Insertion of G-quadruplex sequences resulted in down-regulation of gene expression, and the extent of repression correlated with G-quadruplex stability (161). Furthermore, in a

computational search Chowdhury and co-workers found G-quadruplex motifs enriched in bacterial promoters across more than 140 bacterial species (14,104). In a follow-up study they found them to be enriched in certain organism-specific gene classes. Regarding the radioresistant *Deinococcus radiodurans*, putative G-quadruplex forming sequences were found specifically in correlation to radioresistance genes. Quadruplex stabilization via addition of small molecule compounds led to attenuation of radioresistance *in vivo* (43).

1.2.2 Nucleic acid triplexes in vivo

Based on *in vitro* experiments triplex structures have been suggested to play a role in a range of cellular functions, such as transcriptional or translational regulation, interferences with recombination and replication (*see Chapter 1.2*), post-transcriptional RNA processing and DNA repair. The main focuses of studies investigating triplex structures *in vivo* were H-DNA and TFOs. Different studies have identified triplex motifs in eukaryotes and prokaryotes by means of computation. Most algorithms search for TFO binding sites (162-164), potential triplex target sites (165), or focus on inverted repeats (166,167) and H-DNA (168,169). Evidence for the *in vivo* existence of triplex DNA structures is increasing – immunodetection by triple-helix specific antibodies has been reported (170-172). Those antibodies are able to detect DNA-DNA/DNA and DNA-DNA/RNA (/ indicates Hoogsteen bond) triplex structures (172-175). In addition, different proteins which specifically recognize triplex structures in cells have been identified in human (176), Drosophila (177), yeast (178) and other mammalian cells (179-181). Among those are RecQ helicases (182-184) that actively unwind triplexes in 3'→5' direction, but also heterogeneous ribonucleoproteins (176), intermediate filament proteins (181), high mobility group proteins (182,185,186) and proteins involved in DNA repair (187-189).

Intermolecular triplexes have been used for the artificial regulation of gene expression and may be suitable for therapeutic use (56,190). There are different examples for transcription being influenced by TFO-directed triplex formation *in vivo*. A mechanism where the triplex formation in the 5' untranslated region (UTR) shields DNA from duplex targeting proteins such as transcription factors was shown for the *ets2* gene in prostate cancer cells (191). In that study, TFOs were designed to overlap the binding site of the transcription factor Sp1, thus triplex formation inhibited transcription. The same principle of transcriptional inhibition was shown for the *BCR/ABL* locus in human cells (192). In biomedical applications, intermolecular triplexes have been reported to block protein-DNA interactions (193) and

influence site-directed recombination (194). The possibility of site-specific delivery of target agents via the formation of intermolecular triplexes between the DNA duplex and the TFO has been exploited (195). Using this concept with peptide nucleic acid (PNA) molecules as TFOs different studies showed the introduction of hereditary gene modifications (196,197) and the improvement of the delivery of peptides into the cell nucleus (198). Nucleotide excision repair (NER) factors are able to recognize intermolecular triplex structures (199) and support triplex-induced mutagenesis and recombination events in cells (200,201). Several analysis tools exist for the computational search for TFO binding sites in genomic loci (162,163,168). Putative triplex target sites are over-represented in both prokaryotic and eukaryotic genomes (202,203).

H-DNA is known to induce genetic instability, to have influence on DNA replication and repair and to be involved in transcription (12). Computational studies revealed that natural sequences with the potential to adopt an H-DNA structure are very abundant in mammalian cells (166). Mirror repeats capable of forming H-DNA structures have been found in promoters and coding regions of many genes involved in diseases, such as Friedreich's ataxia, autosomal dominant polycystic kidney disease (ADPK), fragile X syndrome, spinocerebellar ataxia and muscular dystrophy (204). One well-studied example is Friedreich's ataxia: Here, H-DNA structures can be induced by expansion of GAA repeats and lead to stalling of the RNA polymerase, thereby silencing the transcription of the frataxin gene (205). The ADPK disease is associated with mutations in the *TSC2* and *PKD* genes. The proposed mutagenic mechanism involved double strand breaks leading to a replication fork blockade inducing gene conversion by recombination. Interestingly, these genes contain long polypurine/polypyrimidine repeats which are able to form H-DNA and have been shown to be hot-spots for recombination in this region (206). The implication of H-DNA in transcriptional regulation was also studied for the *C-MYC* oncogene. The H-DNA forming sequence of the C-*MYC* promoter serves as a *cis*-acting element downregulating transcription in mammalian cells (207,208). Different studies investigating the role of H-DNA in the regulation of eukaryotic transcription demonstrated either up- (209) or downregulation (210,211) without clarifying specific mechanisms. Genetic instability induced by double-strand breaks adjacent to H-DNA sequences was demonstrated for the *C-MYC* triplex sequence, but also for model H-DNA sequences in mammalian cells (212). Such double-strand breaks could be induced by replication stalling. *In vivo* studies revealed that distinct R-type triplex DNA structures can lead to polymerase arrest during elongation of replication, as proposed for the ADPK disease (213,214). Like intermolecular triplexes, intramolecular structures are also able to induce recombination and repair (94). Furthermore, H-DNA sequences have been mapped at recombination hot-spots in mouse myeloma cells (215).

H-DNA forming sequences inserted into shuttle vectors stimulated recombination events between plasmids in mammalian cells (216). Processes demonstrating recombination between two triplex structures forming at homologous sites have been proposed for H-DNA structures as well (217-219).

Triplex structures in RNA are known to contribute to folding and tertiary structure stability (72), some of them even provide enzymatic or catalytic activity (220,221). Furthermore, they have been reported to cause ribosomal frameshifting during translation. A prominent example is mRNA of the HIV virus which forms an intramolecular triplex inducing -1 ribosomal frameshifting (222,223). Additionally, triplex structures can play a role in chromosomal organization and epigenetics (56). H-DNA formation could provide contact points that interact with non-coding RNAs or cell matrix-associated proteins (224). In addition, the chromatin condensation is influenced by triplex-helices. As triplex structures are less flexible, the nucleosome reconstitution could be affected (225). Schmitz et al. described a TFO-directed triplex which regulates the methylation status of DNA by mediating the recruitment of methyltransferases to promoters (226). DNA methylation plays an important role in epigenetics and is known to influence gene expression and cell differentiation (227,228).

1.2.2.1 Triplexes in prokaryotes

As is the case for G-quadruplexes most of the studies investigating triplexes *in vivo* were performed in eukaryotes. Information about prokaryotic triplex structures and their functions is rare. Indeed, only few sequences with the potential to form triplex structures were found in prokaryotic species (166). However, long (≥12 nt) oligopurine/oligopyrimidine tracts have been discovered in bacterial genomes near regulatory regions (229), suggesting a functional role. Some studies investigating eukaryotic triplex structures were performed in bacterial cells using plasmids, as they are more convenient model systems. Chemical probing of intracellular DNA showed the formation of H-DNA during transcription of long GC stretches upstream of a promoter in an *E. coli* plasmid system (230). Triplex formation via addition of TFOs was demonstrated to inhibit transcriptional initiation by the *E. coli* RNA polymerase *in vitro* (231,232). However, subsequent *in vivo* studies were not performed in prokaryotic cells. The 2.5 kbp long polypyrimidine sequence associated with the ADPK disease, which was found in the human *PKD* gene (*see 1.2.2*), has also been investigated in *E. coli* plasmids. It has been shown to induce double-strand breaks at the H-DNA forming regions which resulted in large scale deletions (233). Furthermore, this sequence activated an SOS

response and NER in *E. coli* (234). Two independent studies demonstrated the dimerization of plasmids containing potential triplex forming sequences in *E. coli*, suggesting a role as interaction point in recombination (235,236). In 1992, Kato et al. showed that triplex DNA inserted in the promoter region of a reporter plasmid expressing β-lactamase resulted in increased *lacZ* gene expression compared to a control plasmid. They suggested that the triplex structure kept the template in a superhelicity state favorable for gene expression (237). However, when an H-DNA sequence was inserted between the promoter and the coding sequence or directly in the coding region, a downregulation of bacterial gene expression was observed, possibly related to transcriptional regulation (238-240). Although these sequences do not originate from bacteria their influence on bacterial systems implies that triplex structures play a role in bacterial gene regulation, genetic stability and repair mechanisms.

In two subsequent studies, Maher and co-workers investigated so-called PIT (potential intrastrand triplex) elements naturally occurring in *E. coli*, *Synechocystis sp. and H. influenza* (10,85). They characterized the PIT motif in *E. coli* and proposed a triplex structure of the corresponding oligonucleotide. In a follow-up study (85) they wanted to elucidate the function of PIT elements. Although they showed that, depending on the processability of the polymerase, PIT elements are able to block DNA polymerase elongation *in vitro*, they found no effect in *in vivo* studies. The PIT elements showed no promoter and terminator activity, had no effect on RNA polymerase and reverse transcriptase and did not interfere with conjugation.

In a different study, a bacterial protein interacting with triplex DNA was described. The protein TnsC, regulating the transposition of transposon Tn7 was observed to detect triplex DNA. Triplex recognition then leads to specific insertion of the transposon adjacent to intra- or intermolecular pyrimidine triplex motifs (241).

1.3 Nucleic acid repeats forming alternative structures

Nucleic acid repeats are certain DNA motifs containing sequence elements which are repeated in several units. The similar units can either occur consecutively (in the same strand direction, e.g. 5'→3') or separated by different numbers of nucleotides (interspersed repeats). Furthermore, they can be located in opposite direction to each other ("mirror" repeats). Nucleic acid repeat sequences appear to be related to the formation of non-canonical structures in genomic DNA (*examples are shown in Figure 1.7*).

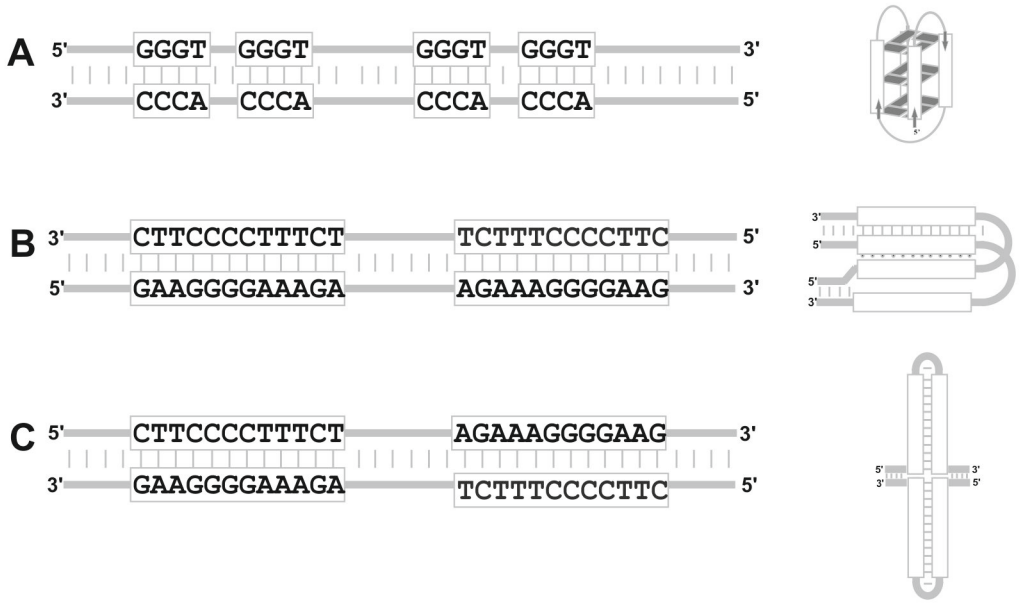


Figure 1.7: Examples of DNA repeats.

Schematics of arbitrary repeat sequences on DNA level and their corresponding secondary structures are shown. Repetitive units are framed in white. **A** Interspersed SSR (GGGT) able to form a G-quadruplex structure. **B** Mirror repeat (CTTCCCCTTTCT-NN-TCTTTCCCCTTC; N represents any nucleotide) which could form H-DNA. **C** Palindromic repeat (CTTCCCCTTTCT-NN-AGAAAGGGGAAG; N represents any nucleotide) which could form stem loop or cruciform structures.

Prokaryotic repeats have been classified according to different criteria like their total size, genomic distribution, coding capability as well as their number of occurrence in the genome. Examples for different categories are simple sequence repeats (SSR), tandem repeats (TR), miniature inverted repeats (MITE), repetitive extragenic palindromic (REP) sequences and clustered regularly interspaced short palindromic repeats (CRISPRs). The 20-48 bp long CRISPR (242) repeats have been shown to play a role in the adaptive immune response of bacteria. REPs (243,244) are palindromic, 20-40 bp long DNA repeats which can occur as single units or in clusters, so-called bacterial interspersed mosaic elements (BIMEs). MITEs are generally less than 200 bp in length and require a transposase for transposition. They can fold into long stem-loop structures on RNA level and frequently carry functional motifs, such as promoter sequences or protein binding sites (245,246). TRs contain multiple units, which are directly repeated in a head-to-tail manner and span from 1-100 base pairs (247,248) (units with a size of 1-9, 10-100 and >100 bp are termed micro-, mini- and macrosatellites, respectively). They are found in a variety of prokaryotic species (249,250) and can show considerable differences even among closely related species (251), suggesting TR to be subject to evolutionary changes (252). Kashi and co-workers investigated tandem iterations in *E. coli* and found them to be under-represented in open reading frames (ORFs) when exceeding a length of 3 bp (253). Microsatellites with a length

of 1-6 bp – also termed SSRs (254) – participate in bacterial adaption (255,256): high mutation rates at repeat sites can lead to an expansion or contraction of the SSRs which is related to bacterial phase variation. Phase variation describes a specific ON- or OFF- switch of the gene expression of a given factor involved in the interaction with the host, such as the invasiveness or the adherence to host cells (257-259). Most repeats occur in intergenic regions up to 200 bp upstream of the start codon, containing proximal regulators of gene expression.

Nucleic acid repeats can have strong effects on the local DNA structure in the genome. They are prone to fold into hairpins or more complex structures. Sequences with the potential to adopt such non-canonical nucleic acid structures are abundant in eukaryotic and prokaryotic genomes. Recently, Huang and Mrázek presented a survey of local sequence patterns that promote non-canonical DNA conformations from 1,424 prokaryotic chromosomes (260): They found that SSR are suppressed, whereas longer TR showed at least a slight over-representation in whole genome analyses across all phyla. Repeat sequences with the potential to form G-quadruplexes and H-DNA structures were found to be normally represented in most prokaryotic genomes with their analysis.

Both repeats and non-B-DNA structures have been associated with genomic instability. Inverted repeats were found to cause deletions in *E. coli* as early as the 1980s (261-263). Instability caused by TR sequences has been attributed to different hereditary diseases (264). Chromosomal plasticity in *Pseudomonas fluorescens* species has been linked to MITE sequences (265). REP sequences have been linked to genetic instability in *E. coli* toxin-antitoxin systems (266), and other repetitive sequences have been described in relation to genomic plasticity in bacteria (267,268). Most repeat sequences have the potential to fold into secondary structures on DNA and/or RNA level, as it has been described for pneumococcal bacteria (269). Also, those non-canonical nucleic acid structures are prone to interfere with translation, transcription, replication or recombination (*see Chapter 1.2*). The exact mechanisms of those influences, however, have not been elucidated to date. The function and role of many repetitive elements occurring in eukaryotes and prokaryotes is still unclear.

2 AIM OF THIS THESIS

Non-canonical nucleic acids have been investigated in detail for decades (*see Chapter 1*). Quadruplex *(see Chapter 1.1.1)* and triplex structures *(see Chapter 1.1.2)* occur in G-rich sequence strains, and the evidence about their *in vivo* existence is increasing. Several studies suggested them to influence regulatory and life cycle states in cells, and many of these structures have been associated with human diseases *(see Chapter 1.2).* Although computational searches provided vast evidence for the occurrence of potential alternative structure motifs across all kingdoms of life, the concrete mechanisms of their influences and functions are unclear. Studies carried out in prokaryotic systems are particularly rare.

In this thesis, two topics – both dealing with G-rich alternative structures in prokaryotes – were covered: 1. Positional effects of G-quadruplexes on *E. coli* gene expression and 2. Investigation of DNA triplex repeats naturally occurring in *E. coli.*

The aim of the first topic was to gain new insights into the secondary structure-mediated regulation of gene expression in *E. coli*. For this purpose, a series of reporter gene constructs containing systematically varied positions of G-quadruplexes were generated. Those sequences were then inserted at several positions within the promoter, 5´-UTR, and 3´-UTR regions. In an engineered system, G-rich sequences in the vicinity of the ribosome binding site were analyzed for gene activating behavior. A possible activation mechanism has been proposed, which makes those designs suitable for the application in addressable systems. Furthermore, potential quadruplex forming sequences occurring naturally in the *E. coli* genome were investigated for their influence on gene expression. In addition, first studies investigating G-quadruplex sequences occurring in the ORF of the *kdpD* and *kefC* genes of *E. coli* and *Salmonella* subspecies were undertaken.

The aim of the second topic was to investigate a particular type of intrastrand triplex which has been described in earlier studies but whose function and exact structure never could be clarified. This motif was characterized by *in silico* and biochemical (CD, NMR, *in vivo* probing) studies. Furthermore, the genomic stability around this motif was investigated, and different mechanisms for its involvement in recombination or replication were proposed. We also investigated whether this motif is involved in the organization of the bacterial nucleoid. This thesis also describes the collaborative design of a database allowing the search for intrastrand triplex motifs in 5,246 genomes of bacterial and archeal species. This way, intrastrand triplex motifs were found to be widely distributed in bacteria.

3 RESULTS AND DISCUSSION

3.1 Positional effects of G-quadruplexes on *E. coli* gene expression

Several studies have proven the influence of G-quadruplexes on eukaryotic gene expression. However, studies investigating the impact of four-stranded structures in prokaryotic genetic systems are rare. This chapter describes a comprehensive survey showing that both the strand orientation and the exact position of a G-quadruplex sequence strongly influence the secondary structure mediated effect on transcription and translation. G-quadruplex structures occurring in UTRs *(see Chapter 3.1.1)* as well as those occurring in ORFs *(see Chapter 3.1.2)* are investigated in artificial setups, but also with natural examples from bacteria.

3.1.1 G-quadruplexes in untranslated regions

As described above *(see Chapter 1.2.1),* sequences with the potential to adopt G-quadruplex structures have been found to be over-represented in certain regulatory regions, especially near promoter sites (43,104,109) in different organisms. The following subsections will systematically describe the influence of G-quadruplex sequences – of different stabilities and at different positions in the UTR – on gene expression in a reporter system in *E. coli* K-12.

3.1.1.1 *In vitro* characterization of the G-rich sequences used

In the following gene-expression studies potential G-quadruplex sequences of different stabilities and their respective non-quadruplex controls were used. To characterize their folding properties, we performed circular dichroism (CD) spectroscopy using synthetic DNA oligonucleotides *(sequences listed in Figure 3.1 A and Table 7.1)*. The spectra were measured in 10 mM Tris-HCl containing 100 mM KCl *(see Figure 3.1 B)*. K^+ has been reported to be the major cation in prokaryotic cells, cytosolic concentrations of 100 – 200 mM were determined as the physiological range for *E. coli* (270,271). The G_3T, G_3A and G_2T sequences fold into parallel G-quadruplexes, showing the typical maximum signal at 265 nm and a minimum at 240 nm (272). The G_2CT sequence folds into an antiparallel G-quadruplex structure, with a maximum peak around 290 nm and a minimum around 265 nm (272).

A

name	sequence	Tm [°C]
G_3T	GGGTGGGTGGGTGGG	> 95
G_3A	GGGAGGGAGGGAGGG	> 95
G_2T	GGTGGTGGTGG	54.4
G_2CT	GGCTGGCTGGCTGG	51.4
ctrl 1	GGGTGTGTGTGTGTG	< 20
ctrl 2	TGTGGCGTGGGCGGG	52.6
ctrl 3	CACTCACTCACTCCC	< 20

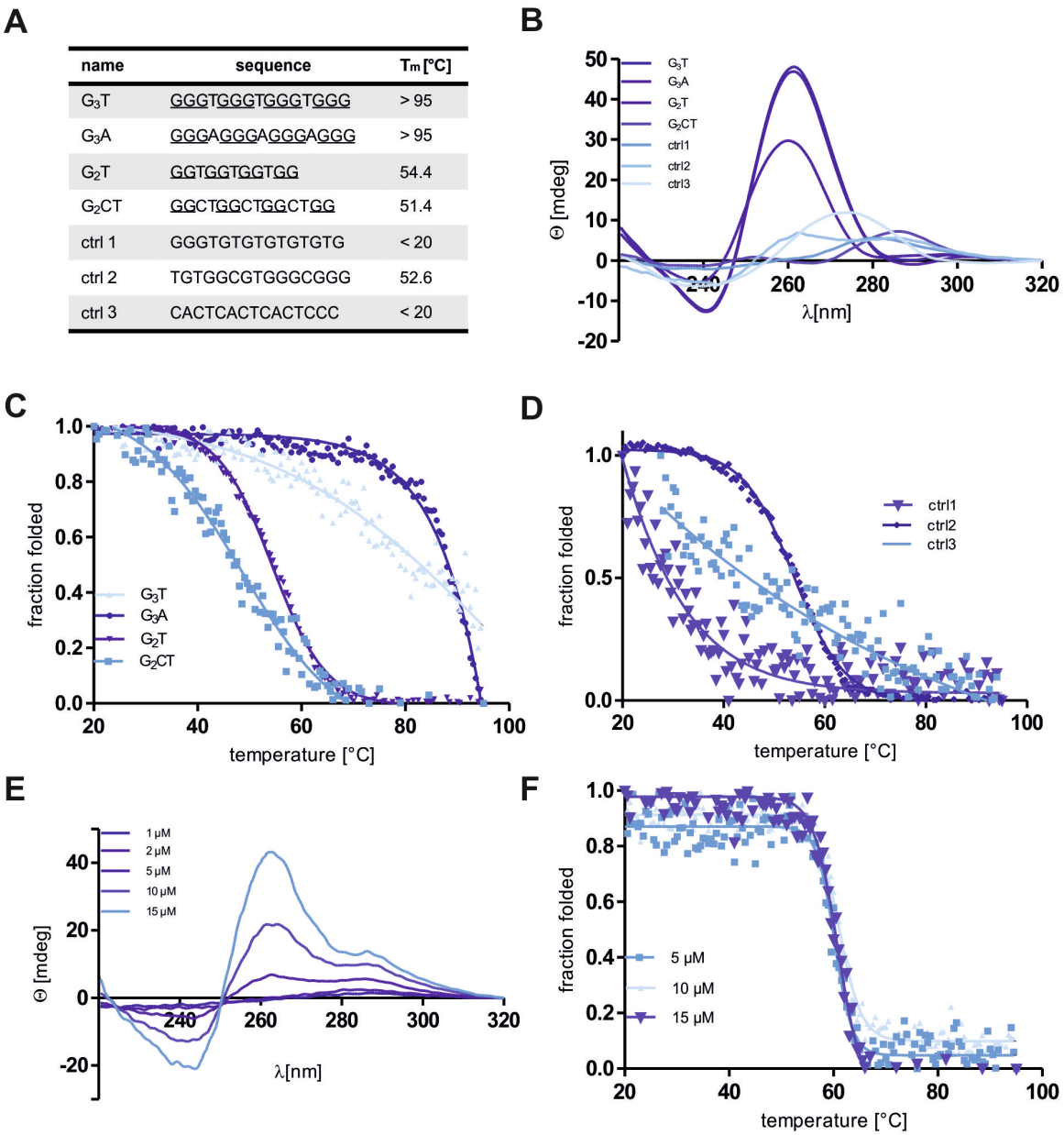


Figure 3.1: *In vitro* characterization of G-quadruplex sequences and controls.

A Name and sequence of quadruplex constructs. Guanines involved in G-quadruplex formation are underlined. The melting temperature of the different oligonucleotides (T_m) is indicated. **B** CD spectra of 5 µM DNA in the presence of 100 mM KCl. **C, D** Melting profiles of sequences shown in **A** at the respective CD maximum (G_3T, G_3A and G_2T: 260 nm; G_2CT: 290 nm; ctrl1:280 nm, ctrl2: 265 nm; ctrl3: 270nm). **C** Quadruplex bearing constructs. **D** Sequences used as controls. Control 2 is able to form a stable secondary structure. **E** CD spectra of control 2 at different oligonucleotide concentrations. **D** Thermal denaturation curves of control 2 at different oligonucleotide concentrations. Melting temperatures and sequences are given in *Figure 3.1 A*. Ctrl is representative for control. © Cell Press.

CD signatures decrease from quadruplexes with three G-tetrads to those comprising two tetrads. The sequence G_2CT shows weak modulation of ellipticity. In order to determine the stabilities of the G-quadruplexes, thermal denaturation was measured at 265 nm and 290 nm

(melting curves and temperatures are shown in Figure 3.1). We determined melting temperatures T_m of 54.37°C and 51.35°C for the G_2T and G_2CT constructs. The T_m of G_3T and G_3A could not be accurately determined as they were very thermostable and only started to denature at temperatures higher than 80°C. Structures formed by controls (ctrl) 1 and 3 melted immediately and had melting temperatures lower than 20°C, indicating that they are not able to form stable structures. However, control 2 was able to form a secondary structure with a maximum occurring at 270 nm and a minimum at 240 nm. This structure showed a surprisingly high T_m of 52.62°C. We were interested in the structural properties of this control sequence and studied CD and thermal denaturation at different oligonucleotide concentrations (5 µM, 10 µM and 15 µM). We observed a CD spectrum suggesting a parallel G-quadruplex fold for higher oligonucleotide concentrations *(see Figure 3.1 E)*. The melting temperature remained constant, indicating the formation of an intramolecular structure *(see Figure 3.1 F)*. G-quadruplex formation with bulged-out nucleotides has been described by Phan and co-workers (273), and such a structure might form here as well.

3.1.1.2 General concept and first constructs

First of all, we designed a series of sequences containing G-rich elements around the promoter and untranslated regions of a reporter gene coding for the enhanced green fluorescent protein (eGFP), schematically shown in *Figure 3.2 A.* In a first set of constructs the G_3T sequence – which is reported and also shown to form a very stable G-quadruplex structure (274) *(see Chapter 3.1.1.1)* – was placed either in the core promoter (between the conserved -10 and -35 promoter regions) or immediately at the 3'-end of the conserved -10 region. For each set G-tracts were placed once in the sense and in the antisense strand. The total numbers of nucleotides between the conserved regions, as well as the conserved sequences themselves were not changed compared to the original ("wildtype") promoter sequence *(see Figure 3.2 B&C)*. The influence of the G-quadruplex forming sequence on the eGFP expression was investigated in two different plasmid systems in *E. coli* K-12: 1. The pQE-J06-eGFP system with the G-rich sequences inserted around the constitutive J06 promoter and 2. the pBAD-eGFP plasmid (based on pBAD-18 (275) with eGFP reporter gene) with the G-quadruplexes inserted around the arabinose-inducible araBAD promoter region. *Figure 3.2 B and C* show the exact sequences which were replaced in contrast to the wildtype promoter of the pQE and pBAD vector system, respectively. Whereas gene expression with the J06 promoter proceeds continuously, gene expression from plasmids under the control of the araBAD promoter can be regulated by the concentration of arabinose in the growth medium. The pBAD promoter originates from the *E. coli* arabinose operon. It is

regulated by the AraC protein (276): In the absence of arabinose interaction of AraC with regions (*araI* and *araO*$_2$) adjacent to the core promoter leads to the formation of a DNA loop which prevents binding of the RNA polymerase to the promoter and results in low transcription levels (277). In the presence of arabinose transcription from the pBAD promoter is turned on due to loop opening and interaction of the AraC protein with arabinose and only one region close to the promoter (*araI*) (276). Two different vector systems were chosen for analysis in order to better be able to differentiate plasmid- or promoter-specific effects, which are not necessarily triggered by secondary structure formation.

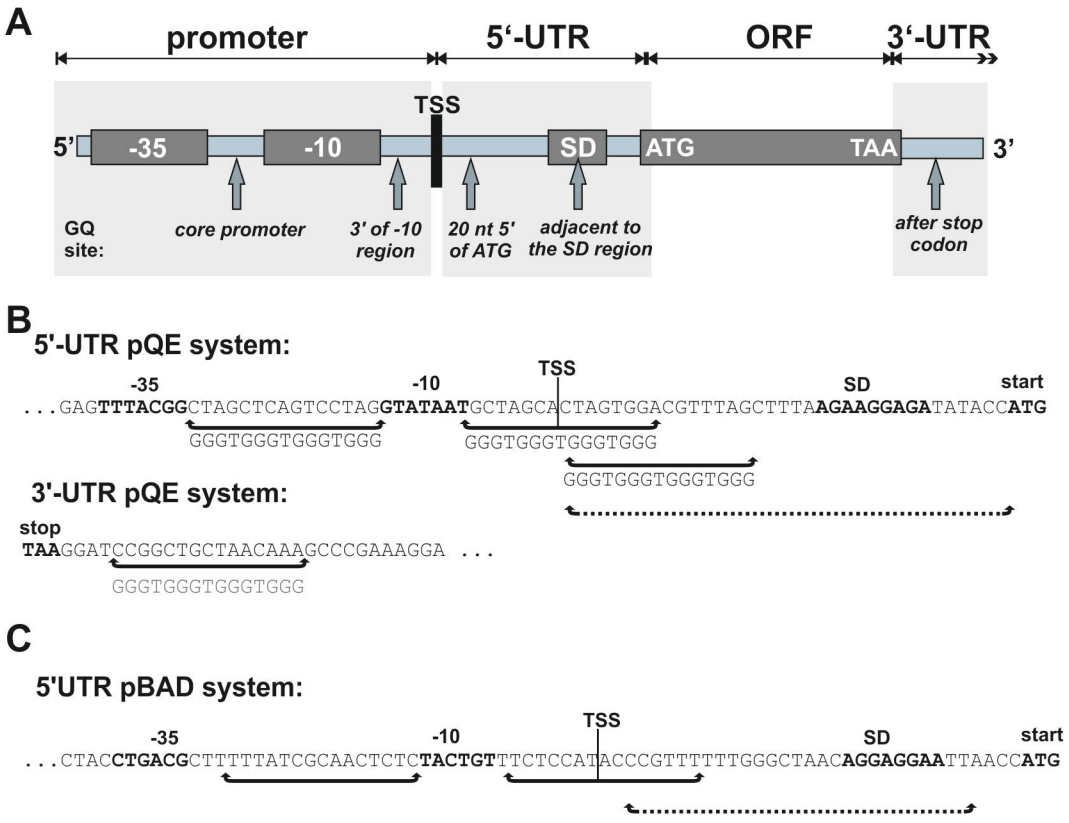


Figure 3.2: G-quadruplex insertion sites.

A Schematic representation of G-quadruplex insertion sites investigated in this study. -35 and -10 represent the conserved promoter regions; TSS indicates the transcription start site and SD stands for the Shine-Dalgarno region. ATG and TAA represent either the start or the stop codon of the reporter gene. Arrows indicate the sites that have been replaced by G-quadruplex forming sequences or their respective controls in this study. G-quadruplexes have been investigated both on the sense or antisense strand. **B** Nucleotide sequence of the sense strand in the 5'-UTR and in the 3'-UTR of the pQE reporter system. **C** Nucleotide sequence of the coding strand (sense strand) occurring in the 5' regulatory region of the pBAD reporter system. Sequences that have been replaced by 15 nt long G-rich elements are indicated by lines, when only 11 nt were replaced (G_2T) the first and the last 2 nt of the indicated sequence were not changed relative to wt. Dotted lines indicate the range which has been sequence-modified for investigation of G-quadruplex influence adjacent to the SD region.

Since for most mammalian systems G-quadruplexes in promoter regions have been reported to silence gene expression, we expected a fluorescence decrease in our reporter systems when comparing G-quadruplex constructs with their respective non-quadruplex forming control 1 (*sequence listed in Figure 3.1 A and Table 7.1*).

Indeed, quadruplex insertion between the -10 and -35 region of the J06 promoter resulted in a decrease in gene expression of 86.0% compared to the control sequence *(see Figure 3.3 A)*. A reduced decrease (42.0%) was observed for the quadruplex construct compared to control 1 at the same position in the araBAD promoter *(Figure 3.3 B)*. In this case the quadruplex-induced effect might be influenced by the arabinose-induced binding of the regulatory protein AraC from position -35 to -51 (277,278). However, in general the same effects can be observed for both promoter systems: Quadruplexes inserted between the -10 and -35 region in the sense strand did not significantly influence gene expression; quadruplexes inserted downstream of the -10 region in the sense strand decreased gene expression and quadruplexes inserted in the antisense strand between the -10 and -35 region showed the highest decrease in gene expression compared to the respective control 1. Interestingly, when G-rich sequences were inserted downstream of the -10 region in the antisense strand we observed a significant increase in gene expression compared to the non-quadruplex control for the J06 promoter construct. The G-quadruplex inserted at the same position in the araBAD promoter region showed no significant change in gene expression when compared to control 1. However, when compared to a second, non-quadruplex-forming control gene expression is increased significantly for the quadruplex inserted downstream of the -10 region in the antisense strand of the pBAD setup as well *(see Figure 3.6 D)*. Still, as the two controls differ, the effect cannot clearly be linked to G-quadruplex formation. In the following experiments we investigated three different quadruplex locations in more detail: 1. G-quadruplexes within the core promoter sequence on the antisense strand; 2. G-quadruplexes in the 5'-UTR located 20 nt upstream of the start codon as well as quadruplexes surrounding the SD region; 3. G-quadruplexes inserted into the 3'-UTR.

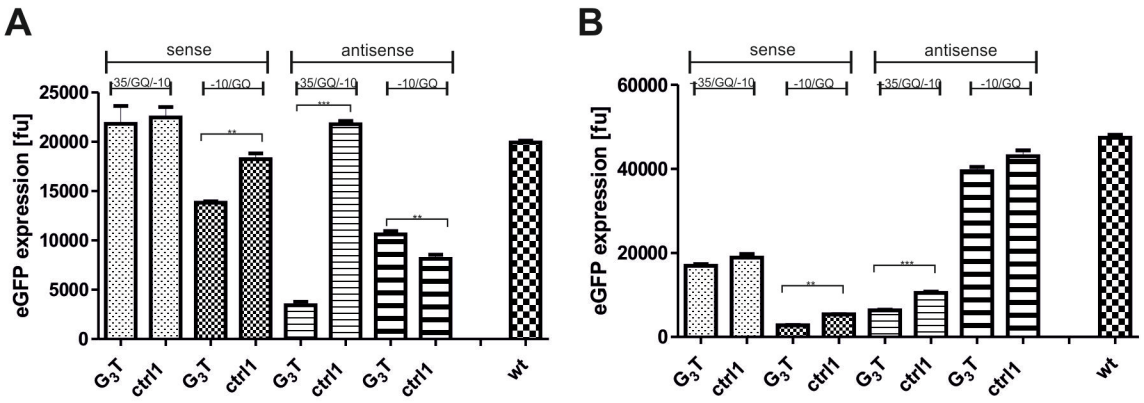


Figure 3.3: Influence of G-quadruplexes in bacterial promoters on gene expression.

A/B Quadruplex forming sequences were placed either between the -10 and -35 region (-35/GQ/-10) or downstream of the -10 region (-10/GQ), each on the sense or antisense strand in an eGFP reporter system under the control of **A** the constitutive pQE promoter and **B** the inducible pBAD promoter. All experiments were performed in triplicates. Error bars represent standard deviations of three independent experiments. According to the unpaired t-test ** indicates P<0.001, *** indicates P<0.0001. Ctrl is representative for control. © Cell Press.

3.1.1.3 Influence of quadruplexes on the antisense strand of the core promoter

The significant decrease in gene expression with the quadruplex located between the conserved promoter regions on the antisense strand (86.0%*, see Figure 3.3 A)* motivated us to study this construct in more detail. Therefore, experiments were carried out with the constitutively expressing J06 promoter. Given that sequence changes in these highly regulatory regions might have a huge influence on gene expression (279), additional non-quadruplex-forming controls were investigated at this position. Furthermore G-quadruplexes of different stabilities including G_3T and less stable quadruplexes comprising two tetrads with longer loops were inserted *(sequences and in vitro characterization are shown in Chapter 3.1.1.1)*.

As expected, less stable G-quadruplexes repressed gene expression less effectively compared to thermodynamically stable ones. However, all tested sequences resulted in reduced gene expression compared to controls 1 and 3 as well as the wildtype system. In comparison to control 1, the constructs G_3A, G_2T and G_2CT repressed gene expression by 72.3%, 59.7% and 32.6%, respectively *(see Figure 3.4 A).* Control 2 showed a repression of 38.3% in gene expression, which is comparable to G_2CT. This is in accordance with the similar T_m of both structures and, again, shows a strong correlation between repression of gene expression and the stability of the formed secondary structures. For further experiments we used only control 1 and 3 as non-structure-forming controls.

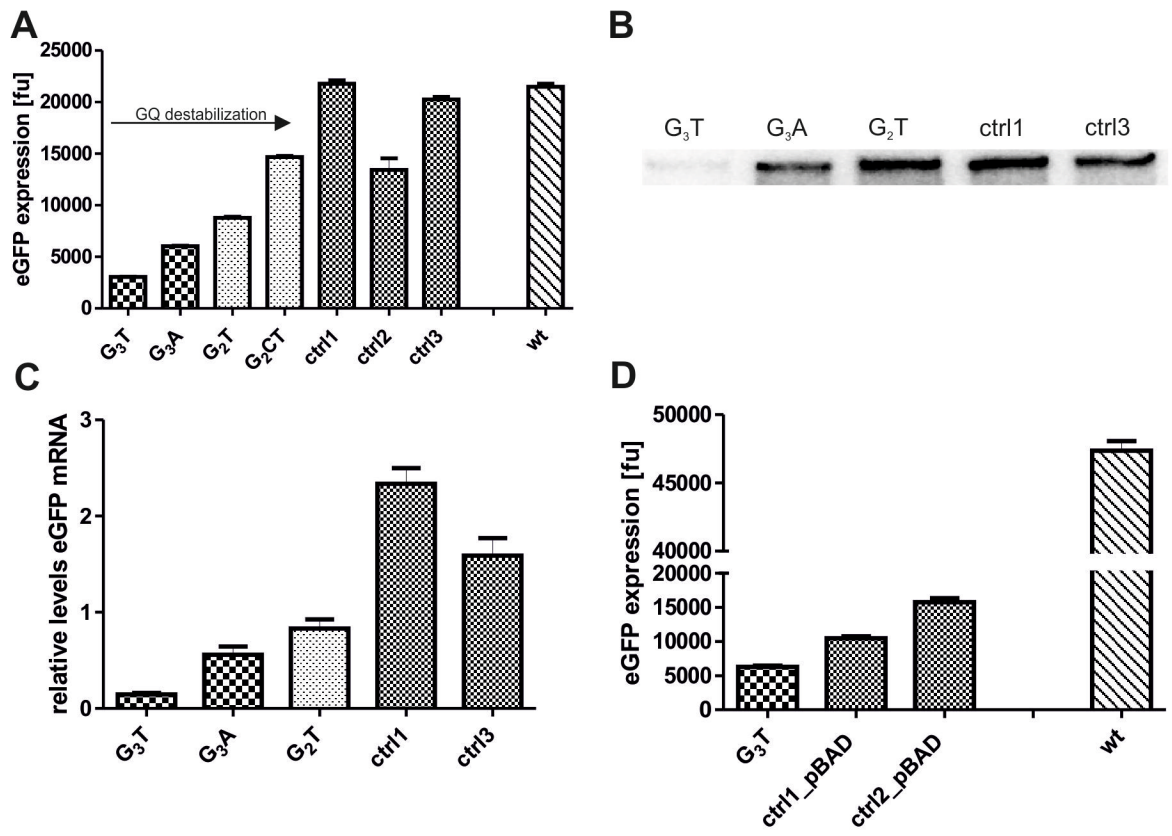


Figure 3.4: Influence of G-quadruplexes in the core promoter region.

A eGFP gene expression of constructs with G-quadruplex sequences placed in the core promoter of the pQE-eGFP system. **B** *In vitro* transcription of different constructs with *E. coli* RNA polymerase. **C** Analysis of eGFP mRNA levels by semi-quantitative RT-PCR relative to the expression of the genomically encoded *ssrA* gene. **D** Gene expression of constructs with G-quadruplex inserted between the -35 and -10 region in the antisense strand of the pBAD-eGFP system (ctrl1_pBAD: GGGTGTGTGTGTGTG; ctrl2_pBAD: GGGTGAGTGAGTGAG). All experiments were performed in triplicates. Error bars represent standard deviations of three independent experiments. Ctrl is representative for control. © Cell Press.

The pBAD vector system also showed reduced gene expression for the G_3T construct in comparison to two different controls: 39.9% and 60.0% compared to crtl1_pBAD and ctrl2_pBAD, respectively (*see Figure 3.4 D*).

As the G-quadruplex is located in front of the TSS, it should not be located on the mRNA; thus, we hypothesized regulation of gene expression on a transcriptional level. In order to assay effects of the quadruplex on transcription, we performed *in vitro* transcription reactions with *E. coli* RNA polymerase *(see Figure 3.4 B)*. For the most stable G-quadruplex, G_3T, we observed almost no full-length transcription product.

Also, the G_3A G-quadruplex construct showed reduced transcription compared to the G_2T construct and the controls. Additionally, *in vivo* eGFP mRNA levels were analyzed via semi-quantitative reverse transcription PCR (RT-PCR) relative to the expression of the genomically encoded *ssrA* gene. mRNA levels of G-quadruplex-containing constructs were

decreased in comparison to controls. For G_3T, G_3A and G_2T constructs, relative RNA levels were decreased by 93.7%, 76.0% and 64% when compared to control 1 *(see Figure 3.4 C)*. Hence, it seems that the quadruplex formation efficiently inhibits transcription at the investigated position.

In an attempt to detect G-quadruplex formation in the living bacterium, *in vivo* footprinting with dimethylsulfate (DMS) was performed. Bacteria were incubated with DMS that selectively methylates guanines at the N7 position. The inserted G-rich stretch should be protected from methylation when engaged in Hoogsteen interactions in the quadruplex structure, whereas N7 should be accessible for methylation in the duplex form. After DMS quenching, DNA isolation and cleavage at methylated positions, the DMS accessible sites were identified via a primer extension reaction. *Figure 3.5* shows the footprinting reaction of the G_3T construct in comparison to control 3. Cleavage is observed at the respective guanine sites, which might be explained by a quadruplex structure forming only temporarily during transcription. Interestingly, a strong band can be observed directly 5' of the G-rich stretch for the G_3T construct, but not for the control 3 construct. This might result from enhanced cleavage at this site or polymerase stop during primer extension. Furthermore, we tested whether the addition of the quadruplex-stabilizing ligand N-Methyl Mesoporphyrin (NMM) could enhance the influence of the quadruplex on gene expression. However, no change in the gene expression pattern was observed upon addition of different concentrations of NMM (2 µM, 20 µM and 100 µM, *see Chapter 3.1.1.8*).

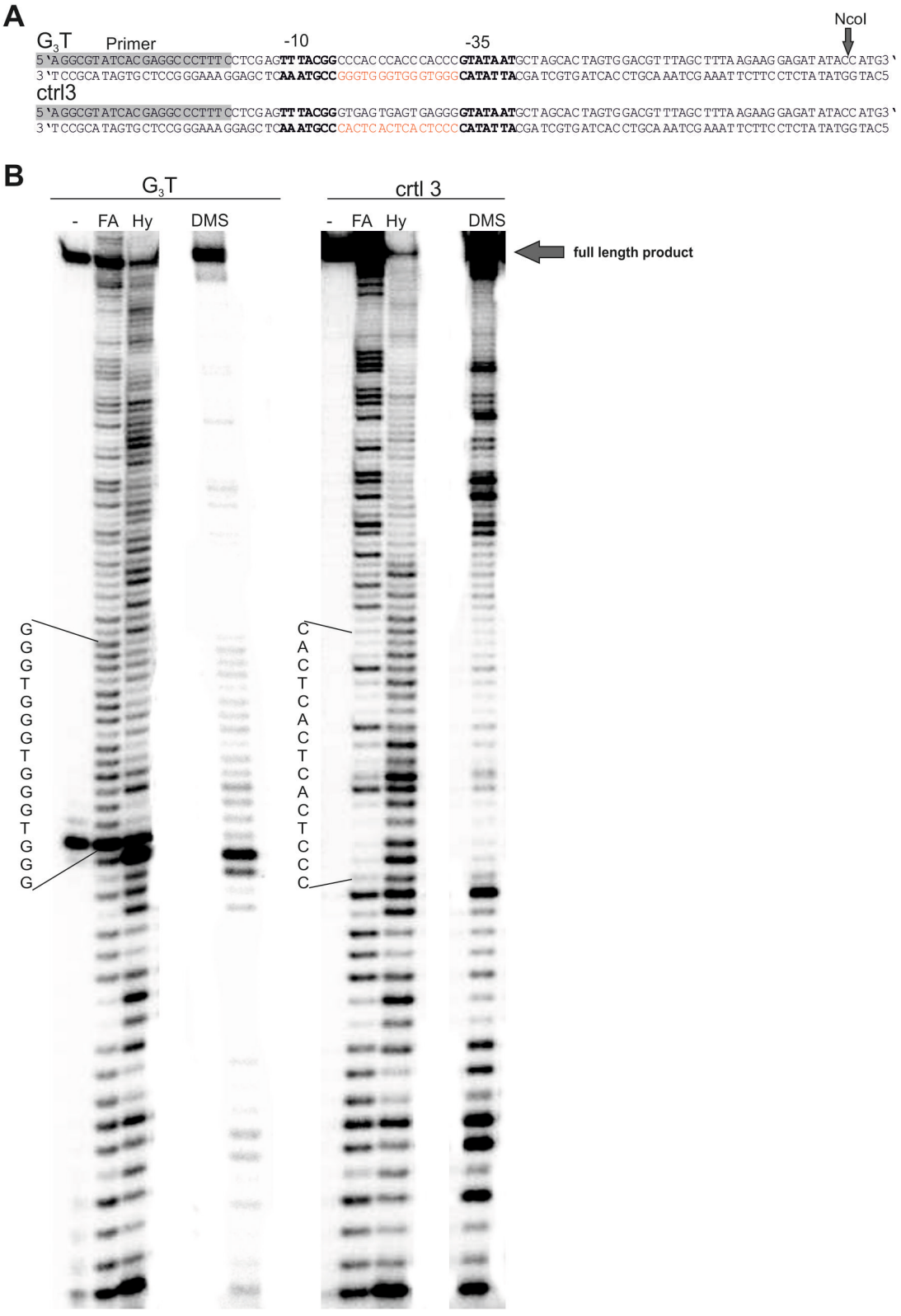


Figure 3.5: *In vivo* footprint with DMS.

A Nucleotide sequences of the investigated promoter region for the G_3T and control 3 plasmids are shown. Gray: Primer binding site; bold black: -10 and -35 promoter regions; red: G_3T and control sequence inserted. **B** Primer extension reaction analyzed on a 10% denaturing PAGE. DMS treated probes (DMS) are shown in comparison to non-treated DNA (-) and the two sequencing reactions for purine sequencing (FA) and pyrimidine sequencing (Hy). Ctrl is representative for control. © Cell Press.

3.1.1.4 Influence of quadruplexes 20 nt in front of the start codon

In the initial experiments a significant modulation of gene expression and strand bias for G-quadruplexes placed within the promoter was observed. The next aim was to investigate whether similar effects could be observed in transcribed regions. Chowdhury and co-workers have reported that most of the G-quadruplex motifs in *E. coli* are found within 100 nt upstream of the start codon. Some of these motifs were identified around 20 base pairs upstream of the start codon, such as in front of genes like *yhiP* or *yabB* and their orthologues in other organisms (14). Therefore, we decided to focus on G-quadruplexes inserted 20 nt upstream of the start codon *(see Figure 3.2).*

The most stable G_3T quadruplex and the less stable G_2CT quadruplex were inserted into the pQE-eGFP reporter system, and their eGFP expression was compared to two different controls *(see Figure 3.6 A)*. Intriguingly, the quadruplex sequences in the antisense strand enhanced gene expression in comparison to controls and wildtype vector. For the G_2CT construct gene expression increased by 37%, whereas for the very stable G_3T gene expression increased more than 100% compared to control 1. When the motif was inserted into the sense strand instead, a decreased gene expression was observed: 60.0% for G_3T and 49% for G_2CT, relative to control 1. When the quadruplex was inserted into the pBAD vector system 20 nt in front of the start codon on the sense strand, a decreased gene expression compared to controls was observed as well (47.5% compared to crtl 1_pBAD and 71.7% compared to ctrl 2_pBAD; *see Figure 3.6. B*). However, insertion of the G-quadruplex into the pBAD vector system 20 nt in front of the start codon on the antisense strand, did not lead to clear effects. In fact, the gene expression of the two controls differed and no conclusive results related to G-quadruplex formation could be derived, accordingly the strong effect following the antisense insertion could also be vector-dependent.

We wanted to clarify whether the modulation of gene expression in the pQE vector system occurs on the transcriptional or the translational level. Determination of mRNA levels revealed an increase of 44% when comparing the eGFP mRNA level of G_3T occurring on the antisense strand to control 1 *(see Figure 3.6 C)*. However, the eGFP mRNA levels of G_3T inserted in the sense strand remained constant compared to the controls *(Figure 3.6 D)*. This finding indicates that a G-quadruplex-related regulation strongly depends on strand orientation. Whereas G-quadruplexes found on the antisense strand seem to regulate gene expression on a transcriptional level, those occurring on the sense strand rather seem to influence translation. In the following studies we focused on mRNA-based influences on gene expression caused by translational regulation.

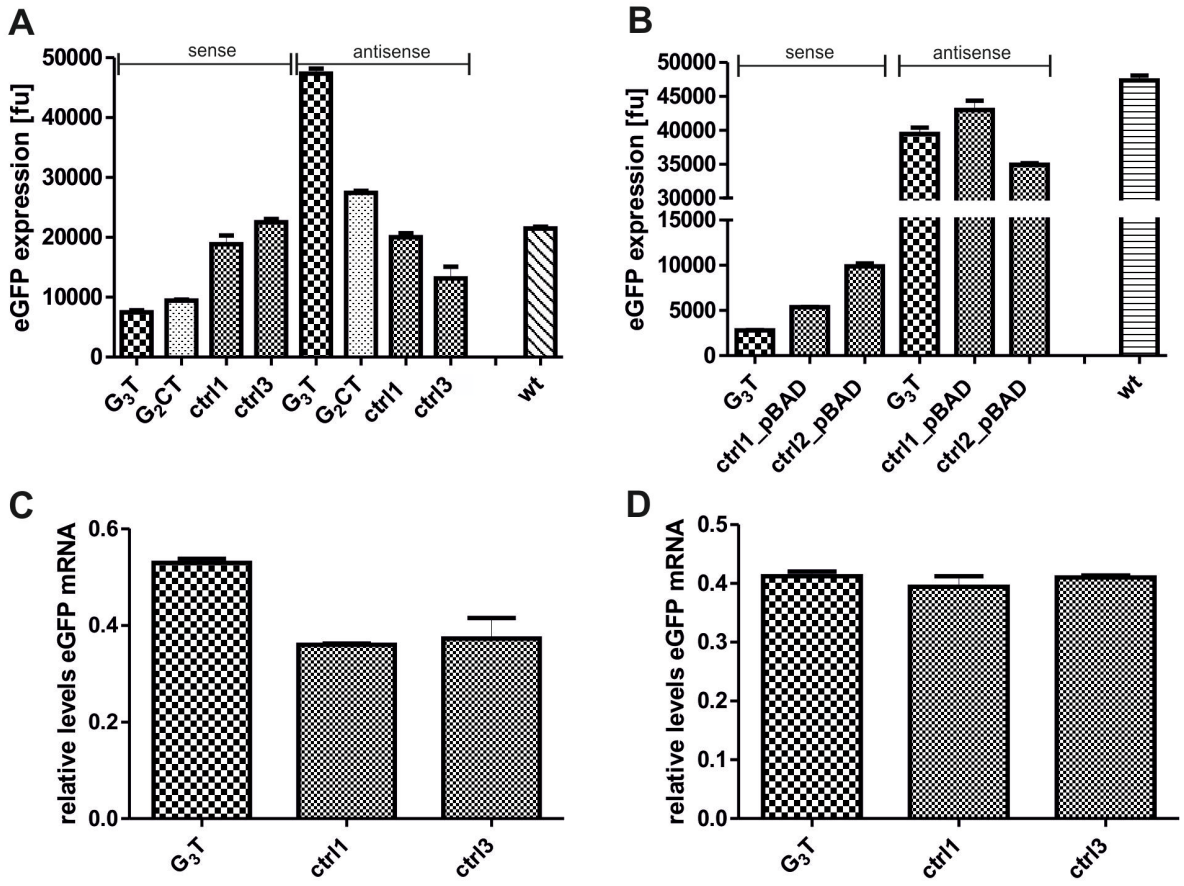


Figure 3.6: Influence of G-quadruplexes inserted 20 nt in front of the start codon.

Gene expression levels of constructs with G-quadruplex forming sequences of different stabilities inserted 20 nt upstream of eGFP start codon, with G-tracts either on the sense or antisense strand: **A** in the pQE vector system and **B** in the pBAD vector system. Respective eGFP mRNA levels of the pQE constructs have been analyzed by RT-PCR and are shown in **B** for the sense and **C** for the antisense strand. RNA levels were calculated relative to the expression of the genomically encoded *ssrA* gene. Ctrl is representative for control. © Cell Press.

3.1.1.5 Engineering of SD-adjacent quadruplexes

Translational modulation of gene expression via G-quadruplexes has been shown earlier by Hartig and co-workers. In that study, artificially designed sequences were placed around the ribosome binding site so that secondary structure formation inhibited interaction of the 16S rRNA with the eGFP-mRNA and hence initiation of translation. Repression of gene expression correlated with the thermodynamic stability of the G-quadruplex (161). A pronounced decrease of up to 96% of gene expression was observed in comparison to the wildtype, non-structured SD region in the artificial quadruplex system. Cis-repression and trans-activation of bacterial ribosome binding sites via secondary structure formation have been described in the context of engineered riboregulators (280). We were curious as to whether the opposite effect – activation of gene expression – could be accomplished by G-quadruplex formation in the 5'-UTR. In nature, e.g. the cold-sensing thermometer relies on RNA-regulatory mechanisms where a secondary structure formation liberates the SD region and hence activates gene expression (281). In order to implement a similar system based on quadruplex formation, another set of 5´-UTRs was designed. In these artificial designs the SD site can be masked by the formation of a long stem-loop structure in the mRNA *(see Figure 3.7 A)*. This stem-loop structure contains a G-rich sequence strain potentially able to form a G-quadruplex. G-quadruplex formation competes with Watson-Crick base pairing in the stem (stem length ranging from 13 to 20 base pairs). An insertion of up to five single-nucleotide mismatches destabilizes the stem and should simplify quadruplex formation. The formed quadruplex ultimately dissolves the stem-loop structure so that the ribosome binding -site should become accessible for the ribosome, thus facilitating translation. These designs were investigated in a pBAD-eGFP reporter system. Predicted mfold structures (282) of the designs showing the mRNA region upstream of the start codon are depicted in *Figure 3.8*. We chose the G_3U quadruplex for our investigations as it is a short sequence that folds into a remarkably stable RNA G-quadruplex structure (283). In the first design the G_3U quadruplex sequence was inserted 21 nt upstream of the eGFP start codon with full base pairing in the stem-loop structure. When comparing the quadruplex-bearing construct (G_3U) to its control which should not be able to form a G-quadruplex (G_3U ctrl) we observed a slight increase in gene expression (15%). Destabilization of the stem-loop structure by insertion of five mismatched base pairs allowed an easier quadruplex formation. For this construct (G_3Umm) we observed a pronounced activation of gene expression of more than 100% compared to the control (G_3Umm ctrl). Destabilization of the G-quadruplex by introduction of longer loops (G_3CUmm) in the aforementioned construct still increased gene expression compared to the respective control (G3CUmm ctrl), but with less efficiency (89%, *see Figure 3.7 B*).

However, addition of different G-quadruplex stabilizing compounds did not influence the gene expression levels *(see Chapter 3.1.1.8)*. It is noteworthy that the mutations preventing G-quadruplex formation in the controls do not interfere with the complementary base pairing in the stem, but rather occur in the loop sequence *(see Figure 3.8)*. Hence, base-pairing interactions within the hairpin structure do not differ between quadruplex constructs and controls. The similarity of eGFP-mRNA levels in G-quadruplex constructs and respective controls confirmed the regulation on the translational level *(see Figure 3.7 C)*. Our results show that freeing the masked ribosome binding -site by the formation of a G-quadruplex in the mRNA could be a possible mechanism of translational regulation.

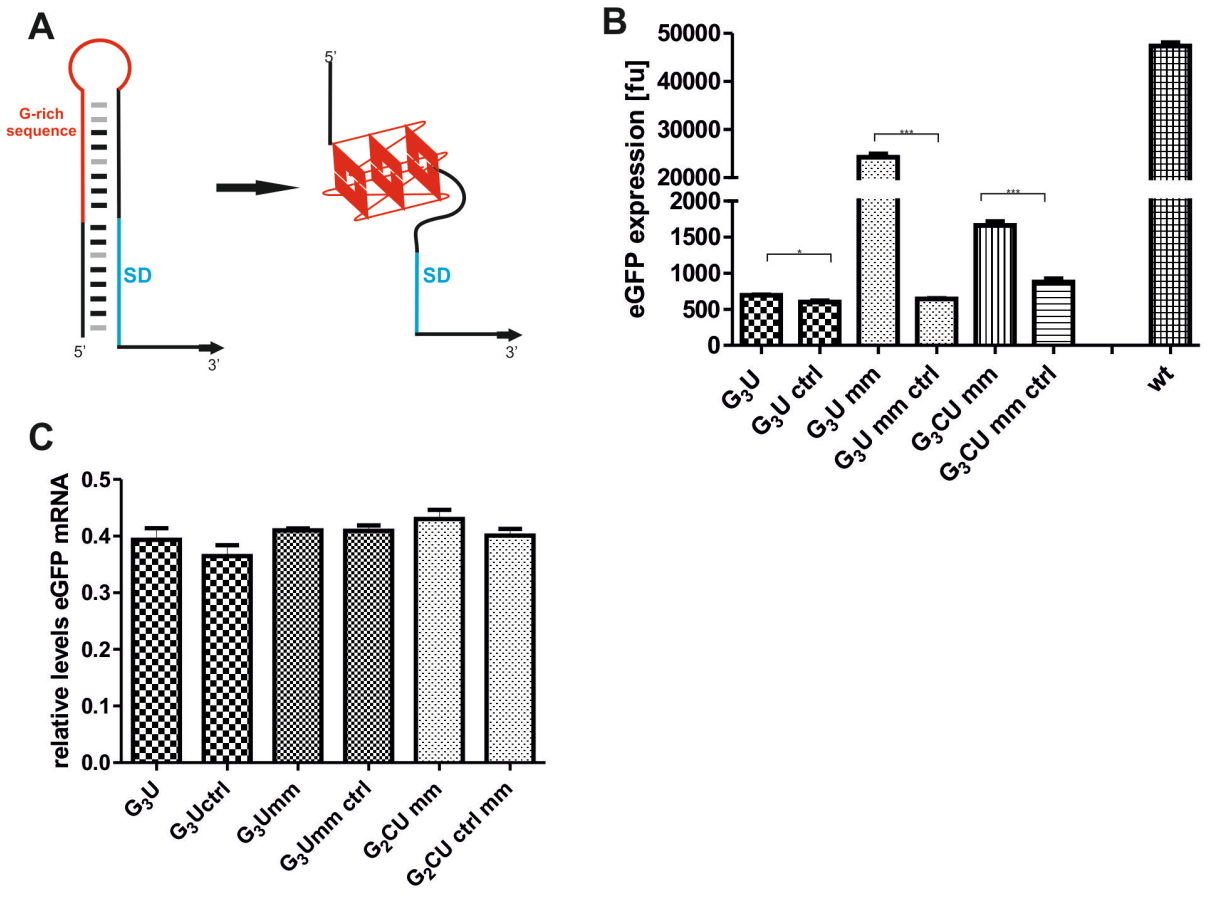


Figure 3.7: Artificial system comprising SD-adjacent quadruplexes.

A Mechanism suggested for enhancing gene expression via G-quadruplex formation. The red sequence is G-rich and able to form a quadruplex, but can also partly pair with the black sequence immediately 5' of the SD region (blue). Access to the SD region can be blocked by formation of a stem loop structure. G-quadruplex formation leads to break up of the stem-loop structure and freeing of the SD site. Grey base pairs indicated mismatches introduced for facilitating quadruplex formation. **B** Modulation of eGFP expression. G_3U: GGGUGGGUGGGUGGG; G_3U ctrl: GGGUGGGUGUGUGUG; G_3CU mm: GGGCUGGGCUGGGCUGGG; G_3CU mm ctrl: GGGCTGGGCTGTGCTGTG. **C** Analysis of eGFP mRNA levels by semi-quantitative RT-PCR for the engineered G-quadruplex constructs compared to their respective control. RNA levels were calculated relative to

the expression of the genomically encoded *ssrA* gene. All experiments were performed in triplicates. Error bars represent standard deviations of three independent experiments, * indicates P<0.05, *** indicates P<0.0001. © Cell Press.

The demonstrated system of activating quadruplexes contradicts the results observed earlier for the G_3T quadruplex inserted into the 5'-UTR on the sense strand 20 nt upstream of the start codon where a repression of gene expression was observed. This decrease was related to translational modulation, as mRNA levels did not change for G-quadruplex-bearing constructs compared to the respective controls in real-time PCR experiments. For the G_3T inserted into the 5'-UTR on the sense strand 20 nt upstream of the start codon, the SD region should be easily accessible for the ribosome (*see Figure 3.8 G* for mfold prediction).

In the engineered system, the opposite effect was observed: G-quadruplexes inserted 21 nt upstream of the start codon activated gene expression. However, in this design the whole 5'-UTR was modified (nt composition and length between the SD site and the start codon) in order to mask the SD region if no quadruplex formation occurs. Hence, these two designs are not comparable and both results reflect possible influences of G-quadruplex sequences located close to the SD site. Furthermore, the overall gene expression of both the controls and the quadruplex constructs decreased in this system in comparison to the wt sequence, indicating an influence of sequence changes. In fact, changes in the 5'-UTR sequence have been found to alter the mRNA translation rate (284), as the 30S ribosomal complex appears to bind the upstream 5'-UTR (285,286). Therefore, the overall sequence and shape of the 5'-UTR has a strong impact on the actual effect of the G-quadruplex on gene expression.

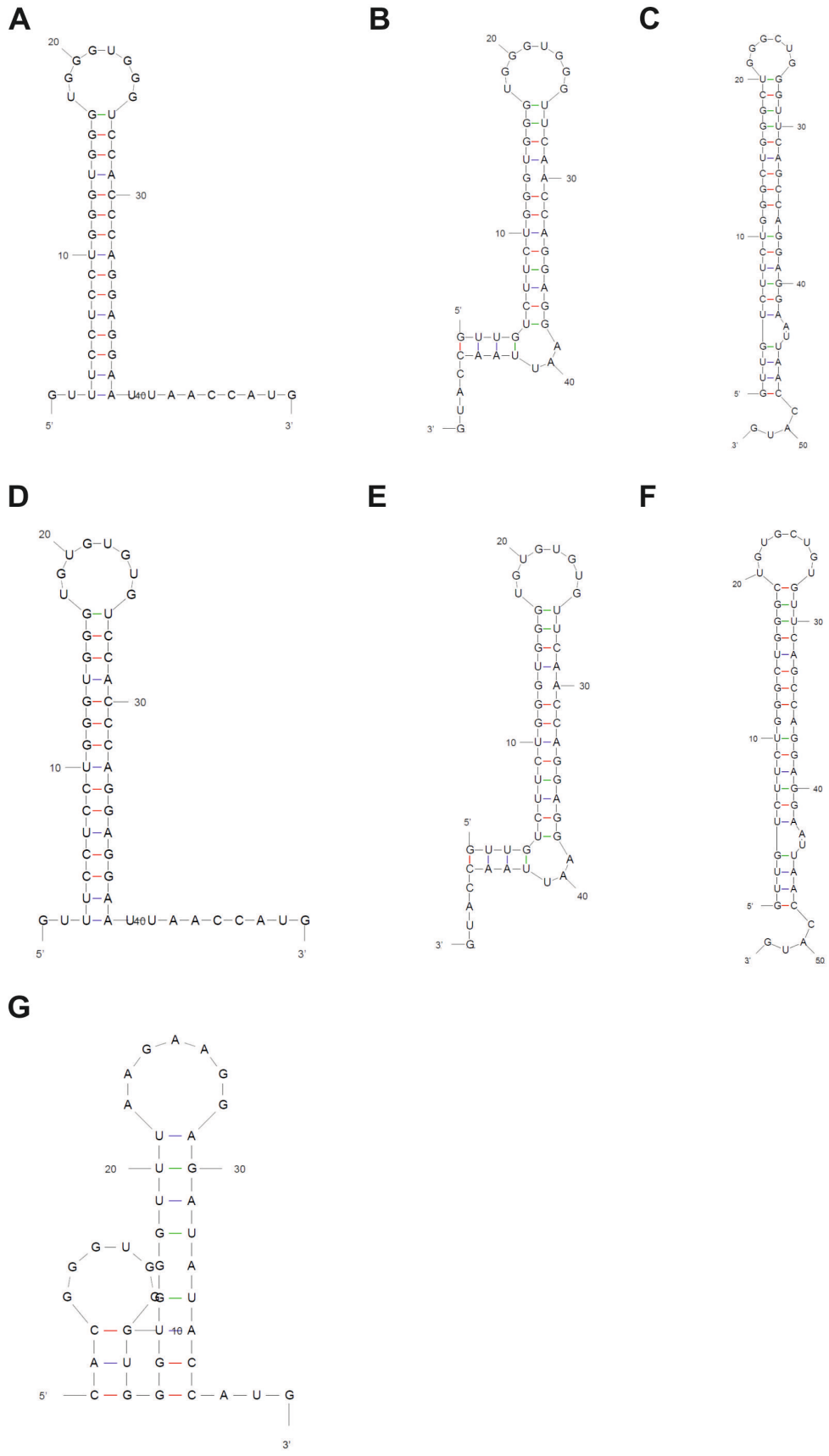


Figure 3.8: Predicted mRNA structures for the 5' region of the artificial constructs.

Predicted mRNA structure for the 5' region of G-quadruplexes investigated for liberating SD site when blocked by a stem-loop structure. **A** G_3U **B** G_3Umm **C** G_3CU **D** G_3U ctrl **E** G_3Umm ctrl **F** G_3CU ctrl **G** Predicted structure of the G-quadruplex (G_3T) inserted 20 nt in front of gene start in the pQE system. The SD site has the typical consensus sequence 5'-AGGAGGA-3'. Structure prediction according to mfold (http://mfold.rna.albany.edu/). © Cell Press.

3.1.1.6 Naturally occurring quadruplexes in the SD region in *E. coli*

In two different artificial systems pronounced modulation of translation via G-quadruplexes has been shown: 1. Repression of gene expression up to 96% compared to the wildtype vector by masking the SD region (161) and 2. More than 100% activation of gene expression compared to a control by liberation of the SD site. The observation of drastic quadruplex-mediated effects in 5´-UTRs raises the question whether quadruplexes in natural genetic contexts exert similar control over gene expression. Hence the occurrence of potential G-quadruplexes surrounding the SD region in genomic sequence data of *E. coli* MG1655 was investigated. Using the ProQuad Pattern Algorithm (104) we searched for G-quadruplexes with 2-5 tetrads and loops of 1-5 nucleotides that overlap with the anticipated SD sequences located approximately 10 – 12 nt upstream of the start codon. 46 potential quadruplexes in the vicinity to SD sequences were identified *(see Table 13.1 in the appendices).* Gene functions were categorized using the KEGG database (287,288). The sequences were widely distributed in all kinds of genes, with most quadruplexes in fundamental functional categories of metabolic pathways, microbial metabolism in diverse environments and biosynthesis of secondary metabolites. Importantly, all identified quadruplex sequences are anticipated to form structures with two tetrads and do not show a conserved sequence pattern. Most of them also occur in other *E. coli* subtypes, e.g. *Escherichia coli* CFT073 (ECC) and *Escherichia coli* O157:H7 str. Sakai (ECS).

Although G-quadruplexes comprising only two tetrads are less stable, such structures could function as regulatory elements as well. Recently, Chowdhury and co-workers showed that a quadruplex motif formed by two tetrads is involved in the regulation of the expression of the human thymidine kinase (289). Wieland et al. also observed pronounced inhibition of gene expression by means of artificially designed sequences masking the SD region with two tetrad-containing quadruplexes (161). In order to investigate some of the naturally occurring G-quadruplex sequences in more detail, their influence on gene expression was studied. For this purpose, we placed the quadruplexes including the whole natural 5'-UTR in front of a β-galactosidase reporter gene under control of the araBAD promoter. We randomly chose

five different genes with putative G-quadruplex sequences overlapping the ribosome binding site from our set of sequences identified in the *E. coli* genome: *oxyR, relA, rseA, napH* and *yadI (see Figure 3.9)*. The G-quadruplex sequences in front of those genes differ in loop length and distance from the start codon *(sequences listed in Figure 3.9 and Table 13.1 in the appendices)*. One sequence (relA) even included the start codon within the possible G-quadruplex sequence. For each construct we designed two controls which should not be able to form a G-quadruplex, *see Figure 3.9*. Mutating guanines outside the core SD region should not alter the efficiency of 16S rRNA interactions with the mRNA, but instead reduce quadruplex-based secondary structure formation. However, in this region – which is crucial for initiating translation in bacteria – it is very likely that even small sequence changes influence gene expression (290). Some of the mutants showed very high gene expression patterns, which might be explained by sequence changes that facilitate ribosomal interactions. Especially yadIm1 showed an unexpectedly high gene expression although it contains only two G to U mutations. However, A/U rich sequences upstream of the SD site have been reported to serve as mRNA stabilizing elements (291). Unfortunately, both controls for the yadI construct behaved very differently in the gene expression studies and thus did not allow conclusions to be drawn with regard to G-quadruplex formation *(see Figure 3.9 B)*. In any case, for three other constructs (relA, oxyR, napH) an effect on gene expression was observed which seems to be related to secondary structure formation. For the sequences upstream of the *E. coli oxyR (see Figure 3.9 D)* and *relA (see Figure 3.9 B)* genes, gene expression significantly increased in both mutants: 87.1% for oxyRm1 and 92.0% for oxyRm2 as well as 59.2% for relAm1 and 85.9% for relAm2 *(see Figure 3.9 B and D)*. Regarding the napH construct, we observed a significant decrease of gene expression for both controls (more than 100%). However, gene expression of napH m1 decreased sevenfold compared to napH m2. As the two mutants differ considerably it is difficult to associate this with secondary structure formation. Also, addition of the quadruplex stabilizing compound NMM did not change the gene expression levels significantly *(see Chapter 3.1.1.8)*. For the *rseA* 5'-UTR G-rich sequence, β-galactosidase expression is increased compared to both mutants *(see Figure 3.9 B)*. In this case, the effect could not be deemed significant according to the unpaired t-test.

Next, the G-quadruplex in front of the *oxyR* gene was investigated in more detail. The stability of the *oxyR* quadruplex RNA sequence *(see Figure 3.9 C oxyR and oxyRm1 and Table 7.1)* was characterized via CD spectroscopy and thermal denaturation *(see Figure 3.9 E and F)*. The oxyR sequence forms a parallel four-stranded structure as expected for an RNA quadruplex with a melting temperature of 56.2°C. The control sequence oxyRm1 *(see Figure 3.9 C and Table 7.1)* – containing two G-to-U mutations – showed a shifted CD signal

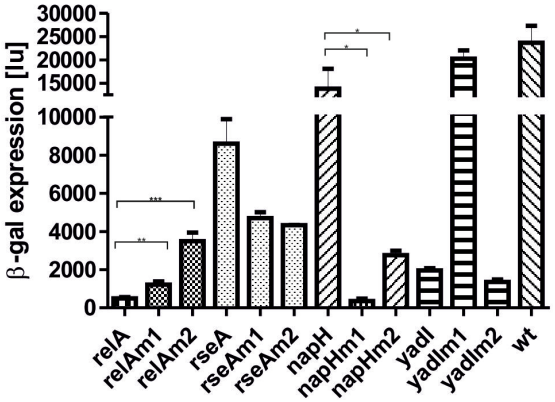
and a much lower melting temperature of 38.6°C. Therefore, we assumed the formation of a G-quadruplex with moderate stability for the oxyR construct and no stable structure formation for the control oxyRm1. As described above, the *oxyR* sequence showed significant reduction of gene expression compared to the controls oxyRm1 and oxyRm2.

In order to analyze the influence of certain nucleotide changes on gene expression in more detail, another set of controls for the *oxyR* G-quadruplex sequence was designed *(see Figure 3.9 C&D)*. Here, we included sequences which were mutated outside of the G-tract, i.e. they should still be able to form a G-quadruplex structure (oxyRm3, oxyRm4 and oxyRm5). With these control constructs we wanted to support our assumption that changes in gene expression result from secondary structure formation and are not only the effect of sequence changes in this regulatory region. Accordingly, we expected reduced gene expression for controls able to form G-quadruplexes with respect to the non-quadruplex controls. For oxyRm3 and oxyRm4 the respective A was changed to U 14 nt and 10 nt in front of the start codon. Gene expression increased significantly (80%) compared to the naturally G-rich oxyR sequence, but still remained repressed in comparison to the mutants that were not able to form a G-quadruplex (oxyRm1 and oxyRm2). Interestingly, when U was changed into A 13 nt upstream of the start codon, gene expression decreased even more than in the natural *oxyR* sequence. In oxyRm6 the last G-tract was mutated, so that no G-quadruplex formation should be possible. In this case gene expression was also repressed when compared to the other mutants, but still significantly increased (77%) when compared to the naturally occurring *oxyR* sequence. Presumably, both effects (the secondary structure formation as well as the single-nucleotide changes in the SD region) contribute to the observed changes in gene expression. To exclude the influence of sequence mutations on mRNA stability or altered transcription rates for the oxyR constructs, we determined oxyR mRNA levels via RT-PCR. We found similar mRNA abundances for G-quadruplex constructs and mutants 1 and 2 *(see Figure 3.9 H)*, suggesting differential translation initiation as the likely cause of the observed differences in gene expression. Furthermore, we showed that this modulation is not selective for a specific plasmid or read-out system as the insertion of the same SD background in front of the eGFP gene in the pQE vector led to comparable results *(see Figure 3.9 G)*.

A

name	sequence
relA	GGAGAGGACG**AUG**GUUGCGG
relAm1	GGAGAGGACG**AUG**GUUGCUG
relAm2	UUAGAGGACG**AUG**GUUGCGG
rseA	GGAUACUGGAUAAGGGUAUUAGGC**AUG**
rseAm1	GGAUACUGGAUAAGGGUAUUAUGC**AUG**
rseAm2	GUAUACUGGAUAAGGGUAUUAGGC**AUG**
napH	GGCUGGCUGGAGGGGAACA**AUG**
napHm1	GGCUGUCUGGAGGGGAACA**AUG**
napHm2	UUCUGGCUGGAGGGGAACA**AUG**
yadI	GGCUAAGGAGGAAGGCG**AUG**
yadIm1	UUCUAAGGAGGAAGGCG**AUG**
yadIm2	GGCUAAUGAGGAAGGCG**AUG**

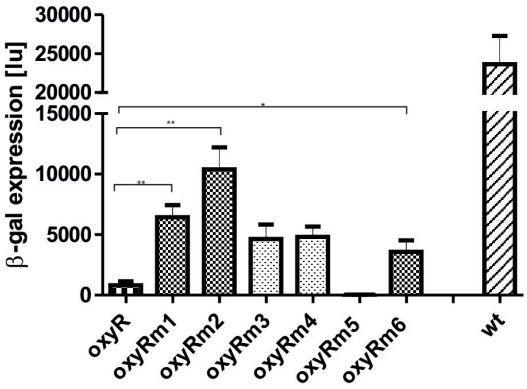
B



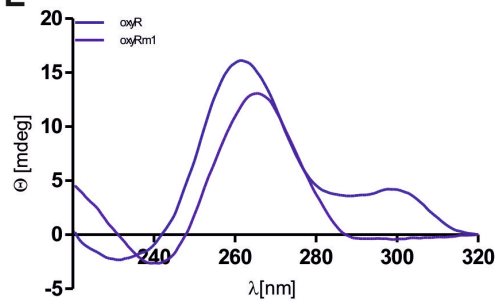
C

name	sequence
oxyR	GGCGAUGGAGGAUGGAUA**AUG**
oxyRm1	GUCGAUGGAGGAUUGAUA**AUG**
oxyRm2	UUCGAUGGAGGAUGGAUA**AUG**
oxyRm3	GGCGUUGGAGGAUGGAUA**AUG**
oxyRm4	GGCGAUGGUGGAUGGAUA**AUG**
oxyRm5	GGCGAAGGAGGAUGGAUA**AUG**
oxyRm6	GGCGAUGGAGGAUGAAUA**AUG**
wU	UAAAGAGGAGAAAUUAACU**AUG**

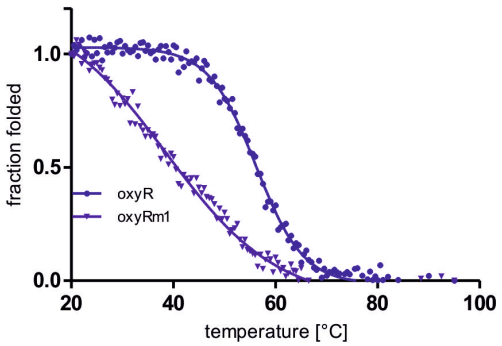
D



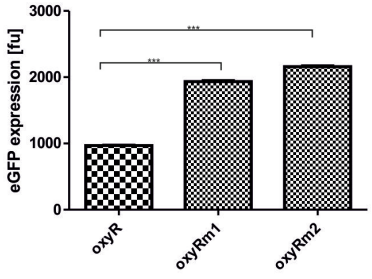
E



F



G



H

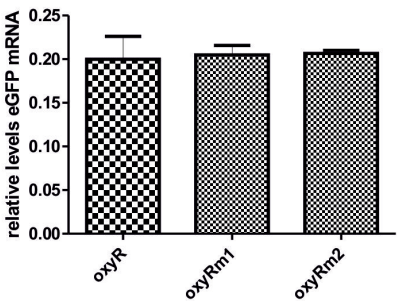


Figure 3.9: Naturally occurring quadruplexes in *E. coli* SD regions.

A Sequences (5' to 3') of quadruplexes occurring in the SD region of the *E. coli relA, rseA, napH* and *yadI* genes with their respective control mutants. Wt stands for the SD sequence in the wildtype pBAD vector. **B** β-galactosidase expression of constructs listed in **A**. **C** Sequence of the G-quadruplex in front of the *oxyR* gene and the respective controls. **D** β-galactosidase expression of oxyR constructs. **E** CD spectra and **F** thermal denaturation curves at 260 nm of the G-quadruplex in front of the *E. coli oxyR* gene compared to the oxyR mutant 1.**G** Gene expression of G-quadruplexes naturally occurring in *E. coli* SD region in front of *oxyR* gene investigated in the pQE vector in front of an eGFP reporter gene. **H** Analysis of eGFP mRNA levels by semi-quantitative RT-PCR for the SD region upstream of the *E. coli* oxyR gene compared to controls. RNA levels were calculated relative to the expression of the genomically encoded *ssrA* gene. Error bars represent standard deviations of three independent experiments, * indicates P<0.05, ** indicates P<0.001, *** indicates P<0.0001. © Cell Press.

3.1.1.7 Influence of quadruplexes in the 3'-UTR

Finally, we analyzed whether a G-quadruplex sequence inserted in the bacterial 3'-UTR influences gene expression as well. For eukaryotes, functions of 3'-UTR G-quadruplexes as *cis*-regulatory elements have been reported (145). In this part of the study the G-rich sequence was inserted 4 nt downstream of the eGFP stop codon in the pQE-eGFP reporter plasmid *(see Figure 3.2).* When inserting the G-quadruplex into the antisense strand, we found no consistent modulation activity by comparison of quadruplexes with different stabilities. In addition, control sequences inserted into the antisense strand also showed different behavior *(see Figure 3.10 A)*.

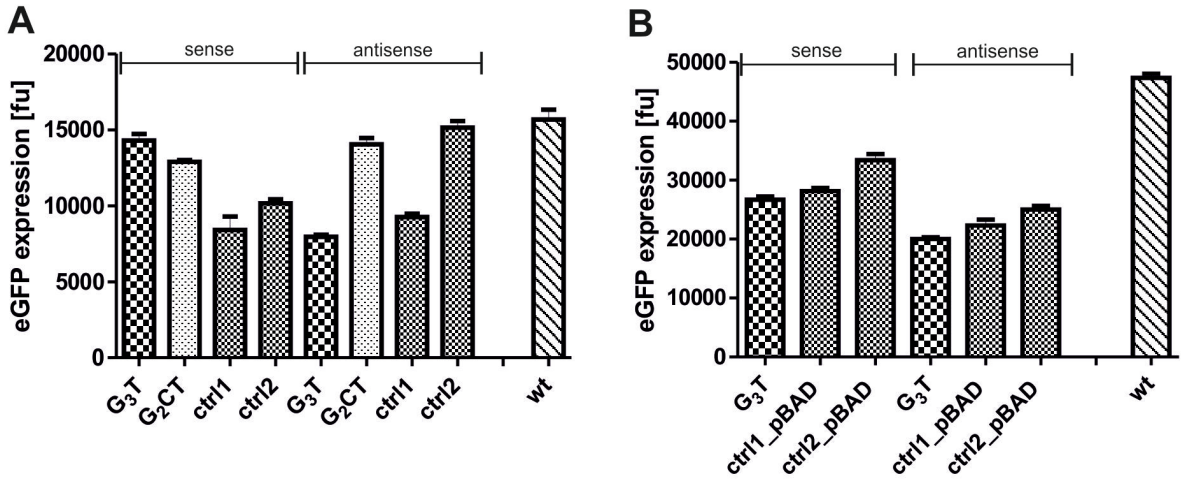


Figure 3.10: Influence of G-quadruplexes inserted into the 3' UTR.

eGFP expression of constructs with G-quadruplex forming sequences of different stabilities inserted 4 nt downstream of eGFP stop codon, with G-tract either on the sense or antisense strand in: **A** pQE vector system and **B** pBAD vector system. Error bars represent standard deviations of three independent experiments. Ctrl is representative for control. © Cell Press.

Placing the G-quadruplex in the sense strand after the stop codon increased gene expression compared to the respective controls (70% increase compared to control 1 and 40% compared to control 3). However, constructs in the pBAD plasmid with eGFP under control of the araBAD promoter did not show significant changes in gene expression when the G-quadruplex was inserted 4 nt after the stop codon in both strands *(see Figure 3.10 B)*. In conclusion, it seems that quadruplexes inserted into the immediate 3'-UTR do not influence gene expression in a consistent manner in *E. coli* K-12.

3.1.1.8 Effects of G-quadruplex stabilizing compounds

In the studies described above, different constructs were designed that changed their gene expression levels when a potential G-quadruplex forming sequence was inserted at distinct positions. We were interested in whether the modulatory effect of the quadruplex on gene expression could be intensified by addition of a G-quadruplex stabilizing compound. In particular, the influence of compounds on the following constructs was investigated: 1. Constructs with the G-quadruplex on the antisense strand of the core promoter (*see Chapter 3.1.1.3)*; 2. Constructs with SD-adjacent G-quadruplexes in the engineered system (*see Chapter 3.1.1.5*); 3. Constructs with naturally occurring G-quadruplexes in the SD region of *E. coli* (*see Chapter 3.1.1.6)*. Applying compounds to these three systems would also show differences between compound influence during the process of transcription and translation, as the first system is transcriptionally controlled and the second and third interfere with translation. For most of these investigations we used the compound NMM *(see Chapter 1.1.1.1)*. However, when incubating transformed bacteria with different concentrations of NMM (2 µM, 20 µM, 100 µM in LB medium) we could not determine an effect on gene expression in our systems *(see Figure 3.11)*. After these first inconclusive results we additionally tested other compounds for the constructs of the engineered system, e.g. 360A, TMPyP4 and Phen-DC_3 or Phen-DC_6 *(see Chapter 1.1.1.1),* and did not observe an intensifying effect on gene expression changes either *(see Figure 3.11 E).* We further tested the application of NMM in different growth phases (exponential and stationary), but to no avail *(see Figure 3.11 A&B).*

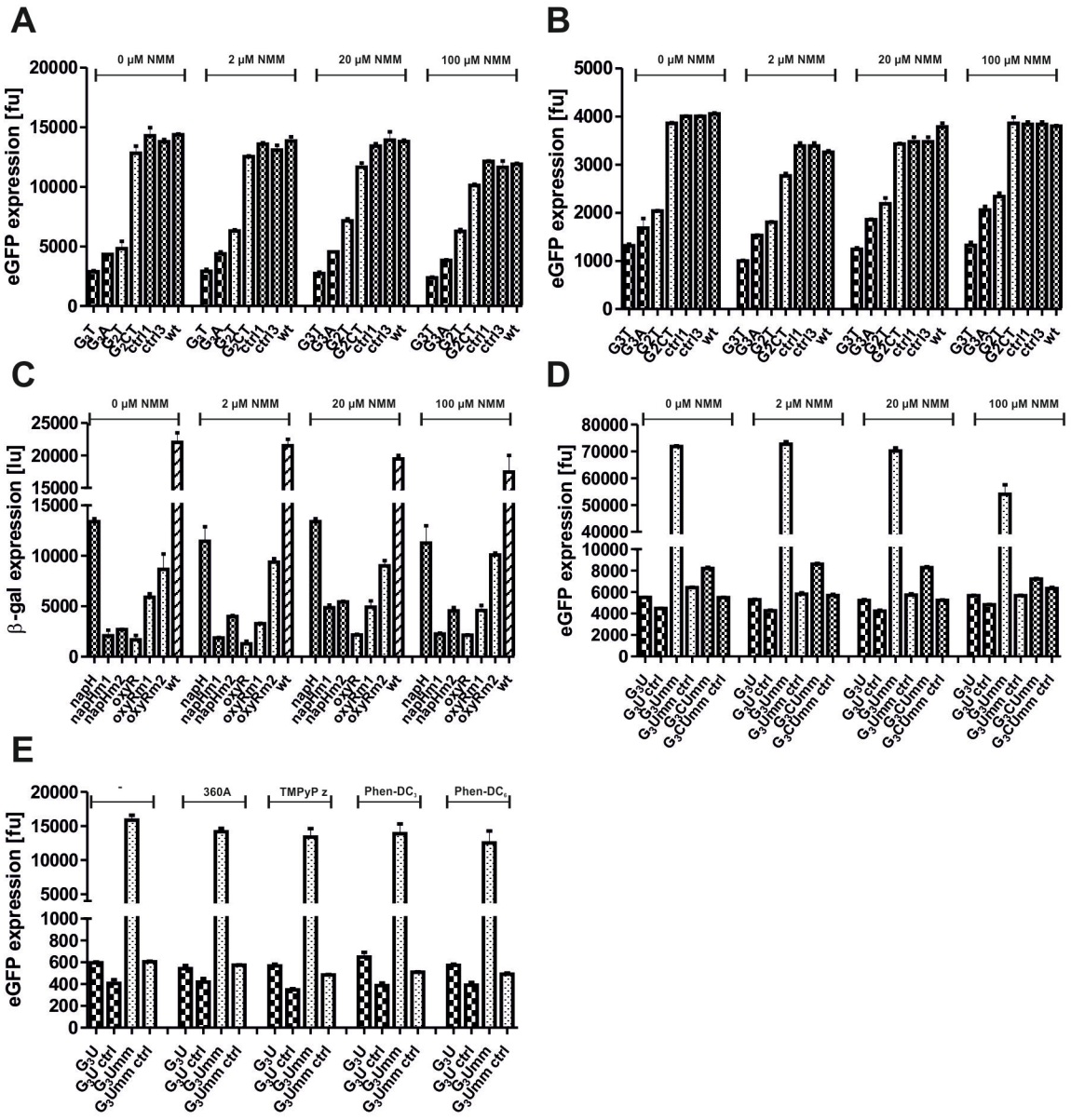


Figure 3.11: Effects of G-quadruplex stabilizing compounds.

Different concentrations of NMM were added to transformed bacteria with plasmid constructs having **A, B** G-quadruplexes at in the core promoter of the antisense strand (*see Chapter 3.1.1.3*); **C** naturally occurring G-quadruplexes in the SD region of *E. coli* (*see Chapter 3.1.1.6)*; **D** SD-adjacent G-quadruplexes in the engineered system (*see Chapter 3.1.1.5*). Cells were grown in LB with the respective amount of NMM until stationary phase in **A, C, D**, **E** and until exponential phase in **B**. **E** Different compounds (360A, TMPyP z, Phen-DC_3, Phen-DC_6; each 100 µM) or no compound (-) were added to bacteria transformed with constructs investigated in the engineered system. eGFP or β-galactosidase gene expression was determined. Error bars represent standard deviations of three independent experiments.

3.1.2 G-quadruplexes in open reading frames (ORFs)

Apart from their occurrence in 5'-UTRs, sequences with the potential to form G-quadruplex structures are present in ORFs as well. As described in *Chapter 1.2* these sequences have been shown to interfere with translation by induction of ribosomal stalling and frameshifting. The last year has seen a re-emergence of studies in this field *(see Chapters 1.2.1 and 1.2.1.1)*. However, most of these studies were, again, performed in eukaryotic cells. Although Sugimoto and co-workers investigated certain G-quadruplex sequences found in ORFs of *E. coli in vitro*, the subsequent *in vivo* experiments were performed in mammalian cells. *In vivo* studies investigating G-quadruplexes in bacterial ORFs are rare. In this chapter two distinct quadruplex motifs occurring in the ORFs of the *kdpD* and *kefC* genes of *Salmonella* subspecies were investigated.

3.1.2.1 Identification of G-quadruplex motifs and *in vitro* characterization

During the course of our investigations we detected a potential quadruplex motif found in the ORF of the *kdpD* gene in the *Salmonella enterica subsp. enterica serovar typhimurium* strain LT2: 5'-GGCGTGGGGCTGGGGCTGGCG-3' *(see Table 3.1 sequence No. 1)*. Interestingly, this particular motif is different from the common quadruplexes as it bears 2 cytosines within a G-tract *(see Figure 3.12 A)*. Such a motif could either fold into a quadruplex with four tetrads, integrating the cytosines (*see Figure 3.12 A&B*), or it could fold into a quadruplex with three tetrads by bulging out the cytosines (*see Figure 3.12 C*). DNA G-quadruplexes bearing Cs in tetrads have been shown to be stable *in vitro* (292). Although we hypothesize the formation of a tetrad comprising 2 cytosines in *Figure 3.12 A*, such a potential structure was not confirmed by modeling and determination of lengths and angles of chemical bonds. Likewise structures with bulged out nucleotides have been reported to fold into stable structures (273). However, 16 different quadruplex structures with bulged out nucleotides are possible from this sequence, *Figure 3.12 B* shows one example. First, we characterized this kdpD motif for its structural properties and stability using CD spectroscopy and thermal denaturation measurements (*see Figure 3.12 B&C and D&E*). When stabilized by monovalent cations (25 mM K^+, 100 mM K^+, 25 mM Na^+, 100 mM Na^+ and 500 mM Na^+) the DNA oligonucleotide folds into an antiparallel G-quadruplex structure, showing the typical maximum at around 290 nm and the minimum at 265 nm. With increasing concentrations of potassium we observed spectral changes to a structure with two maxima (260 nm and

290 nm at 100 and 500 mM K^+) and a minimum at around 240 nm. Possibly, a structural change to a parallel quadruplex takes place at higher potassium concentrations. A structural transition from an antiparallel to a parallel fold has been shown for a quadruplex with a similar sequence (5'-GGGGCTGGGGCTGGGGCTGGGG-3') (20).

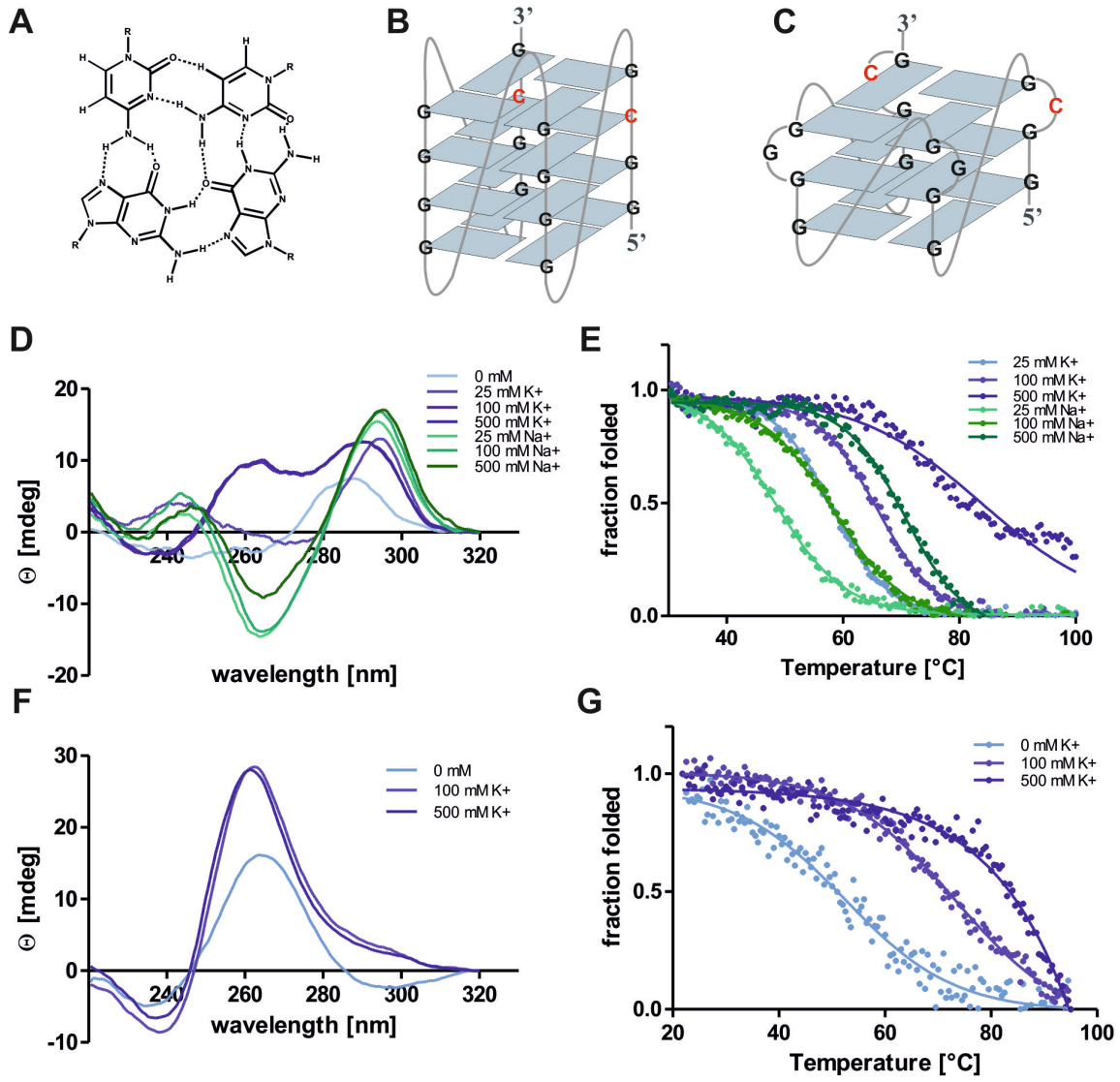


Figure 3.12: *In vitro* characterization of the kdpD quadruplex.

A Hypothetical tetrad of the kdpD motif with two Cs involved in the formation of a parallel structure as shown in **B**. **C** A potential parallel G-quadruplex structure with bulged out nucleotides (other structures are possible as well). **D** CD spectra of kdpD DNA oligonucleotide (5'-GGCGTGGGGCTGGGGCTGGCG-3') at different ion concentrations. **E** Thermal denaturation of **D** at 290 nm. **F** CD spectra of kdpD RNA oligonucleotide (5'-GGCGUGGGGCUGGGGCUGGCG-3') at different K^+ concentrations. **G** Thermal denaturation of **F**.

The melting temperatures for the respective DNA oligonucleotide are: 58±0.1°C, 65.9±0.1°C, 83.8±0.9°C for 25 mM, 100 mM and 500 mM K^+ and 48.3±0.2°C, 58.4±0.1°C, 70.7±0.1°C for 25 mM, 100 mM and 500 mM Na^+, respectively (*see Figure 3.12 E*). The corresponding RNA

oligonucleotide (5'-GGCGUGGGGCUGGGGCUGGCG-3') folds into a parallel G-quadruplex (*see Figure 3.12 F*), that is stabilized with increasing potassium concentrations (0 mM, 100 mM and 500 mM K^+) and has melting temperatures ranging from 54.4±0.8°C over 73.3±0.6°C to ≥ 95°C, respectively *(see Figure 3.12 G)*.

We were interested in whether G-quadruplex motifs occur in other ORFs as well and also whether those motifs are related to specific gene classes or similar genes of different organisms. Therefore, we screened the *E. coli K-12* MG1655 and *Salmonella enterica subsp. enterica serovar typhimurium* strain LT2 genomes for potential G-quadruplexes occurring within the ORFs. Again, we used the ProQuad Pattern Algorithm (104) and searched for G-quadruplexes with 3-5 tetrads and loops of 1-7 nucleotides: 13 potential sequences were found in each, *E. coli* K-12 MG1655 and *Salmonella enterica subsp. enterica serovar typhimurium* strain LT2. However, these sequences were different in lengths and nucleotide composition and did neither belong to certain gene classes nor occur within similar genes *(listed in Table 3.1)*.

Table 3.1: Potential G-quadruplexes within protein coding sequences.

Quadruplexes found in the ORF of *E. coli K-12* MG1655 and *Salmonella enterica subsp. enterica serovar typhimurium* strain LT2, according to the ProQuad pattern search (http://quadbase.igib.res.in/). Gene name and function of the quadruplex-containing gene are given according to NCBI. The respective quadruplex sequence is shown.

No.	Organism	Sequence (5' - 3')	Pattern length	Locus tag	Gene function
1	*Salmonella*	GGCGTGGGGCTGGGGCTGGCG	21	STM0703	*kdpD*: sensory kinase in two-component regulatory system with KdpE
2	*Salmonella*	GGGGAGGGCTGGGAACGGTAGGG	23	STM0332	putative hydrolase or acyltransferase
3	*Salmonella*	GGGAAATGCTGGGCAGCGTCGGGCGGGGG	29	STM0457	putative hydrolase
4	*Salmonella*	GGGTGACCGGGGCGGGGAAAGGG	23	STM0598	*entA*: 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase
5	*Salmonella*	GGGCACATGGGTATTGGGTGTGGGG	25	STM0920	Fels-1 prophage attachment and invasion protein
6	*Salmonella*	GGGTGACGCAGGGGAAGGGCGCGGG	25	STM1307	*torS*: sensory kinase in multi-component regulatory system with TorR (regulator) and TorT (periplasmic sensor)
7	*Salmonella*	GGGATGGGTGCGGGTAGAGGCGGG	24	STM1365	putative oxidase
8	*Salmonella*	GGGTGCCGTGGGTGTCCGGGCGGG	24	STM2515	putative hydrolase
9	*Salmonella*	GGGCCGGGTCCAGCGGGCGCGGG	23	STM3619	glycosyltransferase
10	*Salmonella*	GGGGCGGGACGGGCCTGGGG	20	STM3826	*torS*: sensory kinase in multi-component regulatory system with TorR (regulator) and TorT (periplasmic sensor)
11	*Salmonella*	GGGATGGGGATCGCGGGCGGG	21	STM4065	putative permease of the Na+:galactoside

No.	Organism	Sequence (5' - 3')	Pattern length	Locus tag	Gene function
					symporter family
12	*Salmonella*	GGGTCGGGCGGGAGGAGGG	19	STM4297	*melR*: regulator of melibiose operon
13	*Salmonella*	GGGATGCGGGCCAAAGGGCAGGG	23	STM4400	putative cationic amino acid transporter
14	*E. coli*	GGGGAGTTGGGGGAATAAGGGCGGAGGG	28	b0052	*pdxA*: 4-hydroxy-L-threonine phosphate dehydrogenase, NAD-dependent
15	*E. coli*	GGGCTGGGTGATGGGCTCGCGGG	23	b0089	*ftsW*: lipid II flippase; integral membrane protein involved in stabilizing FstZ ring during cell division
16	*E. coli*	GGGCGCGGGTCTGGGGCTGGTGGG	24	b0153	*fhuB*: fused iron-hydroxamate transporter subunits of ABC superfamily: membrane components
17	*E. coli*	GGGAATGCCAGGGCAGCGGGCATCTGGG	28	b0311	*betA*: choline dehydrogenase, a flavoprotein
18	*E. coli*	GGGTGGGGAGGGGGATGGGG	20	b0869	*ybjB*: putative NAD-dependent oxidoreductase
19	*E. coli*	GGGTTGGGGGCTGGGTTACTTCGGG	25	b1015	*putP*: proline:sodium symporter
20	*E. coli*	GGGTCAAGGGCTGGGCTTCGGG	22	b2630	*rnlA*: CP4-57 prophage; RNase LS
21	*E. coli*	GGGATGGGGTCCGGGTTGGG	20	b2647	*ypjA*: adhesin-like autotransporter
22	*E. coli*	GGGGATGGGAAAATCGGGGCATATTGGG	28	b3197	*kdsD*: D-arabinose 5-phosphate isomerase
23	*E. coli*	GGGCTGGGGCTGGGCGGG	18	b2455	*eutE*: aldehyde oxidoreductase, ethanolamine utilization protein
24	*E. coli*	GGGTGCCTGGGGACTGGCTGGG	22	b3560	*glyQ*: glycine tRNA synthetase, alpha subunit
25	*E. coli*	GGGCATCGGGGCGCTGGGTTGGG	23	b2789	*gudP*: putative D-glucarate transporter
26	*E. coli*	GGGCGGGTTGATGGGAACGGG	21	b1840	*yebZ*: inner membrane protein

Next, we were interested in whether potential quadruplex motifs were similar within genes of different *Salmonella* subspecies. In particular, we focused on motifs with the sequence pattern 5'-GGGGCTGGGG-3', as the quadruplex motif $d(G_4CT)_3G_4$ has been shown to form a stable G-quadruplex which is over-represented in proteobacteria and was also found within the ORF of *Salmonella enterica subsp. enterica serovar Gallinarium* (20). Using NCBI blast we found that these potential G-quadruplexes in different *Salmonella* substrains mainly occurring within the two genes *kdpD* and *kefC (see Table 3.2)*. Both of them are related to the potassium transport in cells. In enteric bacteria the potassium uptake is regulated by the major permeases Trk and Kdp as well as a minor permease Kup. KdpD is part of a two component signal transduction system that acts together with KdpE and controls the expression of the KdpFABC transporter. In *E. coli* and *Salmonella typhimurium* the KdpFABC systems are regulated in a similar manner (293). KdpD is a membrane-bound sensor kinase

whose autophosphorylation is affected by changes in the intracellular K^+ concentration. Phosphorylated KdpD transfers a phosphoryl group to KdpE which induces the KdpFABC operon, upon osmotic upshock and under K^+ limiting conditions (294). Different regulatory models for this system have been proposed (295-298). However, for *Salmonella* it is suggested that high osmolarity acts positively to induce the operon and K^+ functions negatively to repress it (293).

KefC is a glutathione-regulated potassium efflux system protecting the cell from electrophile toxicity. Potassium efflux by KefC in *E. coli* has been reported to be activated by adducts formed by the reaction of gluthatione with electrophylic compounds (299). KefC is a membrane protein consisting of a membrane domain attached to a C-terminal K^+ transport and nucleotide-binding domain via a flexible linker.

We decided to investigate the influence of the G-quadruplex on the expression of these two proteins in more detail. The potential quadruplex forming sequence 5'-GGCGTGGGGCTGGGGCTGGCG-3' of the *Salmonella kdpD* gene encodes for the amino acid sequence GVGLGLA (aa number 839-846, total protein length 894 aa). However, according to the codon usage database (http://www.kazusa.or.jp/codon/) this amino acid sequence is not the most frequent for the corresponding nucleic acid sequence. It is located at the C-terminus in the histidine kinase domain of the protein bearing the catalytic domain, which is responsible for the phosphorylation reaction. Sequence comparison of different histidine kinases for distinct subfamilies revealed that a part of this protein sequence seems to be conserved, namely GNGNGLN, where N stands for non-conserved amino acids (300).

The potential G-quadruplex motif 5'-GGGGCTGGGGCTGGGGCTGGGG-3' in the *kefC* gene of *Salmonella* encodes for the amino acid sequence LGLGLGLG and is located at the C-terminal end of the protein (aa number 533-541, total protein length 620 aa). This sequence does not belong to a conserved protein motif.

Table 3.2: G-quadruplexes occurring in ORFs of different Salmonella subspecies.

Quadruplexes found in the ORF of Salmonella species. Search was performed using the nucleotide BLAST webserver (http://blast.ncbi.nlm.nih.gov/). We applied the following parameters: **Algorithm**: "megablast" **Query sequence**: "GGGGCTGGGG", **Database**: "NCBI Genomes", **Organism:** "Salmonella (taxid:590)". Sequences with the potential to form G-quadruplexes were selected from hits. Gene name and function of the quadruplex-containing gene are given according to NCBI. The respective quadruplex sequence is shown.

Salmonella Strain	Accession number	Sequence (5'-3')	Pattern length	locus tag	Gene function
Salmonella enterica subsp. enterica serovar Gallinarum/pullorum str. RKS5078	CP003047.1	GGGGCTGGGGCTGGGGCTGGGG	22	SPUL_0091	*kefC*: glutathione-regulated potassium-efflux system
		GTGGGGCTGGGGCTGGAAATGGGGCTGGGGCTGGCG	36	SPUL_2259	*KdpD* Pseudo
Salmonella enterica subsp. enterica serovar Gallinarum str. 287/91	AM933173.1	GGGGCTGGGGCTGGGGCTGGGG	22	SG0700	*kdpD*: sensor protein KdpD
		GGGGCTGGGGCTGGGGCTGGGG	22	SG0088	*kefC*: glutathione-regulated potassium-efflux system
Salmonella enterica subsp. enterica serovar Heidelberg str. B182	CP003416.1	GTCCCCGGCGTGGGGCTGGGGCTGGCGATTT	31	SU5_01387	*kdpD*: sensor protein KdpD
Salmonella enterica subsp. enterica serovar Typhimurium str. 798	CP003386.1	GTCCCCGGCGTGGGGCTGGGGCTGGCGATTT	31	UMN798_0762	*kdpD*: sensor protein KdpD
Salmonella enterica subsp. enterica serovar Typhi str. P-stx-12	CP003278.1	CCCCCGGCGTGGGGCTGGGGCTGGCGATTT	30	STBHUCCB_22980	*kdpD*: sensor protein KdpD, hypothetical protein
Salmonella enterica subsp. enterica serovar Typhimurium str. UK-1	CP002614.1	CCGGGATGGGGTTTTCCAATGTGGGGCTGGGGCTGG	36	STMUK_2961	*fucO*: L-1,2-propanediol oxidoreductase
		GGCGTGGGGCTGGGGCTGGCG	21	STMUK_0709	*kdpD*: sensor protein KdpD
Salmonella enterica subsp. enterica serovar Typhimurium str. ST4/74	CP002487.1	GTTGCCGGGTAGGGGTTTTCCAATGTGGGGCTGGGGCTGG	40	STM474_3118	*fucO*: L-1,2-propanediol oxidoreductase
		TCGGGGGCTGGGGCTGTTGCCTGGGCAGGG	30	STM474_2745	gifsy-1 prophage RecE
		GGCGTGGGGCTGGGGCTGGCG	21	STM474_0725	*kdpD*: sensor protein KdpD
Salmonella enterica subsp. enterica serovar Weltevreden str. 2007-60-3289-1	FR775197.1	GGCGTGGGGCTGGGGCTGGCG	21	SENTW_0680	*kdpD*: sensor protein KdpD
Salmonella enterica subsp. enterica serovar Typhimurium str. T000240 DNA	AP011957.1	GTTGCCGGGTAGGGGTTTTCCAATGTGGGGCTGGGGCTGG	40	STMDT12_C30250	L-1,2-propanediol oxidoreductase
		TCGGGGGCTGGGGCTGTTGCCTGGGCAGGG	30	STMDT_C26580	Gifsy-1 prophage RecE
		GGCGTGGGGCTGGGGCTGGCG	21	STMDT12_C07600	*kdpD*: sensor protein KdpD
		GCGGGCGACAGGCGCTGGAGGCGCTGGGGCTGGGGCG	37	STMDT12_C00870	*kefC*: glutathione-regulated potassium-efflux system

We were curious if these potential G-quadruplex motifs could fulfill a regulatory role on mRNA level, e.g. by influencing translation and protein expression. What is especially interesting about these two proteins is that they are both involved in the regulation of the intracellular K^+ concentration. As described in *Chapter 1.1.1,* quadruplexes are stabilized by K^+ ions. Especially quadruplex motifs with moderate stability, such as the one found within the *kdpD* gene, might need stabilizing ions for stable structure formation (*see Figure 3.12*). In this case G-quadruplex formation might act as a potential negative feedback: high K^+ concentrations could trigger the formation of the secondary structure in the mRNA, causing ribosomal stalling (or the translation of a truncated and disfunctional protein) and finally resulting in downregulation of the kdpFABC operon, so that no more K^+ is transported into the cell. This is a daring hypothesis, and in order to explore the possibility of G-quadruplexes exerting such roles *in vivo*, we tested the protein expression of his-tagged *kdpD* and *kefC* genes in a plasmid system by Western blotting to gain first insights *(see Chapter 3.1.2.2).*

3.1.2.2 Construct design and Western Blot analysis

In a first attempt to investigate the function of G-quadruplexes in bacterial ORFs *in vivo* we designed plasmid constructs (*see Table 13.8 in the appendices, constructs 76-91)* containing variants of the *Salmonella kdpD* and *kefC* genes (*listed in Table 3.3*) and expressed them in *E. coli* cells (cloning procedures are described in *Chapter 7.14*).

Table 3.3: Quadruplex sequences and mutants used in different constructs.

Quadruplex sequences and designed mutants cloned with the respective gene into plasmid constructs and investigated via Western blotting. The quadruplex sequence (or mutant) is shown in the second column. Underlined Gs and Cs might participate in quadruplex formation. The column "Origin" describes the organism and the gene in which the quadruplex occurs in nature. The structure potentially adopted by the sequence is listed.

Name	Quadruplex sequence (5'-3') within ORF	Origin	Potential structure
kdpD	GGCGTGGGGCTGGGGCTGGCG	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 kdpD*	G4 with C in tetrad or G3 with bulges
kdpD M1	GGCGTTGGCCTCGGCCTCGCC	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 kdpD mutated*	no quadruplex or G2 with C in tetrad
kdpD *E. coli*	GGGGTAGGGCTTGGACTGGCA	*Escherichia coli MG1655 kdpD*	G2
kdpD M2	GGGGTGGGGCTGGGGCTGGCG	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 kdpD mutated*	G4
kdpD Gal	GGCGTGGGGCTGGGGCTGGAAATGGGGCTGGGGCTGGCG	*Salmonella enterica subsp. Enterica serovar Gallinarium/pullorum str. RKS5078 kdpD*	G4
kdpD Gal M1	GGCGTAGGACTAGGACTGGAAATGGGACTAGGACTGGCG	*Salmonella enterica subsp. Enterica serovar Gallinarium/pullorum str. RKS5078 kdpD mutated*	G2
kefC	GCGCTGGGGCTGGGGCTGGGGCTGGGGCGTTATGAA	*Salmonella enterica subsp. Enterica serovar Gallinarium/pullorum str. RKS5078 kefC*	G4CT
kefC M1	GCGCTAGGACTAGGACTAGGACTAGGACGTTATGAA	*Salmonella enterica subsp. Enterica serovar Gallinarium/pullorum str. RKS5078 kefC mutated*	G2
EutE	GCCGGGCTGGGGCTGGGCGGGGAA	*Escherichia coli MG1655 eutE*	G3
EutE M1	GCCGGACTAGGACTAGGCGGAGAA	*Escherichia coli MG1655 eutE mutated*	G2

According to studies performed in eukaryotic cells (*see Chapter 1.2.1*) we expected G-quadruplexes occurring within bacterial genes to induce ribosomal halt or frameshifting which would result in reduced gene expression or the production of a truncated protein. In our analysis whole sequences of quadruplex-carrying genes were cloned into the pBAD-18 plasmid under control of the arabinose-inducible araBAD promoter (for *kdpD* constructs

sequences of the *kdpD/E* operon were inserted into the vector). The protein of interest was 5' His tagged with a penta-His-linker, which enabled the detection on a Western blot by immunostaining with an anti-His-Antibody. To better assign the effect to G-quadruplex formation we designed different mutants, where the G-quadruplex sequence was mutated in a way that should allow no quadruplex formation or the formation of a less stable G-quadruplex (*see Table 3.3;* kdpD M1, kdpD Gal M1, kefC M1). However, we made sure that the amino acid sequence of the protein remained similar. Furthermore we created a construct containing the G-quadruplex within the *eutE* gene of *E. coli* MG1655 (*see Table 3.1 No. 23*) as this quadruplex was reported to produce a truncated protein product in a synchronized *in vitro* translation assay and to reduce fluorescence levels in a reporter system in mammalian cells (159). To ensure that there are no promoter-specific interactions, we additionally cloned the kdpD constructs under the control of the natural Salmonella promoter (kdpD (S)). For Western Blot analysis, bacteria were inoculated (1:500) from an outgrown culture in either LB or M9 medium supplemented with 1 mM arabinose. Cells were grown to exponential (OD_{600} = 0.3-0.6) or stationary phase (overnight) before proteins were isolated via sonification and separated with a denaturing SDS PAGE prior to blotting (*see Chapters 7.11.4.1 and 7.17*). In a first trial, we wanted to find out if there were differences between the expression levels and products of constructs under control of the different promoters: araBAD and the natural Salmonella promoter. To confirm adequate blotting, we applied the KlenTaq protein (size: 62.8 kDa) as a control. For both kinds of constructs we observed the expression of the correct protein (kdpD with His-Tag: 100.34 kDa) – however, higher expression levels were reached under control of the pBAD promoter (*see Figure 3.13 A*). Thus, all subsequent experiments were carried out with constructs under control of the araBAD promoter. Next, we investigated whether the correct products are achieved from the kdpD Gal, kefC and eutE constructs (*see Figure 3.13 B*). Although we observed a protein with the correct length for the kdpD Gal construct, no obvious change in protein expression level or protein size was observed in comparison to its mutant (which should form a less stable G-quadruplex structure). In addition, the KefC protein (68.23 kDa) was not expressed in our constructs according to the blot, the product of the eutE construct only yielded a faint band on the blot, and no truncated protein was visible. Next, we investigated the expression of the different kdpD mutants in both exponential and stationary phase as well as in LB *(see Figure 3.13 C&D*) and M9 (*see Figure 3.13 E&F*) media.

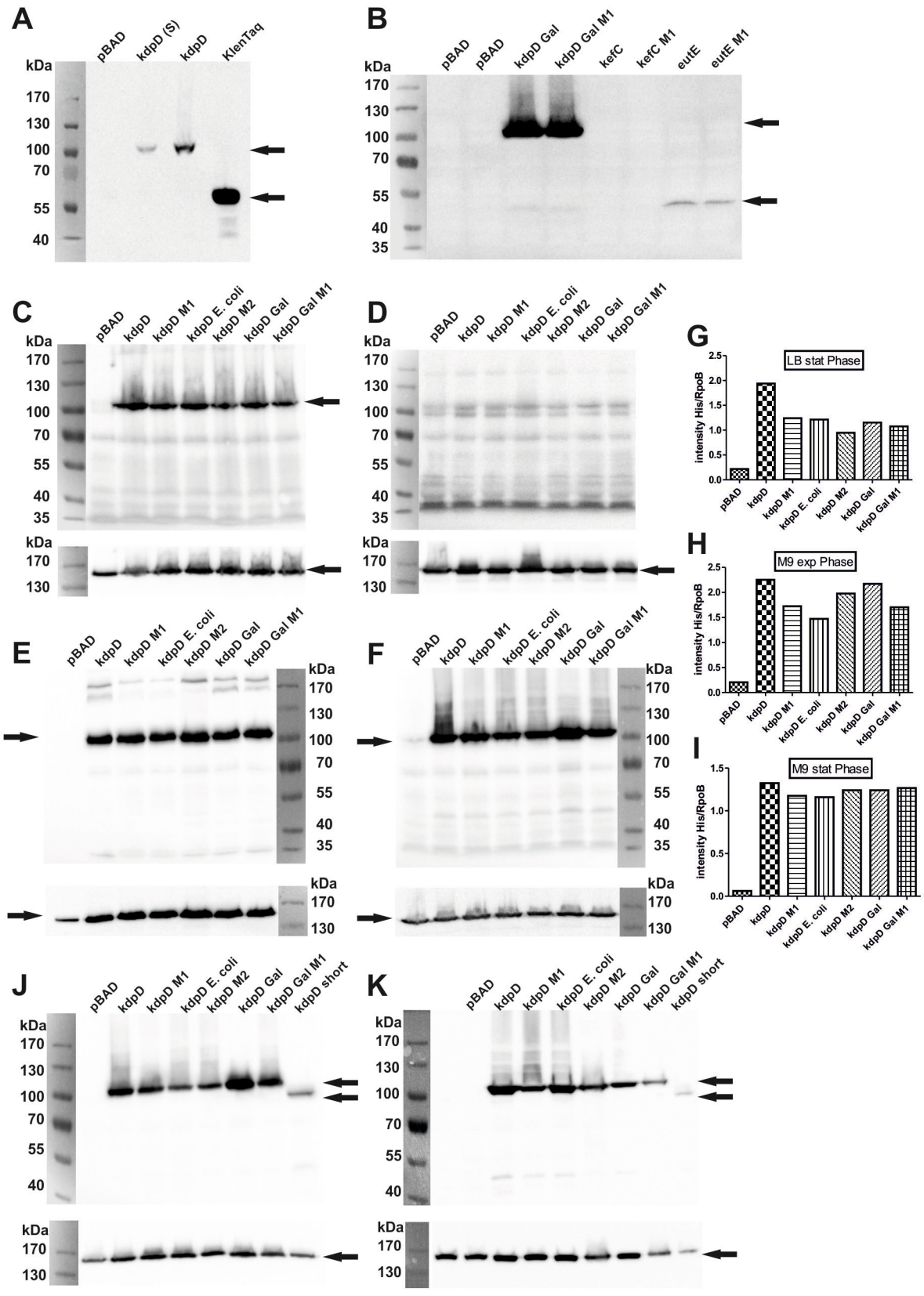


Figure 3.13: Western Blot analysis of kdpD constructs.

A Comparison of kdpD expression of constructs under control of the araBAD (kdpD) and Salmonella (kdpD (S)) promoters. pBAD refers to the empty vector without kdpD gene and KlenTaq refers to the KlenTaq protein used as transfer control.

B Protein expression of kdpD Gal, kefC and eutE constructs. **C, D** Protein levels of different kdpD constructs from cells grown in LB medium to **C** stationary phase and **D** exponential phase. The lower blot shows rpoB expression levels of the corresponding products. **E, F** Protein levels of kdpD constructs from cells grown in M9 medium to **E** stationary phase and **F** exponential phase. The lower blot shows rpoB expression levels of the corresponding products. **G, H, I** Blot evaluation by comparing the intensity of the kdpD detection to the intensity of the corresponding rpoB detection. **J, K** Protein expression of different kdpD constructs in comparison to the expression of a truncated protein, produced by introduction of a stop codon in front of the potential G-quadruplex sequence within the kdpD gene (kdpD short). The lower blot shows rpoB expression levels of the corresponding products. For all blots, pBAD dedicates expression level of the empty plasmid.

To quantify the band intensity and prove equal loading, we immuno-stained the blot after a stripping procedure (*see Chapter 7.18*) with an anti-RpoB-antibody (*rpoB* encodes for the β subunit of the RNA polymerase and is a stably expressed housekeeping gene). However, results were inconclusive. When bacteria were grown in LB medium, the KdpD protein was only expressed, if the culture was incubated overnight and the proteins were isolated from stationary phase cells (*see Figure 3.13 C*). An evaluation of the band intensity (*see Figure 3.13 G*), revealing KdpD expression relative to the RpoB expression showed highest expression level for the construct bearing the natural kdpD gene. Interestingly, expression levels for all KdpD mutants bearing less stable G-quadruplexes (kdpD M1, kdpD *E. coli*) were lower compared to the natural gene. However, even the expression levels of the mutant – which should form a more stable G-quadruplex (kdpD M2) – and for those of the kdpD Gal construct were lower compared to the natural gene. When grown in M9 media cells with all kdpD constructs showed protein expression in stationary as well as exponential phase (*see Figure 3.13 E&F*). Whereas in stationary phase all constructs had almost equal expression levels of KdpD, in exponential phase the constructs with the most stable G-quadruplexes showed the highest protein expression (kdpD, kdpD M2, kdpD Gal) when evaluated relative to RpoB expression (*see Figure 3.13 H&I*). Enhanced protein expression might also result from G-quadruplex formation: The formed G-quadruplex might induce translational halt which could facilitate complex protein folding and thereby contribute to the formation of a stable protein. On the blot we observed very thick bands for the kdpD and the kdpD Gal constructs (*see Figure 3.13 F*), which raised the question whether these bands include a second, lower band that is not properly separated and could result from a truncated protein. Thus, we designed a control construct bearing a stop codon 4 nt in front of the G-quadruplex sequence in the kdpD gene (kdpD short). *Figure 3.13 J and K* show two biological replicates analyzed on two blots. The band of the truncated protein can be clearly distinguished from the kdpD bands, and therefore the production of a truncated protein could be excluded. However, a comparison of the two blots (*see Figure 3.13 J&K*) showed that expression levels differ a lot between different experiments. Thus, the results are ambiguous and no conclusions could be drawn. Further experiments should be performed in order to determine the ability of these potential G-quadruplexes to interfere with translation.

3.1.3 Discussion

Several studies have described quadruplex-mediated alterations in gene expression for eukaryotic systems. Although in computational searches quadruplex-forming sequences have been found enriched in regulatory regions of prokaryotes, studies that systematically investigate their influence *in vivo* are sparse. In 2012, Chowdhury and co-workers reported the influence of G-quadruplex-stabilizing compounds on the radioresistance of the bacterium *Deinococcus radiodurans*. As yet, their study is the only one investigating the influence of G-quadruplexes occurring in bacterial promoter regions on gene expression *in vivo*. However, in their case quadruplexes were enriched in the promoter region of a functional gene class in this respective organism (43). In our studies, we aimed at a more generalized examination by systematically analyzing the influence of G-rich sequences in bacterial gene-regulatory regions. To our knowledge this is the first comprehensive study to show that the effect of G-quadruplexes in the bacterial promoter region on gene expression of the downstream gene is position-dependent. Recently, the influence of strand asymmetry on quadruplex-mediated alteration of transcription was described for eukaryotes by the group of Maiti (301). In their study, a quadruplex sequence in the 5'-UTR only repressed transcription efficiency when placed in the antisense strand. However, translational repression of gene expression was also possible when the G-quadruplex was found in the sense strand (301). The comparison of prokaryotic and eukaryotic systems in this context might prove difficult as genetic mechanisms differ significantly. Hence, conclusions drawn from studies in eukaryotic contexts are not necessarily valid for bacteria; instead separate investigations are necessary.

Our results illustrate that G-quadruplexes can be involved in bacterial gene regulation on both transcriptional and translational levels. We have set up reporter systems based on two different plasmids bearing either eGPF or β-galactosidase reporter genes. First, quadruplex sequences were inserted within the core promoter region (between the unaltered conserved -35 and -10 regions). We found that quadruplex-forming sequences in the core promoter region significantly decreased gene expression by transcriptional modulation when located on the antisense strand, whereas insertion into the sense strand showed much less influence. Introducing different quadruplex sequences into the antisense strand at this position showed that transcriptional repression correlates with quadruplex stability. The non-coding or antisense strand serves as template for the *E. coli* RNA polymerase. Although the transcription start site is located downstream of the promoter region, polymerase binding to the promoter is essential for transcription initiation. Usually, the *E. coli RNA* polymerase core enzyme binds to the σ^{70} factor to form the holoenzyme. The sigma factor is responsible for

promoter recognition. It identifies the -10 and -35 region of the double stranded promoter DNA, forming the closed promoter complex. A σ subunit (σ_2) separates the strands of the DNA at the -10 region and binds to the sense (non-template) strand, forming the open promoter complex. Transcription is then initiated by binding of a nucleotide triphosphate (ATP) to the nucleotide +1 at the TSS. An initial transcript of about 10 nucleotides causes the release of the sigma factor from the core RNA polymerase. Finally, the RNA polymerase leaves the promoter region and enters the elongation phase in which the transcription bubble is enlarged to 17 bp and the polymerase moves along the antisense strand (279,302). Promoter recognition can be strongly influenced by the nucleotide composition of the surrounding 5'-UTR: Upstream elements can enhance the efficiency of RNA polymerase binding. The formation of a secondary structure, like a G-quadruplex, could create a physical barrier that hinders polymerase binding or complicates promoter recognition by σ^{70}. In another scenario, polymerase binding could facilitate quadruplex formation, which ultimately might hamper the initiation of transcription or the elongation phase.

In order to investigate the influence of G-rich sequences downstream of the conserved promoter in the 5'-UTR we inserted quadruplexes into the sense and antisense strands 20 nt upstream of the start codon. An increase of gene expression was observed for G-quadruplexes placed into the antisense strand and a repression of gene expression for those inserted in the sense strand. Again, we investigated quadruplexes with different thermodynamic stabilities and observed a correlation between their effect and the quadruplex stability. Compared to controls, mRNA levels remained constant when G-stretches were placed in the sense strand, but were enhanced in constructs bearing the G-sequence in the antisense strand. This points to translational versus transcriptional control of gene expression. The G-quadruplex on the antisense strand influences transcription and might interfere with polymerase elongation or binding. However, the reason for the enhancement of gene expression due to a quadruplex at this position is unclear. A possible explanation could be that *E. coli* RNA polymerase needs to separate the template and non-template strands for the transcription process. Here, G-quadruplex formation competes with stable G-C Watson-Crick base pairing in the DNA double strand. Formation of a G-quadruplex structure could facilitate strand separation and thereby support the helicase activity of RNA polymerase. Strands are not separated in the core promoter and the polymerase can bind to the double strand, but in the region downstream of the -10 part the polymerase actively separates the double helix (279,302). This might explain why G-quadruplex, when inserted in the promoter, results in transcriptional repression, whereas its insertion downstream of the promoter increases gene expression. *E. coli* RNA polymerase has been shown to interact similarly with different promoters (303). Importantly, we observed a similar behavior in two different

σ^{70}-dependent promoters, the constitutive J06 promoter and the araBAD promoter activated by arabinose-induced binding of araC from position -35 to -51 (277,278). However, we cannot exclude the possibility that G-quadruplexes in other systems behave differently.

After having investigated transcriptional regulation by G-quadruplexes in detail in the discussed constructs, we turned to insertions of quadruplexes into transcribed, mRNA-based regions of the 5'-UTR. Especially the initiation of translation seems to be strongly influenced by quadruplex formation, as previously shown in a series of artificially designed, SD-masking expression constructs resulting in down-regulation of translation (161). Masking of the ribosomal binding site is a common mechanism for translational regulation of gene expression, e.g. in riboswitches, RNA thermometers, sRNA-mediated regulation and in artificially designed riboregulators. Different systems have been described using engineered devices as sensitive switches of gene expression in prokaryotic organisms (304-306). In this study, we successfully constructed a system where the SD was masked by means of a hairpin structure. A quadruplex sequence was incorporated into the loop and part of the stem structure so that quadruplex formation destabilized the hairpin structure and thereby liberated the ribosome binding site, effectively resulting in activation of gene expression.

Having shown that quadruplexes located in the vicinity of the SD site are able to strongly influence translation, we wondered whether quadruplexes occur at these positions in natural genetic sequences. Searching the *E. coli* genome, we found 46 putative G-quadruplexes occurring on the coding strand within the SD region. We investigated the influence of these naturally occurring G-rich stretches within the SD region on gene expression and observed a significant quadruplex-mediated repression for two of the five naturally occurring 5'-UTR regions we investigated. The G-rich region of the 5´-UTR of the *oxyR* gene was examined in detail and compared to several mutants which should not be able to form a G-quadruplex. The upstream regions taken from two further genes did not significantly alter gene expression compared to controls. As suggested earlier (161), the secondary structure of a G-quadruplex might complicate the binding of the ribosome to the SD region and thereby decrease gene expression efficiency. However, for the quadruplex near the SD site occurring in the *napH* 5'-UTR we observed an increase of gene expression compared to two controls. As described earlier, the overall 5'-UTR nucleotide composition has a strong impact on translation (284). Also, in our engineered system, the sequence surrounding the SD region plays an important role in quadruplex-mediated translational modification. Competing secondary structures can be dissolved by quadruplexes and thus enable the access to the SD region. As we have no concrete knowledge of the exact sequence of the SD region in this case, it is difficult to draw mechanistic conclusions.

We also tested the influence of the quadruplex stabilizing compound NMM, but no effects on gene expression were observed in our systems. Although the use of quadruplex-stabilizing compounds is common in this field, *in vivo* most of them have been applied to eukaryotic cells (e.g. yeast (15)). Regarding bacteria, NMM was reported to stabilize G-quadruplex structures in *Neisseria gonorrhoeae* (44) and *Deinococcus radiodurans* (43). However, the effects of NMM on G-quadruplexes in *E. coli* in particular have not been shown so far. Also, the addition of other compounds (TMPyP4, 360A) did not lead to an intensifying effect on gene expression changes. It is possible that those compounds are not properly absorbed in *E. coli* species – prokaryotic uptake mechanisms of said compounds have not been described so far. Furthermore, there is no information on intracellular degradation and half-lives. Consequently, the incubation time of the compound might prove critical as well. Potentially, as fluorescence and luminescence molecules have long half-lives our read-out systems may not be suitable for detecting an enhanced effect in gene expression changes. So far reporter systems used in combination with quadruplex-binding ligands in prokaryotes differed from ours and were also more sensitive (radioactivity (43), phase-variation assay (44)).

Finally, we showed that G-quadruplex sequences inserted 4 nt after the stop codon of a reporter gene did not result in a consistent modulation of gene expression. The findings about G-quadruplex functions in regulatory regions are summarized in *Figure 3.14*.

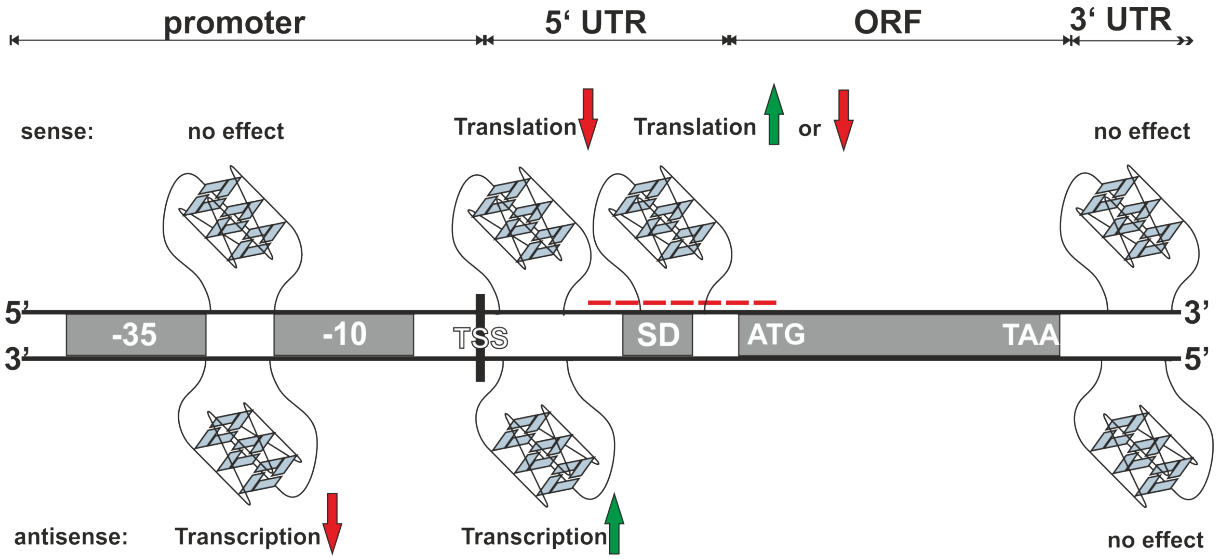


Figure 3.14: Summary of effects mediated by G-quadruplexes in regulatory regions.

Different quadruplex insertion sites and the respective effects are shown. Red arrows pointing up: increased gene expression; red arrows pointing down: decreased gene expression. Dashed red line indicates the sequence range which was modified for investigation of the SD adjacent region.

Apart from G-quadruplexes in regulatory regions, we also conducted first experiments investigating G-quadruplexes within bacterial ORFs. Sugimoto and co-workers published a series of experiments that showed G-quadruplexes at these positions to be responsible for ribosomal halt, frameshifting or the production of a truncated product. Although they did *in vitro* studies with G-quadruplex sequences found in the *E. coli* ORFs, all *in vivo* studies were performed with eukaryotic cells. The aim of our investigations was to allocate an *in vivo* function to G-quadruplexes within bacterial ORFs. We found two different potential G-quadruplex motifs present in K^+ transporters of *Salmonella* subspecies. After *in vitro* characterization we decided to design constructs that express these genes with the potential G-quadruplexes (or mutated versions) and investigate them via Western blotting. However, our results were inconclusive: We did neither observe the formation of a truncated or elongated protein product nor consistent enhancement or repression of protein expression. Although our *in vivo* investigations were conducted in a bacterial system, we investigated a sequence derived from *Salmonella* in *E. coli*. *E. coli* and *Salmonella enterica* species can be considered phylogenetically related (307) and share a large amount of their genomic material (308,309), however, they display lifestyle divergences (310,311) and it can only be assumed that gene regulatory mechanisms are similar. Also, in *Salmonella* subspecies specific proteins could exist that recognize quadruplex structures, but are not present in *E. coli*. To exclude organism-specific effects *in vivo* analysis in *Salmonella* should be conducted. Nevertheless, we have only just begun the investigations in this field, and there are lots of possibilities to improve the experimental procedures. Western blot analysis is not only very time-intensive, but also has a rather low sensitivity, potentially making it impossible to detect small portions of truncated or elongated proteins. The design of an artificial system where the potential G-quadruplex would occur within a read-out gene (e.g. eGFP) would be an option to facilitate the analysis and to test different G-quadruplex sequences for their function within ORFs. One could also try to intensify the G-quadruplex-mediated effect by stabilizing the structure with G-quadruplex-interacting compounds. First hints as to the function in bacteria could also arise from an *in vitro* translation assay.

In conclusion, the comprehensive study presented here gives new insights into quadruplex-mediated regulation of gene expression in *E. coli.* We were able to show that both the strand orientation and the exact position of a G-quadruplex in the 5'-UTR strongly influence its effect on transcription. Translational alterations are also dependent on the position and the surrounding sequence of the G-quadruplex in the 5'-UTR. Although the presented data do not show a direct role of natural quadruplexes in gene regulation, we cannot exclude the possibility of quadruplexes playing functional roles in controlling gene expression. In such a scenario it might be possible that these distinct structures are specifically induced under

certain conditions. It is important to note that intracellular K^+ concentrations increase in response to osmotic upshock and environmental stresses (270,312). Intriguingly, G-quadruplexes are stabilized by monovalent ions showing the highest affinity for K^+. We found several G-quadruplexes in the 5'-UTRs of genes related to stress responses (*see Table 13.1 in the appendices*). OxyR, the oxidative stress regulator, is a transcriptional regulator in the oxidative and nitrosative stress response (313,314). *RelA* encodes for an enzyme involved in the stringent response which activates the synthesis of the regulatory molecules Guanosine-3'-5'-bis(diphosphate) (ppGpp) and Guanosine-3'-diphosphate-5'-triphosphate (pppGpp), both acting as alarmones to amino acid starvation (315). RseA can inhibit and regulate the sigma E factor (316). The Sigma E system is involved in the responses to heat shock, osmotic stress or other stresses on membrane and periplasmic proteins. Other genes related to environmental stress have been identified in our search (*see Table 13.1 in the appendices*). Interestingly, also the genes found in within the ORFs of *Salmonella* species were related to the K^+ transporter (*kdpD*) expressed upon osmotic stress. One could speculate that the identified quadruplex motifs might function as regulatory units responding to stress or other environmental changes. However, stress responses and bacterial lifestyle changes are regulated by several complex and overlaying pathways. This makes it difficult to prove the formation and influence of a quadruplex structure which, in addition, might only form temporarily. We recently carried out initial experiments with osmotic up-shock that should have resulted in temporarily increased intracellular K^+ levels, but found no conclusive influence in reporter gene assays (data not shown). However, further experiments along these lines utilizing even more suited reporter gene assays might be able to shed more light on the possibility of quadruplex formation as a natural mechanism for conditional gene regulation. Also, one should keep in mind that the presence of a G-quadruplex motif in one strand is inevitably tied to the presence of a C-rich pattern, the so called i-motif, in the complementary strand. Although the formation of i-motifs is only reported for lower pH (317), it is possible that they might hold a functional role as well.

3.2 The intrastrand triplex motif "TM" in *E. coli*

Nucleic acid repeat sequences are abundant in prokaryotic and eukaryotic genomes, and most of them are prone to form secondary structures distinct from the B-DNA. Prominent examples for such structures are DNA triplexes, which can form intra- or intermolecularly via Hoogsteen base pairing. This chapter describes a particular type of intrastrand triplex motif, the TM which occurs in different genomes of bacterial and archaeal species and is significantly enriched in *E. coli*. This motif has been described in earlier studies, but its function was never elucidated.

3.2.1 Intrastrand triplex motifs in bacteria

Different studies have identified triplex motifs in eukaryotes and prokaryotes by means of computation. Most algorithms search for TFO binding sites (162-164), potential triplex target sites (165), or focus on inverted repeats (166,167) and H-DNA (168,169). As described above (*see Chapter 1.1.2*), intramolecular triplexes do not necessarily have to form H-DNA within a DNA double strand, but can also occur within one single-stranded DNA oligonucleotide. Databases with selective search functions for such intrastrand triplex motifs are rare. In 2000, Maher and co-workers used the Palingol program (318) to search for intrastrand triplexes, describing them as two hairpins sharing a common homopurine strand. They defined 4 different triplex classes: Class I and II contain purine motif triplexes with reverse Hoogsteen bonding, whereas class III and IV contain pyrimidine motif triplexes stabilized by Hoogsteen bonds (*described in Chapter 1.1.2.1*). Using their search strategy, they found representative intrastrand triplex motifs in the genomes of *E. coli K-12*, *Synechocystis sp.* and *Haemophilus influenza*, with the class II motif being the most abundant (10). However, they did not search other prokaryotic genomes. In our studies we aimed at a more general and simplified identification for intrastrand triplex motifs. Thus, we developed a search algorithm for finding potential intrastrand triplexes among the different triplex classes in prokaryotes. Our Intrastrand Triplex Finder (ITxF) database contains 5,246 different genomes of bacterial and archaeal species, based on fully sequenced genomes and plasmids from the NCBI webserver (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The basic layout performs searches for homopurine-homopyrimidine regions that are either A-, T-, C- or G-rich, defining the three stems of the triplex structure. The regions in-between are defined as loops and can contain any nucleotide. The search identifies potential triplex structures

with a stem size of 5-12 nt and a loop size of 1-6 nt. As occurrence of imperfect triplexes has been described in different studies (58-60), our algorithm allows one mismatched base pair in the triplex if the stem length has a minimal size of 7 nt. Using the ITxF program, we identified large numbers of A/T- and G/C-rich triplexes in different prokaryotes: In total, 2,485,777 triplex sequences were found within the 5,246 analyzed genomes and plasmids (examples for *E. coli* subspecies are shown in *Table 13.2 in the appendices*). When looking for a specific type of triplex only, it is easy to choose the nucleotide composition of the stem region and define certain stem and loop lengths in the database. The program also shows the triplex class (class I to IV) for specific sequences. Analyzing all genomes (including plasmids), we found that class II triplexes are the most abundant: 40.8% of the triplexes found belonged to class I, 46.6% to class II, 7% to class III and 5.6% to classIV. When Hoyne et al. searched for intrastrand triplexes in *E. coli* K-12, *Synechocystis sp.* and *H. influenza* they only found small numbers of triplex motifs in total (25, 18 and 21). When we performed the search in the ITxF database, we found much higher number of triplex sequences: 431 triplexes in *E. coli* K-12 MG1655, 652 triplexes in *Synechocystis* PCC 6803 and 824 triplexes in *Haemoophilus influenzae* Rd KW20. However, our search strategy is completely different to the one of Hoyne et al.: Whereas Hoyne et al. defined their triplexes via a pattern recognition program searching for hairpin structures, we used a new algorithm specifically aimed at intrastrand triplexes. Hoyne et al. searched for triplexes with a stem length of perfectly matched triplets of greater than 7 nt and loops from 0 to 10 nt; whereas we searched for triplex structures having a stem length from 5 to 12 nt allowing a mismatch when larger than 7 nt and having loops with a size of 1-6 nt. In contrast to Hoyne et al., we found sequences with the potential to form class I triplex structures present in all 3 species, but – similar to their findings – class II triplexes were the most abundant. Furthermore, the program shows the orientation of the identified triplex structures within the circular genome. The ITxF database is available online at http://bioinformatics.uni-konstanz.de/utils/showtriplex/. To our knowledge, it is the first database that allows searches for intrastrand triplexes (not necessarily H-DNA) in prokaryotes. The abundance of triplex motifs in bacteria suggests potential regulatory, organizational or adaptive functions, as proposed by others (56). During our searches, our interest was drawn to one particular motif: 5'-CCCTCTCCCCTTTCGGGGAGAGGGTTAGGGTGAGGGG-3', which is the consensus sequence of a purine motif triplex (class II) bearing a C-rich stem (9 nt), a first loop (3-6 nt), a first G-rich stem (9 nt), a second loop (3 nt), a second G-rich stem (9 nt) and one mismatched base pair in the stem (*see Figure 3.15*). This potential triplex motif, in the following named TM, has already been described in an earlier study by Maher and co-workers, but so far no function was assigned. In contrast to earlier publications (10,85) and due to the enormous amount of available prokaryotic genome sequences, we found the TM

in 174 proteobacterial genomes *(see Table 13.3 in the appendices*): With a total number of 192 TM sequences, *Herpetosiphon aurantiacus ATCC 23779* contains the most potential triplex motifs of this type. Closely related genera like *E. coli* and *Shigella* species carry comparable copy numbers, whereas other genera such as *Enterobacter* include far more TM sequences (up to 175). Intriguingly, in some closely related species, such as *Salmonella,* the TM is absent. However, a search in our database yields results for other triplex-forming sequences in these genomes. Comparing all analyzed strains, we found that most bacterial genomes contain less than 30 TMs (*see Figure 3.15 D*). We were interested in the function of this particular TM and decided to characterize it in *E. coli K-12 substrain* MG1655 in more detail.

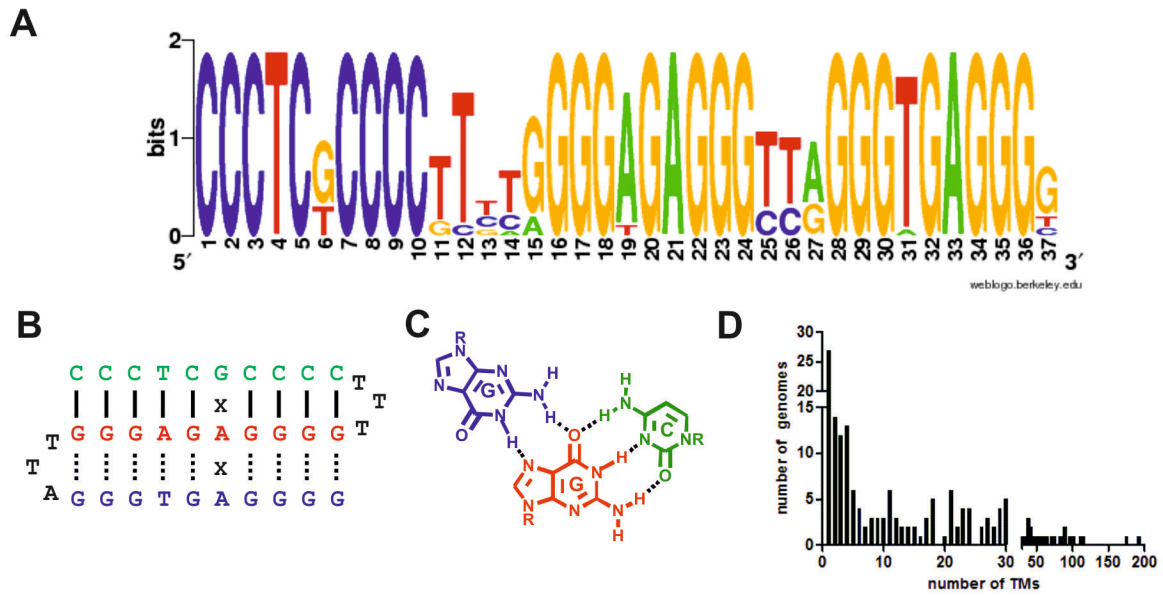


Figure 3.15: The TM sequence in prokaryotes.

A Consensus motif of TM sequences found on the *E. coli K12* MG1655 genome. **B** TM sequence folding into a DNA class II triplex motif. Hoogsteen hydrogen bonds are indicated by dashed lines. (modified from Maher et al. (10)). **C** DNA triplets found in the TM motif. **D** Frequency of TM sequences found in different proteobacterial genomes (*listed in Table 13.3 in the appendices*).

3.2.2 Structural characterization of the "TM"

In order to confirm that the identified TM forms a stable triplex structure, DNA oligonucleotides were characterized via circular dichroism (CD) and nuclear magnetic resonance (NMR) spectroscopy (*sequences listed in Table 7.2*). Characteristic CD spectra of DNA triplex structures differ depending on the sequence of the oligonucleotide (272). However, most of the intramolecular triplex DNA shows a minimum around 240 nm, a maximum around 257 nm and a second minimum at approximately 280 nm (79,319). These peaks were also observed in the CD spectra of the TM oligonucleotides. We measured CD spectra for two types of TM: TM type A with the sequence 5'-TTA-3' and TM type B with the sequence 5'-CCG-3' in the second loop. We investigated both TM types with and without one mismatched base pair (mm), respectively (see *Figure 3.16 A*). Furthermore, we analyzed a control sequence (5'-CCCTCGCCCCTTTGCCGAGAGCGTTAGCGTGAGCGG-3') containing four G-to-C mutations and should not be able to form a triplex – this sequence yielded spectra that resemble duplex (B-DNA) structures (*see Figure 3.16 A and 3.16 G*) (272). We proved the structures to be very stable as CD spectra showed the characteristic peaks up to a temperature of 75°C, although CD signatures decreased with increasing temperature (*see Figure 3.16 C-G*). Next, we determined the stability of the TM oligonucleotides (5 µM) by thermal denaturation studies: melting temperatures of 82±4°C, 78±1°C, 74±1°C and 70±1°C were determined for TM type A, TM type B, TM type A mm and TM type B mm, respectively, as shown in *Figure 3.16 H*. Although CD spectra showed minima and maxima that were observed for triplex structures before, characteristic peaks for parallel G-quadruplex structures are very similar (minimum at 240 nm and maximum at 260 nm). The ability of G-rich triplex sequences to fold into quadruplex structures is well known (63) and competes with triplex formation: the G-rich part of the TM motif could, in principle, form an intramolecular G-quadruplex with 3 tetrads (*see Figure 3.17 B*). To exclude quadruplex formation and prove triplex folding, we carried out NMR measurements. The ^{1}H-NMR spectra of TM oligonucleotides display 18 sharp signals in the imino proton range that clearly demonstrate the formation of well-defined triplex structures (*see Figure 3.16 B*). If an intramolecular G-quadruplex structure formed we would expect much less imino proton signals (3-4 signals). However, when complementary strands were added, CD signals characteristic for duplex structures were observed, and the NMR spectrum of TM type B mm showed less and broader signals in the imino proton range (*see Figure 3.17 C&E*).

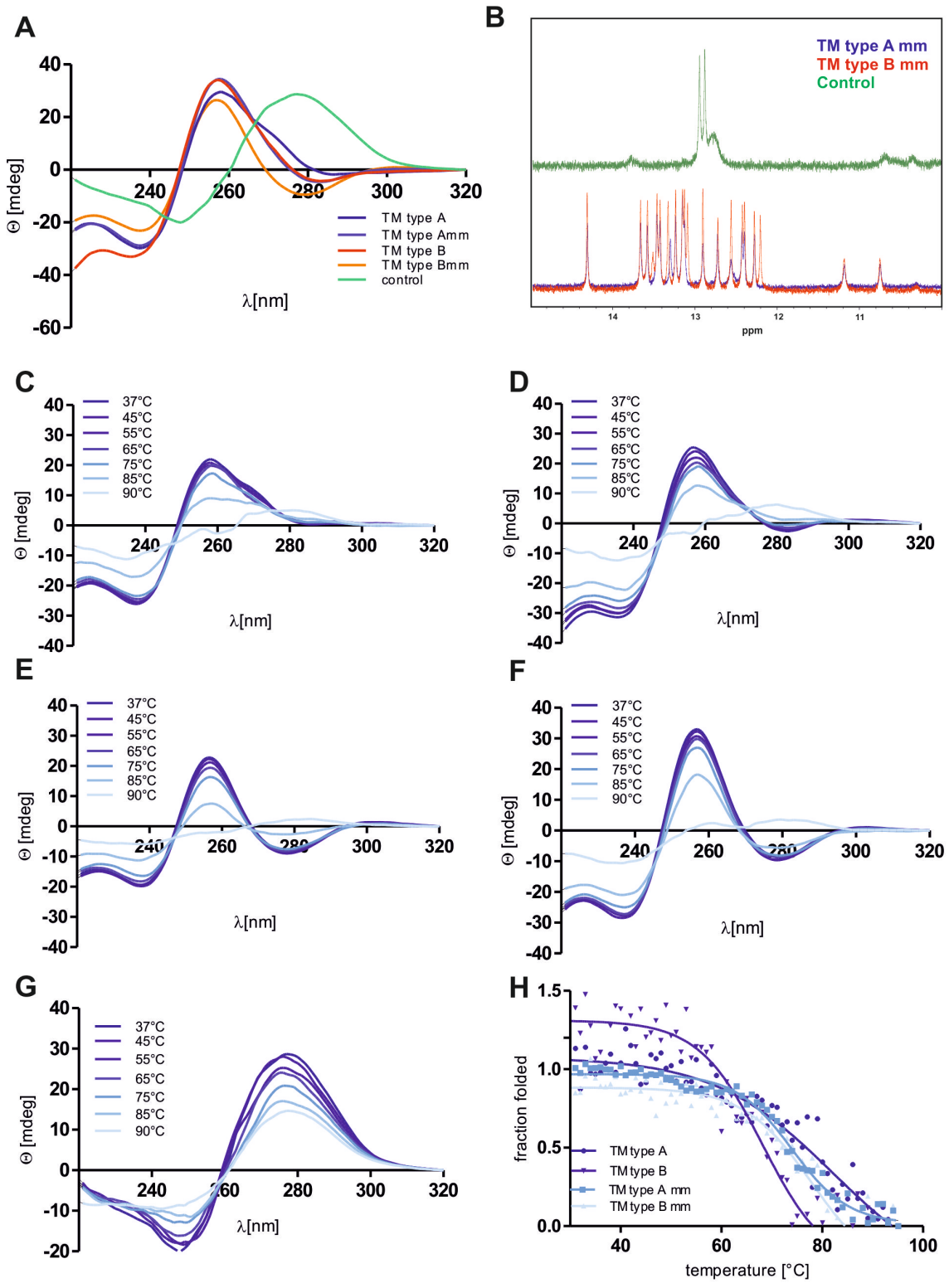


Figure 3.16: Spectroscopic analyses of the TM.

A Circular dichroism spectroscopy of TM (TM type A and A mm in shades of blue, TM type B and Bmm in shades of red) and control (green) oligonucleotides. **B** Imino proton area of ^{1}H-NMR spectra of TM and control oligonucleotides. **C-G** Circular dichroism spectroscopy at different temperatures of **C** TM type A complete match **D** TM type B complete match **E** TM type Amm **F** TM type B mm **G** control. **H** Thermal denaturation studies of TM type A, B, Amm and B m at 257 nm. Some of the data were kindly provided by Malte Sinn and Stefanie Wagner.

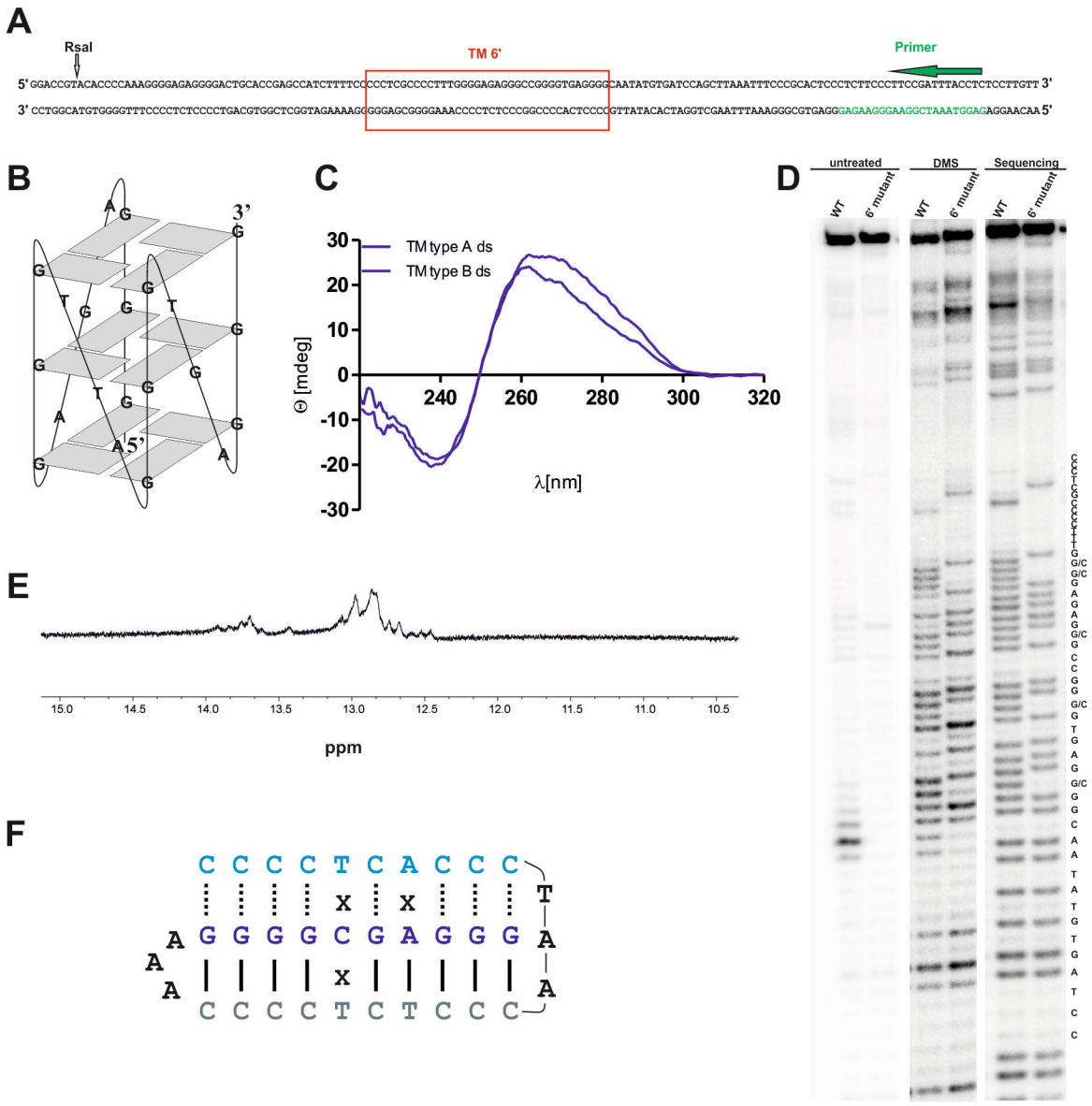


Figure 3.17: Structural characterization of the TM.

A Schematical illustration of primer extension reaction for DMS footprint. Red Box indicates TM motif, primer binding site is marked in green and RsaI restriction site is shown in black. **B** Schematical illustration of the quadruplex motif that could be formed by the G-rich stretches of the TM sequence. **C** Circular dichroism of the double stranded TM type A and TM type B. **D** *In vivo* DMS probing of the TM sequence found at the genomic site 6' in comparison to the 6' mutant: Primer extension reaction is described in **A** and analyzed on a 10% denaturing PAGE. DMS-treated probes (DMS) are shown in comparison to non-treated DNA (untreated) and the sequencing reaction for purine sequencing, according to Maxam and Gilbert (320). **E** Imino proton range of ^{1}H NMR of the double stranded TM type B mm. **F** Conformation of a pyrimidine motif class III triplex that could potentially be formed by the C-rich antisense strand.

In the bacterial cell, triplex conformations do not necessarily have to occur permanently, but could also arise temporarily, e.g. when negative supercoiling favors strand separation and

the formation of non-duplex DNA structures (321). In order to investigate whether chromosomal sites containing TM sequences are double-stranded or fold into alternate conformations in the living bacterium, we performed *in vivo* footprinting. Bacteria were incubated with dimethyl sulfate (DMS) that selectively methylates guanines at N7. The central G-rich stretch of the TM (*shown in red in Figure 3.15 B*) should be protected in a triplex fold, whereas N7 should be methylated in the duplex form at the respective locus. After DNA isolation and cleavage at methylated positions, the DMS-accessible sites were identified via a primer extension reaction (*see Figure 3.17 A*). *Figure 3.17 C* shows the footprinting reaction of the TM site at the 6' position of the *E. coli* chromosome (TM No. 3, a type B triplex containing 1 mm, *see Table 3.4*) in comparison to a genomic mutant containing G to C exchanges (6' mutant). However, we observed cleavage at the respective guanine sites in the WT strain with band intensities comparable to the control. Interestingly, a strong band can be observed directly in front of the triplex sequence which was not present in the mutated chromosome (*see Figure 3.17 C*: untreated WT versus 6' mutant). This might result from enhanced cleavage at this site or polymerase stop during primer extension, possibly resulting from a temporary triplex formation.

3.2.3 The "TM" sequence in E. coli

Larger sequences containing the TM have been described earlier for *E. coli* by means of bioinformatic searches for repeats or hybridization assays, and were named BoxC (244,322-324). BoxC regions have been defined as 56 bp long imperfect palindromes, occurring 32 times in the *E. coli K-12* chromosome. About 10 years later, a shorter version of this sequence (approximately 36 nt) was re-discovered as PIT element, occurring 25 times in *E. coli* K-12 (10) and investigated for potential functions (85). We now characterized the TM, which is partly identical to PIT and BoxC elements. However, those earlier studies were not able to prove triplex formation *in vivo* or to assign a function to the motif *in vivo*. The TM has a length of 33-37 nucleotides and one mismatched base pair at most, and it was found 23 times in the *E. coli* K-12 MG1655 chromosome (*sequences listed in Table 3.4*). The TMs always occur intergenic with no bias to strand orientation (*see Figure 3.18*). Using NCBI BLAST (http://blast.ncbi.nlm.nih.gov) (325), we did not find an association of the motif with high mobility genetic elements such as transposons, phages or plasmids. The consensus sequence of the 23 motifs shows an extraordinary degree of identity (*see Figure 3.15 A*).

Table 3.4: TM sequences found in *E. coli* MG1655.

No	TM sequence (5‘ to 3‘)	Length (nt)	Type	Genome localization	Genome position (')	Strand orientation
1	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	34	B	164547 - 164580	3.5	sense+
2	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	35	A	164631 - 164597	3.5	antisense-
3	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	36	B mm	282101 - 282136	6	sense+
4	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	34	B	289246 - 289279	6.2	sense+
5	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	388664 - 388699	8	sense +
6	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	497843 - 497878	11	sense+
7	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	624579 - 624614	13	sense+
8	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	36	B	624676 - 624641	13	antisense-
9	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	1351239 - 1351204	29	antisense-
10	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	36	B mm	3045989 - 3046024	66	sense+
11	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	3046087 - 3046052	66	antisense-
12	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	36	B	3239599 - 3239634	70	sense+
13	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	37	A mm	3239698 - 3239662	70	antisense-
14	CCCTCGCCCCTTTGGGGTGAGGGTTAGGGTGAGGGG	36	A mm	3390529 - 3390494	73	antisense-
15	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	3504892 - 3504857	75.5	antisense-
16	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	3608684 - 3608719	78	sense+
17	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	34	A	3781061 - 3781028	81.5	antisense-
18	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	36	B mm	3781121 - 3781156	81.5	sense+
19	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	3908495 - 3908530	84	sense+
20	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	34	B	3959491 - 3959458	85	antisense-
21	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGAGAGGGG	36	A mm	4070452 - 4070487	88	sense+
22	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	4314285 - 4314320	93	sense+
23	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	36	A mm	4549883 - 4549848	98	antisense-

When searching the 37 nt long TM consensus sequence in the *E. coli* MG1655 strain we received an E value of $6x10^{-14}$, indicating the number of hits one can "expect" to see by chance when searching the database with the effective sequence space of 256 million nucleotides. A simplified back-of-the-envelope calculation yields a 100% by-chance occurrence of the TM in an arbitrary sequence with a length of approximately 4.3 sextillion nt. Hence, the investigated motif is significantly overrepresented in the MG1655 genome. In general, the loop sequences show less sequence conservation compared to the stem regions of the potential triplex sequence. The putative triplex formed by the TM sequence is a G-/C-rich class II purine motif structure (*see Figure3.15 B&C*). The complementary, C-rich strand might be able to form a pyrimidine motif class III triplex which is stable under acidic

conditions (*see Figure 3.17 F*) that usually do not occur within the bacterial cell. We identified two different TM types: TM type A with the sequence 5'-TTA-3' and TM type B with the sequence 5'-CCG-3' in the second loop. In total, 15 of the 23 TMs found belong to type A while the other 8 can be assigned to type B motifs (*see Table 3.4*). Among the different motifs, type A – having one mismatched base pair – is the most frequent (13 TMs). Regarding the strand orientation of the TMs in the genome, we found 13 motifs located on the sense (plus) strand and 10 motifs located on the antisense (minus) strand of the genome. In five cases, two TMs are located in the same close proximity on the genome, showing a quasi-palindromic arrangement with inverted orientation and one TM on the plus and the other on the minus strand: Here, a type A motif is always combined with a type B motif (TM numbers 1&2, 7&8, 10&11, 12&13, 17&18 in *Table 3.4*). Furthermore, we analyzed the flanking genes around the TM (*see Table 3.5*). The formation of a triplex structure might affect the regulation of the local gene expression and could be related to a general mechanism for a certain gene class. However, by categorizing gene functions using the KEGG database (288,326) we found the motifs located in close proximity to all kind of genes, which does not point at a general functional correlation. Most TMs were located close to genes of general categories like metabolic pathways, biosynthesis of secondary metabolites and amino acids or ABC transporters. It is known that secondary structures in close proximity to gene start can interfere with transcription or translation (*see Chapter 3.1*) (43,161,327). Therefore, we checked the distance of the TMs relative to the open-reading frame (ORF) of the neighboring genes in *E. coli* MG1655 (*see Figure 3.18 A*). The motifs were more often found upstream of an ORF than downstream of an ORF. However, the space between the TM and the ORF ranges from very close (10 nt) to larger distances (310 nt), showing no trend to a specific proximity. As the highly regulatory regions (SD site, promoter) are located less than 100 bp from the ORF, our findings do not give clues to a general regulatory function of the TM on the level of gene expression. Regarding the location of the TMs within operons, we found no bias of operon arrangement relative to the TM (*see Table 3.5*). Apart from that, most of the operons were merely predicted from the ecocyc database (http://ecocyc.org/) lacking experimental evidence.

Table 3.5: Flanking genes of TMs in *E. coli*.

5‘ gene	5‘ gene locus tag	5‘ gene orientation	Separation (nt)	TM No	Separation (nt)	3‘ gene	3‘ gene locus tag	3‘ gene orientation	Operon location
hrpB	b0148	sense +	13	1	150	*mrcB*	b0149	sense +	
hrpB	b0148	sense +	63	2	98	*mrcB*	b0149	sense+	
yagA	b0267	antisense -	117	3	141	*yagE*	b0268	sense+	5' of yagA/B/N (no exp. evidence) and 5' of yagE/F (no exp. evidence)
yagI	b0272	antisense -	83	4	21	*argF*	b0273	antisense-	
tauD	b0368	sense+	17	5	51	*hemB*	b0369	antisense-	3' of tauA/B/C/D, 3'of hemB
adk	b0474	sense+	23	6	176	*hemH*	b0475	sense+	
fepB	b0592	antisense -	67	7	270	*entC*	b0593	sense+	5' of ent C/E/B/A/H
fepB	b0592	antisense -	130	8	208	*entC*	b0593	sense+	5' of ent C/E/B/A/H
fabI	b1288	antisense -	164	9	167	*ycjD*	b1289	antisense-	
ygfF	b2902	antisense-	85	10	144	*gcvP*	b2903	antisense-	3' of gcvT/H/P (no exp. evidence)
ygfF	b2902	antisense-	150	11	80	*gcvP*	b2903	antisense-	3' of gcvT/H/P (no exp. evidence)
alx	b3088	sense+	53	12	310	*sstT*	b3089	sense+	
alx	b3088	sense+	116	13	245	*sstT*	b3089	sense+	
aaeR	b3243	sense+	44	14	53	*tldD*	b3244	antisense-	
frlR	b3375	sense+	73	15	41	*yhfS*	b3376	antisense-	3' of yhfX/W-php- yhfU/T/S (no exp. evidence)
zntA	b3469	sense+	33	16	31	*tusA*	b3470	antisense-	
lldD	b3605	sense+	11	17	153	*tm*L	b3606	sense+	3' of IldP/R/D (no exp. evidence)
lldD	b3605	sense+	102	18	57	*tm*L	b3606	sense+	3' of IldP/R/D (no exp. evidence)
pstB	b3725	antisense-	128	19	19	*pstA*	b3726	antisense-	middle of operon pstS/C/A/B-phoU
ilvC	b3774	sense+	14	20	39	*ppiC*	b3775	antisense-	
yihR	b3879	antisense-	49	21	27	*yhiS*	b3504	antisense-	
yjdP	b4487	sense+	88	22	23	*phnP*	b4092	antisense-	3' of phnC/D/E/F/G/H/I/J/K/L/M/N/O/P (no exp. evidence)
imH	b4320	sense+	137	23	70	*gntP*	b4321	antisense-	3' of fimA/I/C/D/F/G/H

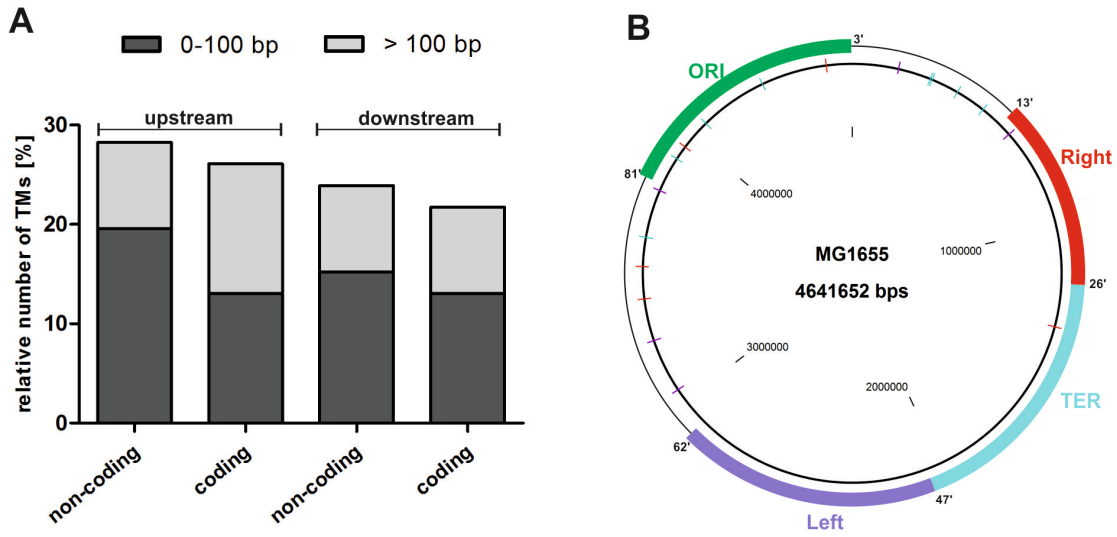


Figure 3.18: Location of the TMs in the *E. coli* MG1655 genome.

A Distance of TMs relative to neighboring ORFs. Two categories are shown: 0-100 bp and >100 bp away from start of the ORF. For both strands (coding and non-coding) the region upstream and downstream of the ORF was analyzed. **B** Map of *E. coli* MG1655 chromosome illustrating TM distribution and macrodomain organization. TM sequences are indicated as lines: TMs on plus strand (blue) TMs on minus strand (purple) and palindromic TM sequences (red) are shown.

In addition, we were interested in the chromosomal distribution of the TMs. We noticed a non-random distribution on the *E. coli* MG1655 chromosome in accordance with new insights into chromosomal macrodomain arrangements. The chromosome has a distinct positioning in the bacterial cell (328), the subcellular positions of genes correlate linearly with their chromosomal position (329) and the *E. coli* chromosome is divided into macrodomains (330,331). Macrodomains (MD) are defined as large regions where interactions occur, whereas between these regions interactions are more restricted (331). Boccard, Espeli and co-workers mapped the Ori MD (ranging from approximately 81' to 3', containing the origin of replication, *oriC*), the Ter MD (ranging from approximately 26' to 47', harboring the *dif* site), the Right and Left MDs, and two non-structured (NS) regions. The 23 TMs are almost exclusively found in the Ori MD and the two adjacent non-structured MDs (*see Figure 3.18 B*). They are regularly distributed with a mean distance of approximately 150 kbp. Moreover, the positions are symmetric with respect to the *oriC* / *dif* axis: TMs only occur in the first half of both left and right replicores.

3.2.3.1 "TMs" and chromosomal organization

The strikingly non-random distribution of TM sequences matching the MD organization in the *E. coli* chromosome led us to speculate whether the putative triplex-forming motif might be involved in organizing features of bacterial chromosomes. Regarding bacteria, few studies obtained insight into chromosomal interactions; those studies were mainly carried out in order to demonstrate co-localization of similarly regulated gene clusters (332). Three-dimensional chromosomal architectures can be elucidated by methods related to chromosome conformation capture (3C) (333).

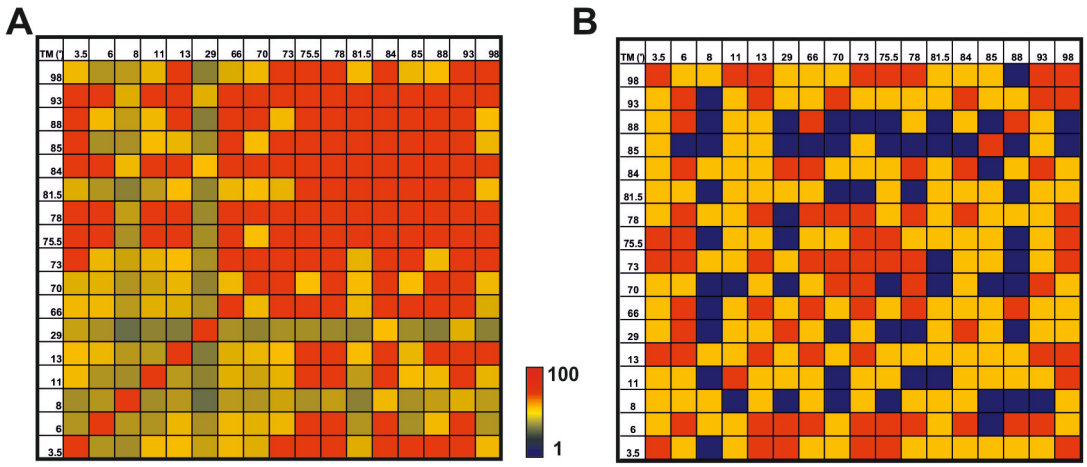


Figure 3.19: Long distance interactions between TM sites.

A Contact matrix of interactions between TMs. Data reanalyzed from Hi-C results of Voss et al. Color code of the contact matrix indicates the number of sequencing reads per interaction. **B** Intensity of sequencing reads at TM sites in comparison to neighboring fragments for the Hi-C data of Voss et al. Interaction frequency of the TM containing fragment relative to neighboring fragments is indicated by color code: red: highest interaction frequency at TM site compared to neighboring sites, yellow: TM fragment is among the highest interaction fragments, blue: TM site is localized on fragment with lowest interaction frequency.

The coupling of 3C-based methods with high throughput sequencing (Hi-C) yields a resolution between several hundred kb and 10 kb (334,335). The Hi-C method identifies all chromosome interactions at once by deep sequencing of a conformation capture library (334). In recent Hi-C studies genome interaction maps of *Caulobacter crescentus* (336) and *E. coli* K-12 (337) (338) have been constructed.

The data sets from *E. coli* grown in exponential phase were very similar for the studies from Voss et al. and Cagliero et al. (complete re-evaluated Hi-C data sets were kindly provided by Dr. Stefanie Wagner and are shown in *Figure 13.1* the appendices). We used the Hi-C results of Voss et al. to backtrack interactions between fragments containing TMs.

Figure 3.19 A shows the heat-map of interactions between TMs, the respective fragments were extracted from the reevaluated Hi-C data set (*see Figure 13.1 in the appendices*). During the exponential growth phase almost all TMs showed interactions with other TMs. The interaction frequencies with other TMs were low only for the TMs at 8’ and 29' (*see Figure 3.19 A*). Interestingly, in contrast to all other TMs the motif at 29’ is located within the Ter domain (*see Figure 3.18 B*). As already suggested (337,338), the observed interactions could occur within the replication bubble forming in the Ori region. Therefore, as TMs are located in the Ori region, the interactions observed between them are expectable. In order to see if the interactions between TMs result from the high interaction frequencies within the Ori region, we compared the interaction frequencies of the TM-neighboring Hi-C fragments, by comparing the total number of reads. *Figure 3.19 B* shows the results from this analysis: We found that in 29.1% of the cases interaction frequencies between two TMs were higher compared to the surrounding fragments (red in *Figure 3.19 B*). In most of the investigated fragments (50.5%) the value of an interaction between two TMs was not the highest, but comparable to neighboring values (*yellow in Figure 3.19 B*). Taken together, both the TMs with highest interaction frequencies and those that are among the highest interaction frequencies of their groups form 79.6% of the investigated fragments. Interaction frequencies of TM fragments are in 20.4% of the cases among the lowest of the surrounding fragments (blue in *Figure 3.19 B).* Although those findings could lead to the speculation that the TMs might indeed participate in chromosome organization, the data do not clearly show this and further experiments would be necessary to verify this hypothesis. In addition, we do not assume that the long-distance interactions between TMs are responsible for stable organization in chromosome folding. If at all, they might rather help structuring the chromosome, possibly even during replication.

3.2.3.2 "TMs" and genomic instability

Repetitive sequences and non-canonical DNA structures are associated with highly variable genetic regions (245,327,339). We were curious as to whether TMs could be involved in genomic rearrangements, recombination or bacterial evolution. Therefore, we screened 56 different *E. coli* strains (*see Table 3.6)* from 40 distinct genome groups (http://www.ncbi.nlm.nih.gov/genome/genomegroups/) for TM elements and compared their genetic variability around those regions: We found 823 TMs in total (*see Table 13.4 in the appendices*). For a better recognition of homologous regions between different strains we split the genomes into aligned locally collinear blocks (LCB) where we detected the different TM sequences (see *Table 13.5 in the appendices* for LCB assignment). We identified 62 conserved TM loci in which the TMs have homologous surrounding sequences but may be located at different positions in the genomes of different *E. coli* substrains. The TM locus 63 contains two TMs found in *E. coli* strain DH 10B (NC_010473) which could not be assigned to any LCB (TM numbers 135 and 136). *Figure 3.20* shows the distribution of the 823 TM sequences in these 63 TM loci in the different strains. Regarding the phylogenetic origin, we observed 27 strains that separated in the third generation (*shades of green in Figure 3.20*), containing less TMs (approximately 5 TMs each (121 of 823 TMs)) than the other 29 strains (approximately 24 TMs each (703 of 823 TMs); *shades of red in Figure 3.20*). This made us consider the possibility of an evolutionary role of the TM sequences. However, it does not seem that the 27 substrains share a common feature (e.g. living in another environment, adaptive functions) that clearly separates them from the other 29 substrains. A long-term experiment by Lenski and co-workers reported genomic evolution through 40,000 generations in *E. coli REL606* (340): We could not find any correlation of the evolutionary mutations to TM sites in *E. coli REL606* when backtracking their data (data not shown). Next, we investigated the size of the variable region surrounding each TM (indicated in *Table 13.6 in the appendices*) by analyzing point mutations and deletions in the alignment files. Therefore, we calculated the range of sequence variability around the particular TM within an LCB by splitting each LCB into windows of 11 nucleotides and defining a sequence variability value v_j for each window (*see Chapter 7.20 for calculation*). We observed that an average length of 2966 nucleotides is variable around each TM locus. For better evaluation of our data we analyzed the genetic instability in four different and randomly chosen control groups. In most investigated control regions we observed no genetic instability, although on average 9 of 48 control regions of each group showed sequence variability as well (*see Table 13.6 in the appendices*)

.

Table 3.6: Description of the analyzed *E. coli* genomes.

56 different *E. coli* strains were used for our genomic instability analysis. The number of TMs is shown for each particular strain. Reference numbers for chromosomal and plasmid sequences according to the NCBI.

Genome No	Organism/Name	Chromosomes	Plasmids	Number of TMs
1	*Escherichia coli* O157:H7 str. Sakai	NC_002695.1	NC_002127.1 NC_002128.1	11
2	*Escherichia coli* Xuzhou21	NC_017906.1	NC_017903.1 NC_017907.1	11
3	*Escherichia coli* O157:H7 str. TW14359	NC_013008.1	NC_013010.1	11
4	*Escherichia coli* O157:H7 str. EC4115	NC_011353.1	NC_011350.1 NC_011351.1	11
5	*Escherichia coli* O55:H7 str. CB9615	NC_013941.1	NC_013942.1	10
6	*Escherichia coli* O55:H7 str. RM12579	NC_017656.1	NC_017658.1 NC_017653.1 NC_017654.1 NC_017657.1 NC_017655.1	10
7	*Escherichia coli* SE11	NC_011415.1	NC_011407.1 NC_011408.1 NC_011411.1 NC_011413.1 NC_011416.1 NC_011419.1	26
8	*Escherichia coli* IAI1	NC_011741.1	-	29
9	*Escherichia coli* W	NC_017664.1	NC_017665.1 NC_017662.1	30
10	*Escherichia coli* KO11FL	NC_016902.1	NC_016903.1 NC_016904.1	30
11	*Escherichia coli* LY180	NC_022364.1	-	30
12	*Escherichia coli* APEC O78	NC_020163.1	-	23
13	*Escherichia coli* E24377A	NC_009801.1	NC_009786.1 NC_009787.1 NC_009788.1 NC_009789.1 NC_009790.1 NC_009791.1	27
14	*Escherichia coli* CFT073	NC_004431.1	-	2
15	*Escherichia coli* str. 'clone D i2'	NC_017651.1	-	2
16	*Escherichia coli* str. 'clone D i14'	NC_017652.1	-	2
17	*Escherichia coli* ABU 83972	NC_017631.1	NC_017629.1	2
18	*Escherichia coli* 536	NC_008253.1	-	
19	*Escherichia coli* LF82	NC_011993.1	-	2
20	*Escherichia coli* O83:H1 str. NRG 857C	NC_017634.1	NC_017659.1	2
21	*Escherichia coli* 042	NC_017626.1	NC_017627.1	13
22	*Escherichia coli* UTI89	NC_007946.1	NC_007941.1	1
23	*Escherichia coli* UM146	NC_017632.1	NC_017630.1	1
24	*Escherichia coli* IHE3034	NC_017628.1	-	1
25	*Escherichia coli* PMV-1	NC_022370.1	NC_022371.1	1

Genome No	Organism/Name	Chromosomes	Plasmids	Number of TMs
26	*Escherichia coli* S88	NC_011742.1	NC_011747.1	2
27	*Escherichia coli* APEC O1	NC_008563.1	NC_009838.1 NC_009837.1	2
28	*Escherichia coli* str. K-12 substr. MG1655	NC_000913.3	-	23
29	*Escherichia coli* str. K-12 substr. W3110	NC_007779.1	-	23
30	*Escherichia coli* DH1	NC_017625.1	-	23
31	*Escherichia coli* BW2952	NC_012759.1	-	21
32	*Escherichia coli* str. K-12 substr. DH10B	NC_010473.1	-	24
33	*Escherichia coli* str. K-12 substr. MDS42	NC_020518.1	-	21
34	*Escherichia coli* ATCC 8739	NC_010468.1	-	20
35	*Escherichia coli* HS	NC_009800.1	-	19
36	*Escherichia coli* 55989	NC_011748.1	-	29
37	*Escherichia coli* O104:H4 str. 2009EL-2071	NC_018661.1	NC_018662.1 NC_018663.1	29
38	*Escherichia coli* O104:H4 str. 2011C-3493	NC_018658.1	NC_018659.1 NC_018660.1 NC_018666.1	29
39	*Escherichia coli* O104:H4 str. 2009EL-2050	NC_018650.1	NC_018652.1 NC_018654.1 NC_018651.1	29
40	*Escherichia coli* SE15	NC_013654.1	NC_013655.1	2
41	*Escherichia coli* JJ1886	NC_022648.1	NC_022649.1 NC_022650.1 NC_022651.1 NC_022661.1 NC_022662.1	1
42	*Escherichia coli* NA114	NC_017644.1	-	1
43	*Escherichia coli* O26:H11 str. 11368	NC_013361.1	NC_013363.1 NC_013362.1 NC_013369.1 NC_014543.1	29
44	*Escherichia coli* O111:H- str. 11128	NC_013364.1	NC_013366.1 NC_013367.1 NC_013368.1 NC_013365.1 NC_013370.1	28
45	*Escherichia coli* ETEC H10407	NC_017633.1	NC_017722.1 NC_017724.1 NC_017721.1 NC_017723.1	20
46	*Escherichia coli* O103:H2 str. 12009	NC_013353.1	NC_013354.1	25
47	*Escherichia coli* O127:H6 str. E2348/69	NC_011601.1	NC_011602.1 NC_011603.1	1
48	*Escherichia coli* P12b	NC_017663.1	-	18
49	*Escherichia coli* B str. REL606	NC_012967.1	-	20
50	*Escherichia coli* BL21(DE3)	NC_012971.2	-	20
51	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG'	NC_012947.1	-	21
52	*Escherichia coli* SMS-3-5	NC_010498.1	NC_010485.1 NC_010486.1 NC_010487.1 NC_010488.1	7
53	*Escherichia coli* IAI39	NC_011750.1	-	6
54	*Escherichia coli* O7:K1 str. CE10	NC_017646.1	NC_017649.1 NC_017650.1 NC_017648.1 NC_017647.1	6

Genome No	Organism/Name	Chromosomes	Plasmids	Number of TMs
55	*Escherichia coli* UMNK88	NC_017641.1	NC_017642.1 NC_017639.1 NC_017640.1 NC_017643.1 NC_017645.1	13
56	*Escherichia coli* UMNF18	NZ_AGTD01000001.1	NZ_AGTD01000002.1 NZ_AGTD01000003.1 NZ_AGTD01000004.1 NZ_AGTD01000005.1 NZ_AGTD01000006.1	23

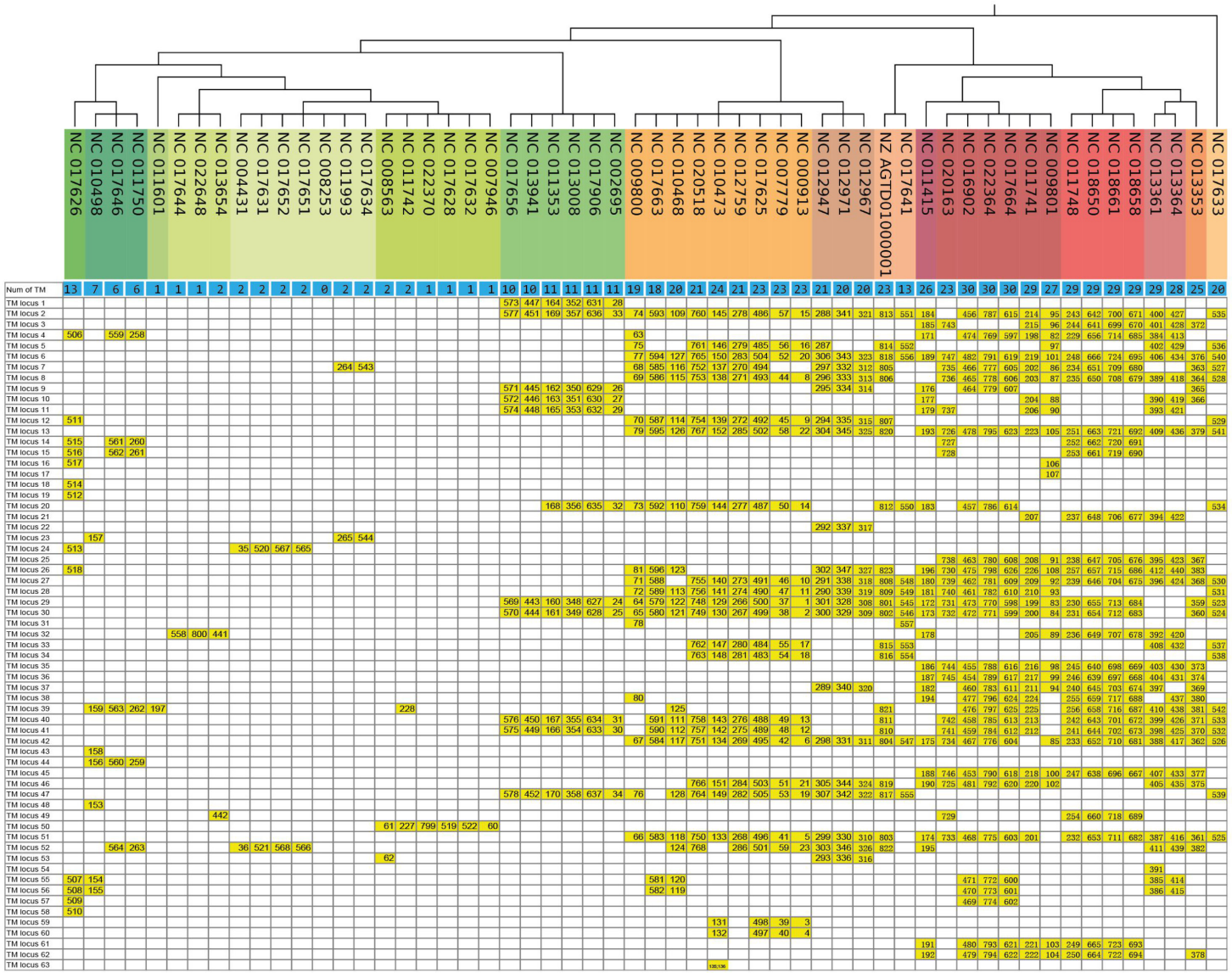


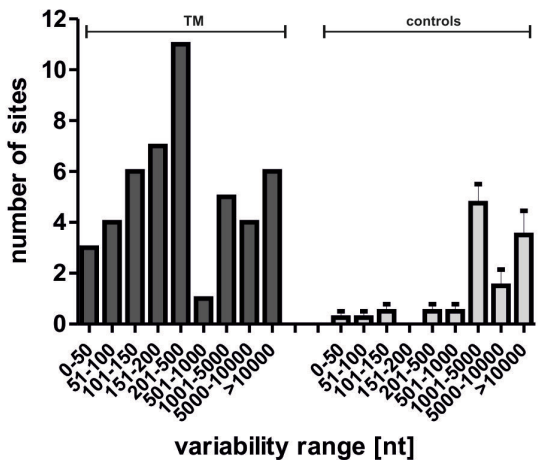
Figure 3.20: TM motifs in different strains of *E. coli*.

A Distribution of TM sequences in the 63 TM loci. Each row stands for a different *E. coli* substrain (genome number listed). The phylogeny of the strains according to the NCBI web server is shown. The number of TM sequences found in the respective strain is listed in the first line (blue). Each of the other lines represents one TM locus. TM motifs occurring in a specific locus within a specific strain are indicated in yellow (TM number according to *Table 13.4 in the appendices*). Data kindly provided by Peiwen Xiong.

Especially for the regions between 1 and 500 nt around the TM loci the sequence variability was observed to be much higher than in the control groups (*see Figure 3.21 A*). Taken together, our findings strongly suggest TMs as a source for genetic instability.

However, the accurate mechanism could not be deducted from the data presented so far. To gain a deeper insight, we focused on the 23 TM sequences found in the *E. coli* MG1655 genome. We picked a region comprising 500 nt upstream and 500 nt downstream of a TM sequence and used NCBI megaBLAST to analyze the sequence similarity of the region around the triplex motif compared to the other *E. coli* substrains. In our analysis we defined 5 different categories: 1. No change – the TM and flanking sequence are similar in the compared genomes; 2. Region missing – a large region (more than 300 bp) containing either non-coding or coding sequences shows no homology; 3. Intergenic changes – the intergenic (non-coding) region is less homologous (completely/partly deleted or sequence insertions), but the flanking coding regions remain similar in the aligned strains; 4. No homology – the entire region cannot be found at all in the aligned strain. In our examination, palindromic sequences were investigated within one region: we compared 1008 regions in total (56 strains multiplied with 18 TM sites). *Figure 3.21 B and C* show the distribution of those categories when comparing the 18 genomic sites in the 56 genomes (*see Table 3.6*). We observed that in 38% of the analyzed regions no considerable change in the genomic sequence can be found. In 39% of the cases we observed intergenic changes. In almost 10% of the analyzed regions a large sequence part (> 300 bp) was not homologous ("region missing") and about 13% of the TMs occurring in *E. coli* MG1655 were not homologous to other *E. coli* substrains at all. In our analysis, we recognized that in many cases of the categories "region missing" and "intergenic changes" triplex motifs with lower stability are present, which results from a mutated TM (examples are shown in *Figure 3.22*). Therefore, we further organized those two classes into the subcategories "TM missing" and "TM mutated". For the category "region missing" the result was well balanced: In 57% of the genomes the TM was missing and in 43% of the genomes the TM was mutated. In the category "intergenic changes" we found 33% of the genomes with TM missing and 59% with TM mutated regions (detailed evaluation is shown in *Table 3.7*). Interestingly, regarding the palindromic sites more closely we observed that either the TM sequences were completely missing or a part of both sequences was missing so that stable stem loop structures might be able to form (*see Figure 3.22*). This effect was listed as palindromic effect and made up 8% of the "intergenic changes". For a better evaluation of our results we again compared three sets of control sequences regarding the same criteria (*see Figure 3.21 and Table 13.7 in the appendices*).

A



B

		no change	region missing	intergenic change	no homology
TM	No.	382	100	398	128
	%	37.89	9.92	39.48	12.69
Random control	No.	952	56	6	0
	No.	947	97	2	18
	No.	998	7	12	47
	%	90.76 ± 2.64	5.01 ± 4.23	0.63 ± 0.47	3.60 ± 1.66

C

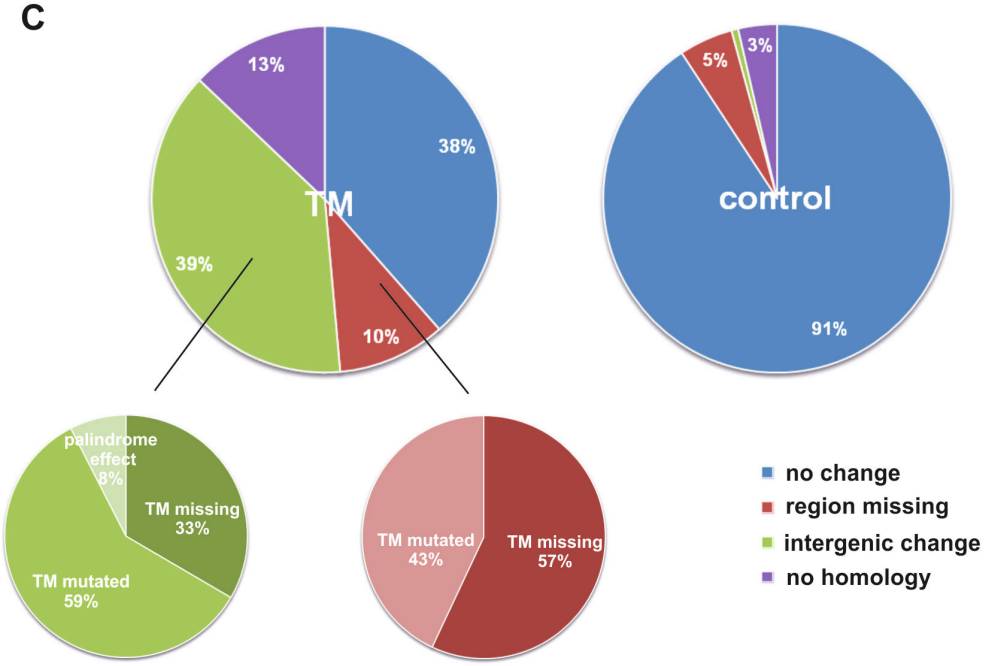


Figure 3.21: Genomic instability around the TM sequences of *E. coli* subspecies.

A Variability range in nucleotides around the TM motifs occurring in the 63 TM loci of the 56 *E. coli* strains compared to random controls. Details on variability calculation are described in the experimental part. Data kindly provided by Peiwen Xiong. **B** and **C** show results of megaBLAST sequence comparison of the region around (500 bp up- and downstream) the 23 TMs found in *E. coli* MG1655 to the other 55 *E. coli* genomes. 4 different categories were defined: 1. No change – sequences are identical in the different strains; 2. Region missing – a region larger than 300 bp is missing in the compared strain; 3. Intergenic change – less homology occurs in the intergenic region; 4. No homology – the respective motif is not found in the compared strain. Absolute numbers of strains and percentages for each category are listed for the investigated TM sites and 3 random control groups in

B. C illustrates the distribution of the different categories when comparing TM or random control regions. The categories “region missing” and “intergenic deletions” were further subdivided if the TM motif was missing or mutated. For the palindromic regions an effect generating a potential hairpin structure was observed. This feature was labeled “palindrome effect”.

Table 3.7: Evaluation of the genomic instability studies.

The regions around the 23 TMs of *E. coli* MG1655 were analyzed for genomic instability. Each motif is assigned to one of the different categories for each particular genome (*listed in Table 3.6*).

TM No.	no change	region missing		intergenic deletions			not found
		TM missing	TM mutated	TM missing	TM mutated	palindrome effect	
TM 1/2	33			21		2	
TM 3	4						52
TM 4	4						52
TM 5	27		1	22	6		
TM 6	28			11	17		
TM 7/8	26			17	13		
TM 9	16	9	17		13		1
TM 10/11	28	28					
TM 12/13	27				11	18	
TM 14	20			1	35		
TM 15	28	13	15				
TM 16	14				42		
TM 17/18	11			21	14	10	
TM 19	20			9	27		
TM 20	28			19	9		
TM 21	20				14		22
TM 22	28			8	20		
TM 23	20	7	10	4	14		1

These control regions had an average length of 1070 bp and were randomly chosen from *E. coli MG1655;* they always contained an intergenic (non-coding) region, carrying no TM, framed by two coding regions. Intriguingly, we found that for more than 90% of the regions in the controls no sequence change occurs. A larger region of the analyzed genomic parts was missing in 5% of the investigated strains, and in 3% of the strains the analyzed region was not found at all. In less than 1% of the investigated sites a short intergenic sequence part was missing. These findings indicate a significantly higher genomic variability around the TM motifs compared to control regions: Both the number of sites missing a larger region and those bearing intergenic deletions showed a considerable increase compared to the controls.

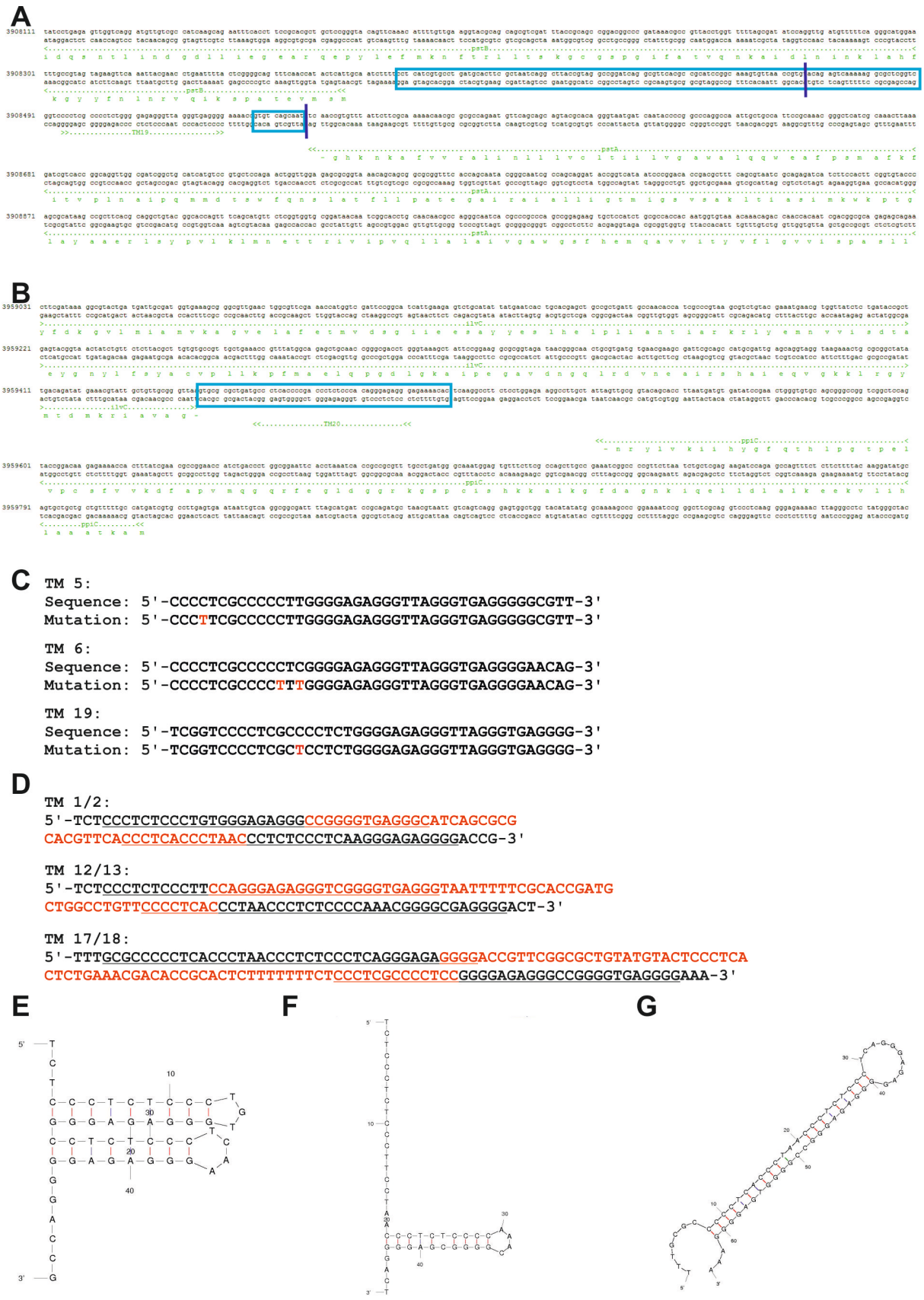


Figure 3.22: Representative modifications of the TM sites with "intergenic change"

A and B Blue boxes mark regions that are deleted in some strains at those specific sites. Dark blue lines mark region where mutations can occur. **A** TM19 – subcategory: TM mutated **B** TM 20 – subcategory: TM deleted. **C** Examples of less stable triplex motifs occurring through mutation in the subcategory TM mutated (red nucleotides indicate mutations). **D** TM sequences at the indicated palindromic regions with intergenic spacers and 3 nt flanking region. TM sites are underlined; deleted sequences are indicated in red. Potential stem-loop structures that could be formed when the red sequence is deleted at the particular site: **E** TM1/2, **F** TM 12/13 and **G** TM 17/18. Structure prediction according to the mfold webserver

3.2.4 Discussion

The relation between DNA sequence repeats and genomic instability has been described in different studies: Instability caused by TR sequences has been attributed to different hereditary diseases (264); chromosomal plasticity in *Pseudomonas fluorescens* species has been associated to MITE sequences (265); REP sequences have been linked to genetic instability in *E. coli* toxin-antitoxin systems (266), and other repetitive sequences have been described in correlation to genomic plasticity in bacteria (267,268). Most of the repeat sequences have the potential to fold into non-canonical secondary structures on DNA and/or RNA level, as it has been described for pneumococcal bacteria (269). Also, those non-canonical nucleic acid structures are prone to interfere with translation, transcription, replication or recombination. The exact mechanisms of those influences, however, have not been elucidated to date. The function and role of many repetitive elements occurring in eukaryotes and prokaryotes is still unclear. In this chapter we focused on intrastrand triplex DNA repeats in prokaryotes. We generated the ITxF database which enables an easy search for intrastrand triplex structures of different structural classes within 5,246 prokaryotic genomes. Although different computational tools allowing the search for potential triplex sites in genomes have been reported, the ITxF database is – to our knowledge – the first one defining intrastrand triplex structures that are not necessarily H-DNA or TFO binding sites. Data extracted from our database search showed the abundance of triplex motifs in bacterial chromosomes, suggesting that they play an important role in the bacterium instead of being randomly distributed. For functional characterization, we focused on the particular repetitive TM sequence which has been investigated earlier but whose function and exact structure were never clarified. Whereas earlier studies only showed this motif to be enriched in 5 different bacterial strains we could prove its high occurrence in several prokaryotes and the significant enrichment in *E. coli* MG1655. Using CD and NMR spectroscopy, we demonstrated the stable triplex formation of TM oligonucleotides. Maher and co-workers also suggested triplex formation for their PIT elements based on studies involving UV spectroscopy and electrophoretic mobility shift assays. However, NMR measurements are necessary to exclude the formation of a competitive intramolecular G-quadruplex structure. In the double-stranded genome the complementary strand always competes with triplex formation. CD and NMR studies of the double-stranded TM sequence did not show characteristic triplex signals and rather suggested double stranded DNA. Furthermore, in *in vivo* DMS footprinting experiments methylation was not blocked at specific guanines involved in triplex formation. We therefore concluded that within the *E. coli* chromosome the double stranded structure is preferred to triplex formation as well. It might be possible that the triplex

structure only forms temporarily or during a specific bacterial growth phase, which would serve as an explanation for the observed cleavage at these sites. Obviously the DNA needs to be single-stranded for the triplex formation – conditions during the cell cycle that favor ssDNA might arise during replication, recombination, transcription or translation. As cells are not synchronized, it would also be possible that only in a small amount of the investigated chromosomes a triplex structure is formed at a given time point but this fraction cannot be detected with our assay due to heterogeneity of the overall population. Intriguingly, we observed a strong band just 5' of the triplex sequence, which could result either from enhanced cleavage or polymerase stop during primer extension. Maher and co-workers reported that the Taq polymerase stops when a PIT element is located on the template strand during PCR *in vitro* (10). Therefore, it is very likely that in our *in vivo* footprinting experiment, the polymerase stops in front of the TM during primer extension, which hints at the formation of a triplex structure.

In silico characterization showed the TM sequence not to be related to certain functional gene classes or within a certain distance and orientation to flanking genes – similar to results for PIT and BoxC elements. In addition, PIT motifs did not show any promoter or terminator activity in earlier studies (85). However, our investigations showed that the TM sequences are not randomly distributed within the *E. coli* chromosome: They are found in the Ori MD and the two non-structured MDs. These facts allowed us to speculate that the TM motif might not be involved in gene regulatory mechanisms, but could rather be related to chromosome organization. Architecture-shaping repetitive sequences that are highly distributed in the chromosome – such as DnaA Boxes, GATC motifs or Ter sites (summarized in (341)) – have been identified in different bacteria. We speculated that distinct interactions between the triplex motifs could help to organize the bacterial chromosome and potentially bring certain genes into territorial proximity. By analyzing Hi-C data of Voss et al. and focusing on fragments bearing TM sequences, we could indeed detect enhanced interaction frequencies between TM sites. Even though the interactions observed in the Ori region of *E. coli* genomes seem to be enhanced due to the replication bubble and the presence of multiple Ori domains (337,338), TMs might play a role in the structuring of bacterial genomes. Analyzing raw data derived from Hi-C experiments by Voss et al. we observed similar or even higher interaction frequencies between the majority of TM sites in comparison to their neighboring sites. This led us to suggest that TM sites might co-localize at least temporarily during the exponential growth phase in *E. coli*. This could help structuring the bacterial chromosome during replication. However, our data do not clearly prove this hypothesis, and further experiments (such as 3C studies) should be conducted to verify this suggestion.

Finally, we investigated the genetic stability around the TMs: We analyzed 823 TMs occurring in 56 different *E. coli* genomes and found the genomic sites around those sequences to be quite variable when compared to control regions. Our first assumption was that the TM could be related to recombination, as e.g. H-DNA sequences have been mapped to recombination hotspots in mouse myeloma cells (215). When investigating the genomic changes between the 23 TM sequences found in *E. coli* MG1655 and the other strains in more detail, we found that only in 10% of the strains a larger region was missing around a TM. If induction of recombination was the function of TMs we should have seen a higher percentage of large regions changing in our analysis. However, in 13% of the analyzed strains no TM sequences or their 500 bp flanking regions could be found. Large deletions occur at those sites, which might also be related to recombination. The formation of the intrastrand triplex motif would provide the complementary strand in a single-stranded status which could be used as a locus for homology searching and trigger recombination events. Alternatively, recombination could occur between two TMs forming at homologous regions. Such mechanisms have been proposed for H-DNA structures as well (217-219). Furthermore, TM-related recombination events at non-homologous sites, such as illegitimate recombination, could be induced by DNA breakage or strand slippage near the TM structure. However, most of the TMs could be assigned to a particular TM locus for all of the 56 genomes. In comparison to random control groups we found the genetic instability increased in flanking regions adjacent to the TM (around 500 bp flanking sites). In addition, most of the changes observed around the TMs were small intergenic changes (39%), where the TM sequence was either completely deleted or mutated. Furthermore, for palindromic regions we observed that in most cases intergenic parts were deleted, so that a secondary structure, like a hairpin, might still be able to form. Interferences with bacterial replication might serve as a better explanation for the observed genomic instabilities near the TM sites. Replication of the circular bacterial chromosome proceeds bidirectionally, starting at the ORI where the replisome is recruited. A replication fork acts on each chromosomal arm, and its progression can be blocked when meeting obstacles such as bound proteins, bulky adducts, interruptions in the DNA template or non B-DNA structures. An inactivation of the replication fork may lead to double-strand breaks, polymerase stalling or replication slippage. Replication slippage could serve as an explanation for the deletion of the TM motif (91,92). As described before, stalled replication can be reinitiated by primase, which creates a new primer that binds after the obstacle and leaves a gap in the DNA sequence (93). This mechanism could also explain the deletion of the TM including its flanking region. Furthermore, a stalled replication fork may cause double strand breaks, which may be processed in a mutagenic fashion. In mammalian cells H-DNA has been reported to induce double strand breaks (212). Usually, the proceeding of replication after stalling or the processing of gaps occurring during

replication is initiated by DNA repair enzymes. Non B-DNA structures themselves have been suggested to be identified by DNA repair proteins as they represent distortions of the DNA double helix (94,95). A consequence of DNA repair is the introduction of mutations or small deletions, which could explain the high amount of intergenic deletions and mutated TMs we found in our analysis. In another scenario, DNA repair in proximity to the TM sequence may generate superhelical stress and thereby facilitate the transition from the DNA double strand to the triplex structure. Non-B-DNA structures forming during DNA repair could alter the repair process and have been suggested to contribute to error-generating repair and genomic instability, when analyzed in a plasmid system in mammalian cells (96). Although such mechanisms are known to be related to repeat sequences and non-canonical DNA structures, to our knowledge genetic instability related to an intrastrand triplex motif has not been reported so far. The majority of the studies reporting genome instability associated with triplex motifs focused on H-DNA (reviewed in (95,327,342)), less studies investigated bacterial triplexes (234). In this study we clearly demonstrate that the TM sequence, which is present in a variety of bacteria and able to fold into an intrastrand triplex structure *in vitro*, is related to genomic instability. However, the question about the benefits for the bacterium remains. Structure-induced genomic plasticity via the introduction of mutations or deletions and chromosomal rearrangements potentially contribute to evolutionary functions, such as the adaption to rapid changes. It might be possible that TM sequences are involved in genomic evolution and organismal adaption. However, the origin of the motif remains unclear: Has it been inserted or deleted during evolution? In our analysis we found less TM sequences in 27 strains that separated in the third generation. However, those strains do not seem to differ in adaptive or environmental functions compared to the 29 substrains bearing more TM sequences. Data from a long-term experiment of evolutionary mutations by Lenski and co-workers in *E. coli B REL600* (340) showed no correlation to TM sequences either. This study was done with a laboratory strain, which is generally not exposed to drastic environmental changes. Possibly adaptive mutations need harsher conditional changes or a longer time period to evolve. A daring hypothesis as to how the TM sequence could have been distributed in the different bacterial genomes is via horizontal gene transfer (HGT). HGT describes a process that brings non-parental genetic information into a cell. Usually, HGT introduces new genes to the host bacterium, which confer a novel pathway for cell survival or encode highly efficient proteins. TM sequences could be leftovers from such processes. However, analyzing the genes around TMs, we could not find a tendency for enrichment of genes encoding for enzymes involved in reactions central to cellular survival or peripheral cellular mechanisms. More intensive analyses of triplexes in different bacteria (using the ITxF database) and their surrounding genes, combined with instability and evolutionary studies, could give insights as to whether such processes might be related to

intrastrand triplex motifs in general. Furthermore one has to keep in mind, the possibility that triplex motifs might interact with certain proteins. Proteins could recognize either the sequence or the triplex structure. However, first experiments in this area (protein fishing and EMSA, Malte Sinn's Master's Thesis) did not give any hints that the TM serves as recognition site for proteins. Therefore, this topic was not addressed further in this thesis.

Altogether, we found it impressive that such a variety of bacterial genomes possess several copies of TM sequences with significant consensus identity. We were able to show that the TMs – forming stable triplex DNA structures *in vitro* – seem to induce high genetic instability in *E. coli* subspecies and might be involved in structuring the bacterial chromosome. Although different forms of this particular triplex have been studied before (10,85,244,322), our findings provide a concrete function of the TM for the first time. Nevertheless, we could not prove a distinct mechanism for the induction of genome plasticity via the TM. Thus, questions about the evolutionary origin, expansion or retention of the motif remain intriguing.

4 SUMMARY AND OUTLOOK

Guanine-rich nucleic acids are prone to fold into secondary structures: Hairpins, cruciforms, G-quadruplexes, triplexes and other alternative conformations are known (343). These structures have the potential to interfere with transcription, translation, replication and recombination, which has been demonstrated in different artificial systems (159,161,301) (*see also Chapter 1.2*). Furthermore, they can be stabilized by a variety of compounds and serve as contact-points for several drugs (128). Guanine-rich repeats and their associated secondary structures have been linked to different human diseases (139,207,214). Antigenic and phase variation of multiple pathogenic bacteria was also related to G-rich repeats (44,157,158). However, only few of these functions and their mechanisms are clarified in detail. Especially research investigating the role of alternative nucleic acid structures in prokaryotic contexts is sparse. Therefore it is of great importance to gain insights into secondary structure-mediated mechanisms, their function and occurrence in bacterial genetic contexts.

This thesis focused on two specific conformations formed within guanine repeat sequences: G-quadruplexes and triplexes. Both are stabilized by Hoogsteen-base pairing and are known to have the potential to exert regulatory functions in eukaryotes (*see Chapter 1.2*). However, studies describing their functional properties in prokaryotic organisms are rare.

In the first part of the thesis the aim was to assign modulatory effects on bacterial gene expression to G-quadruplexes at distinct genetic positions. Using reporter gene constructs, we presented a systematic analysis of the influence of quadruplex sequences in bacterial gene-regulatory regions: G-quadruplexes of different stabilities and corresponding control sequences were inserted at several positions within the core promoter, 5'-UTR, and 3'-UTR regions. We demonstrated drastic effects on gene expression which depended on the strand and the position of the insertion. G-rich sequences on the antisense strand located in the core promoter and those on the sense strand in the vicinity of the ribosome binding site located in the 5´-UTR showed pronounced inhibitory effects and were analyzed in detail. In general, we observed that G-quadruplexes located on the antisense strand influence gene expression by modulation of transcription; whereas those located on the sense strand modulate translation. Furthermore, we designed a model system that caused gene activation via quadruplex formation adjacent to a hairpin that masks the SD region. In addition, we studied the influence of natural quadruplex sequences of *E. coli* occurring in crucial positions and demonstrated that some of these quadruplexes seem to evoke significant modulation of gene expression as well. In a first attempt, we developed a system for the investigation of G-

quadruplexes occurring in ORFs: We computationally searched for such structures in the genomes of *Salmonella* and *E. coli* subspecies and designed plasmids expressing the His-tagged proteins – containing G-quadruplex sequences or respective controls – for analysis on Western blots. However, the results obtained were inconclusive and this part of the project should be resumed in further studies. Nevertheless, we showed strong position-dependent effects of quadruplex secondary structures on gene expression. Theses findings provoke the question whether nature makes use of such simple measures for controlling gene expression in a conditional manner. Investigations of naturally occurring G-quadruplexes at such specific locations and within the genomic context could shed some light on this question. Although this has been done for some G-quadruplex sequences (43), deeper analysis of the genes in close proximity of potential G-quadruplexes could offer valuable clues to general or organism-specific mechanisms. An interesting approach could be the investigation of the relation between potential G-quadruplex formation and environmental stress, especially osmotic shock and salt stress. Most of the potential G-quadruplexes we found close to the *E. coli* SD region were located in front of stress-related genes. Also, the G-quadruplexes we found in ORFs of *Salmonella* and *E. coli* were associated to stress induced K^+ transporters. Interestingly, K^+ is known to stabilize such structures, which makes it likely that a functional correlation exists. In general, the presented study significantly broadens the insights into the effects of nucleic acid secondary structure formation on gene expression in bacteria.

In the second part of this thesis we focused on triplex structures. The aim was to identify and characterize intrastrand triplexes in prokaryotes. Intrastrand triplex structures form within one DNA oligonucleotide and are stabilized via Hoogsteen base pairing. In a collaborative approach, we designed the ITxF web server which allows the search for intrastrand triplex motifs in 5246 different genomes and plasmids of bacterial and archaeal species. Using this web server we demonstrated that more than 2 millions potential triplexes among all triplex classes could form within the investigated species. The ITxF website allows the search for distinct sequence patterns. Using this tool, we identified one particular triplex motif – the TM – which is significantly enriched in *E. coli* and other proteobacteria. This motif has been described in earlier studies as BoxC and PIT elements, but its function was never elucidated. Since the distribution of this element in bacterial genomes was observed to be non-random, we re-investigated the TM. We characterized the motif from *E. coli* MG1655 by circular dichroism and NMR and showed that the TM forms a stable triplex structure *in vitro*. *In vivo* footprinting and spectroscopic studies in the presence of the complementary strand suggested that the triplex is not formed within the chromosomal DNA. However, we thought that it might form temporarily during situations in which the strands are separated, such as

transcription, translation, replication or recombination. Earlier studies by Maher and co-workers investigated regulatory functions of such triplexes on gene expression, leading to no conclusive results (85). We recognized a chromosomal distribution restricted to the Ori and the two adjacent non-structured macrodomains. In order to elucidate a putative function of the TM in chromosome organization, we analyzed the interactions between fragments containing the TMs of a Hi-C study. Interestingly, we found that almost all TMs showed interactions with other TMs, and by comparing the interaction frequencies to the neighboring fragments we discovered that in most of the cases these frequencies are among the highest of their groups. Therefore, TMs might help structuring the chromosome, e.g. during replication. Another possible role of TMs could be that the triplex is responsible for genomic rearrangements or instabilities. We performed comparative blast analyzes of 56 sequenced *E. coli* substrains based on their phylogenetic origin and showed that the genomic instability around TMs is enriched compared to control sequences. To our knowledge, this is the first study describing intrastrand triplexes as inducers of genomic instability. However, although we could dedicate functional properties to the TM for the first time, some questions remain open. The origin of the motif is still unclear. Deeper evolutionary studies and sequence comparisons could be performed, taking into account other species than *E. coli* as well. In addition, 3C experiments specifically investigating TM interactions could shed more light on interactions between TMs and organizational features (first experiments in this area were performed by Dr. Stefanie Wagner). Stabilization of the triplex by interacting compounds could “freeze” a certain conformation and give more insights into genomic structures and possible mechanisms. Furthermore, more intensive studies with the new ITxF database could give insights as to the existence of other conserved motifs, such as the TM, in different species, how the total number of triplexes varies between close related species and whether the phenomenon observed for *E. coli* could be valid for other gram-negative bacteria as well.

Altogether, in two independent studies we proved the importance of non-canonical nucleic acid structures in bacteria. Although extensive studies in this field have been performed with eukaryotic cells, conclusions drawn from eukaryotes cannot simply deemed valid for prokaryotes as the genetic mechanisms differ significantly. Hence, separate investigations are necessary. This thesis demonstrated that both regulatory roles influencing gene expression and organizational features as well as genomic instability can be evoked by non-canonical nucleic acids in bacteria.

5 ZUSAMMENFASSUNG UND AUSBLICK

Guaninreiche Nukleinsäuren sind für die Ausbildung von Sekundärstrukturen prädestiniert: Haarnadelstrukturen, 3- oder 4-Wege-Kreuzungen, G-Quadruplexe, Triplexe und andere Konformationen sind bekannt (343). In verschiedenen artifiziellen Systemen wurde gezeigt, dass solche Strukturen regulatorische Mechanismen wie Transkription, Translation, Replikation oder Rekombination beeinflussen können (159,161,301). Außerdem können sie mit diversen chemischen Verbindungen interagieren und sind somit potentielle Angriffspunkte für Medikamente (128). Repetitive Guaninreiche Sequenzen und die mit ihnen verbundenen Sekundärstrukturen wurden mit verschiedenen humanen Krankheiten (139,207,214) in Verbindung gebracht und sollen auch mit der antigenischen und Phasen-Variation in pathogenen Bakterien in Zusammenhang stehen (44,157,158). Nur wenige dieser Funktionen und die damit verbundenen Mechanismen konnten bisher im Detail geklärt werden. Es sind wenige Untersuchungen vorhanden, die sich im Speziellen mit alternativen Nukleinsäurestrukturen im bakteriellen Kontext auseinandersetzen. Deshalb ist es von großem Interesse, Einblicke in das Vorhandensein von potenziellen Sekundärstrukturen und die durch sie ausgelösten Mechanismen und Funktionen in Bakterien zu erhalten.

Diese Arbeit beschäftigt sich mit zwei speziellen Konformationen, die sich innerhalb Guaninreicher Sequenzen ausbilden können und durch Hoogsteen Basenpaarung stabilisiert sind: G-Quadruplexe und Triplexe. Obwohl beide Strukturtypen potentielle regulatorische Funktionen in Eukaryoten ausüben (*siehe Chapter 1.2*), sind nur wenige Studien bekannt, die diese Strukturen und ihre Funktionen in Bakterien beschreiben.

Das Ziel des ersten Teils dieser Arbeit bestand darin, den Einfluss von G-Quadruplexen in bestimmten genetischen Positionen auf die bakterielle Genexpression zu untersuchen. Hierzu haben wir verschiedene Reportergen-Konstrukte mit einer systematischen Anordnung der G-Quadruplexe in verschiedenen regulatorischen Regionen konstruiert. Die Quadruplexsequenzen und entsprechenden Kontrollen wurden an verschiedenen Positionen innerhalb des Promoters, der 5‘-UTR und der 3‘-UTR eingebracht. Je nach Position der potentiellen Quadruplexsequenz und ihrer Orientierung auf dem Plus- oder Minusstrang des Plasmids konnten drastische Änderungen der Genexpression beobachtet werden. Guanin-reiche Sequenzen auf dem Minusstrang innerhalb des Promoters und solche auf dem Plusstrang in unmittelbarer Nähe zu der ribosomalen Bindestelle (in der 5’-UTR) zeigten deutliche inhibierende Effekte und wurden detaillierter untersucht. Im Allgemeinen konnte gezeigt werden, dass Quadruplexe auf dem Minusstrang die Transkription beeinflussen,

während jene auf dem Plusstrang Auswirkungen auf die Translation haben. Des Weiteren konnten wir in einem artifiziellen System zeigen, dass G-Quadruplexe in der Nähe der Ribosomenbindestelle — in einem speziellen Design — die Genexpression auch aktivieren können. In diesen Konstrukten ermöglicht die Ausbildung der Quadruplexstruktur den Zugang zur ribosomale Bindestelle, die vorher durch eine Haarnadelstruktur blockiert wurde. Daraufhin haben wir natürlich vorkommende potentielle Quadruplexsequenzen im Bereich der Ribosomenbindestelle aus *E. coli* untersucht und konnten auch hier einen Einfluss auf die Genexpression nachweisen. Zudem beschäftigten wir uns, in einem ersten Ansatz, mit natürlichen Quadruplex-Sequenzen, die im Protein-kodierenden Bereich (open reading frame, ORF) vorkommen. Wir durchsuchten die Genomsequenzen von *E. coli* und *Salmonella* nach potentiellen Quadruplexen im ORF und beschlossen, zwei dieser potentiellen Sequenzen innerhalb eines Plasmidsystems zu untersuchen. Wir fusionierten die Quadruplex enthaltenden Proteinsequenzen mit einem Histidin-Tag und untersuchten die Proteinexpression auf einem Western Blot. Aufgrund uneindeutiger Ergebnisse konnte keine Aussage über den Einfluss der Quadruplexsequenz getroffen werden, dieser Teil der Studie sollte in weiterführenden Experimenten vertieft werden. Zusammengefasst konnten wir starke positionsabhängige Einflüsse von Quadruplexstrukturen auf die bakterielle Genexpression zeigen. Unklar bleibt, ob solche einfachen Methoden zur Kontrolle der Genexpression in natürlichen Organismen angewandt werden. Daher sollten natürlich vorkommende Quadruplexsequenzen an solchen definierten Positionen sowie die umgebenden Genfunktionen genauer untersucht werden. Wie in wenigen Fällen geschehen (43), könnten dadurch allgemeine oder Organismus-spezifische Mechanismen aufgeklärt werden. Ein interessanter Ansatz ist die Untersuchung von Quadruplexen im Zusammenhang mit Stress-assoziierten Genen. Hierbei spielen vor allem osmotischer Schock und Salzstress eine interessante Rolle: Viele der potentiellen Quadruplex Sequenzen, die wir in der Nähe der Ribosomenbindestelle und im bakteriellen ORF fanden, standen mit Stress-induzierten Genen und K^+ Transportern in Verbindung. Da K^+ Quadruplexe stabilisiert, könnte hier ein funktioneller Zusammenhang bestehen. Die in dieser Arbeit präsentierte Studie stellt eine erhebliche Erweiterung der Kenntnisse von den, durch Sekundärstrukturen ausgelösten, Auswirkungen auf die Genexpression dar.

Im zweiten Teil dieser Arbeit beschäftigten wir uns mit Triplexstrukturen. Ziel war es, Triplexe, die sich innerhalb eines Nukleinsäure Strangs ausbilden können (intrastrand Triplexe), in bakteriellen Genomen zu identifizieren und zu charakterisieren. In einer kooperativen Studie entwickelten wir den ITxF Web-Server. Damit können intrastrand Triplexstrukturen in 5246 verschiedenen bakteriellen Genomen und Plasmiden gesucht werden. Wir konnten über 2 Millionen solcher Triplexstrukturen – aus allen vier

Triplexklassen — innerhalb dieser Spezies finden. Die ITxF-Datenbank ermöglicht zudem die Suche nach speziellen Sequenzmustern und definierten Triplextypen. Auf diese Weise identifizierten wir ein spezielles Triplexmotiv, das TM, das innerhalb *E. coli* und anderen Proteobakterien gehäuft auftritt. Dieses Motiv wurde in früheren Studien als BoxC oder PIT beschrieben, seine Funktion konnte allerdings nie geklärt werden. Zunächst charakterisierten wir dieses Motiv mittels spektroskopischer Methoden und konnten zeigen, dass es *in vitro* eine stabile Triplexstruktur ausbildet. Nach Zugabe des komplementären DNA-Strangs und in „*in vivo* footprinting"-Experimenten konnte keine Triplexstruktur nachgewiesen werden, was die Ausbildung dieses Motivs innerhalb des bakteriellen Chromosoms in Frage stellt. Dennoch könnte sich die intrastrand-Triplexstruktur temporär ausbilden und zwar in Situationen in denen der DNA-Doppelstrang ohnehin in Einzelstränge separiert ist, etwa während der Transkription, Translation, Replikation oder Rekombination. In früheren Studien untersuchten Hoyne et al. den Einfluss dieser Triplexstruktur auf die Genexpression und konnten keine regulatorischen Auswirkungen feststellen (85). In unseren Untersuchungen stellten wir fest, dass das TM nicht willkürlich auf dem *E. coli* Chromosom verteilt ist, sondern hauptsächlich in der Ori und den beiden angrenzenden nicht-strukturierten Makrodomänen vorkommt. Dies legt eine Funktion im Rahmen der strukturellen Organisation des Chromosoms nahe. Um dies zu überprüfen, analysierten wir die Interaktionen zwischen TM-enthaltenden Fragmenten eines Hi-C Experiments. Wir konnten zeigen, dass nahezu alle TM-enthaltenden Fragmente mit anderen TM-enthaltenden Fragmenten interagieren und schlussfolgerten, dass TMs an der Strukturierung des Chromosoms beteiligt sein könnten, zum Beispiel während der Replikation. Eine andere mögliche Funktion von TMs könnte die Induktion genomischer Umstrukturierung oder genomischer Instabilität sein. In Sequenzanalysen von 56 sequenzierten *E. coli* Untergruppen konnten wir zeigen, dass die Instabilität (z.B. Mutationen oder Deletionen) um die TMs herum im Vergleich zu Kontrollsequenzen erhöht ist. Die in dieser Arbeit beschriebene Studie zeigt erstmalig, dass intrastrand-Triplexmotive Ursache genomischer Instabilität sein können. Obwohl wir diesem Motiv zum ersten Mal funktionelle Eigenschaften zuordnen konnten, bleiben viele Fragen offen. Die evolutionäre Herkunft des Motivs bleibt ungeklärt, weswegen vertiefende, evolutionäre Untersuchungen und Sequenzvergleiche, auch mit anderen bakteriellen Stämmen, durchgeführt werden sollten. Zudem könnten 3C-Experimente Aufschluss über die Interaktionen zwischen TMs geben (erste Experimente in diese Richtung wurden bereits von Dr. Stefanie Wagner durchgeführt). Die Stabilisierung des Triplexes durch spezifische chemische Verbindungen könnte die Konformation über einen längeren Zeitraum „einfrieren" und strukturelle Untersuchungen erleichtern. Weiterhin sollten intensivere Untersuchungen anhand der neuen ITxF-Datenbank durchgeführt werden: Gibt es vergleichbare Motive in anderen Bakterien? Wie sind die Varianzen zwischen eng verwandten Spezies in Bezug auf

die Anzahl der Triplexmotive? Interessant wäre in diesem Zusammenhang außerdem, ob die Beobachtungen in Bezug auf *E. coli* auf andere gramnegative Bakterien übertragen werden können.

Zusammenfassend kann gesagt werden, dass diese Arbeit in zwei unabhängigen Studien die vielfältigen Funktionen nicht-kanonischer Nukleinsäure Strukturen in Bakterien darstellt. Wir konnten zeigen, dass in Bakterien sowohl regulatorische Einflüsse auf die Genexpression als auch Auswirkungen auf die bakterielle Chromosomen-Organisation und die genomische Instabilität durch nicht-kanonische Nukleinsäuren induziert werden können.

6 MATERIALS

6.1 Chemicals and reagents

All chemicals were of p.a. or molecular biology quality grade. They were bought from Roth, Karlsruhe or Sigma-Aldrich, Hamburg, except of those listed in *Table 6.1* or if stated elsewhere. Water was drawn from a combined reverse osmosis/ultrapure water system.

Table 6.1: Chemicals and reagents.

Reagent	Supplier
$FeSO_4$	Merck, Darmstadt
$MgCl_2$	Merck, Darmstadt
$MgSO_4$	Merck, Darmstadt
NaCl	VWR, Darmstadt
Sodium Acetate	Merck, Darmstadt
Tryptone	MP Biomedicals, Eschwege
Yeast extract	MP Biomedicals, Eschwege

6.2 Nucleotides and radiochemicals

ATP, GTP, CTP, UTP	Fermentas, St. Leon-Roth
Carrier DNA	Salmon testis DNA, Roth, Karlsruhe
dATP, dGTP, dCTP, dTTP	Fermentas, St. Leon-Rot
γ-^{32}P-ATP	Hartmann Analytic, Braunschweig

6.3 Oligonucleotides and primers

All oligonucleotides and primers were synthesized by Metabion GmbH, Martinsried or Sigma Aldrich, Hamburg. Oligonucleotide sequences for *in vitro* experiments are described in *Chapters 7.1 and 7.2.* Primers used for cloning procedures are listed in *Table 13.8* the appendices.

6.4 Bacterial strains

Experiments investigating positional effects of G-quadruplexes *(see Chapter 3.1.1)* were performed with the *E. coli* XL10gold *strain* (Invitrogen, Darmstadt; endA1 glnV44 recA1 thi-1 gyrA96 relA1 lac Hte Δ(mcrA)183 Δ(mcrCB-hsdSMR-mrr)173 tet^R F'[proAB $lacI^qZ\Delta M15$ Tn10(Tet^R Amy Cm^R)]).

Experiments investigating G-quadruplex sequences occurring in ORFs *(see Chapter 3.1.2)* were conducted in *E. coli* BL21 (DE21) gold (Invitrogen, Darmstadt; F^- ompT gal dcm lon $hsdS_B(r_B^- m_B^-)$ λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])).

Experiments investigating the intrastrand triplex motif "TM" were conducted with the *E. coli* K12 MG1655 strain (Invitrogen, Darmstadt; F^- λ^- ilvG- rfb-50 rph-1).

6.5 Enzymes, kits and compounds

Table 6.2: Enzymes and kits.

Enzyme/Kit	Supplier
Antarctic phosphatase	NEB, Frankfurt
BSA	NEB, Frankfurt
DNAse I	NEB, Frankfurt
DNeasy Blood and Tissue kit	Qiagen, Hilden
DpnI	NEB, Frankfurt
E. coli RNA polymerase	NEB, Frankfurt
Gal-Screen™ β-galactosidase reporter gene assay system	Life Technologies, Darmstadt
GeneRuler™ DNA ladder (1kb, 100bp, Ultra low range)	NEB, Frankfurt
HindIII	NEB, Frankfurt
LumiGLO	NEB, Frankfurt
NcoI	NEB, Frankfurt
Phusion HotStart II DNA-polymerase	Thermo Scientific, Massachusetts
Proteinase K (50 µg/mL)	NEB, Frankfurt
Quick ligation kit	NEB, Frankfurt
RNAeasy Mini Kit	Qiagen, Hilden
RNAse A	NEB, Frankfurt
Roti ® -Aqua-Phenole	Roth, Karlsruhe
Roti ® -Phenole	Roth, Karlsruhe

Enzyme/Kit	Supplier
Rotiphorese acrylamide 25% (19:1)	Roth, Karlsruhe
ssRNA ladder	NEB, Frankfurt
Superscript III reverse transcriptase	Invitrogen, Darmstadt
T4 Polynucleotidekinase (PNK)	NEB, Frankfurt
VENT exo- polymerase	NEB, Frankfurt
XhoI	NEB, Frankfurt
Zuppy™ Plasmid Miniprep Kit	Zymo Research, Freiburg
Zymo DNA clean & concentrator kit	Zymo Research, Freiburg
Zymoclean Gel DNA Recovery Kit	Zymo Research, Freiburg

Table 6.3: Compounds.

Compound	Supplier
360A	Corinne Guetta, Institute Curie-UMR176
Anti-His-Antibody	AD1.1.10, Santa Cruz Biotechnology
Anti-RNA-polβ-Antibody	8RB13, Santa Cruz Biotechnology
Arabinose	VWR, Darmstadt
Goat-anti-mouse IgG-HRP	SC-2005, Santa Cruz Biotechnology
KlenTaq protein	AG Marx, Universität Konstanz
NMM	Sigma Aldrich, Hamburg
Phen DC_3	Corinne Guetta, Institute Curie-UMR176
Phen DC_6	Corinne Guetta, Institute Curie-UMR176
SYBRgreen	Sigma Aldrich, Hamburg
TMPyP z	Diana Gonçalves

6.6 Solutions, buffers and media

Table 6.4: General solutions and buffers.

Solution or buffer	Components
Agarose gel staining solution	Ethidium bromide (0.5 µg/mL in 0.5xTBE)
Agarose loading buffer	Glycerol (30% v/v) Bromphenolblue (0,25% w/v) Xylencyanol
Blocking buffer	Non-fat milk powder (5%) PBST (0.1%)
Denaturing PAGE loading dye	Formamide (80% v/v) EDTA (2 mM)
DMS stop solution	2-Mercaptoethanole (1 M) Sodium acetate (0.5 M) carrier DNA (50 µg/µl)

Solution or buffer	Components
DNA sequencing stop solution	Sodium acetate (0.5 M) carrier DNA (50 µg/µl) pH 6.0
PBS buffer (1x)	NaCl (137 mM) KCl (2.7 mM) Na_2HPO_4 (4.3 mM) KH_2PO_4 (1.47 mM) pH 7.4
PBST buffer (1x)	Tween 20 (20 g/L) 1 x PBS
RNA loading buffer	Formamide (80%) Bomphenolblue (0.025%) Xylenecyanol (0.025%) EDTA (50 mM)
SDS loading buffer (6x)	Tris-HCL (0.5M, pH 6.8, 1.2 mL) Glycerol (4.7 mL) SDS (1.2 g) Bromphenol blue (6 mg) DTT (0.93g) Total volume of 10 mL
SDS running buffer (10x)	Roth, Karlsruhe Tris base (30.3 g/L) Glycine (144 g/L) SDS (1g/L) pH 8.3
Sequencing dye	Formamide (80% v/v) EDTA (2 mM) Bromphenolblue (0.25% w/v) Xylenecyanol (0.025%)
SET buffer (1x)	NaCl (150mM) EDTA (15 mM) Tris-HCL (60 mM) pH 8.3
TBE buffer (1x)	Tris Base (89 mM) Boric acid (89 mM) EDTA pH 8.0 (2 mM)
Transfer buffer	Tris (25 mM) Glycin (192 mM) Ethanol (20% v/v)
Tris buffer	Tris-HCl (10mM) pH 7.5

Table 6.5: Media.

Medium	Components
LB Agar	Tryptone (10 g/L) Yeast extract (5 g/L) NaCl (10 g/L) Agar (20 g/L)

Medium	Components
	pH 7.0
LB medium	Tryptone (10 g/L) Yeast extract (5 g/L) NaCl (10 g/L) pH 7.0
M9 medium	M9 salts (1x) $MgSO_4$ (2 mM) Glucose or glycerol (0.4% (w/v)) $CaCl_2$ (100 µM) Vitamin mix (1x) or Casamino acids (0.1% (w/v))
M9 salts (5x)	Na_2HPO_4 (64 g/L) KH_2PO_4 (15 g/L) NaCl (2.5 g/L) NH_4Cl (5 g/L)
SOC medium	Yeast extract (0.5% (w/v)) Tryptone (2% (w/v)) NaCl (10 mM) KCl (2.5 mM) $MgCl_2$ (10 mM) $MgSo_4$ (10 mM) Glucose (20 mM)
Vitamin mix	Cyanocobalamin (100 mg/L) Aminobenzoic acid (80 mg/L) D(+)-biotin (20 mg/L) Niacin (100 mg/L) Ca-D(+)-pantothenic acid (100 mg/L) Pyridoxamine chloride (300 mg/L)

6.7 Laboratory consumables

Table 6.6: Laboratory consumables.

Consumable	Supplier
96-well plates (black)	Greiner, Sigma Aldrich, Hamburg
96-well plates (clear)	Greiner, Sigma Aldrich, Hamburg
Conical tubes (15 mL, 50 mL)	VWR, Darmstadt
Electroporation cuvettes	Biorad, München
Glass wool	VWR, Darmstadt
Gloves latex	MaiMed, Neuenkirchen
Gloves nitril	VWR, Darmstadt
PCR tubes	VWR, Darmstadt
Petri dishes	Roth, Karlsruhe
Pipette tips	Süd-Laborbedarf GmbH, Gauting
Razorblades	Schneider

Consumable	Supplier
Reaction tubes	VWR, Darmstadt
Sephadex G-25 column	GE Healthcare, München
UV cuvettes	Eppendorf, Hamburg
Western Blot Millipore Immobilon P	Roth, Karlsruhe
Whatmanpaper	GE Healthcare, München

6.8 Equipment

Table 6.7: Equipment.

Equipment	Supplier
96-well plate incubator	Heidolph Inkubator 1000 and Titramax 1000
CCD-Imager	Peqlab Fusion Xpress
Circular dichroism spectrometer	Jasco J-815, Groß-Umstadt
Culture incubation shaker	HT infors Ecotron,
Electroporator	Eppendorf Elektroporator 2510, Hamburg
Fluorescence plate reader	Tecan infinite M200, Männedorf, Switzerland
Gel documentation device	Biometra, Göttingen
Gel drier	Biorad, München
Heating block	Stuart, Staffordshire, UK
Laboratory pipettes	Eppendorf, Hamburg
Microwave	Privileg
NanoQuant plate	Tecan, Männedorf, Switzerland
PAGE gadget	Biorad SequiGel GT, München
PCR cycler Thermocycler Gradient	Biometra Thermocycler, Göttingen
pH meter	Metler Toledo, Seven easy
qPCR cycler	TOptical thermocycler, Biometra, Göttingen
Sonicator	Bender&Hobein, Transsonic T310/H
Table top centrifuge	Eppendorf mini spin, Hamburg
Thermomixer	Eppendorf, Hamburg
UV light table	Biometra, Göttingen
UV spectrometer	Tecan infinite M200, Männedorf, Switzerland
Voltage generator	Consort E833, Sigma Aldrich, Hamburg
Vortexer	VWR, Darmstadt
X-ray screen cassette	Fujifilm BAS Cassette 2 2048, Düsseldorf
X-ray screen reader	Biorad, München
X-ray screens	Fuji, Düsseldorf

6.9 Software

Table 6.8: Software and web servers.

Software	Producer
Aj qPCR 2.1	Analytik Jena
CorelDraw Suite 16	Corel
GraphPad Prism 5	GraphPad
i-Control 1.4.9	Tecan
mfold web server	Micheal Zuker http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi
MS Office Excel 2007	Microsoft
MS Office Word 2007	Microsoft
NCBI blastn web server	http://www.ncbi.nlm.nih.gov/BLAST
Quadbase Proquad web server	http://quadbase.igib.res.in/
Quantity one 4	Biorad

7 METHODS

7.1 Oligonucleotide design

Oligonucleotides for *in vitro* analyses have been designed either according to their folding properties or according to sequences naturally occurring in genomes, respectively. *Table 7.1* lists sequences used for the investigation of positional effects of G-quadruplex motifs *(see Chapter 3.1),* and *Table 7.2* lists oligonucleotides used for the investigation of the TM *(see Chapter 3.2).*

Table 7.1: Oligonucleotide sequences (5' to 3').

These oligonucleotide sequences were used for *in vitro* experiments investigating positional effects of G-quadruplexes on gene expression in *E. coli*.

Name	Sequence 5‘ to 3‘
G_3T	GGGTGGGTGGGTGGG
G_3A	GGGAGGGAGGGAGGG
G_2T	GGTGGTGGTGG
G_2CT	GGCTGGCTGGCTGG
ctrl 1	GGGTGTGTGTGTGTG
ctrl 2	CACTCACTCACTCCC
oxyR	GGCGAUGGAGGAUGGAUA
oxyR M1	GUCGAUGGAGGATUUGAUA
kdpD RNA	GGCGUGGGGCUGGGGCUGGCG
kdpD DNA	GGCGTGGGGCTGGGGCTGGCG

Table 7.2: Oligonucleotide sequences (5‘ to 3‘).

These oligonucleotide sequences were used for the investigation of the TM in *E. coli*.

Name	Sequence 5' to 3'
type A	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT
type A mm	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG
type B	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT
type B mm	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG
control	CCCTCGCCCCTTTGCCGAGAGCGTTAGCGTGAGCGG

7.2 Radioactive labeling of oligonucleotides

Primers used for *in vivo* footprinting experiments were 5'-end-labeled with γ-^{32}P-ATP prior to usage in primer extension assays. 5'-end-labeling is performed by phosphate exchange reaction through T4-polynucleotide kinase (T4-PNK). *Table 7.3* shows the composition of the labeling reaction.

Table 7.3: Reagents and mixture composition for 5'-end-labeling with γ-^{32}P-ATP.

Reagent	Volume [µL]	Final concentration
water	39.5	
T4-PNK-buffer A (10x)	5	1 x
DNA (50 µM)	2	2 µM
γ-^{32}P-ATP	1.5	
T4-PNK	2	

7.3 DNA quantification

Oligonucleotide concentrations were determined using the Tecan Infinite M200 NanoQuant plate. Quantification of DNA and RNA was conducted by UV absorption measurements at 260 nm using the UV/VIS photometer. Water or the respective buffer was used as a reference. Based on Lambert-Beer's law the DNA concentration could be calculated. Extinction coefficients as well as molecular weight and melting behavior were determined using the web application "IDT Oligo Analyzer" (http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/).

7.4 Ethanol precipitation

Via ethanol precipitation unwanted salts were removed and DNA samples are concentrated. One tenth volume of Sodium Acetate buffer (1 M, pH 5.4) was added to equalize ion concentrations. After addition of three volumes of ethanol (100%) the solution was incubated for at least 15 min at -80°C. The sample was centrifuged at 14,000 rpm for 15 min. The supernatant was discarded and the DNA pellet was dissolved in a suitable volume of water.

7.5 *In vitro* transcription

In vitro transcription was performed to analyze effects on transcription of the G-quadruplex located in the core promoter *(see Chapter 3.1).* PCR products were used as templates for *in vitro* transcription. PCR products were generated as described below *(see Chapter 7.6),* using constructs (1) pQE-J06-coreG_3T and (5) pQE-J06-coreG_3Tctrl1 *(for plasmid sequences see appendices)* as templates and the following primers: Forward: 5'-AGTGCCACCTGACGTCTAAGAAACC-3' and Reverse: 5'-GATGATGATGATGATGATGGC-3'. PCR products had a length of 187 base pairs containing the pQE-J06 promoter and the eGFP gene start. They were purified via agarose gel purification and 0.5 µg were used for the *in vitro* transcription. Prior to *in vitro* transcription templates were folded by heating to 95°C and cooling down to 4°C within 50 min. *In vitro* transcription was performed with α-^{32}P-GTP and *E. coli* RNA polymerase according to the manufacturers protocol (NEB). *Table 7.4* shows the composition of the *in vitro* transcription reaction.

Table 7.4: Reagents and mixture composition for *in vitro* transcription.

Reagent	Volume [µL]	Final concentration
PCR product	12.5	0.01 µg/mL
CTP, ATP, UTP mix (25 mM)	3.75	1x
GTP (5 mM)	1	2 µM
α-^{32}P-GTP	1	
E. coli RNA polymerase buffer (5x)	10	1x
RNAse Inhibitor (40 U/µL)	0.5	
Pyrophosphatase (0.1 U/µL)	0.5	
E. coli RNA Polymerase (1 U/µL)	0.6	
water	20.15	

In vitro transcribed samples were ethanol precipitated and analyzed by means of a 10% denaturing polyacrylamide gel electrophoresis (PAGE).

7.6 Polymerase chain reaction (PCR)

PCR was performed to construct of templates for *in vitro* transcription and to obtain the desired constructs via the cloning procedures. The general program used for PCR with Phusion Hot Start II polymerase using the BioMetra thermal cycler is depicted in *Table 7.5* and the composition of the reaction mixture is described in *Table 7.6.*

Table 7.5: Standard PCR program for Phusion DNA polymerase.

Cycles 2 to 4 were repeated 30 times in total.

PCR cycle	Temperature [°C]	Time [s]
Lid preheating	98	-
1. Initial denaturation	98	120
2. Denaturation	98	50
3. Annealing	58	60
4. Elongation	72	120
5. Final elongation	72	300
6. Cooling	4	-

Table 7.6: Reagents and reaction mixture for standard PCR.

Reagent	Volume [µL]	Final concentration
HF buffer (5x)	10	1x
dNTPs (25 mM)	1	0.5 µM
Template	1	0.6 ng/µL
Primer fw (5 µM)	4	0.4 µM
Primer rev (5 µM)	4	0.4 µM
DMSO (100%)	3 (1)	6% (2%)
Phusion Hot Start II Polymerase (2U/µL)	0.5	
water	26.5 (24.5)	

The PCR reaction was evaluated by means of agarose gel electrophoresis. Where required, PCR products were purified via agarose gel electrophoresis as described in *Chapter 7.11.1.*

7.7 Circular Dichroism (CD)

Circular dichroism (CD) spectroscopy measures differences in the absorption of left-handed polarized light versus right-handed polarized light, which arise due to structural asymmetry. By comparing the CD spectra to empiric data secondary structure of DNA can be determined.
For sample preparation oligonucleotides were prepared as a 5 µM solution in Tris buffer (*see Table 6.4)* supplemented with 100 mM KCl or NaCl in a reaction volume of 600 µL. DNA folding was facilitated by heating to 98°C for 5 min, followed by slow cooling to 20°C over night.

CD spectra were measured on a JASCO-J815 spectropolarimeter equipped with a MPTC-490S/15 multicell temperature unit using a 1 cm optical path. 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 were zero-corrected at 320 nm.

7.8 Thermal denaturation

For thermal denaturation oligonucleotides were prepared as for CD measurements *(see Chapter 7.7)*. Folded samples were heated from 20°C to 100°C with a heating rate of 0.5°C min-1. The CD signal at the respective maximum was recorded every 0.5°C (290 nm; 265 nm). The temperature of the half-maximal decay of ellipticity T_m was obtained from the normalized ellipticity decrease.

7.9 NMR measurements

NMR spectra were acquired at 278 K on a Bruker Avance III 600 MHz spectrometer equipped with a TCI-H/C/N triple resonance cryoprobe. 100 µM of the respective oligonucleotide was dissolved in 1x PBS buffer 5% Vol. D_2O as field lock. Triplex structures were folded by heating up to 98°C and slowly cooling down to room temperature. ^{1}D-proton spectra were acquired with 32,000 data points using 10k accumulated scans due to low sample concentration and processed with an exponential line broadening window function.

Solvent suppression was achieved by excitation sculpting (344). Acquired data were processed and analyzed using Bruker Topspin and MestReNova software.

7.10 Phenol/chloroform extraction

During *in vivo* footprinting and mRNA level determination procedures nucleic acids were purified from protein contamination using phenol/chloroform extraction, with Rotiphenol for DNA and AquaPhenol for RNA *(see Table 6.2).* For this purpose, 1 volume phenol/chloroform mixture was added to the nucleic acid sample and mixed by vortexing for 5 s. Phases were separated by centrifugation at 15,000 rpm for 5 min, and the aqueous phase was carefully transferred into a clean reaction tube. Phenol/chloroform extraction was repeated and further purification and concentration of the sample was achieved by ethanol precipitation (*see Chapter 7.4)*.

7.11 Electrophoretic studies

7.11.1 Oligonucleotide purification via agarose gel electrophoresis

PCR products used for cloning or *in vitro* transcription were purified via agarose gel electrophoresis. If not described otherwise, 2.5% (w/v) agarose was dissolved in 0.5x TBE by heating and poured into an appropriate gel tray. After cooling down to 25°C the agarose gel was placed in the appropriate gel chamber filled with 0.5x TBE, the nucleotide samples mixed with agarose loading buffer *(see Table 6.4)* were loaded and run with a voltage of 5-10 V/cm (distance between electrodes). Gels were stained in 0.01% (w/v) ethidiumbromide in water for 15 min and briefly washed in water. To avoid DNA degradation by UV light, preparative gels were cut in two pieces. One piece was irradiated with UV light in order to identify and mark the position of the desired band. After reunion of the two agarose gel pieces on a plastic tray, the appropriate piece was cut out of the untreated gel piece.

7.11.2 Oligonucleotide purification via preparative PAGE

Oligonucleotides were purified via preparative PAGE prior to their usage in footprinting experiments. A denaturing polyacrylamide-urea gel electrophoresis was used to separate the full-length oligonucleotide from other contaminants (see *Table 7.7*).

Table 7.7: Reagents and mixture composition for preparative PAGE.

Reagent	Volume [mL]	Final concentration
Acrylamide 25% (19:1)	80	10%
TBE (10x) containing Urea (9M)	20	1 x
Urea (9M)	100	4.5 M
APS	1.6	
TEMED	0.08	

Samples were mixed with denaturing PAGE loading dye (*see Table 6.4*) in 1:1 ratio and heated at 95°C for 3 min. The gel was pre-run and warmed in 1 x TBE buffer for at least 30 min at 700 V. The samples were loaded on the gel and run at 700 V until the desired resolution was obtained as determined empirically.
After the completion of electrophoresis the gel was wrapped with plastic film and analyzed in UV light. Excessive UV exposure was avoided. The oligonucleotide bands were marked, cut out and placed in a 1.5 mL Eppendorf tube. About 1 mL water was added before incubation with rotation at room temperature over night. The solution was filtered with glass wool and ethanol precipitated (*see Chapter 7.4*).

7.11.3 Agarose gel electrophoresis

Agarose was dissolved in 0.5 x TBE by heating and poured into an appropriate gel tray. After cooling down to 25°C, the agarose gel was placed in the appropriate gel chamber filled with 0.5x TBE, the nucleotide samples were mixed with agarose loading buffer *(see Table 6.4),* loaded and run with a voltage of 5-10 V/cm (distance between electrodes). Gels were stained in 0.01% (w/v) ethidiumbromide in water for 15 min and briefly washed in water. Analytical gels were visualized and photographed by a Biometra GelDoc using UV light with a wavelength of 260 nm.

7.11.4 Denaturing, analytical PAGE

Analytical PAGE was performed to analyze *in vitro* transcription products and *in vivo* footprinting experiments. DNA fragments were separated on a 10% denaturing PAGE (*see Table 7.7*). Samples were prepared by resuspending DNA pellets in 20 µl sequencing dye *(see Table 6.4)* and heating up to 95°C for 3 min. The gel was pre-run and warmed at 1600 V for at least 30 min before the samples were loaded. The gel was run in 1x TBE buffer for 3 hours at 1600V. After the run the gel was dried in the gel vacuum-dried and placed in an X-ray film cassette over night. The next day the film was screened on Bio-Rad Phosphorimager.

7.11.4.1 SDS polyacrylamide gel electrophoresis

SDS polyacrylamide gel electrophoresis (SDS-PAGE) was used prior to Western blotting for protein separation according to size. For this purpose, proteins were isolated from cell extracts: 20 mL of exponentially grown cells (OD_{600} = 0.4) were centrifuged at 4000 rpm for 10 min. The pellet was dissolved in 100 µL 1x SDS PAGE loading dye prior to sonification (2 times for 30 s) and centrifugation at 13,400 rpm for 20 min. The supernatant was used for the SDS-PAGE. Equal amounts of cell protein were loaded on the SDS-PAGE (determined either by means of a Bradford assay (100 µg) or via protein extraction from an equal number of cells ($1x10^7$ cells)). The SDS-PAGE gels were composed of a stacking and a separating gel which were prepared as described in *Table 7.8*. The proteins of interest were KdpD and KefC, both were fused to a His-Tag and had sizes of 99.7 and 68.2 kDa – they were analyzed on a 10% SDS PAGE in a BioRAD MINI protean gel apparatus. Prior to loading, samples were heated to 95°C for 5 min. The prestained PageRuler protein ladder was used as a protein size reference. The gel was run in 1x SDS Running buffer at 80 V for 2 h and 45 min.

Table 7.8: Reagent composition for SDS PAGE.

Reagent	Volume [mL]	Final concentration
	Stacking gel	
water	1.46	
Rotiphorese 40 (37.5:1)	0.25	10%
Tris-HCl (1 M, pH 8.8)	0.25	
10% SDS	0.02	
10% APS	0.02	

Reagent	Volume [mL]	Final concentration
TEMED	0.02	
	Separating gel	
water	2.4	
Rotiphorese 40 (37.5:1)	1.25	10%
Tris-HCl (1 M, pH 8.8)	1.25	
10% SDS	0.05	
10% APS	0.05	
TEMED	0.05	

7.12 *In vivo* DMS footprinting

In vivo DMS footprinting was performed according to the *in vitro* method described by Sun and Hurley (345) which is based on the method of Maxam and Gilbert (320). In each case DNA probing was performed in the living bacterium: On plasmid level for G-quadruplex constructs and on chromosomal DNA for investigation of the TM. Probed DNA was isolated and analyzed by primer extension with a radioactively labeled primer (labeling procedure as described in *Chapter 7.2)*. *Chapters 7.12.1 and 7.12.2,* describe exact procedures for plasmid and chromosomal DNA, respectively.

7.12.1 In vivo footprinting of plasmid DNA

Bacteria were grown in LB medium supplemented with carbenicilline (100 µg/µL) over night. 50 µL of a 10% DMS solution was added to 5 mL overnight bacterial culture in LB medium and incubated for 5 min at 37°C before placing on ice. 2 mL of the bacterial culture were pelleted at 4°C and plasmid DNA was extracted with the Zuppy ™ Plasmid Miniprep Kit. DNA was digested with NcoI *(digestion procedure as described in Chapter 7.14.1)* to generate a full-length product in the primer extension assay and purified via phenol/chloroform extraction (*see Chapter 7.10)*. Sequencing controls were generated by treating isolated plasmid DNA with formic acid (FA) for purine sequencing or with hydrazine (Hy) for pyrimidine sequencing, as described by Maxam and Gilbert (320). Cleavage at the modified sites was performed by addition of 10% piperidine at 94°C for 30 min. Piperidine was removed in a vacuum concentrator. For primer extension the primer (5'-AGGCGTATCACGAGGCCCTTTC-3') was radioactively 5'-end-labeled with γ-^{32}P-ATP *(see Chapter 7.2)*. Primer extension was performed with VENTexo- polymerase *(see Table 7.9 and 7.10)*.

Table 7.9: Reagents and mixture composition for footprint reaction.

Reagent	Volume [µL]	Final concentration
10 x Thermo Pol Puffer	10	1x
$MgSO_4$ (0.1 M)	1	1 µM
dNTPs (10 mM)	2	0.2 mM
primer	4	
template	50	
Vent exo- polymerase (2 U/µL)	3	
water	30	

Table 7.10: Primer extension program.

Primer extension	Temperature [°C]	Time [s]
Lid preheating	98	-
1. Initial denaturation	94	600
3. Annealing	64	120
4. Elongation	75	180
5. Final elongation	75	600
6. Cooling	4	-

Primer extension products were purified by phenol/chloroform extraction *(see Chapter 7.10)* and ethanol precipitation *(see Chapter 7.4)* and analyzed on a 10% denaturing PAGE gel *(see Chapter 7.11.4)*

7.12.2 In vivo footprinting of genomic DNA

50 mL *E. coli MG1655* cells were grown until exponential phase (OD_{600} = 0.1) in M9 minimal medium at 37°C *(see Table 6.5)*. 50 µL of a 10% DMS solution were added to the bacterial culture and incubated for 5 min at 37°C before placing on ice. The whole 50 mL of the bacterial culture were pelleted at 4°C and washed with 1 x PBS buffer. Next, cell pellets were dissolved in 480 µL SET buffer (346) and cell lysis was performed by addition of 20 µL of 20% SDS for 30 min at 37°C. Finally, 1.5 µL of Proteinase K was added and phenol/chloroform purification *(see Chapter 7.10)* was used for extraction of chromosomal DNA. The DNA was digested with RsaI to generate a full-length product in the primer extension assay and purified via phenol/chloroform extraction. Sequencing controls were generated by treating isolated chromosomal DNA with formic acid (FA) for purine sequencing

or hydrazine (Hy) for pyrimidine sequencing, as described by Maxam and Gilbert (320). Cleavage at the modified sites was performed by addition of 10% piperidine at 94°C for 30 min. The piperidine was removed in a vacuum concentrator. For primer extension the primer (5'-GAGGTAAATCGGAAGGGAAGAGG-3') was radioactively 5'-end-labeled with γ-^{32}P-ATP *(see Chapter 7.2)*. Primer extension was performed with VENT(exo-) polymerase, as described in *Table 7.9* and *Table 7.10* (cycles 2 to 4 were repeated 40 times in total). Primer extension products were purified by phenol/chloroform extraction *(see Chapter 7.10)* and ethanol precipitation *(see Chapter 7.4)* and analyzed on a 10% denaturing PAGE gel *(see Chapter 7.11.4)*

7.13 Determination of RNA levels

Bacteria were grown in LB medium supplemented with carbenicilline (100 µg/µL) over night or until late exponential phase (OD_{600} = 0.3) at 37°C. Total RNA was extracted using RNAeasy Mini Kit. Isolated RNA was digested with DNAseI (2 times for 10 min) and further purified by phenol/chloroform extraction using Aqua Phenol *(see Chapter 7.10)*. The reverse transcription reaction was performed with 1 µg total RNA and random hexamer priming using the Superscript III reverse transcriptase (SSIIIRT) in a total volume of 20 µL for 60min at 50°C. First, RNA, primer and dNTPs were incubated before adding the reverse transcriptase was added *(see Table 7.11)*.

Table 7.11: Reverse transcription reaction.

Reagent	Volume [µL]	Final concentration
RNA	6	1 µg
N6 Primer (214 ng/µL)	1	10.7 ng/µL
dNTPs (10 mM)	1	0.5 mM
water	10	
Incubated at 65°C for 5 min and placed on ice for 1 min		
FS buffer (5x)	4 µL	1 x
DTT (0.1 M)	1	5 mM
SSIIIRT	1	
water	1	
Incubated at 25°C for 5 min, followed by 50°C for 60 min and 70°C for 15 min		

Real-time PCR analysis was performed on a TOptical thermocycler. Each reaction mixture was prepared using Phusion Hot-Start Polymerase II for amplification and SYBRgreen for

detection in a total volume of 13 µL *(see Table 7.12)* with the standard PCR program *(see Table 7.5)*. The following primers were used for the amplification reaction of the eGFP mRNA (fw: 5'-AAGCTGACCCTGAAGTTCATCTGC-3'; rev: 5'-TTCACCTCGGCGCGGGTCTTGTAG-3'), the β-galactosidase mRNA (fw:5'-ATGACCATGATTACGGATTCACTG-3'; rev: 5'-GCGATCGGTGCGGGCCTCTTC-3') and the *ssrA* mRNA (reference gene; fw: 5'-ACGGGGATCAAGAGAGGTCAAAC-3'; rev: 5'-GGACGGACACGCCACTAAC-3'). RNA levels were calculated assuming a static PCR efficiency of two for each primer pair and determined relative to the expression of the genomically encoded *ssrA* gene.

Table 7.12: Reaction mixture for semiquantitative RT-PCR.

Reagent	Volume [µL]	Final concentration
HF buffer (5x)	2.6	1x
cDNA template	0.6	0.5-1 µg
dNTPs (2 mM)	2	0.2 mM
Primer fw (10 µM)	0.65	0.5 µM
Primer rev (10µM)	0.65	0.5 µM
Phusion Hot Start II polymerase (2 U/µL)	0.078	
SYBRgreen (100x)	0.078	0.6x
water	7.044	

7.14 Cloning procedures

The plasmids used were either based on the pQE-TriSystem (Qiagen) or the pBAD18a (275) vector systems. Promoters used were the J06 promoter (modified from the Anderson promoter library (http://parts.igem.org/Promoters/Catalog/Anderson) and the araBAD promoter (275) *(promoter sequences are shown in Figure 3.2)*. Standard molecular cloning procedures were performed as described in literature (347) and below.

7.14.1 Restriction endonuclease digest

During cloning procedures, DNA was digested with restriction enzymes to generate the correct fragments and vectors for subsequent ligation. In most cases, digestion of plasmids mostly was performed with two restriction endonucleases (High Fidelity) in parallel. The standard reaction mixture which was incubated for 3 h at 37°C is shown in *Table 7.13*. If only

one restriction endonuclease was used the amount of water was adjusted accordingly. Depending on the restriction enzyme, the reaction conditions were adapted according to the manufacturer's suggestion.

Table 7.13: Standard reaction mixture for digestion with restriction endonucleases.

Reagent	Volume [µL]	Final concentration
Template (e.g. plasmid DNA (30 ng/µL))	30	18 ng/µL
Restriction enzyme 1 (20 U/µL)	1	
Restriction enzyme 2 (20 U/µL)	1	
NEB CutSmart buffer (10x)	5	1 x
BSA (20 mg/mL)	0.5	0.2 mg/mL
water	12.5	

Following a PCR the template DNA was removed from the reaction by digestion with the *DpnI* restriction endonuclease.

7.14.2 Ligation

Digestion products were ligated to obtain the desired plasmid constructs. For this purpose, double-stranded DNA inserts and linear plasmids were ligated in a 3 to 1 molar ratio using the Quick ligation kit at 25°C for 20 min. The ligation products were purified using the DNA clean & concentrator kit.

7.14.3 Electro-transformation of plasmids in E. coli

Transformation of plasmid DNA into *E. coli* cells was performed by electroporation. 80 µL of electrocompetent *E. coli* cells were thawed on ice, mixed with 2 µL plasmid DNA (approximately 60 ng) and transferred into a pre-chilled electroporation cuvette. After transformation (Voltage = 1800 V, Time constant (τ) = 5 ms), cells were incubated in 1 mL SOC at 37°C for 1 h and eventually plated on agarose plates supplemented with the appropriate antibiotic.

7.14.4 Whole plasmid PCR

For cloning via whole plasmid PCR, primers with 5' overhang were used and PCR was performed as described in *Chapter 7.6.* These overhangs contained the sequence to be inserted or changed in the plasmid. Secondary structure forming sequences were split and partly attached to each of the primers in order to prevent secondary structure formation within one primer. Primers and respective plasmids are listed in *Chapter 7.14.6 and Table 13.8.* PCR products were purified with the DNA clean & concentrator kit and digested with the DpnI restriction endonuclease *(see Chapter 7.14.1).* Digestion products were purified via agarose gel electrophoresis *(see Chapter 7.11.1)* and ligated as described in *Chapter 7.14.2* before being transformed into the appropriate cells *(see Chapter 7.14.3).* Successful cloning was verified by DNA sequencing (GATC).

7.14.5 Introduction of a DNA insert

Some vectors were cloned by introduction of an insert. For this purpose, the plasmid DNA was digested with XhoI and NcoI *(see Chapter 7.14.1)*. After digestion parental plasmid DNA was dephosphorylized with Antarctic phosphatase to avoid relegation of the parental vector. Dephosphorylized vectors were purified via agarose gel electrophoresis *(see Chapter 7.11.1)*. Oligonucleotides used as inserts were hybridized by heating to 95°C followed by cooling to 4°C within 50 min. Hybridized oligonucleotides were digested with XhoI and NcoI and purified with the clean & concentrate kit prior to ligation *(see Chapter 7.14.2)* with the plasmid. Successful cloning was verified by DNA sequencing (GATC).

7.14.6 Design of Plasmid constructs

Constructs based on the pQE-J06-eGFP plasmid were created either by whole plasmid PCR or by ligation of an insert containing the modified 5'-UTR with the digested vector.

pBAD-eGFP-based constructs used to investigate positional effects of G-quadruplexes in regulatory regions were created by whole plasmid PCR.

Constructs with natural G-quadruplexes from *E. coli* MG1655 surrounding the SD region were investigated with a β-galactosidase reporter in the pBAD-18 backbone. To introduce the lacZ reporter gene in the pBAD-18 vector, we performed PCR on the MG1655 genomic DNA (isolated via Qiagen DNeasy Blood and Tissue kit) using elongated primers containing the respective SD surrounding and restriction sites. *LacZ* insert and pBAD-18 plasmid were both digested with HindIII and SalI, purified and ligated. Phusion Hot Start II Polymerase was utilized for PCR amplifications, and ligations were performed with the Quick ligation kit.

pBAD-18 constructs with G-quadruplexes occurring in *E. coli* and *Salmonella* ORFs were generated by whole plasmid PCR.

The chromosomal TM mutant was generated and kindly provided by Dr. Stefanie Wagner: *E. coli* BW19610 (348) was used for cloning, plasmid purification and maintenance. To generate an *E. coli* strain carrying a mutated 6' TM, bp 281075-281622 of *E. coli* K-12 MG1655 were amplified introducing pointmutations by overlapping PCR using primer pairs SW09(5'-ACGCGTCGACAGCCGGTGGCAGGTG-3') / SW10 (5'-CGCTCACGCCGGCGCTCTCGGCAAAGGGGCGAGGGGGAAAAGATG-3') and SW11 (5'-TTTGCCGAGAGCGCCGGCGTGAGCGGCAATATGTGATCCAGC-3') / SW12 (5'-GCTCTAGACGCCTGCTTTGATC-3') and cloned into pKNG101 (349) using restriction sites SalI/XbaI. The resulting plasmid pSW05 was verified by sequencing (GATC biotech) and transferred into *E. coli K12* MG1655 by electroporation. Allelic exchange was selected by plating on 5% sucrose according to Kaniga et al (349). The *E. coli* 6' mutant carrying the mutated 6' TM was verified by sequencing (GATC biotech).

All constructs with respective cloning procedures and primers are listed in *Table 13.8* in the appendices. Bacteria were routinely grown in LB or M9 medium supplemented with 50 µg/mL streptomycin or carbenicillin for plasmid selection if necessary.

7.15 Determination of eGFP expression levels

Gene expression of constructs with G-quadruplexes in the promoter region and constructs with G-quadruplexes liberating the SD site was examined by eGFP readout for both pBAD and pQE vector systems. For this purpose, bacterial cultures were grown aerobically in LB Medium *(see Table 6.5)* over night or until late exponential phase (OD_{600} = 0.3) at 37°C. 100 µL of each culture were transferred into 96-well-microplates, and the eGFP fluorescence

was determined with a TECAN Infinite M200 plate reader (excitation wavelength = 288 nm, emission wavelength = 535 nm). Fluorescence values were OD_{600} corrected.

7.16 Determination of β-galactosidase expression levels

Gene expression of constructs bearing the naturally occurring SD-G-quadruplexes was determined by β-galactosidase assay. For this purpose, bacterial cultures were grown aerobically in LB Medium *(see Table 6.5)* over night at 37°C. Outgrown cultures were diluted 1:2000 with LB medium and induced by addition of 1 mM arabinose. When grown to OD_{600} = 0.5, *lacZ* expression was determined through the Gal-Screen β-galactosidase reporter gene assay system and luminescence was measured with a TECAN Infinite M200 plate reader. Luminescence values were corrected by dividing by the OD_{600} values.

7.17 Western Blot

Western Blot was performed prior to immunodetection with His-Tag specific antibodies. For this purpose, the proteins separated by SDS-PAGE were transferred onto a PVDF membrane (0.45 µM pore size). After incubation of the membrane in 100% Ethanol for 1 min, the transfer was performed using a constant electric field of 350 mA and 100 V for 60 min. Subsequently, the membrane was washed with 1x PBS for 5-10 min and incubated in blocking buffer over night. After washing in PBST (twice for 5-10 min) the membrane was incubated with the anti-His antibody (1:1000 dilution in PBST) with constant shaking for 2h. After removal of the anti-His antibody, the membrane was washed six times (10 min each) with PBST before the secondary antibody (goat-anti-mouse IgG-HRP) was added for 1h (1: 5000 dilution in PBST). Again the membrane was washed six times (10 min each) with PBST and once with PBS in order to remove the Tween. The protein was detected by chemiluminescence read-out on a CCD-Imager after incubation with the LumiGLO reagent as described by the manufacturer. As an internal standard to prove equal sample loading and for quantification, incubation with the Anti-rpoB-antibody instead of the anti-His was performed after stripping of the membrane (*see Chapter 7.18*).

7.18 Stripping procedure

Stripping was performed to eliminate an antibody interaction before the proteins on the Western blot should be detected with a secondary antibody against the RpoB protein. For this purpose, the membrane was washed three times for 10 min in PBST. Then the membrane was incubated for 1 h at 55°C in stripping buffer before it was washed again twice with PBST for 20 min. Afterwards the membrane can be incubated with another primary antibody (*see Chapter 7.17*).

7.19 Identification of long range interactions using Hi-C data

Raw data from Hi-C experiments of *E. coli* K-12 in the exponential growth phase taken from Voss et al (replicate 3 with cross linker) were re-analyzed with the Hi-C KNIME workflow (338). The data sets chosen from Voss et al. correspond to the data set from Cagliero et al. because the experimental setups were comparable (see Voss et al.). The *E. coli K12* MG1655 genome was divided into 20 kb fragments (bins) after mapping the raw read data to the genome using RazerS3 (350),(338)).The read frequencies were generated as .csv files without further normalization of the data. Fragments carrying TMs were highlighted and interactions between TM carrying fragments were extracted from the table and shown as heat map with a maximum of 100 reads for data from Voss et al. The different maxima reads are due to different sample sizes.

7.20 Genomic instability studies around the "TM" sequences

To identify TM motifs in different bacterial strains we defined the following search pattern: 5'-CCCTC[TG]CCCNNNNNNGGG[AT]GAGGGNNNGGG[TA]GAGGG[GTC-]-3', where N represents any nucleotide (A,T,G,C) and nucleotides in brackets stand for the different possibilities at the respective position, - means no nucleotide. TM motifs in accordance with this pattern were determined using the Perl program. Whole genome sequences of 56 *E .coli* strains were downloaded from the National Center for Biotechnology Information (NCBI) (updated on June 25th, 2014). Multiple whole genome alignment of all 56 *E. coli* genomes was performed via Mugsy 1.2.3 after which the locally collinear blocks (LCBs) were determined (351). The LCBs containing TM sequences were realigned by MAFFT v7 (352).

After all alignments the TMs sharing similar surrounding sequences were categorized into homologous loci. To calculate the sequence variability v_j around a particular TM within the LCBs, we scanned all the aligned sequences using the following formula: $v_j = \frac{1}{l}\sum_{j=j-(l-1)/2}^{j+(l-1)/2}(n_j - 1)$, where n_j describes the nucleotide status at the aligned site j (either A,T,C,G or gap) and l stands for the length of the regarded window, which here is 11 nucleotides. For the j^{th} site of an aligned sequence, v_j represents the average variability of the surrounding l nucleotides. Each aligned site has a corresponding v_j. To identify the dimension of the variable regions in the LCBs, we scanned the measured v_j values within one LCB. The start of a variable region was defined with $v_j > 0.9$ continuing for 10 consecutive nucleotides. A $v_j < 0.5$ for 10 continuous nucleotides defined the end of a variable region. That way we defined the variable sequence range for each LCB. We applied the same strategy for 3 control groups, containing random sites from the 56 *E. coli* genomes. Although analyzing 63 TM loci we determined only 48 v_j in total, because some TMs are located in close proximity and they share one variable region.

Categorization (no change, region missing, intergenic deletions, not found) of sequence changes for the 23 TM sequences of *E. coli* MG1655 was done using the nucleotide BLAST webserver (http://blast.ncbi.nlm.nih.gov/) (325) and by comparison of nucleotide BLAST (algorithm: megablast) results for each of the 23 TM sequences found in *E. coli* MG1655 with the 55 different *E. coli* genomes. Palindromic sites were regarded within one region, therefore 18 regions were aligned. We applied the following parameters: **Query sequence**: respective TM sequence (*see Table 3.4*), **Database**: “NCBI genomes (chromosome)”, **Organism:** all 56 genomes listed in *Table 3.6*.

8 ABBREVIATIONS

°C	degree celsius
µg	microgram
µL	microliter
µM	micromolar
360A	2,6-pyridine-dicarboxamide bisquinolinium
APS	ammonium persulfate solution
bp	base pair
CD	circular dichroism
DMS	dimethylsulfate
DNA	deoxyribonucleic acid
E. coli	*Escherichia coli*
eGFP	enhanced green fluorescent protein
g	gram
h	hour
L	liter
M	molar
min	minute
mL	milliliter
mm	millimeter
mRNA	messenger RNA
mRNA	messenger RNA
nm	nanometer
nM	nanomolar
NMM	N-methyl mesoporphyrin IX
NMR	nuclear magnetic resonance
nt	nucleotide

OD	optical density
ORF	open reading frame
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
Phen DC_3	6,6'-disubstituted-2,2'-bipyridine
Phen DC_6	2,9-disubstituted-1,10-phenanthroline
RBS	ribosome binding site
RNA	ribonucleic acid
SD	Shine-Dalgarno region
T4-PNK	T4 Polynucleotide kinase
TEMED	N,N,N',N'-tetramethyleneethylenediamine
TM	particular triplex motif:
TMPyP	5,10,15,20-tetrakis(Nmethyl-4-pyridyl)porphyrin
UTR	untranslated region
V	Volt

9 RECORD OF CONTRIBUTION

Chapter 3.1.1

I designed, conducted and analyzed all of the experiments.

Chapter 3.1.2

I designed and analyzed all of the experiments. Cloning procedures and Western Blotting was performed together with Astrid Joachimi (Technical Assistant, AG Hartig).

Chapter 3.2

This project was performed in collaboration with Dr. Stefanie Wagner (former Post Doc, AG Hartig), Peiwen Xiong (PhD Student, AG Meyer) and Prof. Dr. Tancred Frickey (Applied Bioinformatics, University of Konstanz).

Data shown in *Figure 3.16* were partly provided by Dr. Stefanie Wagner and Malte Sinn (Bachelor's Thesis).

Raw data (*Figure 13.1*) for analysis shown in *Figure 3.19* were provided by Dr. Stefanie Wagner.

Data for *Figure 3.21 A, Table 13.4, Table 13.5* and *Table 13.6* were provided by Peiwen Xiong. Analysis shown in *Figure 3.21* was performed in collaboration with Peiwen Xiong.

The ITxF website was designed in collaboration with Prof. Dr. Tancred Frickey and Dr. Stefanie Wagner. The programming was performed by Prof. Dr. Tancred Frickey.

Data shown in *Table 13.3* was provided by Dr. Kangkan Halder.

10 BIBLIOGRAPHY

1. Watson, J.D. and Crick, F.H. (1953) The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, **18**, 123-131.
2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.
3. Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737-738.
4. Bochman, M.L., Paeschke, K. and Zakian, V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nature reviews. Genetics*, **13**, 770-780.
5. Pearson, C.E., Zorbas, H., Price, G.B. and Zannis-Hadjopoulos, M. (1996) Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *Journal of cellular biochemistry*, **63**, 1-22.
6. Singleton, M.R., Scaife, S., Raven, N.D. and Wigley, D.B. (2001) Crystallization and preliminary X-ray analysis of RecG, a replication-fork reversal helicase from Thermotoga maritima complexed with a three-way DNA junction. *Acta crystallographica. Section D, Biological crystallography*, **57**, 1695-1696.
7. Yuan, C., Rhoades, E., Heuer, D.M. and Archer, L.A. (2005) Mismatch-induced DNA unbending upon duplex opening. *Biophysical journal*, **89**, 2564-2573.
8. Lobachev, K.S., Rattray, A. and Narayanan, V. (2007) Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells. *Frontiers in bioscience : a journal and virtual library*, **12**, 4208-4220.
9. Palecek, E. (1991) Local supercoil-stabilized DNA structures. *Critical reviews in biochemistry and molecular biology*, **26**, 151-226.
10. Hoyne, P.R., Edwards, L.M., Viari, A. and Maher, L.J., 3rd. (2000) Searching genomes for sequences with the potential to form intrastrand triple helices. *Journal of molecular biology*, **302**, 797-809.
11. Hoyne, P.R., Gacy, A.M., McMurray, C.T. and Maher, L.J., 3rd. (2000) Stabilities of intrastrand pyrimidine motif DNA and RNA triple helices. *Nucleic acids research*, **28**, 770-775.
12. Jain, A., Wang, G. and Vasquez, K.M. (2008) DNA triple helices: biological consequences and therapeutic potential. *Biochimie*, **90**, 1117-1130.
13. Capra, J.A., Paeschke, K., Singh, M. and Zakian, V.A. (2010) G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS computational biology*, **6**, e1000861.
14. 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 research*, **16**, 644-655.
15. Hershman, S.G., Chen, Q., Lee, J.Y., Kozak, M.L., Yue, P., Wang, L.S. and Johnson, F.B. (2008) Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic acids research*, **36**, 144-156.
16. Gellert, M., Lipsett, M.N. and Davies, D.R. (1962) Helix formation by guanylic acid. *Proceedings of the National Academy of Sciences of the United States of America*, **48**, 2013-2018.
17. Holder, I.T., Drescher, M. and Hartig, J.S. (2013) Structural characterization of quadruplex DNA with in-cell EPR approaches. *Bioorg Med Chem*, **21**, 6156-6161.
18. Hardin, C.C., Henderson, E., Watson, T. and Prosser, J.K. (1991) Monovalent cation induced structural transitions in telomeric DNAs: G-DNA folding intermediates. *Biochemistry*, **30**, 4460-4472.
19. Shiber, M.C., Braswell, E.H., Klump, H. and Fresco, J.R. (1996) Duplex-tetraplex equilibrium between a hairpin and two interacting hairpins of d(A-G)10 at neutral pH. *Nucleic acids research*, **24**, 5004-5012.
20. Rehm, C., Holder, I.T., Groß, A., Wojciechowski, F., Urban, M., Sinn, M., Drescher, M. and Hartig, J.S. (2014) A bacterial DNA quadruplex with exceptional K^+ selectivity and unique structural polymorphism. *Chemical Science*, **5**, 2809-2818.
21. Joachimi, A., Benz, A. and Hartig, J.S. (2009) A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorg Med Chem*, **17**, 6811-6815.
22. Sacca, B., Lacroix, L. and Mergny, J.L. (2005) The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides. *Nucleic acids research*, **33**, 1182-1192.
23. Arora, A. and Maiti, S. (2009) Differential biophysical behavior of human telomeric RNA and DNA quadruplex. *The journal of physical chemistry. B*, **113**, 10515-10520.
24. Zhang, D.H., Fujimoto, T., Saxena, S., Yu, H.Q., Miyoshi, D. and Sugimoto, N. (2010) Monomorphic RNA G-quadruplex and polymorphic DNA G-quadruplex structures responding to cellular environmental factors. *Biochemistry*, **49**, 4554-4563.
25. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. and Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic acids research*, **34**, 5402-5415.
26. Huppert, J.L. (2010) Structure, location and interactions of G-quadruplexes. *The FEBS journal*, **277**, 3452-3458.
27. Halder, R., Riou, J.F., Teulade-Fichou, M.P., Frickey, T. and Hartig, J.S. (2012) Bisquinolinium compounds induce quadruplex-specific transcriptome changes in HeLa S3 cell lines. *BMC research notes*, **5**, 138.
28. Alberti, P., Ren, J., Teulade-Fichou, M.P., Guittat, L., Riou, J.F., Chaires, J., Helene, C., Vigneron, J.P., Lehn, J.M. and Mergny, J.L. (2001) Interaction of an acridine dimer with DNA quadruplex structures. *Journal of biomolecular structure & dynamics*, **19**, 505-513.
29. Nielsen, M.C. and Ulven, T. (2010) Macrocyclic G-quadruplex ligands. *Current medicinal chemistry*, **17**, 3438-3448.
30. Wong, H.M., Payet, L. and Huppert, J.L. (2009) Function and targeting of G-quadruplexes. *Current opinion in molecular therapeutics*, **11**, 146-155.
31. Schultes, C.M., Guyen, B., Cuesta, J. and Neidle, S. (2004) Synthesis, biophysical and biological evaluation of 3,6-bis-amidoacridines with extended 9-anilino substituents as potent G-quadruplex-binding telomerase inhibitors. *Bioorganic & medicinal chemistry letters*, **14**, 4347-4351.

32. Koeppel, F., Riou, J.F., Laoui, A., Mailliet, P., Arimondo, P.B., Labit, D., Petitgenet, O., Helene, C. and Mergny, J.L. (2001) Ethidium derivatives bind to G-quartets, inhibit telomerase and act as fluorescent probes for quadruplexes. *Nucleic acids research*, **29**, 1087-1096.
33. Keating, L.R. and Szalai, V.A. (2004) Parallel-stranded guanine quadruplex interactions with a copper cationic porphyrin. *Biochemistry*, **43**, 15891-15900.
34. Range, K., Mayaan, E., Maher, L.J., 3rd and York, D.M. (2005) The contribution of phosphate-phosphate repulsions to the free energy of DNA bending. *Nucleic acids research*, **33**, 1257-1268.
35. Clark, G.R., Pytel, P.D., Squire, C.J. and Neidle, S. (2003) Structure of the first parallel DNA quadruplex-drug complex. *Journal of the American Chemical Society*, **125**, 4066-4067.
36. Brassart, B., Gomez, D., De Cian, A., Paterski, R., Montagnac, A., Qui, K.H., Temime-Smaali, N., Trentesaux, C., Mergny, J.L., Gueritte, F. *et al.* (2007) A new steroid derivative stabilizes g-quadruplexes and induces telomere uncapping in human tumor cells. *Molecular pharmacology*, **72**, 631-640.
37. Baker, E.S., Lee, J.T., Sessler, J.L. and Bowers, M.T. (2006) Cyclo[n]pyrroles: size and site-specific binding to G-quadruplexes. *Journal of the American Chemical Society*, **128**, 2641-2648.
38. Piazza, A., Boule, J.B., Lopes, J., Mingo, K., Largy, E., Teulade-Fichou, M.P. and Nicolas, A. (2010) Genetic instability triggered by G-quadruplex interacting Phen-DC compounds in Saccharomyces cerevisiae. *Nucleic acids research*, **38**, 4337-4348.
39. De Cian, A. and Mergny, J.L. (2007) Quadruplex ligands may act as molecular chaperones for tetramolecular quadruplex formation. *Nucleic acids research*, **35**, 2483-2493.
40. De Cian, A., Cristofari, G., Reichenbach, P., De Lemos, E., Monchaud, D., Teulade-Fichou, M.P., Shin-Ya, K., Lacroix, L., Lingner, J. and Mergny, J.L. (2007) Reevaluation of telomerase inhibition by quadruplex ligands and their mechanisms of action. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 17347-17352.
41. De Cian, A., Delemos, E., Mergny, J.L., Teulade-Fichou, M.P. and Monchaud, D. (2007) Highly efficient G-quadruplex recognition by bisquinolinium compounds. *Journal of the American Chemical Society*, **129**, 1856-1857.
42. Shi, D.F., Wheelhouse, R.T., Sun, D. and Hurley, L.H. (2001) Quadruplex-interactive agents as telomerase inhibitors: synthesis of porphyrins and structure-activity relationship for the inhibition of telomerase. *Journal of medicinal chemistry*, **44**, 4509-4523.
43. Beaume, N., Pathak, R., Yadav, V.K., Kota, S., Misra, H.S., Gautam, H.K. and Chowdhury, S. (2013) 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*, **41**, 76-89.
44. Cahoon, L.A. and Seifert, H.S. (2009) An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. *Science*, **325**, 764-767.
45. Han, H., Langley, D.R., Rangan, A. and Hurley, L.H. (2001) Selective interactions of cationic porphyrins with G-quadruplex structures. *Journal of the American Chemical Society*, **123**, 8902-8913.
46. Mita, H., Ohyama, T., Tanaka, Y. and Yamamoto, Y. (2006) Formation of a complex of 5,10,15,20-tetrakis(N-methylpyridinium-4-yl)-21H,23H-porphyrin with G-quadruplex DNA. *Biochemistry*, **45**, 6765-6772.
47. Nicoludis, J.M., Miller, S.T., Jeffrey, P.D., Barrett, S.P., Rablen, P.R., Lawton, T.J. and Yatsunyk, L.A. (2012) Optimized end-stacking provides specificity of N-methyl mesoporphyrin IX for human telomeric G-quadruplex DNA. *Journal of the American Chemical Society*, **134**, 20446-20456.
48. Wei, C., Jia, G., Yuan, J., Feng, Z. and Li, C. (2006) A spectroscopic study on the interactions of porphyrin with G-quadruplex DNAs. *Biochemistry*, **45**, 6681-6691.
49. Lubitz, I., Borovok, N. and Kotlyar, A. (2007) Interaction of monomolecular G4-DNA nanowires with TMPyP: evidence for intercalation. *Biochemistry*, **46**, 12925-12929.
50. Monchaud, D., Yang, P., Lacroix, L., Teulade-Fichou, M.P. and Mergny, J.L. (2008) A metal-mediated conformational switch controls G-quadruplex binding affinity. *Angewandte Chemie*, **47**, 4858-4861.
51. Goncalves, D.P., Rodriguez, R., Balasubramanian, S. and Sanders, J.K. (2006) Tetramethylpyridiniumporphyrazines--a new class of G-quadruplex inducing and stabilising ligands. *Chemical communications*, 4685-4687.
52. Felsenfeld, G. and Rich, A. (1957) Studies on the formation of two- and three-stranded polyribonucleotides. *Biochimica et biophysica acta*, **26**, 457-468.
53. Shchyolkina, A.K., Borisova, O.F., Minyat, E.E., Timofeev, E.N., Il'icheva, I.A., Khomyakova, E.B. and Florentiev, V.L. (1995) Parallel purine-pyrimidine-purine triplex: experimental evidence for existence. *FEBS letters*, **367**, 81-84.
54. Dervan, P.B. (1986) Design of sequence-specific DNA-binding molecules. *Science*, **232**, 464-471.
55. Lyamichev, V.I., Mirkin, S.M. and Frank-Kamenetskii, M.D. (1986) Structures of homopurine-homopyrimidine tract in superhelical DNA. *Journal of biomolecular structure & dynamics*, **3**, 667-669.
56. Buske, F.A., Mattick, J.S. and Bailey, T.L. (2011) Potential in vivo roles of nucleic acid triple-helices. *RNA biology*, **8**, 427-439.
57. Rusling, D.A., Brown, T. and Fox, K.R. (2006) DNA triple-helix formation at target sites containing duplex mismatches. *Biophysical chemistry*, **123**, 134-140.
58. Sun, J.S., Mergny, J.L., Lavery, R., Montenay-Garestier, T. and Helene, C. (1991) Triple helix structures: sequence dependence, flexibility and mismatch effects. *Journal of biomolecular structure & dynamics*, **9**, 411-424.
59. Xodo, L.E., Alunni-Fabbroni, M., Manzini, G. and Quadrifoglio, F. (1993) Sequence-specific DNA-triplex formation at imperfect homopurine-homopyrimidine sequences within a DNA plasmid. *European journal of biochemistry / FEBS*, **212**, 395-401.
60. Roberts, R.W. and Crothers, D.M. (1991) Specificity and stringency in DNA triplex formation. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 9397-9401.
61. Gowers, D.M. and Fox, K.R. (1997) DNA triple helix formation at oligopurine sites containing multiple contiguous pyrimidines. *Nucleic acids research*, **25**, 3787-3794.
62. Mergny, J.L., Sun, J.S., Rougee, M., Montenay-Garestier, T., Barcelo, F., Chomilier, J. and Helene, C. (1991) Sequence specificity in triple-helix formation: experimental and theoretical studies of the effect of mismatches on triplex stability. *Biochemistry*, **30**, 9791-9798.
63. Floris, R., Scaggiante, B., Manzini, G., Quadrifoglio, F. and Xodo, L.E. (1999) Effect of cations on purine.purine.pyrimidine triple helix formation in mixed-valence salt solutions. *European journal of biochemistry / FEBS*, **260**, 801-809.
64. Malkov, V.A., Voloshin, O.N., Soyfer, V.N. and Frank-Kamenetskii, M.D. (1993) Cation and sequence effects on stability of intermolecular pyrimidine-purine-purine triplex. *Nucleic acids research*, **21**, 585-591.

65. Paris, C., Geinguenaud, F., Gouyette, C., Liquier, J. and Lacoste, J. (2007) Mechanism of copper mediated triple helix formation at neutral pH in Drosophila satellite repeats. *Biophysical journal*, **92**, 2498-2506.
66. Sen, D. and Gilbert, W. (1990) A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature*, **344**, 410-414.
67. Hampel, K.J., Crosson, P. and Lee, J.S. (1991) Polyamines favor DNA triplex formation at neutral pH. *Biochemistry*, **30**, 4455-4459.
68. Potaman, V.N. and Sinden, R.R. (1998) Stabilization of intramolecular triple/single-strand structure by cationic peptides. *Biochemistry*, **37**, 12952-12961.
69. Nixon, P.L., Rangan, A., Kim, Y.G., Rich, A., Hoffman, D.W., Hennig, M. and Giedroc, D.P. (2002) Solution structure of a luteoviral P1-P2 frameshifting mRNA pseudoknot. *Journal of molecular biology*, **322**, 621-633.
70. Su, L., Chen, L., Egli, M., Berger, J.M. and Rich, A. (1999) Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nature structural biology*, **6**, 285-292.
71. Appasamy, S.D., RamLan, E.I. and Firdaus-Raih, M. (2013) Comparative sequence and structure analysis reveals the conservation and diversity of nucleotide positions and their associated tertiary interactions in the riboswitches. *PloS one*, **8**, e73984.
72. Holland, J.A. and Hoffman, D.W. (1996) Structural features and stability of an RNA triple helix in solution. *Nucleic acids research*, **24**, 2841-2848.
73. Carmona, P. and Molina, M. (2002) Binding of oligonucleotides to a viral hairpin forming RNA triplexes with parallel G*G*C triplets. *Nucleic acids research*, **30**, 1333-1337.
74. Devi, G., Zhou, Y., Zhong, Z., Toh, D.F. and Chen, G. (2014) RNA triplexes: from structural principles to biological and biotech applications. *Wiley interdisciplinary reviews. RNA*.
75. Haner, R. and Dervan, P.B. (1990) Single-strand DNA triple-helix formation. *Biochemistry*, **29**, 9761-9765.
76. Sklenar, V. and Feigon, J. (1990) Formation of a stable triplex from a single DNA strand. *Nature*, **345**, 836-838.
77. Chen, F.M. (1991) Intramolecular triplex formation of the purine.purine.pyrimidine type. *Biochemistry*, **30**, 4472-4479.
78. Radhakrishnan, I. and Patel, D.J. (1993) Solution structure of a purine.purine.pyrimidine DNA triplex containing G.GC and T.AT triples. *Structure*, **1**, 135-152.
79. Gondeau, C., Maurizot, J.C. and Durand, M. (1998) Circular dichroism and UV melting studies on formation of an intramolecular triplex containing parallel T*A:T and G*G:C triplets: netropsin complexation with the triplex. *Nucleic acids research*, **26**, 4996-5003.
80. Gondeau, C., Maurizot, J.C. and Durand, M. (1998) Spectroscopic investigation of an intramolecular DNA triplex containing both G.G:C and T.A:T triads and its complex with netropsin. *Journal of biomolecular structure & dynamics*, **15**, 1133-1145.
81. Pilch, D.S., Brousseau, R. and Shafer, R.H. (1990) Thermodynamics of triple helix formation: spectrophotometric studies on the d(A)10.2d(T)10 and d(C+3T4C+3).d(G3A4G3).d(C3T4C3) triple helices. *Nucleic acids research*, **18**, 5743-5750.
82. Durand, M., Peloille, S., Thuong, N.T. and Maurizot, J.C. (1992) Triple-helix formation by an oligonucleotide containing one (dA)12 and two (dT)12 sequences bridged by two hexaethylene glycol chains. *Biochemistry*, **31**, 9197-9204.
83. Volker, J., Botes, D.P., Lindsey, G.G. and Klump, H.H. (1993) Energetics of a stable intramolecular DNA triple helix formation. *Journal of molecular biology*, **230**, 1278-1290.
84. Phipps, A.K., Tarkoy, M., Schultze, P. and Feigon, J. (1998) Solution structure of an intramolecular DNA triplex containing 5-(1-propynyl)-2'-deoxyuridine residues in the third strand. *Biochemistry*, **37**, 5820-5830.
85. Hoyne, P.R. and Maher, L.J., 3rd. (2002) Functional studies of potential intrastrand triplex elements in the *Escherichia coli* genome. *Journal of molecular biology*, **318**, 373-386.
86. Giedroc, D.P. and Cornish, P.V. (2009) Frameshifting RNA pseudoknots: structure and mechanism. *Virus research*, **139**, 193-208.
87. Brierley, I. and Dos Ramos, F.J. (2006) Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus research*, **119**, 29-42.
88. Brierley, I., Meredith, M.R., Bloys, A.J. and Hagervall, T.G. (1997) Expression of a coronavirus ribosomal frameshift signal in *Escherichia coli*: influence of tRNA anticodon modification on frameshifting. *Journal of molecular biology*, **270**, 360-373.
89. Chen, C., Zhang, H., Broitman, S.L., Reiche, M., Farrell, I., Cooperman, B.S. and Goldman, Y.E. (2013) Dynamics of translation by single ribosomes through mRNA secondary structures. *Nature structural & molecular biology*, **20**, 582-588.
90. Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V. and Gottesman, M.M. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, **315**, 525-528.
91. Canceill, D. and Ehrlich, S.D. (1996) Copy-choice recombination mediated by DNA polymerase III holoenzyme from *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 6647-6652.
92. Viguera, E., Canceill, D. and Ehrlich, S.D. (2001) Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO journal*, **20**, 2587-2595.
93. Heller, R.C. and Marians, K.J. (2006) Replication fork reactivation downstream of a blocked nascent leading strand. *Nature*, **439**, 557-562.
94. Wang, G. and Vasquez, K.M. (2009) Models for chromosomal replication-independent non-B DNA structure-induced genetic instability. *Molecular carcinogenesis*, **48**, 286-298.
95. Zhao, J., Bacolla, A., Wang, G. and Vasquez, K.M. (2010) Non-B DNA structure-induced genetic instability and evolution. *Cellular and molecular life sciences : CMLS*, **67**, 43-62.
96. Marcadier, J.L. and Pearson, C.E. (2003) Fidelity of primate cell repair of a double-strand break within a (CTG).(CAG) tract. Effect of slipped DNA structures. *The Journal of biological chemistry*, **278**, 33848-33856.
97. Qin, Y. and Hurley, L.H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. *Biochimie*, **90**, 1149-1171.
98. Benham, C.J. (2001) Stress-induced DNA duplex destabilization in transcriptional initiation. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 103-114.
99. Benham, C.J. and Bi, C. (2004) The analysis of stress-induced duplex destabilization in long genomic DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology*, **11**, 519-543.
100. Wang, H., Noordewier, M. and Benham, C.J. (2004) Stress-induced DNA duplex destabilization (SIDD) in the *E. coli* genome: SIDD sites are closely associated with promoters. *Genome research*, **14**, 1575-1584.

101. Todd, A.K., Johnston, M. and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic acids research*, **33**, 2901-2907.
102. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic acids research*, **33**, 2908-2916.
103. Du, Z., Zhao, Y. and Li, N. (2009) Genome-wide colonization of gene regulatory elements by G4 DNA motifs. *Nucleic acids research*, **37**, 6784-6798.
104. 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 research*, **36**, D381-385.
105. Verma, A., Halder, K., Halder, R., Yadav, V.K., Rawal, P., Thakur, R.K., Mohd, F., Sharma, A. and Chowdhury, S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *Journal of medicinal chemistry*, **51**, 5641-5649.
106. Huppert, J.L. and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic acids research*, **35**, 406-413.
107. Eddy, J. and Maizels, N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic acids research*, **36**, 1321-1333.
108. Nakken, S., Rognes, T. and Hovig, E. (2009) The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts. *Nucleic acids research*, **37**, 5749-5756.
109. Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M.A. and Gemmell, N. (2013) Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PloS one*, **8**, e54710.
110. Bacolla, A., Larson, J.E., Collins, J.R., Li, J., Milosavljevic, A., Stenson, P.D., Cooper, D.N. and Wells, R.D. (2008) Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. *Genome research*, **18**, 1545-1553.
111. Hanakahi, L.A., Sun, H. and Maizels, N. (1999) High affinity interactions of nucleolin with G-G-paired rDNA. *The Journal of biological chemistry*, **274**, 15908-15912.
112. Wang, Y. and Patel, D.J. (1993) Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure*, **1**, 263-282.
113. Schaffitzel, C. and Pluckthun, A. (2001) Protein-fold evolution in the test tube. *Trends in biochemical sciences*, **26**, 577-579.
114. Yang, Q., Xiang, J., Yang, S., Zhou, Q., Li, Q., Tang, Y. and Xu, G. (2009) Verification of specific G-quadruplex structure by using a novel cyanine dye supramolecular assembly: I. recognizing mixed G-quadruplex in human telomeres. *Chemical communications*, 1103-1105.
115. Chang, C.C., Kuo, I.C., Lin, J.J., Lu, Y.C., Chen, C.T., Back, H.T., Lou, P.J. and Chang, T.C. (2004) A novel carbazole derivative, BMVC: a potential antitumor agent and fluorescence marker of cancer cells. *Chemistry & biodiversity*, **1**, 1377-1384.
116. Biffi, G., Tannahill, D., McCafferty, J. and Balasubramanian, S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature chemistry*, **5**, 182-186.
117. Lam, E.Y., Beraldi, D., Tannahill, D. and Balasubramanian, S. (2013) G-quadruplex structures are stable and detectable in human genomic DNA. *Nature communications*, **4**, 1796.
118. Biffi, G., Di Antonio, M., Tannahill, D. and Balasubramanian, S. (2014) Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nature chemistry*, **6**, 75-80.
119. Makarov, V.L., Hirose, Y. and Langmore, J.P. (1997) Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. *Cell*, **88**, 657-666.
120. McElligott, R. and Wellinger, R.J. (1997) The terminal DNA structure of mammalian chromosomes. *The EMBO journal*, **16**, 3705-3714.
121. Harley, C.B., Futcher, A.B. and Greider, C.W. (1990) Telomeres shorten during ageing of human fibroblasts. *Nature*, **345**, 458-460.
122. Sfeir, A.J., Chai, W., Shay, J.W. and Wright, W.E. (2005) Telomere-end processing the terminal nucleotides of human chromosomes. *Molecular cell*, **18**, 131-138.
123. Healy, K.C. (1995) Telomere dynamics and telomerase activation in tumor progression: prospects for prognosis and therapy. *Oncology research*, **7**, 121-130.
124. Zahler, A.M., Williamson, J.R., Cech, T.R. and Prescott, D.M. (1991) Inhibition of telomerase by G-quartet DNA structures. *Nature*, **350**, 718-720.
125. Oganesian, L., Moon, I.K., Bryan, T.M. and Jarstfer, M.B. (2006) Extension of G-quadruplex DNA by ciliate telomerase. *The EMBO journal*, **25**, 1148-1159.
126. Oganesian, L., Graham, M.E., Robinson, P.J. and Bryan, T.M. (2007) Telomerase recognizes G-quadruplex and linear DNA as distinct substrates. *Biochemistry*, **46**, 11279-11290.
127. Shay, J.W. and Wright, W.E. (2011) Role of telomeres and telomerase in cancer. *Seminars in cancer biology*, **21**, 349-353.
128. Hurley, L.H. (2002) DNA and its associated processes as targets for cancer therapy. *Nature reviews. Cancer*, **2**, 188-200.
129. Neidle, S. and Parkinson, G. (2002) Telomere maintenance as a target for anticancer drug discovery. *Nature reviews. Drug discovery*, **1**, 383-393.
130. Harley, C.B. (1991) Telomere loss: mitotic clock or genetic time bomb? *Mutation research*, **256**, 271-282.
131. Bodnar, A.G., Ouellette, M., Frolkis, M., Holt, S.E., Chiu, C.P., Morin, G.B., Harley, C.B., Shay, J.W., Lichtsteiner, S. and Wright, W.E. (1998) Extension of life-span by introduction of telomerase into normal human cells. *Science*, **279**, 349-352.
132. Hackett, J.A., Feldser, D.M. and Greider, C.W. (2001) Telomere dysfunction increases mutation rate and genomic instability. *Cell*, **106**, 275-286.
133. Paeschke, K., Capra, J.A. and Zakian, V.A. (2011) DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell*, **145**, 678-691.
134. Simonsson, T., Pecinka, P. and Kubista, M. (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic acids research*, **26**, 1167-1172.

135. Rankin, S., Reszka, A.P., Huppert, J., Zloh, M., Parkinson, G.N., Todd, A.K., Ladame, S., Balasubramanian, S. and Neidle, S. (2005) Putative DNA quadruplex formation within the human c-kit oncogene. *Journal of the American Chemical Society*, **127**, 10584-10589.
136. Cogoi, S. and Xodo, L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic acids research*, **34**, 2536-2549.
137. Kumari, S., Bugaut, A., Huppert, J.L. and Balasubramanian, S. (2007) An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. *Nature chemical biology*, **3**, 218-221.
138. Dai, J., Dexheimer, T.S., Chen, D., Carver, M., Ambrus, A., Jones, R.A. and Yang, D. (2006) An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution. *Journal of the American Chemical Society*, **128**, 1096-1098.
139. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. and Hurley, L.H. (2002) 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*, **99**, 11593-11598.
140. Bejugam, M., Sewitz, S., Shirude, P.S., Rodriguez, R., Shahid, R. and Balasubramanian, S. (2007) Trisubstituted isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of c-kit oncogene expression. *Journal of the American Chemical Society*, **129**, 12926-12927.
141. Zhang, C., Liu, H.H., Zheng, K.W., Hao, Y.H. and Tan, Z. (2013) DNA G-quadruplex formation in response to remote downstream transcription activity: long-range sensing and signal transducing in DNA double helix. *Nucleic acids research*, **41**, 7144-7152.
142. Perrone, R., Nadai, M., Frasson, I., Poe, J.A., Butovskaya, E., Smithgall, T.E., Palumbo, M., Palu, G. and Richter, S.N. (2013) A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter. *Journal of medicinal chemistry*, **56**, 6521-6530.
143. Broxson, C., Beckett, J. and Tornaletti, S. (2011) Transcription arrest by a G quadruplex forming-trinucleotide repeat sequence from the human c-myb gene. *Biochemistry*, **50**, 4162-4172.
144. Huppert, J.L., Bugaut, A., Kumari, S. and Balasubramanian, S. (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic acids research*, **36**, 6260-6268.
145. Beaudoin, J.D. and Perreault, J.P. (2013) Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic acids research*, **41**, 5898-5911.
146. Bugaut, A. and Balasubramanian, S. (2012) 5'-UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic acids research*, **40**, 4727-4741.
147. Azzalin, C.M., Reichenbach, P., Khoriauli, L., Giulotto, E. and Lingner, J. (2007) Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science*, **318**, 798-801.
148. Schoeftner, S. and Blasco, M.A. (2010) Chromatin regulation and non-coding RNAs at mammalian telomeres. *Seminars in cell & developmental biology*, **21**, 186-193.
149. Halder, K., Benzler, M. and Hartig, J.S. (2012) Reporter assays for studying quadruplex nucleic acids. *Methods*, **57**, 115-121.
150. Halder, K., Wieland, M. and Hartig, J.S. (2009) Predictable suppression of gene expression by 5'-UTR-based RNA quadruplexes. *Nucleic acids research*, **37**, 6811-6817.
151. Horsburgh, B.C., Kollmus, H., Hauser, H. and Coen, D.M. (1996) Translational recoding induced by G-rich mRNA sequences that form unusual structures. *Cell*, **86**, 949-959.
152. Yu, C.H., Teulade-Fichou, M.P. and Olsthoorn, R.C. (2014) Stimulation of ribosomal frameshifting by RNA G-quadruplex structures. *Nucleic acids research*, **42**, 1887-1892.
153. Endoh, T. and Sugimoto, N. (2013) Unusual -1 ribosomal frameshift caused by stable RNA G-quadruplex in open reading frame. *Analytical chemistry*, **85**, 11435-11439.
154. Endoh, T., Kawasaki, Y. and Sugimoto, N. (2013) Stability of RNA quadruplex in open reading frame determines proteolysis of human estrogen receptor alpha. *Nucleic acids research*, **41**, 6222-6231.
155. Murat, P., Zhong, J., Lekieffre, L., Cowieson, N.P., Clancy, J.L., Preiss, T., Balasubramanian, S., Khanna, R. and Tellam, J. (2014) G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. *Nature chemical biology*, **10**, 358-364.
156. Cahoon, L.A. and Seifert, H.S. (2011) Focusing homologous recombination: pilin antigenic variation in the pathogenic Neisseria. *Molecular microbiology*, **81**, 1136-1143.
157. Walia, R. and Chaconas, G. (2013) Suggested role for G4 DNA in recombinational switching at the antigenic variation locus of the Lyme disease spirochete. *PloS one*, **8**, e57792.
158. Giacani, L., Brandt, S.L., Puray-Chavez, M., Reid, T.B., Godornes, C., Molini, B.J., Benzler, M., Hartig, J.S., Lukehart, S.A. and Centurion-Lara, A. (2012) Comparative investigation of the genomic regions involved in antigenic variation of the TprK antigen among treponemal species, subspecies, and strains. *J Bacteriol*, **194**, 4208-4225.
159. Endoh, T., Kawasaki, Y. and Sugimoto, N. (2013) Suppression of gene expression by G-quadruplexes in open reading frames depends on G-quadruplex stability. *Angewandte Chemie*, **52**, 5522-5526.
160. Endoh, T., Kawasaki, Y. and Sugimoto, N. (2013) Translational halt during elongation caused by G-quadruplex formed by mRNA. *Methods*, **64**, 73-78.
161. Wieland, M. and Hartig, J.S. (2007) RNA quadruplex-based modulation of gene expression. *Chemistry & biology*, **14**, 757-763.
162. Gaddis, S.S., Wu, Q., Thames, H.D., DiGiovanni, J., Walborg, E.F., MacLeod, M.C. and Vasquez, K.M. (2006) A web-based search engine for triplex-forming oligonucleotide target sequences. *Oligonucleotides*, **16**, 196-201.
163. Buske, F.A., Bauer, D.C., Mattick, J.S. and Bailey, T.L. (2013) Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci. *Bioinformatics*, **29**, 1895-1897.
164. Buske, F.A., Bauer, D.C., Mattick, J.S. and Bailey, T.L. (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome research*, **22**, 1372-1381.
165. Jenjaroenpun, P. and Kuznetsov, V.A. (2009) TTS mapping: integrative WEB tool for analysis of triplex formation target DNA sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome. *BMC genomics*, **10 Suppl 3**, S9.
166. Schroth, G.P. and Ho, P.S. (1995) Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic acids research*, **23**, 1977-1983.
167. Cer, R.Z., Bruce, K.H., Mudunuri, U.S., Yi, M., Volfovsky, N., Luke, B.T., Bacolla, A., Collins, J.R. and Stephens, R.M. (2011) Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic acids research*, **39**, D383-391.

168. Lexa, M., Martinek, T., Burgetova, I., Kopecek, D. and Brazdova, M. (2011) A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics*, **27**, 2510-2517.
169. Hon, J., Martinek, T., Rajdl, K. and Lexa, M. (2013) Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences. *Bioinformatics*, **29**, 1900-1901.
170. Ohno, M., Fukagawa, T., Lee, J.S. and Ikemura, T. (2002) Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. *Chromosoma*, **111**, 201-213.
171. Agazie, Y.M., Lee, J.S. and Burkholder, G.D. (1994) Characterization of a new monoclonal antibody to triplex DNA and immunofluorescent staining of mammalian chromosomes. *The Journal of biological chemistry*, **269**, 7019-7023.
172. Agazie, Y.M., Burkholder, G.D. and Lee, J.S. (1996) Triplex DNA in the nucleus: direct binding of triplex-specific antibodies and their effect on transcription, replication and cell growth. *The Biochemical journal*, **316 (Pt 2)**, 461-466.
173. Lee, J.S., Burkholder, G.D., Latimer, L.J., Haug, B.L. and Braun, R.P. (1987) A monoclonal antibody to triplex DNA binds to eucaryotic chromosomes. *Nucleic acids research*, **15**, 1047-1061.
174. Kanak, M., Alseiari, M., Balasubramanian, P., Addanki, K., Aggarwal, M., Noorali, S., Kalsum, A., Mahalingam, K., Pace, G., Panasik, N. *et al.* (2010) Triplex-forming MicroRNAs form stable complexes with HIV-1 provirus and inhibit its replication. *Applied immunohistochemistry & molecular morphology : AIMM / official publication of the Society for Applied Immunohistochemistry*, **18**, 532-545.
175. Bagasra, O., Stir, A.E., Pirisi-Creek, L., Creek, K.E., Bagasra, A.U., Glenn, N. and Lee, J.S. (2006) Role of micro-RNAs in regulation of lentiviral latency and persistence. *Applied immunohistochemistry & molecular morphology : AIMM / official publication of the Society for Applied Immunohistochemistry*, **14**, 276-290.
176. Guillonneau, F., Guieysse, A.L., Le Caer, J.P., Rossier, J. and Praseuth, D. (2001) Selection and identification of proteins bound to DNA triple-helical structures by combination of 2D-electrophoresis and MALDI-TOF mass spectrometry. *Nucleic acids research*, **29**, 2427-2436.
177. Jimenez-Garcia, E., Vaquero, A., Espinas, M.L., Soliva, R., Orozco, M., Bernues, J. and Azorin, F. (1998) The GAGA factor of Drosophila binds triple-stranded DNA. *The Journal of biological chemistry*, **273**, 24640-24648.
178. Nelson, L.D., Musso, M. and Van Dyke, M.W. (2000) The yeast STM1 gene encodes a purine motif triple helical DNA-binding protein. *The Journal of biological chemistry*, **275**, 5573-5581.
179. Ciotti, P., Van Dyke, M.W., Bianchi-Scarra, G. and Musso, M. (2001) Characterization of a triplex DNA-binding protein encoded by an alternative reading frame of loricrin. *European journal of biochemistry / FEBS*, **268**, 225-234.
180. Kusic, J., Tomic, B., Divac, A. and Kojic, S. (2010) Human initiation protein Orc4 prefers triple stranded DNA. *Molecular biology reports*, **37**, 2317-2322.
181. Li, G., Tolstonog, G.V. and Traub, P. (2002) Interaction in vitro of type III intermediate filament proteins with triplex DNA. *DNA and cell biology*, **21**, 163-188.
182. Plyler, J., Jasheway, K., Tuesuwan, B., Karr, J., Brennan, J.S., Kerwin, S.M. and David, W.M. (2009) Real-time investigation of SV40 large T-antigen helicase activity using surface plasmon resonance. *Cell biochemistry and biophysics*, **53**, 43-52.
183. Maine, I.P. and Kodadek, T. (1994) Efficient unwinding of triplex DNA by a DNA helicase. *Biochemical and biophysical research communications*, **204**, 1119-1124.
184. Kopel, V., Pozner, A., Baran, N. and Manor, H. (1996) Unwinding of the third strand of a DNA triple helix, a novel activity of the SV40 large T-antigen helicase. *Nucleic acids research*, **24**, 330-335.
185. Suda, T., Mishima, Y., Takayanagi, K., Asakura, H., Odani, S. and Kominami, R. (1996) A novel activity of HMG domains: promotion of the triple-stranded complex formation between DNA containing (GGA/TCC)11 and d(GGA)11 oligonucleotides. *Nucleic acids research*, **24**, 4733-4740.
186. Jain, A., Akanchha, S. and Rajeswari, M.R. (2005) Stabilization of purine motif DNA triplex by a tetrapeptide from the binding domain of HMGBI protein. *Biochimie*, **87**, 781-790.
187. Lange, S.S., Reddy, M.C. and Vasquez, K.M. (2009) Human HMGB1 directly facilitates interactions between nucleotide excision repair proteins on triplex-directed psoralen interstrand crosslinks. *DNA repair*, **8**, 865-872.
188. Thoma, B.S., Wakasugi, M., Christensen, J., Reddy, M.C. and Vasquez, K.M. (2005) Human XPC-hHR23B interacts with XPA-RPA in the recognition of triplex-directed psoralen DNA interstrand crosslinks. *Nucleic acids research*, **33**, 2993-3001.
189. Zhao, J., Jain, A., Iyer, R.R., Modrich, P.L. and Vasquez, K.M. (2009) Mismatch repair and nucleotide excision repair proteins cooperate in the recognition of DNA interstrand crosslinks. *Nucleic acids research*, **37**, 4420-4429.
190. Vasquez, K.M. and Wilson, J.H. (1998) Triplex-directed modification of genes and gene activity. *Trends in biochemical sciences*, **23**, 4-9.
191. Carbone, G.M., McGuffie, E.M., Collier, A. and Catapano, C.V. (2003) Selective inhibition of transcription of the Ets2 gene in prostate cancer cells by a triplex-forming oligonucleotide. *Nucleic acids research*, **31**, 833-843.
192. Rapozzi, V., Cogoi, S., Spessotto, P., Risso, A., Bonora, G.M., Quadrifoglio, F. and Xodo, L.E. (2002) Antigene effect in K562 cells of a PEG-conjugated triplex-forming oligonucleotide targeted to the bcr/abl oncogene. *Biochemistry*, **41**, 502-510.
193. Giovannangeli, C. and Helene, C. (1997) Progress in developments of triplex-based strategies. *Antisense & nucleic acid drug development*, **7**, 413-421.
194. Chin, J.Y., Schleifman, E.B. and Glazer, P.M. (2007) Repair and recombination induced by triple helix DNA. *Frontiers in bioscience : a journal and virtual library*, **12**, 4288-4297.
195. Svinarchuk, F., Nagibneva, I., Cherny, D., Ait-Si-Ali, S., Pritchard, L.L., Robin, P., Malvy, C., Harel-Bellan, A. and Chern, D. (1997) Recruitment of transcription factors to the target site by triplex-forming oligonucleotides. *Nucleic acids research*, **25**, 3459-3464.
196. Chin, J.Y., Kuan, J.Y., Lonkar, P.S., Krause, D.S., Seidman, M.M., Peterson, K.R., Nielsen, P.E., Kole, R. and Glazer, P.M. (2008) Correction of a splice-site mutation in the beta-globin gene stimulated by triplex-forming peptide nucleic acids. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 13514-13519.
197. Rogers, F.A., Hu, R.H. and Milstone, L.M. (2013) Local delivery of gene-modifying triplex-forming molecules to the epidermis. *The Journal of investigative dermatology*, **133**, 685-691.
198. Rogers, F.A., Lin, S.S., Hegan, D.C., Krause, D.S. and Glazer, P.M. (2012) Targeted gene modification of hematopoietic progenitor cells in mice following systemic administration of a PNA-peptide conjugate. *Molecular therapy : the journal of the American Society of Gene Therapy*, **20**, 109-118.
199. Vasquez, K.M., Christensen, J., Li, L., Finch, R.A. and Glazer, P.M. (2002) Human XPA and RPA DNA repair proteins participate in specific recognition of triplex-induced helical distortions. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5848-5853.

200. Faruqi, A.F., Datta, H.J., Carroll, D., Seidman, M.M. and Glazer, P.M. (2000) Triple-helix formation induces recombination in mammalian cells via a nucleotide excision repair-dependent pathway. *Molecular and cellular biology*, **20**, 990-1000.
201. Wang, G., Seidman, M.M. and Glazer, P.M. (1996) Mutagenesis in mammalian cells induced by triple helix formation and transcription-coupled repair. *Science*, **271**, 802-805.
202. Ussery, D., Soumpasis, D.M., Brunak, S., Staerfeldt, H.H., Worning, P. and Krogh, A. (2002) Bias of purine stretches in sequenced chromosomes. *Computers & chemistry*, **26**, 531-541.
203. Behe, M.J. (1995) An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. *Nucleic acids research*, **23**, 689-695.
204. Bissler, J.J. (2007) Triplex DNA and human disease. *Frontiers in bioscience : a journal and virtual library*, **12**, 4536-4546.
205. Rajeswari, M.R. (2012) DNA triplex structures in neurodegenerative disorder, Friedreich's ataxia. *Journal of biosciences*, **37**, 519-532.
206. Callen, D.F., Lane, S.A., Kozman, H., Kremmidiotis, G., Whitmore, S.A., Lowenstein, M., Doggett, N.A., Kenmochi, N., Page, D.C., Maglott, D.R. *et al.* (1995) Integration of transcript and genetic maps of chromosome 16 at near-1-Mb resolution: demonstration of a "hot spot" for recombination at 16p12. *Genomics*, **29**, 503-511.
207. Davis, T.L., Firulli, A.B. and Kinniburgh, A.J. (1989) Ribonucleoprotein and protein factors bind to an H-DNA-forming c-myc DNA element: possible regulators of the c-myc gene. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 9682-9686.
208. Ponzielli, R., Katz, S., Barsyte-Lovejoy, D. and Penn, L.Z. (2005) Cancer therapeutics: targeting the dark side of Myc. *European journal of cancer*, **41**, 2485-2501.
209. Rustighi, A., Tessari, M.A., Vascotto, F., Sgarra, R., Giancotti, V. and Manfioletti, G. (2002) A polypyrimidine/polypurine tract within the Hmga2 minimal promoter: a common feature of many growth-related genes. *Biochemistry*, **41**, 1229-1240.
210. Xu, G. and Goodridge, A.G. (1996) Characterization of a polypyrimidine/polypurine tract in the promoter of the gene for chicken malic enzyme. *The Journal of biological chemistry*, **271**, 16008-16019.
211. Maiti, A.K. and Brahmachari, S.K. (2001) Poly purine.pyrimidine sequences upstream of the beta-galactosidase gene affect gene expression in Saccharomyces cerevisiae. *BMC molecular biology*, **2**, 11.
212. Wang, G. and Vasquez, K.M. (2004) Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 13448-13453.
213. Samadashwily, G.M., Dayn, A. and Mirkin, S.M. (1993) Suicidal nucleotide sequences for DNA polymerization. *The EMBO journal*, **12**, 4975-4983.
214. Krasilnikov, A.S., Panyutin, I.G., Samadashwily, G.M., Cox, R., Lazurkin, Y.S. and Mirkin, S.M. (1997) Mechanisms of triplex-caused polymerization arrest. *Nucleic acids research*, **25**, 1339-1346.
215. Weinreb, A., Collier, D.A., Birshtein, B.K. and Wells, R.D. (1990) Left-handed Z-DNA and intramolecular triplex formation at the site of an unequal sister chromatid exchange. *The Journal of biological chemistry*, **265**, 1352-1359.
216. Rooney, S.M. and Moore, P.D. (1995) Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 2141-2144.
217. Biet, E., Sun, J.S. and Dutreix, M. (2003) Stimulation of D-loop formation by polypurine/polypyrimidine sequences. *Nucleic acids research*, **31**, 1006-1012.
218. Rao, B.J. and Radding, C.M. (1994) Formation of base triplets by non-Watson-Crick bonds mediates homologous recognition in RecA recombination filaments. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 6161-6165.
219. Zhurkin, V.B., Raghunathan, G., Ulyanov, N.B., Camerini-Otero, R.D. and Jernigan, R.L. (1994) A parallel DNA triplex as a model for the intermediate in homologous recombination. *Journal of molecular biology*, **239**, 181-200.
220. Toor, N., Keating, K.S. and Pyle, A.M. (2009) Structural insights into RNA splicing. *Current opinion in structural biology*, **19**, 260-266.
221. Qiao, F. and Cech, T.R. (2008) Triple-helix structure in telomerase RNA contributes to catalysis. *Nature structural & molecular biology*, **15**, 634-640.
222. Chou, M.Y. and Chang, K.Y. (2010) An intermolecular RNA triplex provides insight into structural determinants for the pseudoknot stimulator of -1 ribosomal frameshifting. *Nucleic acids research*, **38**, 1676-1685.
223. Belew, A.T., Meskauskas, A., Musalgaonkar, S., Advani, V.M., Sulima, S.O., Kasprzak, W.K., Shapiro, B.A. and Dinman, J.D. (2014) Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. *Nature*, **512**, 265-269.
224. Zheng, R., Shen, Z., Tripathi, V., Xuan, Z., Freier, S.M., Bennett, C.F., Prasanth, S.G. and Prasanth, K.V. (2010) Polypurine-repeat-containing RNAs: a novel class of long non-coding RNA in mammalian cells. *Journal of cell science*, **123**, 3734-3744.
225. Westin, L., Blomquist, P., Milligan, J.F. and Wrange, O. (1995) Triple helix DNA alters nucleosomal histone-DNA interactions and acts as a nucleosome barrier. *Nucleic acids research*, **23**, 2184-2191.
226. Schmitz, K.M., Mayer, C., Postepska, A. and Grummt, I. (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes & development*, **24**, 2264-2269.
227. Vaillant, I. and Paszkowski, J. (2007) Role of histone and DNA methylation in gene regulation. *Current opinion in plant biology*, **10**, 528-533.
228. Geiman, T.M. and Muegge, K. (2010) DNA methylation in early development. *Molecular reproduction and development*, **77**, 105-113.
229. Bucher, P. and Yagil, G. (1991) Occurrence of oligopurine.oligopyrimidine tracts in eukaryotic and prokaryotic genes. *DNA sequence : the journal of DNA sequencing and mapping*, **1**, 157-172.
230. Kohwi, Y. and Panchenko, Y. (1993) Transcription-dependent recombination induced by triple-helix formation. *Genes & development*, **7**, 1766-1778.
231. Duval-Valentin, G., Thuong, N.T. and Helene, C. (1992) Specific inhibition of transcription by triple helix-forming oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 504-508.
232. Maher, L.J., 3rd. (1992) Inhibition of T7 RNA polymerase initiation by triple-helical DNA complexes: a model for artificial gene repression. *Biochemistry*, **31**, 7587-7594.
233. Bacolla, A., Jaworski, A., Larson, J.E., Jakupciak, J.P., Chuzhanova, N., Abeysinghe, S.S., O'Connell, C.D., Cooper, D.N. and Wells, R.D. (2004) Breakpoints of gross deletions coincide with non-B DNA conformations. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14162-14167.

234. Bacolla, A., Jaworski, A., Connors, T.D. and Wells, R.D. (2001) Pkd1 unusual DNA conformations are recognized by nucleotide excision repair. *The Journal of biological chemistry*, **276**, 18597-18604.
235. Kato, M. (1993) Polypyrimidine/polypurine sequence in plasmid DNA enhances formation of dimer molecules in *Escherichia coli*. Dimerization of plasmid DNA in *Escherichia coli*. *Molecular biology reports*, **18**, 183-187.
236. Hampel, K.J., Burkholder, G.D. and Lee, J.S. (1993) Plasmid dimerization mediated by triplex formation between polypyrimidine-polypurine repeats. *Biochemistry*, **32**, 1072-1077.
237. Kato, M. and Shimizu, N. (1992) Effect of the potential triplex DNA region on the in vitro expression of bacterial beta-lactamase gene in superhelical recombinant plasmids. *Journal of biochemistry*, **112**, 492-494.
238. Sarkar, P.S. and Brahmachari, S.K. (1992) Intramolecular triplex potential sequence within a gene down regulates its expression in vivo. *Nucleic acids research*, **20**, 5713-5718.
239. Brahmachari, S.K., Sarkar, P.S., Raghavan, S., Narayan, M. and Maiti, A.K. (1997) Polypurine/polypyrimidine sequences as cis-acting transcriptional regulators. *Gene*, **190**, 17-26.
240. Duval-Valentin, G., de Bizemont, T., Takasugi, M., Mergny, J.L., Bisagni, E. and Helene, C. (1995) Triple-helix specific ligands stabilize H-DNA conformation. *Journal of molecular biology*, **247**, 847-858.
241. Rao, J.E. and Craig, N.L. (2001) Selective recognition of pyrimidine motif triplexes by a protein encoded by the bacterial transposon Tn7. *Journal of molecular biology*, **307**, 1161-1170.
242. Cerullo, V., Pesonen, S., Diaconu, I., Escutenaire, S., Arstila, P.T., Ugolini, M., Nokisalmi, P., Raki, M., Laasonen, L., Sarkioja, M. *et al.* (2010) Oncolytic adenovirus coding for granulocyte macrophage colony-stimulating factor induces antitumoral immunity in cancer patients. *Cancer research*, **70**, 4297-4309.
243. Stern, M.J., Ames, G.F., Smith, N.H., Robinson, E.C. and Higgins, C.F. (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell*, **37**, 1015-1026.
244. Bachellier, S., Clement, J.M. and Hofnung, M. (1999) Short palindromic repetitive DNA elements in enterobacteria: a survey. *Research in microbiology*, **150**, 627-639.
245. Delihas, N. (2008) Small mobile sequences in bacteria display diverse structure/function motifs. *Molecular microbiology*, **67**, 475-481.
246. Siguier, P., Filee, J. and Chandler, M. (2006) Insertion sequences in prokaryotic genomes. *Current opinion in microbiology*, **9**, 526-531.
247. Lopes, J., Ribeyre, C. and Nicolas, A. (2006) Complex minisatellite rearrangements generated in the total or partial absence of Rad27/hFEN1 activity occur in a single generation and are Rad51 and Rad52 dependent. *Molecular and cellular biology*, **26**, 6675-6689.
248. Richard, G.F., Kerrest, A. and Dujon, B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and molecular biology reviews : MMBR*, **72**, 686-727.
249. Yeramian, E. and Buc, H. (1999) Tandem repeats in complete bacterial genome sequences: sequence and structural analyses for comparative studies. *Research in microbiology*, **150**, 745-754.
250. Mrazek, J., Guo, X. and Shah, A. (2007) Simple sequence repeats in prokaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 8472-8477.
251. Mrazek, J. (2006) Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. *Molecular biology and evolution*, **23**, 1370-1385.
252. Kassai-Jager, E., Ortutay, C., Toth, G., Vellai, T. and Gaspari, Z. (2008) Distribution and evolution of short tandem repeats in closely related bacterial genomes. *Gene*, **410**, 18-25.
253. 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 research*, **10**, 62-71.
254. Vogt, P. (1990) Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code". *Human genetics*, **84**, 301-336.
255. van Belkum, A., Scherer, S., van Alphen, L. and Verbrugh, H. (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiology and molecular biology reviews : MMBR*, **62**, 275-293.
256. Zhou, K., Aertsen, A. and Michiels, C.W. (2014) The role of variable DNA tandem repeats in bacterial adaptation. *FEMS microbiology reviews*, **38**, 119-141.
257. van der Woude, M.W. and BaumLer, A.J. (2004) Phase and antigenic variation in bacteria. *Clinical microbiology reviews*, **17**, 581-611, table of contents.
258. Henderson, I.R., Owen, P. and Nataro, J.P. (1999) Molecular switches--the ON and OFF of bacterial phase variation. *Molecular microbiology*, **33**, 919-932.
259. Bayliss, C.D. and Palmer, M.E. (2012) Evolution of simple sequence repeat-mediated phase variation in bacterial genomes. *Annals of the New York Academy of Sciences*, **1267**, 39-44.
260. Huang, Y. and Mrazek, J. (2014) 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*, **21**, 285-297.
261. Collins, J., Volckaert, G. and Nevers, P. (1982) Precise and nearly-precise excision of the symmetrical inverted repeats of Tn5; common features of recA-independent deletion events in *Escherichia coli*. *Gene*, **19**, 139-146.
262. Oliveira, P.H., Prather, K.J., Prazeres, D.M. and Monteiro, G.A. (2010) Analysis of DNA repeats in bacterial plasmids reveals the potential for recurrent instability events. *Applied microbiology and biotechnology*, **87**, 2157-2167.
263. Wojciechowska, M., Bacolla, A., Larson, J.E. and Wells, R.D. (2005) The myotonic dystrophy type 1 triplet repeat sequence induces gross deletions and inversions. *The Journal of biological chemistry*, **280**, 941-952.
264. Wells, R.D., Dere, R., Hebert, M.L., Napierala, M. and Son, L.S. (2005) Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic acids research*, **33**, 3785-3798.
265. Silby, M.W., Cerdeno-Tarraga, A.M., Vernikos, G.S., Giddens, S.R., Jackson, R.W., Preston, G.M., Zhang, X.X., Moon, C.D., Gehrig, S.M., Godfrey, S.A. *et al.* (2009) Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. *Genome biology*, **10**, R51.
266. Mine, N., Guglielmini, J., Wilbaux, M. and Van Melderen, L. (2009) The decay of the chromosomally encoded ccdO157 toxin-antitoxin system in the *Escherichia coli* species. *Genetics*, **181**, 1557-1566.
267. Ogier, J.C., Calteau, A., Forst, S., Goodrich-Blair, H., Roche, D., Rouy, Z., Suen, G., Zumbihl, R., Givaudan, A., Tailliez, P. *et al.* (2010) Units of plasticity in bacterial genomes: new insight from the comparative genomics of two bacteria interacting with invertebrates, Photorhabdus and Xenorhabdus. *BMC genomics*, **11**, 568.
268. Kristoffersen, S.M., Tourasse, N.J., Kolsto, A.B. and Okstad, O.A. (2011) Interspersed DNA repeats bcr1-bcr18 of Bacillus cereus group bacteria form three distinct groups with different evolutionary and functional patterns. *Molecular biology and evolution*, **28**, 963-983.

269. Croucher, N.J., Vernikos, G.S., Parkhill, J. and Bentley, S.D. (2011) Identification, variation and transcription of pneumococcal repeat sequences. *BMC genomics*, **12**, 120.
270. Epstein, W. and Schultz, S.G. (1965) Cation Transport in *Escherichia coli*: V. Regulation of cation content. *The Journal of general physiology*, **49**, 221-234.
271. Shabala, L., Bowman, J., Brown, J., Ross, T., McMeekin, T. and Shabala, S. (2009) Ion transport and osmotic adjustment in *Escherichia coli* in response to ionic and non-ionic osmotica. *Environmental microbiology*, **11**, 137-148.
272. Kypr, J., Kejnovska, I., Renciuk, D. and Vorlickova, M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic acids research*, **37**, 1713-1725.
273. Mukundan, V.T. and Phan, A.T. (2013) Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *Journal of the American Chemical Society*, **135**, 5017-5028.
274. Bugaut, A. and Balasubramanian, S. (2008) A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry*, **47**, 689-697.
275. Guzman, L.M., Belin, D., Carson, M.J. and Beckwith, J. (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol*, **177**, 4121-4130.
276. Schleif, R. (1992) DNA looping. *Annual review of biochemistry*, **61**, 199-223.
277. Lee, N., Francklyn, C. and Hamilton, E.P. (1987) Arabinose-induced binding of AraC protein to araI2 activates the araBAD operon promoter. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 8814-8818.
278. Niland, P., Huhne, R. and Muller-Hill, B. (1996) How AraC interacts specifically with its target DNAs. *Journal of molecular biology*, **264**, 667-674.
279. Roberts, C.W. and Roberts, J.W. (1996) Base-specific recognition of the nontemplate strand of promoter DNA by *E. coli* RNA polymerase. *Cell*, **86**, 495-501.
280. Isaacs, F.J., Dwyer, D.J., Ding, C., Pervouchine, D.D., Cantor, C.R. and Collins, J.J. (2004) Engineered riboregulators enable post-transcriptional control of gene expression. *Nature biotechnology*, **22**, 841-847.
281. Giuliodori, A.M., Di Pietro, F., Marzi, S., Masquida, B., Wagner, R., Romby, P., Gualerzi, C.O. and Pon, C.L. (2010) The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. *Molecular cell*, **37**, 21-33.
282. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research*, **31**, 3406-3415.
283. Zhang, A.Y., Bugaut, A. and Balasubramanian, S. (2011) A sequence-independent analysis of the loop length dependence of intramolecular RNA G-quadruplex stability and topology. *Biochemistry*, **50**, 7251-7258.
284. Espah Borujeni, A., Channarasappa, A.S. and Salis, H.M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic acids research*, **42**, 2646-2659.
285. Simonetti, A., Marzi, S., Myasnikov, A.G., Fabbretti, A., Yusupov, M., Gualerzi, C.O. and Klaholz, B.P. (2008) Structure of the 30S translation initiation complex. *Nature*, **455**, 416-420.
286. Yusupova, G.Z., Yusupov, M.M., Cate, J.H. and Noller, H.F. (2001) The path of messenger RNA through the ribosome. *Cell*, **106**, 233-241.
287. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*, **42**, D199-205.
288. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27-30.
289. Basundra, R., Kumar, A., Amrane, S., Verma, A., Phan, A.T. and Chowdhury, S. (2010) A novel G-quadruplex motif modulates promoter activity of human thymidine kinase 1. *The FEBS journal*, **277**, 4254-4264.
290. Salis, H.M., Mirsky, E.A. and Voigt, C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology*, **27**, 946-950.
291. Komarova, A.V., Tchufistova, L.S., Dreyfus, M. and Boni, I.V. (2005) AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*. *J Bacteriol*, **187**, 1344-1349.
292. Kettani, A., Kumar, R.A. and Patel, D.J. (1995) Solution structure of a DNA quadruplex containing the fragile X syndrome triplet repeat. *Journal of molecular biology*, **254**, 638-656.
293. Frymier, J.S., Reed, T.D., Fletcher, S.A. and Csonka, L.N. (1997) Characterization of transcriptional regulation of the kdp operon of Salmonella typhimurium. *J Bacteriol*, **179**, 3061-3063.
294. Jung, K., Veen, M. and Altendorf, K. (2000) K+ and ionic strength directly influence the autophosphorylation activity of the putative turgor sensor KdpD of *Escherichia coli*. *The Journal of biological chemistry*, **275**, 40142-40147.
295. Asha, H. and Gowrishankar, J. (1993) Regulation of kdp operon expression in *Escherichia coli*: evidence against turgor as signal for transcriptional control. *J Bacteriol*, **175**, 4528-4537.
296. Hamann, K., Zimmann, P. and Altendorf, K. (2008) Reduction of turgor is not the stimulus for the sensor kinase KdpD of *Escherichia coli*. *J Bacteriol*, **190**, 2360-2367.
297. Epstein, W. (1992) Kdp, a bacterial P-type ATPase whose expression and activity are regulated by turgor pressure. *Acta physiologica Scandinavica. Supplementum*, **607**, 193-199.
298. Laermann, V., Cudic, E., Kipschull, K., Zimmann, P. and Altendorf, K. (2013) The sensor kinase KdpD of *Escherichia coli* senses external K+. *Molecular microbiology*, **88**, 1194-1204.
299. Ferguson, G.P., Munro, A.W., Douglas, R.M., McLaggan, D. and Booth, I.R. (1993) Activation of potassium channels during metabolite detoxification in *Escherichia coli*. *Molecular microbiology*, **9**, 1297-1303.
300. Wolanin, P.M., Thomason, P.A. and Stock, J.B. (2002) Histidine protein kinases: key signal transducers outside the animal kingdom. *Genome biology*, **3**, REVIEWS3013.
301. Agarwal, T., Roy, S., Kumar, S., Chakraborty, T.K. and Maiti, S. (2014) In the Sense of Transcription regulation by G-quadruplexes: Asymmetric effects in sense and anti-sense strand. *Biochemistry*.
302. Mekler, V., Minakhin, L. and Severinov, K. (2011) A critical role of downstream RNA polymerase-promoter interactions in the formation of initiation complex. *The Journal of biological chemistry*, **286**, 22600-22608.
303. Siebenlist, U., Simpson, R.B. and Gilbert, W. (1980) *E. coli* RNA polymerase interacts homologously with two different promoters. *Cell*, **20**, 269-281.
304. Klauser, B. and Hartig, J.S. (2013) An engineered small RNA-mediated genetic switch based on a ribozyme expression platform. *Nucleic acids research*, **41**, 5542-5552.
305. Wieland, M., Benz, A., Klauser, B. and Hartig, J.S. (2009) Artificial ribozyme switches containing natural riboswitch aptamer domains. *Angewandte Chemie*, **48**, 2715-2718.

306. Winkler, W., Nahvi, A. and Breaker, R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952-956.
307. Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G. and Little, E. (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, **271**, 470-477.
308. McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F. *et al.* (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature*, **413**, 852-856.
309. Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C., Samuelson, M., Svanborg, C., Gottschalk, G., Karch, H. and Hacker, J. (2003) Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. *J Bacteriol*, **185**, 1831-1840.
310. Callister, S.J., McCue, L.A., Turse, J.E., Monroe, M.E., Auberry, K.J., Smith, R.D., Adkins, J.N. and Lipton, M.S. (2008) Comparative bacterial proteomics: analysis of the core genome concept. *PloS one*, **3**, e1542.
311. Cooper, M.B., Loose, M. and Brookfield, J.F. (2009) The evolutionary influence of binding site organisation on gene regulatory networks. *Bio Systems*, **96**, 185-193.
312. Huo, Y.X., Rosenthal, A.Z. and Gralla, J.D. (2008) General stress response signalling: unwrapping transcription complexes by DNA relaxation via the sigma38 C-terminal domain. *Molecular microbiology*, **70**, 369-378.
313. Seth, D., Hausladen, A., Wang, Y.J. and StamLer, J.S. (2012) Endogenous protein S-Nitrosylation in *E. coli*: regulation by OxyR. *Science*, **336**, 470-473.
314. Anjem, A., Varghese, S. and ImLay, J.A. (2009) Manganese import is a key element of the OxyR response to hydrogen peroxide in *Escherichia coli*. *Molecular microbiology*, **72**, 844-858.
315. Magnusson, L.U., Farewell, A. and Nystrom, T. (2005) ppGpp: a global regulator in *Escherichia coli*. *Trends in microbiology*, **13**, 236-242.
316. Ades, S.E., Connolly, L.E., Alba, B.M. and Gross, C.A. (1999) The *Escherichia coli* sigma(E)-dependent extracytoplasmic stress response is controlled by the regulated proteolysis of an anti-sigma factor. *Genes & development*, **13**, 2449-2461.
317. Day, H.A., Pavlou, P. and Waller, Z.A. (2014) i-Motif DNA: structure, stability and targeting with ligands. *Bioorg Med Chem*, **22**, 4407-4418.
318. Billoud, B., Kontic, M. and Viari, A. (1996) Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic acids research*, **24**, 1395-1403.
319. Khomyakova, E.B., Gousset, H., Liquier, J., Huynh-Dinh, T., Gouyette, C., Takahashi, M., Florentiev, V.L. and Taillandier, E. (2000) Parallel intramolecular DNA triple helix with G and T bases in the third strand stabilized by Zn(2+) ions. *Nucleic acids research*, **28**, 3511-3516.
320. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 560-564.
321. Benham, C.J. (1993) Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory loci. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 2999-3003.
322. Bergler, H., Hogenauer, G. and Turnowsky, F. (1992) Sequences of the envM gene and of two mutated alleles in *Escherichia coli*. *Journal of general microbiology*, **138**, 2093-2100.
323. Kunisawa, T. and Nakamura, M. (1991) Identification of regulatory building blocks in *Escherichia coli* genome. *Protein sequences & data analysis*, **4**, 43-47.
324. Leung, M.Y., Blaisdell, B.E., Burge, C. and Karlin, S. (1991) An efficient algorithm for identifying matches with errors in multiple long molecular sequences. *Journal of molecular biology*, **221**, 1367-1378.
325. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**, 3389-3402.
326. Kanehisa, M. (2000) Pathway databases and higher order function. *Advances in protein chemistry*, **54**, 381-408.
327. Wang, G. and Vasquez, K.M. (2006) Non-B DNA structure-induced genetic instability. *Mutation research*, **598**, 103-119.
328. Thanbichler, M., Viollier, P.H. and Shapiro, L. (2005) The structure and function of the bacterial chromosome. *Current opinion in genetics & development*, **15**, 153-162.
329. Wiggins, P.A., Cheveralls, K.C., Martin, J.S., Lintner, R. and Kondev, J. (2010) Strong intranucleoid interactions organize the *Escherichia coli* chromosome into a nucleoid filament. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 4991-4995.
330. Espeli, O. and Boccard, F. (2006) Organization of the *Escherichia coli* chromosome into macrodomains and its possible functional implications. *Journal of structural biology*, **156**, 304-310.
331. Valens, M., Penaud, S., Rossignol, M., Cornet, F. and Boccard, F. (2004) Macrodomain organization of the *Escherichia coli* chromosome. *The EMBO journal*, **23**, 4330-4341.
332. Wang, W., Li, G.W., Chen, C., Xie, X.S. and Zhuang, X. (2011) Chromosome organization by a nucleoid-associated protein in live bacteria. *Science*, **333**, 1445-1449.
333. de Wit, E. and de Laat, W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes & development*, **26**, 11-24.
334. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
335. Rodley, C.D., Bertels, F., Jones, B. and O'Sullivan, J.M. (2009) Global identification of yeast chromosome interactions using Genome conformation capture. *Fungal genetics and biology : FG & B*, **46**, 879-886.
336. Le, T.B., Imakaev, M.V., Mirny, L.A. and Laub, M.T. (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, **342**, 731-734.
337. Cagliero, C., Grand, R.S., Jones, M.B., Jin, D.J. and O'Sullivan, J.M. (2013) Genome conformation capture reveals that the *Escherichia coli* chromosome is organized by replication and transcription. *Nucleic acids research*, **41**, 6058-6071.
338. Voss, P., Wagner, S., Horn, M., Felletti, M., Zinsmaier, M., Berthold, M. and Hartig, J.S. (2014) A user-friendly, open-source software for analyzing high throughput chromosome conformation capture (Hi-C) data. *Nucleic acids research*, **submitted**.
339. Bichara, M., Wagner, J. and Lambert, I.B. (2006) Mechanisms of tandem repeat instability in bacteria. *Mutation research*, **598**, 144-163.

340. Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., Lenski, R.E. and Kim, J.F. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, **461**, 1243-1247.
341. Touzain, F., Petit, M.A., Schbath, S. and El Karoui, M. (2011) DNA motifs that sculpt the bacterial chromosome. *Nature reviews. Microbiology*, **9**, 15-26.
342. Rogers, F.A. and Tiwari, M.K. (2013) Triplex-induced DNA damage response. *The Yale journal of biology and medicine*, **86**, 471-478.
343. Svozil, D., Kalina, J., Omelka, M. and Schneider, B. (2008) DNA conformations and their sequence preferences. *Nucleic acids research*, **36**, 3690-3706.
344. Hwang, T.L. and Shaka, A.J. (1995) Water Suppression That Works. Excitation Sculpting Using Arbitrary Waveforms and Pulsed Field Gradients. *J Magn Reson*, **112**, 275-279.
345. Sun, D. and Hurley, L.H. (2010) Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay. *Methods in molecular biology*, **608**, 65-79.
346. Owen, R.J. and Borman, P. (1987) A rapid biochemical method for purifying high molecular weight bacterial chromosomal DNA for restriction enzyme analysis. *Nucleic acids research*, **15**, 3631.
347. Sambrook, J.F. and Russel, D.W. (2001) Molecular Cloning: A Laboratory Manual. *Cold Spring Harbor Laboratory Press*, **Cold Spring Harbor, NY**.
348. Metcalf, W.W., Jiang, W. and Wanner, B.L. (1994) Use of the rep technique for allele replacement to construct new *Escherichia coli* hosts for maintenance of R6K gamma origin plasmids at different copy numbers. *Gene*, **138**, 1-7.
349. Kaniga, K., Delor, I. and Cornelis, G.R. (1991) A wide-host-range suicide vector for improving reverse genetics in gram-negative bacteria: inactivation of the blaA gene of Yersinia enterocolitica. *Gene*, **109**, 137-141.
350. Weese, D., Holtgrewe, M. and Reinert, K. (2012) RazerS 3: faster, fully sensitive read mapping. *Bioinformatics*, **28**, 2592-2599.
351. Angiuoli, S.V. and Salzberg, S.L. (2011) Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, **27**, 334-342.
352. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**, 772-780.

11 LIST OF FIGURES

12 LIST OF TABLES

13 APPENDICES

Table 13.1: Quadruplexes found in SD regions of *E. coli* MG1655 genes.

Quadruplexes found in SD regions of *E. coli* MG1655 genes, according to the Proquad pattern search (http://quadbase.igib.res.in/). Gene name and function are given according to NCBI blast and Ecocyc (http://ecocyc.org/). The respective quadruplex sequence is shown in the second column, capitals indicate gene start (AUG). Furthermore, it is indicated whether the quadruplex sequence occurs within any other gene (intra) or in a non-coding region (inter). For quadruplex sequences not including the gene start, the distance to the gene start is given. Occurrence of quadruplexes upstream of the same genes in other *E. coli* subtypes or γ-proteobacteria is given for *Escherichia coli CFT073 (ECC), Escherichia coli O157:H7 str. Sakai (ECS), Shigella flexneri 2a str. 301 (sf301)*, *Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 (STM)* and *Xanthomonas campestris pv. campestris str. ATCC 33913 (Xcc).*The last column indicates whether putative quadruplex related genes are organized in an operon. Potential operons without experimental evidence according to the Ecocyc homepage (http://ecocyc.org/).

Gene	Sequence (5'-3')	Intra-/intergenic	Distance to gene start	Function	Other *E. coli* subtypes	Other organisms	Within operon
astE	gggaggggcaATGG	intra	includes start	succinylglutamate desuccinylase	ECC, ECS	STM (gggggcggctgATGG)	astCADBE
bssR/ yliH	ggctgggaagaggagg	inter	8	regulator of biofilm formation			no
cysW	ggtcggcgtgtggtagg	intra	6	sulfate/thiosulfate ABC transporter	ECC, ECS		w/o exp. Evidence: cysPUWA
ddpC/yddQ	ggacgtggaggtgg	intra	3	membrane component of ABC transporter	ECS		w/o exp-evidence: ddpXABCDF
fadJ	ggttctggaggcgg	intra	4	component of anaerobic fatty acid oxidation complex	ECS	STM	w/o exp.evidence: fadIJ
fliH	ggtaattggcagcggcgagg	intra	6	flagellar biosynthesis protein	ECS (ggtaaTTGCAGCGGCGAGG)	Sf301 (ggtgattggcagcggcgagg), STM (ggtgattggcagcggcgagg), XCC (ggtgcggaggcgatggtATG)	w/o exp. Evidence: fliFGHIJK
ftsB/ygbQ	gggggcaggATGGG	inter	includes start	cell division protein	ECC, ECS	sf301, STM	no
glgX	ggctggttcgggagg	intra	4	glycogen debranching enzyme	ECC, ECS	sf301	glgBXCAP

Gene	Sequence (5‘-3‘)	Intra-/intergenic	Distance to gene start	Function	Other *E. coli* subtypes	Other organisms	Within operon
hofN/yrfC	ggggctggcgcttgggaagg	intra	8	protein involved in utilization of DNA as a carbon source	ECC	sf301	w/o exp. Evidence: hofMNOP
hofQ	ggttgtggcaacggcaagg	intra	8	protein involved in utilization of DNA as a carbon source	ECC, ECS	sf301	no
kefC	ggaatggcaggagg	intra	6	glutathione-regulated potassium efflux system	ECC, ECS	sf301, STM (ggcatggcaggaggtgatcATG)	w/o exp. Evidence: kefFC
mioC	ggtggcggtttATGG	inter	includes start	flavoprotein involved in biotin synthesis	ECC, ECS	sf301	asnC-mioc-mnmG
mqsR/ygiU	gggaggcgggggttATGG	inter	includes start	mRNA interferase, toxin of the MqsR-YgiT toxin-antitoxin system			mqsRA
mraZ	ggaataaggggtgaggctgg	inter	1	conserved protein	ECS	sf301	mraZW-ftsLI-murEF-mraY-murD-ftsW-murGC-ddlB-ftsQAZ-lpxC
murE	ggcgaggggacagGTGG	intra	includes start	cell wall synthesis	ECS	sf301	mraZW-ftsLI-murEF-mraY-murD-ftsW-murGC-ddlB-ftsQAZ-lpxC
napH	ggctggctggagggg	intra	4	non-haem iron-sulfur protein	ECC, ECS	sf301, STM	w/o experimental evidence: napFDAGHBC-ccmABCDEFGH

Gene	Sequence (5'-3')	Intra-/intergenic	Distance to gene start	Function	Other *E. coli* subtypes	Other organisms	Within operon
oxyR	ggcgatggaggatggataATG	inter	includes start	oxidative stress regulator	ECS	sf301	no
panE/ apbA	ggagtggttgcgggtgagg	inter	7	2-Dehydropantoate 2-reductase	ECS	sf301	w/o exp. Evidence: panE-yajL
prc	ggaggccgggccagg	inter	1	ATP-independent periplasmic protease	ECC, ECS	sf301, STM (ggaggccaggcctggcATG)	w/o exp. Evidence: proQ-prc
proC	ggcaggagtgaggcaATGG	inter	includes start	pyrroline- 5-carboxylic acid reductase (PCA reductase)	ECS	sf301	no
relA	ggagaggacgATGGTTGCGG	inter	includes start	GDP pyrophosphokinase / GTP pyrophosphokinase	ECC, ECS	sf301, STM (ggagaggacgATGGTCGCGG)	relA-mazEF
rplA	gggcctggtagtggagg	intra	8	50S ribosomal subunit protein L1	ECC, ECS	STM	rplKAJL-rpoBC
rseA	ggatactggataagggtattagg	inter	1	anti-sigma factor	ECC, ECS	sf301, STM (ggatactggaaaaggtattaggcATG)	rpoE-rseABC
serA	ggattgggtaaATGGCAAAGG	inter	includes start	α-ketoglutarate reductase / D-3-phosphoglycerate dehydrogenase	ECC, ECS	sf301, STM (ggatcggggaaATGGCAAAGG)	no
sgcE	ggtatcaggttaacggagg	intra	5	KpLE2 phage like element, predicted epimerase			ryjB operon (sgcAER)
sufD	ggaggagcagggttATGG	intra	includes start	component of SufBCD Fe-S cluster scaffold complex	ECC, ECS	sf301, STM	sufABCDSE
ulaF/yjfX/sgaE	gggcatggtggaggcgg	intra	5	L-ribulose 5-	ECS		ulaABCDEF

Gene	Sequence (5'-3')	Intra-/intergenic	Distance to gene start	Function	Other *E. coli* subtypes	Other organisms	Within operon
				phosphate 4-epimerase			
uspB/yhiO	gggcaggtcgccggggagg	inter	4	predicted universal stress (ethanol tolerance) protein B	ECC	sf301	no
wcaB	ggctgatggtattcggggg	intra	8	predicted colanic acid biosynthesis acyl transferase	ECC, ECS	sf301, STM (ggccgggctgcggggg)	w/o exp. Evidence: wcaCDEF-gmd-fcl-gmm-wcaI-cpsBG-wcaJ-wzxC
yadI	ggctaaggaggaagg	inter	2	N-acetylgalactosamine-transporting PEP-dependent phosphotransferase system	ECC, ECS	sf301,	no
yagQ/paoD	ggatgtggtttaaggagg	intra	4	molybdenum modification and involved into the aldehyde ferredoxin oxidoreductase YagTSR	ECS		w/o exp. Evidence: paoABCD
ybbA	ggaccaggaacaggaaggg	intra	3	ATP-binding component of a predicted ABC superfamily metabolite uptake transporter	ECC	sf301	w/o exp. Evidence: ybbAP

Gene	Sequence (5'-3')	Intra-/intergenic	Distance to gene start	Function	Other *E. coli* subtypes	Other organisms	Within operon
ydcU	ggttgaggagaggtgaATGG	intra	includes start	membrane component of a predicted spermidine/putrescine ABC transporter	ECC, ECS		w/o exp. Evidence: ydcSTUV
yeeL	gggaggcgaggttagg	inter	1	glycosyltransferase			w/o exp. Evidence: yeeL1_L2
yfdN	ggcaggattcaggggG	intra	includes start	CPS-53 (kpLE1) prophage	ECC		w/o exp. Evidence: yfdONMLK
yfgJ	gggaatggctaccggagg	inter	7	putative membrane protein			w/o exp. Evidence: yfgHI
yfgM	gggttaaggaaggagaagg	intra	5	conserved protein	ECC, ECS	sf301	no
ygfI	ggacaaggatgcgggggagggg	inter	1	DNA-binding transcriptional regulator LYSR-type	ECC (ggatgcgggagaggggggATG)	sf301 (ggatgcggggaggggggATG)	no
ygfK	ggagagggttATGGGGG	inter	includes start	iron-sulfur flavoprotein with NADPH:O_2 oxidoreductase activity	ECC, ECS		w/o exp. Evidence: ygfK-ssnA
yghZ	ggtggaaaggagagg	inter	2	L-glyceraldehyde 3-phosphate (L-GAP) reductase	ECC (ggaaaggagaggtcATGG), ECS	sf301	no
yhcF	ggtcctggtggagg	intra	4	predicted transcriptional regulator	ECC, ECS	sf301	w/o exp. Evidence: ychE-2-yhcF
yhiP/dtpB	ggatggtattggaagg	inter	9	member of the POT family of peptide			no

Gene	Sequence (5‘-3‘)	Intra-/intergenic	Distance to gene start	Function	Other *E. coli* subtypes	Other organisms	Within operon
				transporters			
yjdI	gggaaggtcATGGATCAGG	inter	includes start	conserved protein	ECC, ECS		w/o exp. Evidence: ydjdIJ
yjgR	ggcaggaaactggagg	inter	7	predicted ATPase	ECC, ECS		no
yjjU	ggaaggcagagGTGGGG	inter	includes start	predicted Esterase	ECS	STM (ggaaggcagagGTGGG)	w/o exp. Evidence: yjjUV
mzrA/ecfM/yqjB	ggaaatcgggggtaaggg	intra	0	EnvZ/OmpR osmoregulatory two-component signal transduction system	ECC, ECS		w/o exp. Evidence: yqjA-mzrA

Table 13.2: Examples of triplexes found in *E. coli* subspecies using the ITxF database.

Number of triplexes corresponding to the different triplex classes for the different *E. coli* subspecies found in the ITxF database. The total number of triplexes per strain is given in the last column.

Genome/Plasmid name	Genome/Plasmid size (nt)	ClassI	ClassII	ClassIII	ClassIV	Sum of triplexes per genome/plasmid
>gi|387887350|ref|NC_017910.1| Escherichia blattae DSM 4481 chromosome, complete genome	4158725	118	118	12	20	268
>gi|387604868|ref|NC_017627.1| *Escherichia coli* 042 plasmid pAA, complete sequence	113346	12	9	2	0	23
>gi|387605479|ref|NC_017626.1| *Escherichia coli* 042, complete genome	5241977	207	220	37	29	493
>gi|110640213|ref|NC_008253.1| *Escherichia coli* 536, complete genome	4938920	197	189	25	28	439
>gi|218693476|ref|NC_011748.1| *Escherichia coli* 55989 chromosome, complete genome	5154862	226	213	44	39	522
>gi|386637348|ref|NC_017629.1| *Escherichia coli* ABU 83972 plasmid pABU, complete sequence	1564	0	0	0	0	0
>gi|386637352|ref|NC_017631.1| *Escherichia coli* ABU 83972 chromosome, complete genome	5131397	193	181	27	29	430
>gi|157418083|ref|NC_009837.1| *Escherichia coli* APEC O1 plasmid pAPEC-O1-ColBM, complete sequence	174241	16	9	4	2	31
>gi|117622295|ref|NC_008563.1| *Escherichia coli* APEC O1 chromosome, complete genome	5082025	193	199	35	33	460
>gi|157412014|ref|NC_009838.1| *Escherichia coli* APEC O1 plasmid pAPEC-O1-R, complete sequence	241387	33	25	3	3	64
>gi|443615330|ref|NC_020163.1| *Escherichia coli* APEC O78, complete genome	4798435	178	204	39	24	445
>gi|170018061|ref|NC_010468.1| *Escherichia coli* ATCC 8739 chromosome, complete genome	4746218	192	174	30	31	427
>gi|387825439|ref|NC_012971.2| *Escherichia coli* BL21(DE3) chromosome, complete genome	4558953	169	183	30	30	412
>gi|387823261|ref|NC_012892.2| *Escherichia coli* BL21(DE3), complete genome	4558947	169	183	30	30	412
>gi|238899406|ref|NC_012759.1| *Escherichia coli* BW2952 chromosome, complete genome	4578159	173	177	32	31	413
>gi|254160123|ref|NC_012967.1| *Escherichia coli* B str. REL606 chromosome, complete genome	4629812	177	190	31	30	428
>gi|26245917|ref|NC_004431.1| *Escherichia coli* CFT073 chromosome, complete genome	5231428	203	185	27	30	445
>gi|386593590|ref|NC_017625.1| *Escherichia coli* DH1 chromosome, complete genome	4630707	175	186	33	33	427
>gi|387619774|ref|NC_017638.1| *Escherichia coli* DH1, complete genome	4621430	174	186	33	32	425
>gi|157154711|ref|NC_009801.1| *Escherichia coli* E24377A chromosome, complete genome	4979619	193	210	41	31	475

Genome/Plasmid name	Genome/Plasmid size (nt)	ClassI	ClassII	ClassIII	ClassIV	Sum of triplexes per genome/plasmid
>gi|157149504|ref|NC_009790.1| *Escherichia coli* E24377A plasmid pETEC_74, complete sequence	74224	6	9	1	0	16
>gi|157149498|ref|NC_009789.1| *Escherichia coli* E24377A plasmid pETEC_6, complete sequence	6199	0	0	0	0	0
>gi|157149330|ref|NC_009786.1| *Escherichia coli* E24377A plasmid pETEC_80, complete sequence	79237	9	7	1	1	18
>gi|157149429|ref|NC_009788.1| *Escherichia coli* E24377A plasmid pETEC_73, complete sequence	70609	3	3	0	0	6
>gi|157149574|ref|NC_009791.1| *Escherichia coli* E24377A plasmid pETEC_5, complete sequence	5033	2	0	0	0	2
>gi|157149399|ref|NC_009787.1| *Escherichia coli* E24377A plasmid pETEC_35, complete sequence	34367	1	4	0	0	5
>gi|218687878|ref|NC_011745.1| *Escherichia coli* ED1a chromosome, complete genome	5209548	197	186	26	36	445
>gi|387610477|ref|NC_017633.1| *Escherichia coli* ETEC H10407, complete genome	5153435	183	218	34	28	463
>gi|387615175|ref|NC_017721.1| *Escherichia coli* ETEC H10407 plasmid p52, complete sequence	5175	1	0	0	0	1
>gi|387615182|ref|NC_017723.1| *Escherichia coli* ETEC H10407 plasmid p58, complete sequence	5800	1	1	0	0	2
>gi|387610385|ref|NC_017724.1| *Escherichia coli* ETEC H10407 plasmid p948, complete sequence	94797	10	17	3	0	30
>gi|387610311|ref|NC_017722.1| *Escherichia coli* ETEC H10407 plasmid p666, complete sequence	66681	6	7	1	1	15
>gi|157159467|ref|NC_009800.1| *Escherichia coli* HS, complete genome	4643538	174	176	29	25	404
>gi|218552585|ref|NC_011741.1| *Escherichia coli* IAI1 chromosome, complete genome	4700560	190	217	43	30	480
>gi|218698419|ref|NC_011750.1| *Escherichia coli* IAI39 chromosome, complete genome	5132068	210	189	23	28	450
>gi|386597751|ref|NC_017628.1| *Escherichia coli* IHE3034 chromosome, complete genome	5108383	186	180	30	30	426
>gi|556555082|ref|NC_022649.1| *Escherichia coli* JJ1886 plasmid pJJ1886_2, complete sequence	5167	0	0	0	0	0
>gi|556550243|ref|NC_022648.1| *Escherichia coli* JJ1886, complete genome	5129938	190	171	25	29	415
>gi|556579581|ref|NC_022661.1| *Escherichia coli* JJ1886 plasmid pJJ1886_1, complete sequence	1552	0	1	0	0	1
>gi|556555098|ref|NC_022650.1| *Escherichia coli* JJ1886 plasmid pJJ1886_4, complete sequence	55956	4	6	2	0	12
>gi|556579591|ref|NC_022662.1| *Escherichia coli* JJ1886 plasmid pJJ1886_3, complete sequence	5631	1	3	1	0	5
>gi|556555179|ref|NC_022651.1| *Escherichia coli* JJ1886 plasmid pJJ1886_5, complete sequence	110040	5	2	1	0	8
>gi|386698504|ref|NC_017660.1| *Escherichia coli* KO11FL chromosome, complete genome	5021812	203	197	39	51	490

Genome/Plasmid name	Genome/Plasmid size (nt)	ClassI	ClassII	ClassIII	ClassIV	Sum of triplexes per genome/plasmid
>gi|386703202|ref|NC_017661.1| *Escherichia coli* KO11FL plasmid pRK2, complete sequence	5360	1	1	0	0	2
>gi|378715370|ref|NC_016903.1| *Escherichia coli* KO11FL plasmid pEKO1102, complete sequence	5360	1	1	0	0	2
>gi|378710836|ref|NC_016902.1| *Escherichia coli* KO11FL chromosome, complete genome	4920168	194	203	40	35	472
>gi|378715377|ref|NC_016904.1| *Escherichia coli* KO11FL plasmid pEKO1101, complete sequence	103795	8	7	2	0	17
>gi|170079663|ref|NC_010473.1| *Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome	4686137	183	188	33	32	436
>gi|471332236|ref|NC_020518.1| *Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome	3976195	139	147	28	25	339
>gi|556503834|ref|NC_000913.3| *Escherichia coli* str. K-12 substr. MG1655, complete genome	4641652	177	188	34	32	431
>gi|388476123|ref|NC_007779.1| *Escherichia coli* str. K-12 substr. W3110, complete genome	4646332	173	190	34	31	428
>gi|222154829|ref|NC_011993.1| *Escherichia coli* LF82, complete genome	4773108	166	165	27	28	386
>gi|544388862|ref|NC_022364.1| *Escherichia coli* LY180, complete genome	4835601	191	204	39	32	466
>gi|386617516|ref|NC_017644.1| *Escherichia coli* NA114 chromosome, complete genome	4971461	163	144	24	25	356
>gi|260718930|ref|NC_013354.1| *Escherichia coli* O103:H2 str. 12009 plasmid pO103, complete sequence	75546	10	6	1	2	19
>gi|260842239|ref|NC_013353.1| *Escherichia coli* O103:H2 str. 12009, complete genome	5449314	227	246	38	30	541
>gi|410480052|ref|NC_018654.1| *Escherichia coli* O104:H4 str. 2009EL-2050 plasmid pAA-09EL50, complete sequence	74213	6	9	1	0	16
>gi|410480049|ref|NC_018652.1| *Escherichia coli* O104:H4 str. 2009EL-2050 plasmid pG-09EL50, complete sequence	1549	0	0	0	0	0
>gi|410485110|ref|NC_018651.1| *Escherichia coli* O104:H4 str. 2009EL-2050 plasmid p09EL50, complete sequence	109274	7	8	0	2	17
>gi|410480139|ref|NC_018650.1| *Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome	5253138	214	213	40	30	497
>gi|407471978|ref|NC_018663.1| *Escherichia coli* O104:H4 str. 2009EL-2071 plasmid pG-09EL71, complete sequence	1549	0	0	0	0	0
>gi|407466711|ref|NC_018661.1| *Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome	5312586	215	214	40	30	499
>gi|407471876|ref|NC_018662.1| *Escherichia coli* O104:H4 str. 2009EL-2071 plasmid pAA-09EL71, complete sequence	75573	4	9	1	0	14
>gi|407484805|ref|NC_018666.1| *Escherichia coli* O104:H4 str. 2011C-3493 plasmid pAA-EA11, complete sequence	74217	4	10	1	0	15
>gi|407484675|ref|NC_018659.1| *Escherichia coli* O104:H4 str. 2011C-3493 plasmid pESBL-EA11, complete sequence	88544	5	1	1	1	8
>gi|407484773|ref|NC_018660.1| *Escherichia coli* O104:H4 str. 2011C-3493 plasmid pG-EA11, complete sequence	1549	0	0	0	0	0

Genome/Plasmid name	Genome/Plasmid size (nt)	ClassI	ClassII	ClassIII	ClassIV	Sum of triplexes per genome/plasmid
>gi|407479587|ref|NC_018658.1| *Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome	5273097	215	216	41	30	502
>gi|260751835|ref|NC_013366.1| *Escherichia coli* O111:H- str. 11128 plasmid pO111_3, complete sequence	77690	10	4	2	2	18
>gi|260752012|ref|NC_013365.1| *Escherichia coli* O111:H- str. 11128 plasmid pO111_1, complete sequence	204604	23	17	0	5	45
>gi|260751919|ref|NC_013368.1| *Escherichia coli* O111:H- str. 11128 plasmid pO111_5, complete sequence	6673	0	0	0	0	0
>gi|260871126|ref|NC_013370.1| *Escherichia coli* O111:H- str. 11128 plasmid pO111_2, complete sequence	97897	2	3	0	0	5
>gi|260866153|ref|NC_013364.1| *Escherichia coli* O111:H- str. 11128, complete genome	5371077	227	242	42	37	548
>gi|260751908|ref|NC_013367.1| *Escherichia coli* O111:H- str. 11128 plasmid pO111_4, complete sequence	8140	0	0	0	0	0
>gi|215274578|ref|NC_011602.1| *Escherichia coli* O127:H6 str. E2348/69 plasmid pE2348-2, complete sequence	6147	1	0	0	1	2
>gi|215485161|ref|NC_011601.1| *Escherichia coli* O127:H6 str. E2348/69 chromosome, complete genome	4965553	191	188	24	34	437
>gi|215276192|ref|NC_011603.1| *Escherichia coli* O127:H6 str. E2348/69 plasmid pMAR2, complete sequence	97978	16	4	1	3	24
>gi|209395529|ref|NC_011350.1| *Escherichia coli* O157:H7 str. EC4115 plasmid pO157, complete sequence	94644	2	7	1	0	10
>gi|209395638|ref|NC_011351.1| *Escherichia coli* O157:H7 str. EC4115 plasmid pEC4115, complete sequence	37452	9	11	2	1	23
>gi|209395693|ref|NC_011353.1| *Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome	5572075	230	246	36	34	546
>gi|16445223|ref|NC_002655.2| *Escherichia coli* O157:H7 str. EDL933 chromosome, complete genome	5528445	230	233	32	30	525
>gi|75994447|ref|NC_007414.1| *Escherichia coli* O157:H7 EDL933 plasmid pO157, complete sequence	92077	2	7	1	0	10
>gi|254667448|ref|NC_013010.1| *Escherichia coli* O157:H7 str. TW14359 plasmid pO157, complete sequence	94601	2	7	1	0	10
>gi|254791136|ref|NC_013008.1| *Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome	5528136	228	245	36	33	542
>gi|10955266|ref|NC_002128.1| *Escherichia coli* O157:H7 str. Sakai plasmid pO157, complete sequence	92721	2	7	1	0	10
>gi|15829254|ref|NC_002695.1| *Escherichia coli* O157:H7 str. Sakai chromosome, complete genome	5498450	227	235	34	32	528
>gi|10955262|ref|NC_002127.1| *Escherichia coli* O157:H7 str. Sakai plasmid pOSAK1, complete sequence	3306	0	0	0	0	0
>gi|260763802|ref|NC_013369.1| *Escherichia coli* O26:H11 str. 11368 plasmid pO26_1, complete sequence	85167	5	4	0	0	9
>gi|260751930|ref|NC_013362.1| *Escherichia coli* O26:H11 str. 11368 plasmid pO26_2, complete sequence	63365	6	4	1	1	12

Genome/Plasmid name	Genome/Plasmid size (nt)	ClassI	ClassII	ClassIII	ClassIV	Sum of triplexes per genome/plasmid
>gi|307950758|ref|NC_014543.1| *Escherichia coli* O26:H11 str. 11368 plasmid pO26_4, complete sequence	4073	1	1	0	1	3
>gi|260751828|ref|NC_013363.1| *Escherichia coli* O26:H11 str. 11368 plasmid pO26_3, complete sequence	5686	1	0	0	1	2
>gi|260853213|ref|NC_013361.1| *Escherichia coli* O26:H11 str. 11368 chromosome, complete genome	5697240	236	261	43	33	573
>gi|291280824|ref|NC_013941.1| *Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome	5386352	232	238	34	33	537
>gi|291285839|ref|NC_013942.1| *Escherichia coli* O55:H7 str. CB9615 plasmid pO55, complete sequence	66001	6	5	2	1	14
>gi|387504721|ref|NC_017653.1| *Escherichia coli* O55:H7 str. RM12579 plasmid p12579_1, complete sequence	94015	1	4	0	0	5
>gi|387504828|ref|NC_017654.1| *Escherichia coli* O55:H7 str. RM12579 plasmid p12579_3, complete sequence	12068	2	0	0	0	2
>gi|387504844|ref|NC_017657.1| *Escherichia coli* O55:H7 str. RM12579 plasmid p12579_2, complete sequence	66078	6	5	2	1	14
>gi|387504934|ref|NC_017656.1| *Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome	5263980	224	230	32	31	517
>gi|387504713|ref|NC_017658.1| *Escherichia coli* O55:H7 str. RM12579 plasmid p12579_4, complete sequence	6211	1	0	0	1	2
>gi|387504924|ref|NC_017655.1| *Escherichia coli* O55:H7 str. RM12579 plasmid p12579_5, complete sequence	5954	2	2	0	0	4
>gi|386622390|ref|NC_017649.1| *Escherichia coli* O7:K1 str. CE10 plasmid pCE10C, complete sequence	4197	1	2	0	0	3
>gi|386622393|ref|NC_017650.1| *Escherichia coli* O7:K1 str. CE10 plasmid pCE10D, complete sequence	1549	0	0	0	0	0
>gi|386627436|ref|NC_017647.1| *Escherichia coli* O7:K1 str. CE10 plasmid pCE10A, complete sequence	54289	2	1	0	0	3
>gi|386622414|ref|NC_017646.1| *Escherichia coli* O7:K1 str. CE10 chromosome, complete genome	5313531	222	199	24	32	477
>gi|386627431|ref|NC_017648.1| *Escherichia coli* O7:K1 str. CE10 plasmid pCE10B, complete sequence	5163	0	0	0	0	0
>gi|387615344|ref|NC_017634.1| *Escherichia coli* O83:H1 str. NRG 857C chromosome, complete genome	4747819	171	170	27	30	398
>gi|387615190|ref|NC_017659.1| *Escherichia coli* O83:H1 str. NRG 857C plasmid pO83_CORR, complete sequence	147060	9	7	1	1	18
>gi|386703215|ref|NC_017663.1| *Escherichia coli* P12b chromosome, complete genome	4935294	191	191	29	32	443
>gi|544574430|ref|NC_022370.1| *Escherichia coli* PMV-1 main chromosome, complete genome	4984940	184	174	27	30	415
>gi|544579032|ref|NC_022371.1| *Escherichia coli* PMV-1 pHUSEC411like plasmid, complete sequence	98864	2	2	2	0	6
>gi|218556939|ref|NC_011742.1| *Escherichia coli* S88 chromosome, complete genome	5032268	183	190	31	30	434
>gi|218534477|ref|NC_011747.1| *Escherichia coli* S88 plasmid pECOS88, complete	133853	13	11	5	3	32

Genome/Plasmid name	Genome/Plasmid size (nt)	ClassI	ClassII	ClassIII	ClassIV	Sum of triplexes per genome/plasmid
sequence						
>gi|209917191|ref|NC_011415.1| *Escherichia coli* SE11 chromosome, complete genome	4887515	188	196	36	26	446
>gi|209921952|ref|NC_011419.1| *Escherichia coli* SE11 plasmid pSE11-1, complete sequence	100021	6	3	1	0	10
>gi|209921875|ref|NC_011416.1| *Escherichia coli* SE11 plasmid pSE11-3, complete sequence	60555	7	6	1	0	14
>gi|209916829|ref|NC_011413.1| *Escherichia coli* SE11 plasmid pSE11-2, complete sequence	91158	8	7	2	1	18
>gi|209916825|ref|NC_011411.1| *Escherichia coli* SE11 plasmid pSE11-6, complete sequence	4082	0	0	0	0	0
>gi|209916806|ref|NC_011407.1| *Escherichia coli* SE11 plasmid pSE11-4, complete sequence	6929	1	1	0	0	2
>gi|209916817|ref|NC_011408.1| *Escherichia coli* SE11 plasmid pSE11-5, complete sequence	5366	1	1	0	0	2
>gi|281427817|ref|NC_013655.1| *Escherichia coli* SE15 plasmid pECSF1, complete sequence	122345	11	4	2	1	18
>gi|387828053|ref|NC_013654.1| *Escherichia coli* SE15, complete genome	4717338	161	168	29	25	383
>gi|170650756|ref|NC_010487.1| *Escherichia coli* SMS-3-5 plasmid pSMS35_3, complete sequence	3565	0	2	0	0	2
>gi|170650751|ref|NC_010486.1| *Escherichia coli* SMS-3-5 plasmid pSMS35_4, complete sequence	4074	0	0	0	0	0
>gi|170650740|ref|NC_010485.1| *Escherichia coli* SMS-3-5 plasmid pSMS35_8, complete sequence	8909	0	3	1	0	4
>gi|170650760|ref|NC_010488.1| *Escherichia coli* SMS-3-5 plasmid pSMS35_130, complete sequence	130440	8	6	1	1	16
>gi|170679574|ref|NC_010498.1| *Escherichia coli* SMS-3-5 chromosome, complete genome	5068389	197	198	26	30	451
>gi|386602643|ref|NC_017632.1| *Escherichia coli* UM146 chromosome, complete genome	4993013	185	179	26	30	420
>gi|386602509|ref|NC_017630.1| *Escherichia coli* UM146 plasmid pUM146, complete sequence	114550	7	4	1	1	13
>gi|218703261|ref|NC_011751.1| *Escherichia coli* UMN026 chromosome, complete genome	5202090	207	200	31	33	471
>gi|218692794|ref|NC_011749.1| *Escherichia coli* UMN026 plasmid p1ESCUM, complete sequence	122301	9	4	1	1	15
>gi|218454959|ref|NC_011739.1| *Escherichia coli* UMN026 plasmid p2ESCUM, complete sequence	33809	8	4	2	1	15
>gi|386617281|ref|NC_017643.1| *Escherichia coli* UMNK88 plasmid pUMNK88_Hly, complete sequence	65549	4	5	1	0	10
>gi|386612163|ref|NC_017641.1| *Escherichia coli* UMNK88 chromosome, complete genome	5186416	200	196	31	35	462
>gi|386611993|ref|NC_017639.1| *Escherichia coli* UMNK88 plasmid pUMNK88_K88, complete sequence	81883	7	6	0	0	13

Genome/Plasmid name	Genome/Plasmid size (nt)	ClassI	ClassII	ClassIII	ClassIV	Sum of triplexes per genome/plasmid
>gi|386612081|ref|NC_017640.1| *Escherichia coli* UMNK88 plasmid pUMNK88_Ent, complete sequence	81475	9	8	1	1	19
>gi|386611903|ref|NC_017642.1| *Escherichia coli* UMNK88 plasmid pUMNK88_91, complete sequence	90868	4	3	1	0	8
>gi|386617342|ref|NC_017645.1| *Escherichia coli* UMNK88 plasmid pUMNK88, complete sequence	160573	7	6	1	1	15
>gi|91206245|ref|NC_007941.1| *Escherichia coli* UTI89 plasmid pUTI89, complete sequence	114230	10	3	1	1	15
>gi|91209055|ref|NC_007946.1| *Escherichia coli* UTI89 chromosome, complete genome	5065741	191	178	27	31	427
>gi|386611788|ref|NC_017637.1| *Escherichia coli* W plasmid pRK1, complete sequence	102536	8	7	2	0	17
>gi|386607294|ref|NC_017636.1| *Escherichia coli* W plasmid pRK2, complete sequence	5360	1	1	0	0	2
>gi|386607309|ref|NC_017635.1| *Escherichia coli* W chromosome, complete genome	4900968	195	207	39	34	475
>gi|386707734|ref|NC_017664.1| *Escherichia coli* W chromosome, complete genome	4897452	195	207	39	34	475
>gi|386707622|ref|NC_017665.1| *Escherichia coli* W plasmid pRK1, complete sequence	102535	8	7	2	0	17
>gi|386707609|ref|NC_017662.1| *Escherichia coli* W plasmid pRK2, complete sequence	5360	1	1	0	0	2
>gi|387885599|ref|NC_017907.1| *Escherichia coli* Xuzhou21 plasmid pO157, complete sequence	92728	2	7	1	0	10
>gi|387873357|ref|NC_017903.1| *Escherichia coli* Xuzhou21 plasmid pO157_Sal, complete sequence	37785	6	4	1	1	12
>gi|387880559|ref|NC_017906.1| *Escherichia coli* Xuzhou21 chromosome, complete genome	5386223	218	227	33	30	508
>gi|253771435|ref|NC_012947.1| *Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome	4570938	172	184	29	32	417
>gi|386632422|ref|NC_017652.1| *Escherichia coli* str. 'clone D i14' chromosome, complete genome	5038386	193	179	27	29	428
>gi|386627502|ref|NC_017651.1| *Escherichia coli* str. 'clone D i2' chromosome, complete genome	5038386	193	179	27	29	428

Table 13.3: TM sequences in proteobacteria.

The Number of TMs is given for the respective bacterial strain. Data kindly provided by Kangkan Halder.

Organism	Number of TMs
Herpetosiphon_aurantiacus_ATCC_23779	192
Enterobacter_cloacae_subsp._cloacae_ATCC_13047	175
Enterobacter_sp._638	115
Sphaerobacter_thermophilus_DSM_20745	111
Roseiflexus_sp._RS-1	102
Methylobacterium_nodulans_ORS_2060	97
Rhodopseudomonas_palustris_HaA2	89
Cupriavidus_taiwanensis	89
Pseudomonas_fluorescens_Pf0-1	88
Geobacter_sp._M21	83
Bradyrhizobium_japonicum_USDA_110	75
Polaromonas_sp._JS666	72
Geobacter_bemidjiensis_Bem	64
Synechococcus_sp._JA-3-3Ab	63
Sideroxydans_lithotrophicus_ES-1	59
Rhodopseudomonas_palustris_BisA53	55
Azotobacter_vinelandii_DJ	50
Methylobacterium_radiotolerans_JCM_2831	47
Rhizobium_leguminosarum_bv._viciae_3841	43
Rhizobium_leguminosarum_bv._trifolii_WSM2304	42
Klebsiella_pneumoniae_342	41
Rhodopseudomonas_palustris_CGA009	41
Rhodopseudomonas_palustris_TIE-1	40
Variovorax_paradoxus_S110	40
Xanthomonas_campestris_pv._campestris_str._8004	38
Xanthomonas_campestris_pv._campestris_str._ATCC_33913	38
Sinorhizobium_meliloti_1021	38
Xanthomonas_campestris_pv._campestris_str._B100	37
Pseudomonas_mendocina_ymp	33
Rhodopseudomonas_palustris_BisB5	31
Escherichia_coli_55989	30
Escherichia_coli_IAI1	30
Escherichia_coli_O26-H11_str._11368	30
Methylococcus_capsulatus_str._Bath	30
Ralstonia_eutropha_H16	30
Klebsiella_variicola_At-22	29
Escherichia_coli_E24377A	29
Escherichia_coli_O111-H-_str._11128	29
Oligotropha_carboxidovorans_OM5	29
Escherichia_coli_str._K-12_substr._DH10B	28

Organism	Number of TMs
Neisseria_gonorrhoeae_NCCP11945	28
Shigella_sonnei_Ss046	27
Escherichia_coli_SE11	27
Rhizobium_sp._NGR234	27
Escherichia_coli_O103-H2_str._12009	26
Shigella_flexneri_5_str._8401	24
Shigella_flexneri_2a_str._2457T	24
Shigella_flexneri_2a_str._301	24
Escherichia_coli_BW2952	24
Xanthomonas_axonopodis_pv._citri_str._306	23
Serratia_proteamaculans_568	23
Escherichia_coli_str._K-12_substr._MG1655	23
Escherichia_coli_ATCC_8739	23
Escherichia_coli_'BL21-Gold|DE3|pLysS_AG'	23
Escherichia_coli_B_str._REL606	22
Sphingopyxis_alaskensis_RB2256	22
Xanthomonas_campestris_pv._vesicatoria_str._85-10	21
Shigella_boydii_CDC_3083-94	21
Shigella_boydii_Sb227	21
Escherichia_coli_HS	21
Rhizobium_etli_CIAT_652	21
Azospirillum_sp._B510	21
Geobacter_uraniireducens_Rf4	20
Aeromonas_hydrophila_subsp._hydrophila_ATCC_7966	18
Rhizobium_leguminosarum_bv._trifolii_WSM1325	18
Caulobacter_crescentus_NA1000	18
Caulobacter_crescentus_CB15	18
Neisseria_gonorrhoeae_FA_1090	18
Enterobacter_cloacae_SCF1	17
Escherichia_coli_UMN026	17
Roseiflexus_castenholzii_DSM_13941	17
Neisseria_meningitidis_053442	16
Shigella_dysenteriae_Sd197	15
Neisseria_meningitidis_alpha14	15
Maricaulis_maris_MCS10	14
Neisseria_meningitidis_FAM18	14
Klebsiella_pneumoniae_NTUH-K2044	13
Neisseria_meningitidis_MC58	13
Escherichia_coli_O157-H7_str._Sakai	12
Escherichia_coli_O157-H7_str._EC4115	12
Escherichia_coli_O157-H7_str._TW14359	12
Escherichia_coli_O55-H7_str._CB9615	11
Escherichia_coli_O157-H7_str._EDL933	11

Organism	Number of TMs
Rhodoferax_ferrireducens_T118	11
Neisseria_meningitidis_Z2491	11
Planctomyces_limnophilus_DSM_3776	11
Synechococcus_sp._JA-2-3B'a|2-13|	11
Escherichia_coli_SMS-3-5	10
Bradyrhizobium_sp._BTAi1	10
Rhizobium_etli_CFN_42	10
Geobacter_sp._FRC-32	9
Klebsiella_pneumoniae_subsp._pneumoniae_MGH_78578	9
Thioalkalivibrio_sp._HL-EbGR7	9
Geobacter_metallireducens_GS-15	8
Escherichia_coli_IAI39	8
Azorhizobium_caulinodans_ORS_571	8
Starkeya_novella_DSM_506	7
Chloroflexus_aggregans_DSM_9485	7
Chlorobium_tepidum_TLS	6
Nitrosococcus_oceani_ATCC_19707	6
Bradyrhizobium_sp._ORS278	6
Methylibium_petroleiphilum_PM1	6
Chlorobium_limicola_DSM_245	5
Xanthomonas_oryzae_pv._oryzae_PXO99A	5
Pseudomonas_stutzeri_A1501	5
Hyphomicrobium_denitrificans_ATCC_51888	5
Acaryochloris_marina_MBIC11017	5
Thermus_thermophilus_HB27	5
Stigmatella_aurantiaca_DW4/3-1	4
Xanthomonas_oryzae_pv._oryzae_KACC10331	4
Xanthomonas_oryzae_pv._oryzae_MAFF_311018	4
Nitrosococcus_halophilus_Nc4	4
Halothiobacillus_neapolitanus_c2	4
Sphingomonas_wittichii_RW1	4
Parvibaculum_lavamentivorans_DS-1	4
Xanthobacter_autotrophicus_Py2	4
Hyphomonas_neptunium_ATCC_15444	4
Cupriavidus_metallidurans_CH34	4
Chloroflexus_aurantiacus_J-10-fl	4
Chloroflexus_sp._Y-400-fl	4
Thermus_thermophilus_HB8	4
Desulfatibacillum_alkenivorans_AK-01	3
Geobacter_lovleyi_SZ	3
Pseudomonas_aeruginosa_PA7	3
Aeromonas_salmonicida_subsp._salmonicida_A449	3

Organism	Number of TMs
Nitrobacter_winogradskyi_Nb-255	3
Rhodopseudomonas_palustris_BisB18	3
Agrobacterium_radiobacter_K84	3
Ralstonia_eutropha_JMP134	3
Polaromonas_naphthalenivorans_CJ2	3
Rhodopirellula_baltica_SH_1	3
Meiothermus_ruber_DSM_1279	3
Truepera_radiovictrix_DSM_17093	3
Chlorobium_phaeobacteroides_DSM_266	2
Escherichia_coli_S88	2
Escherichia_coli_APEC_O1	2
Escherichia_coli_CFT073	2
Pseudomonas_aeruginosa_UCBPP-PA14	2
Pseudomonas_aeruginosa_LESB58	2
Pseudomonas_aeruginosa_PAO1	2
Erythrobacter_litoralis_HTCC2594	2
Methylobacterium_populi_BJ001	2
Sinorhizobium_medicae_WSM419	2
Caulobacter_segnis_ATCC_21756	2
Thiomonas_intermedia_K12	2
Microcystis_aeruginosa_NIES-843	2
Streptosporangium_roseum_DSM_43021	2
Thermanaerovibrio_acidaminovorans_DSM_6589	1
Spirochaeta_thermophila_DSM_6192	1
Acidaminococcus_fermentans_DSM_20731	1
Prevotella_melaninogenica_ATCC_25845	1
Rhodothermus_marinus_DSM_4252	1
Bdellovibrio_bacteriovorus_HD100	1
Desulfurivibrio_alkaliphilus_AHT2	1
Desulfococcus_oleovorans_Hxd3	1
Syntrophobacter_fumaroxidans_MPOB	1
Myxococcus_xanthus_DK_1622	1
Legionella_pneumophila_str._Paris	1
Cronobacter_turicensis_z3032	1
Escherichia_coli_UTI89	1
Escherichia_coli_O127-H6_str._E2348/69	1
Marinomonas_sp._MWYL1	1
Aeromonas_caviae	1
Caulobacter_sp._K31	1
Magnetospirillum_magneticum_AMB-1	1
Nitrosospira_multiformis_ATCC_25196	1
Achromobacter_xylosoxidans_A8	1
Leptothrix_cholodnii_SP-6	1

Organism	Number of TMs
Propionibacterium_freudenreichii_subsp._shermanii_CIRM-BIA1	1
Rhodococcus_jostii_RHA1	1
Rhodococcus_erythropolis_PR4	1
Geodermatophilus_obscurus_DSM_43160	1
Clavibacter_michiganensis_subsp._sepedonicus	1
Deinococcus_radiodurans_R1	1

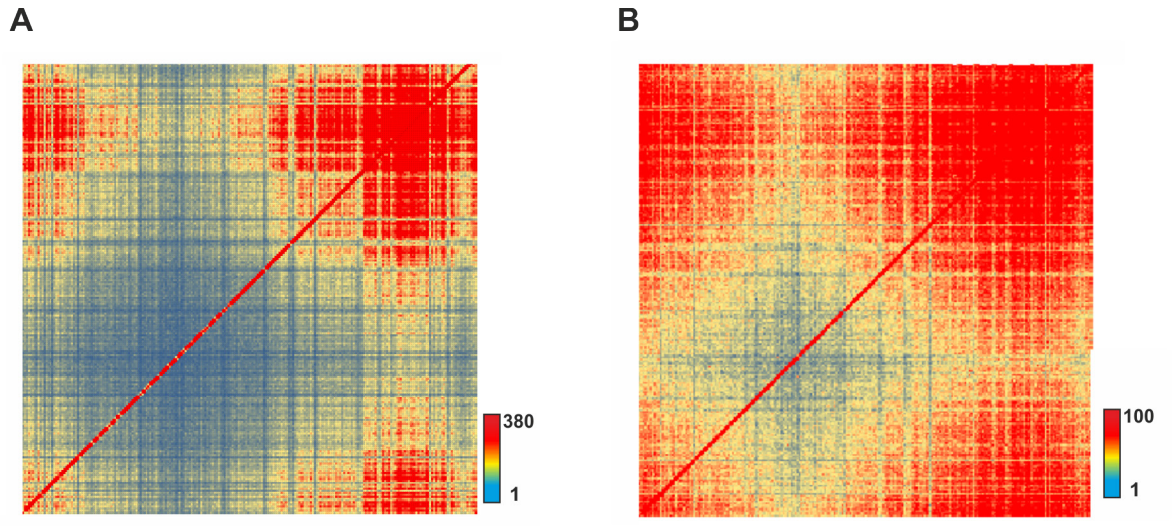


Figure 13.1: Hi-C data.

Hi–C re-evaluation using the KNIME workflow as described in Voss et al. Original data used from **A** Cagliero et al. and **B** Voss et al. Data kindly provided by Dr. Stefanie Wagner.

Table 13.4: Different TM sequences identified in the 56 *E. coli* genomes.

TM No	Code	Start	End	Sequence (5'-3')	Strain
1	NC_000913	164547	164580	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
2	NC_000913	164631	164597	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
3	NC_000913	282101	282136	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
4	NC_000913	289246	289279	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
5	NC_000913	388664	388699	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
6	NC_000913	497843	497878	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
7	NC_000913	624579	624614	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
8	NC_000913	624676	624641	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
9	NC_000913	1351239	1351204	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
10	NC_000913	3045989	3046024	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
11	NC_000913	3046087	3046052	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
12	NC_000913	3239599	3239634	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
13	NC_000913	3239698	3239662	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
14	NC_000913	3390529	3390494	CCCTCGCCCCTTTGGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
15	NC_000913	3504892	3504857	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
16	NC_000913	3608684	3608719	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
17	NC_000913	3781061	3781028	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
18	NC_000913	3781121	3781156	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
19	NC_000913	3908495	3908530	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
20	NC_000913	3959491	3959458	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
21	NC_000913	4070452	4070487	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGAGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
22	NC_000913	4314285	4314320	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
23	NC_000913	4549883	4549848	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MG1655, complete genome.
24	NC_002695	168869	168902	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
25	NC_002695	168953	168919	CCCTCTCCCTTAAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
26	NC_002695	713556	713521	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
27	NC_002695	1493741	1493776	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
28	NC_002695	2959442	2959408	CCCTCGCCCCTTCGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
29	NC_002695	3383056	3383021	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
30	NC_002695	3979231	3979266	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
31	NC_002695	3979330	3979294	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
32	NC_002695	4125763	4125728	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
33	NC_002695	4224808	4224773	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
34	NC_002695	4695203	4695238	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. Sakai chromosome, complete genome.
35	NC_004431	2852458	2852423	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* CFT073 chromosome, complete genome.
36	NC_004431	5144612	5144577	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* CFT073 chromosome, complete genome.
37	NC_007779	164547	164580	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. W3110, complete genome.
38	NC_007779	164631	164597	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* str. K-12 substr. W3110, complete genome.
39	NC_007779	281325	281360	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
40	NC_007779	288470	288503	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. W3110, complete genome.
41	NC_007779	387888	387923	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
42	NC_007779	497067	497102	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
43	NC_007779	623802	623837	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
44	NC_007779	623899	623864	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. W3110, complete genome.
45	NC_007779	1352953	1352918	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
46	NC_007779	3044645	3044680	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. W3110, complete genome.
47	NC_007779	3044743	3044708	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
48	NC_007779	3238255	3238290	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. W3110, complete genome.
49	NC_007779	3238354	3238318	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
50	NC_007779	3390384	3390349	CCCTCGCCCCTTTGGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
51	NC_007779	3566229	3566194	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGAGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
52	NC_007779	3677190	3677223	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. W3110, complete genome.
53	NC_007779	3728186	3728151	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
54	NC_007779	3859294	3859259	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
55	NC_007779	3859354	3859387	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
56	NC_007779	4031731	4031696	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
57	NC_007779	4135524	4135559	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
58	NC_007779	4318963	4318998	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
59	NC_007779	4554563	4554528	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. W3110, complete genome.
60	NC_007946	562046	562081	CCCTCGCCCCTGTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UTI89 chromosome, complete genome.
61	NC_008563	563644	563679	CCCTCGCCCCTGTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O1 chromosome, complete genome.
62	NC_008563	2205660	2205694	CCCTCGCCCCCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O1 chromosome, complete genome.
63	NC_009800	68479	68444	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* HS, complete genome.
64	NC_009800	163595	163628	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* HS, complete genome.
65	NC_009800	163679	163645	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* HS, complete genome.
66	NC_009800	452609	452644	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
67	NC_009800	563908	563943	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
68	NC_009800	660660	660695	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
69	NC_009800	660757	660722	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* HS, complete genome.
70	NC_009800	1398799	1398764	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
71	NC_009800	3073501	3073536	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* HS, complete genome.
72	NC_009800	3073599	3073564	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
73	NC_009800	3434117	3434082	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
74	NC_009800	3549237	3549202	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
75	NC_009800	3654954	3654989	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
76	NC_009800	3935675	3935710	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
77	NC_009800	3987904	3987871	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* HS, complete genome.
78	NC_009800	4259915	4259950	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
79	NC_009800	4338855	4338890	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* HS, complete genome.
80	NC_009800	4520554	4520587	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* HS, complete genome.
81	NC_009800	4623889	4623856	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* HS, complete genome.
82	NC_009801	68414	68379	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
83	NC_009801	166013	166046	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* E24377A chromosome, complete genome.
84	NC_009801	166097	166063	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* E24377A chromosome, complete genome.
85	NC_009801	532990	533025	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
86	NC_009801	647253	647288	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
87	NC_009801	647350	647315	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* E24377A chromosome, complete genome.
88	NC_009801	1212064	1212099	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
89	NC_009801	2526240	2526275	CCCTCTCCCCTTGCGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
90	NC_009801	2812474	2812439	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
91	NC_009801	2983331	2983296	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
92	NC_009801	3249330	3249365	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* E24377A chromosome, complete genome.
93	NC_009801	3249428	3249393	CCCTCCCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
94	NC_009801	3532844	3532879	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
95	NC_009801	3819600	3819565	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
96	NC_009801	3827299	3827264	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
97	NC_009801	3926496	3926531	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
98	NC_009801	3987972	3987937	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* E24377A chromosome, complete genome.
99	NC_009801	3988073	3988108	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
100	NC_009801	4170416	4170381	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
101	NC_009801	4281333	4281300	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* E24377A chromosome, complete genome.
102	NC_009801	4395515	4395550	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
103	NC_009801	4422243	4422208	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
104	NC_009801	4422296	4422331	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* E24377A chromosome, complete genome.
105	NC_009801	4646171	4646206	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
106	NC_009801	4655133	4655166	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* E24377A chromosome, complete genome.
107	NC_009801	4655217	4655183	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* E24377A chromosome, complete genome.
108	NC_009801	4959919	4959886	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* E24377A chromosome, complete genome.
109	NC_010468	390037	390072	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
110	NC_010468	505168	505203	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
111	NC_010468	656220	656256	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
112	NC_010468	656319	656284	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* ATCC 8739 chromosome, complete genome.
113	NC_010468	866657	866692	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
114	NC_010468	2575392	2575427	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
115	NC_010468	3332560	3332595	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* ATCC 8739 chromosome, complete genome.
116	NC_010468	3332657	3332622	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
117	NC_010468	3443717	3443682	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
118	NC_010468	3569813	3569778	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
119	NC_010468	3694841	3694806	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
120	NC_010468	3694882	3694915	CCCTCTCCCTGTGGGAGAGGGTTGGGGTGAGGGC	*Escherichia coli* ATCC 8739 chromosome, complete genome.
121	NC_010468	3835162	3835196	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* ATCC 8739 chromosome, complete genome.
122	NC_010468	3835246	3835213	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* ATCC 8739 chromosome, complete genome.
123	NC_010468	4021997	4022030	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* ATCC 8739 chromosome, complete genome.
124	NC_010468	4095316	4095351	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
125	NC_010468	4102216	4102181	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
126	NC_010468	4319676	4319641	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
127	NC_010468	4670493	4670526	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* ATCC 8739 chromosome, complete genome.
128	NC_010468	4721492	4721457	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ATCC 8739 chromosome, complete genome.
129	NC_010473	138651	138684	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
130	NC_010473	138735	138701	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
131	NC_010473	255429	255464	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
132	NC_010473	264020	264053	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
133	NC_010473	327219	327254	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
134	NC_010473	436398	436433	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
135	NC_010473	563134	563169	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
136	NC_010473	563231	563196	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
137	NC_010473	676394	676429	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
138	NC_010473	676491	676456	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
139	NC_010473	1438659	1438624	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
140	NC_010473	3137881	3137916	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
141	NC_010473	3137979	3137944	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
142	NC_010473	3335366	3335401	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
143	NC_010473	3335465	3335429	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
144	NC_010473	3486296	3486261	CCCTCGCCCCTTTGGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
145	NC_010473	3600659	3600624	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
146	NC_010473	3704452	3704487	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
147	NC_010473	3876661	3876628	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
148	NC_010473	3876721	3876756	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
149	NC_010473	4004102	4004137	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
150	NC_010473	4056434	4056401	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
151	NC_010473	4168172	4168207	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGAGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
152	NC_010473	4412670	4412705	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. DH10B chromosome, complete genome.
153	NC_010498	261597	261633	CCCTCGCCCCATCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SMS-3-5 chromosome, complete genome.
154	NC_010498	272864	272831	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* SMS-3-5 chromosome, complete genome.
155	NC_010498	272905	272940	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGGGAGGGG	*Escherichia coli* SMS-3-5 chromosome, complete genome.
156	NC_010498	695776	695809	CCCTCGCCCCTCGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* SMS-3-5 chromosome, complete genome.
157	NC_010498	2652608	2652573	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SMS-3-5 chromosome, complete genome.
158	NC_010498	4815701	4815735	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SMS-3-5 chromosome, complete genome.
159	NC_010498	4843415	4843450	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SMS-3-5 chromosome, complete genome.
160	NC_011353	168850	168883	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
161	NC_011353	168934	168900	CCCTCTCCCTTAAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
162	NC_011353	716831	716796	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
163	NC_011353	1437327	1437362	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
164	NC_011353	3000804	3000770	CCCTCGCCCCTTCGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
165	NC_011353	3487173	3487138	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete

TM No	Code	Start	End	Sequence (5'-3')	Strain
					genome.
166	NC_011353	4082314	4082349	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
167	NC_011353	4082413	4082377	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
168	NC_011353	4227530	4227495	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
169	NC_011353	4327888	4327853	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
170	NC_011353	4798298	4798333	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. EC4115 chromosome, complete genome.
171	NC_011415	68408	68373	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
172	NC_011415	165911	165944	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* SE11 chromosome, complete genome.
173	NC_011415	165995	165961	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* SE11 chromosome, complete genome.
174	NC_011415	431716	431751	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
175	NC_011415	541673	541708	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
176	NC_011415	717092	717057	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
177	NC_011415	1217464	1217499	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
178	NC_011415	2592880	2592915	CCCTCTCCCCTTGCGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
179	NC_011415	2921462	2921427	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
180	NC_011415	3291534	3291569	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* SE11 chromosome, complete genome.
181	NC_011415	3291632	3291597	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
182	NC_011415	3465928	3465963	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
183	NC_011415	3653589	3653554	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
184	NC_011415	3752687	3752652	CCCTCGCCCCTATGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
185	NC_011415	3760386	3760351	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
186	NC_011415	3921304	3921269	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* SE11 chromosome, complete genome.
187	NC_011415	3921405	3921440	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
188	NC_011415	4102670	4102635	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
189	NC_011415	4213585	4213552	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* SE11 chromosome, complete genome.
190	NC_011415	4330638	4330673	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
191	NC_011415	4356069	4356034	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
192	NC_011415	4356122	4356157	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* SE11 chromosome, complete genome.
193	NC_011415	4585567	4585602	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
194	NC_011415	4763726	4763759	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* SE11 chromosome, complete genome.
195	NC_011415	4798844	4798809	CCCTCGCCCCTCCGGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE11 chromosome, complete genome.
196	NC_011415	4867867	4867834	CCCTCTCCCTGAGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* SE11 chromosome, complete genome.
197	NC_011601	4844357	4844392	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O127:H6 str. E2348/69 chromosome, complete genome.
198	NC_011741	67131	67096	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
199	NC_011741	164790	164823	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* IAI1 chromosome, complete genome.
200	NC_011741	164874	164840	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* IAI1 chromosome, complete genome.
201	NC_011741	401280	401315	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
202	NC_011741	627412	627447	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
203	NC_011741	627509	627474	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* IAI1 chromosome, complete genome.
204	NC_011741	1176418	1176453	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
205	NC_011741	2398328	2398363	CCCTCTCCCCTTGCGGGAGAGGGTAAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
206	NC_011741	2686262	2686227	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
207	NC_011741	2839043	2839010	CCCTCTCCCTTAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* IAI1 chromosome, complete genome.
208	NC_011741	2867445	2867410	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
209	NC_011741	3104063	3104098	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* IAI1 chromosome, complete genome.
210	NC_011741	3104161	3104126	CCCTCCCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
211	NC_011741	3284648	3284683	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
212	NC_011741	3320622	3320657	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* IAI1 chromosome, complete genome.
213	NC_011741	3320721	3320685	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
214	NC_011741	3570845	3570810	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
215	NC_011741	3578544	3578509	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
216	NC_011741	3740147	3740112	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* IAI1 chromosome, complete genome.
217	NC_011741	3740248	3740283	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
218	NC_011741	3924941	3924906	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
219	NC_011741	4035851	4035818	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* IAI1 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
220	NC_011741	4150212	4150247	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
221	NC_011741	4175674	4175639	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
222	NC_011741	4175727	4175762	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* IAI1 chromosome, complete genome.
223	NC_011741	4398197	4398232	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
224	NC_011741	4568675	4568708	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* IAI1 chromosome, complete genome.
225	NC_011741	4572746	4572781	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI1 chromosome, complete genome.
226	NC_011741	4679581	4679548	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* IAI1 chromosome, complete genome.
227	NC_011742	551430	551465	CCCTCGCCCCTGTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* S88 chromosome, complete genome.
228	NC_011742	4817879	4817914	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* S88 chromosome, complete genome.
229	NC_011748	67144	67109	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
230	NC_011748	164648	164681	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* 55989 chromosome, complete genome.
231	NC_011748	164732	164698	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* 55989 chromosome, complete genome.
232	NC_011748	414383	414418	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
233	NC_011748	525154	525189	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
234	NC_011748	640142	640177	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
235	NC_011748	640239	640204	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* 55989 chromosome, complete genome.
236	NC_011748	2553406	2553441	CCCTCTCCCCTTGCGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
237	NC_011748	3004823	3004790	CCCTCTCCCTTAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* 55989 chromosome, complete genome.
238	NC_011748	3033667	3033632	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
239	NC_011748	3271109	3271144	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* 55989 chromosome, complete genome.
240	NC_011748	3551911	3551946	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
241	NC_011748	3587888	3587923	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* 55989 chromosome, complete genome.
242	NC_011748	3587987	3587951	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
243	NC_011748	3839425	3839390	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
244	NC_011748	3847124	3847089	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
245	NC_011748	4041270	4041235	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* 55989 chromosome, complete genome.
246	NC_011748	4041371	4041406	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
247	NC_011748	4228816	4228781	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
248	NC_011748	4339786	4339753	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* 55989 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
249	NC_011748	4476324	4476289	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
250	NC_011748	4476377	4476412	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* 55989 chromosome, complete genome.
251	NC_011748	4698973	4699008	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
252	NC_011748	4702663	4702628	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
253	NC_011748	4702704	4702737	CCCTCTCCCTCCGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
254	NC_011748	4897791	4897826	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
255	NC_011748	4932212	4932245	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* 55989 chromosome, complete genome.
256	NC_011748	4936283	4936318	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 55989 chromosome, complete genome.
257	NC_011748	5134004	5133971	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* 55989 chromosome, complete genome.
258	NC_011750	82649	82614	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI39 chromosome, complete genome.
259	NC_011750	665251	665284	CCCTCGCCCCTCGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI39 chromosome, complete genome.
260	NC_011750	4726243	4726208	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI39 chromosome, complete genome.
261	NC_011750	4726283	4726318	CCCTCTCCCCTCTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* IAI39 chromosome, complete genome.
262	NC_011750	4942649	4942684	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI39 chromosome, complete genome.
263	NC_011750	4997343	4997308	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IAI39 chromosome, complete genome.
264	NC_011993	604718	604683	CCCTCGCCCCACCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LF82, complete genome.
265	NC_011993	2553920	2553885	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LF82, complete genome.
266	NC_012759	164546	164579	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* BW2952 chromosome, complete genome.
267	NC_012759	164630	164596	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* BW2952 chromosome, complete genome.
268	NC_012759	290647	290682	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
269	NC_012759	399826	399861	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
270	NC_012759	526562	526597	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
271	NC_012759	526659	526624	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* BW2952 chromosome, complete genome.
272	NC_012759	1240112	1240077	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
273	NC_012759	2931159	2931194	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* BW2952 chromosome, complete genome.
274	NC_012759	2931257	2931222	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
275	NC_012759	3124769	3124804	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* BW2952 chromosome, complete genome.
276	NC_012759	3124868	3124832	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
277	NC_012759	3275699	3275664	CCCTCGCCCCTTTGGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
278	NC_012759	3390071	3390036	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
279	NC_012759	3495201	3495236	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
280	NC_012759	3667410	3667377	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
281	NC_012759	3667470	3667505	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
282	NC_012759	3794851	3794886	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
283	NC_012759	3847183	3847150	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* BW2952 chromosome, complete genome.
284	NC_012759	3958144	3958179	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGAGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
285	NC_012759	4251043	4251078	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
286	NC_012759	4486391	4486356	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BW2952 chromosome, complete genome.
287	NC_012947	301804	301769	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
288	NC_012947	409311	409346	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
289	NC_012947	704545	704510	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
290	NC_012947	862958	862993	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
291	NC_012947	863056	863021	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
292	NC_012947	1105994	1105959	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
293	NC_012947	1722277	1722243	CCCTCGCCCCCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
294	NC_012947	2444776	2444811	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
295	NC_012947	3189821	3189856	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
296	NC_012947	3197221	3197256	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
297	NC_012947	3197318	3197283	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
298	NC_012947	3315495	3315460	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
299	NC_012947	3427833	3427798	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
300	NC_012947	3616239	3616273	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
301	NC_012947	3616323	3616290	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
302	NC_012947	3803422	3803455	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
303	NC_012947	3883311	3883346	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
304	NC_012947	4121001	4120966	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
305	NC_012947	4365433	4365398	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
306	NC_012947	4494077	4494110	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
307	NC_012947	4546513	4546478	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 'BL21-Gold(DE3)pLysS AG' chromosome, complete genome.
308	NC_012967	167398	167431	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* B str. REL606 chromosome, complete genome.
309	NC_012967	167482	167448	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* B str. REL606 chromosome, complete genome.
310	NC_012967	357572	357607	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
311	NC_012967	469910	469945	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
312	NC_012967	607287	607322	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
313	NC_012967	607384	607349	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* B str. REL606 chromosome, complete genome.
314	NC_012967	614784	614749	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
315	NC_012967	1349608	1349573	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
316	NC_012967	2051730	2051764	CCCTCGCCCCCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
317	NC_012967	2688803	2688838	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
318	NC_012967	2931741	2931776	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* B str. REL606 chromosome, complete genome.
319	NC_012967	2931839	2931804	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
320	NC_012967	3135994	3136029	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
321	NC_012967	3432797	3432762	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
322	NC_012967	3868862	3868897	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
323	NC_012967	3921293	3921260	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* B str. REL606 chromosome, complete genome.
324	NC_012967	4048599	4048634	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
325	NC_012967	4293107	4293142	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
326	NC_012967	4530797	4530762	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* B str. REL606 chromosome, complete genome.
327	NC_012967	4610112	4610079	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* B str. REL606 chromosome, complete genome.
328	NC_012971	167399	167432	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* BL21(DE3) chromosome, complete genome.
329	NC_012971	167483	167449	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* BL21(DE3) chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
330	NC_012971	354551	354586	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
331	NC_012971	466889	466924	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
332	NC_012971	585066	585101	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
333	NC_012971	585163	585128	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* BL21(DE3) chromosome, complete genome.
334	NC_012971	592563	592528	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
335	NC_012971	1337606	1337571	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
336	NC_012971	2017531	2017565	CCCTCGCCCCCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
337	NC_012971	2633910	2633945	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
338	NC_012971	2876848	2876883	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* BL21(DE3) chromosome, complete genome.
339	NC_012971	2876946	2876911	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
340	NC_012971	3069259	3069294	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
341	NC_012971	3364493	3364458	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
342	NC_012971	3797297	3797332	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
343	NC_012971	3849733	3849700	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* BL21(DE3) chromosome, complete genome.
344	NC_012971	3977039	3977074	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
345	NC_012971	4221472	4221507	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
346	NC_012971	4459162	4459127	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* BL21(DE3) chromosome, complete genome.
347	NC_012971	4539253	4539220	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* BL21(DE3) chromosome, complete genome.
348	NC_013008	168850	168883	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
349	NC_013008	168934	168900	CCCTCTCCCTTAAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
350	NC_013008	718424	718389	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
351	NC_013008	1437614	1437649	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
352	NC_013008	2955675	2955641	CCCTCGCCCCTTCGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
353	NC_013008	3442044	3442009	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
354	NC_013008	4038355	4038390	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
355	NC_013008	4038454	4038418	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
356	NC_013008	4183571	4183536	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
357	NC_013008	4283929	4283894	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
358	NC_013008	4754341	4754376	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O157:H7 str. TW14359 chromosome, complete genome.
359	NC_013353	165945	165978	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O103:H2 str. 12009, complete genome.
360	NC_013353	166029	165995	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O103:H2 str. 12009, complete genome.
361	NC_013353	396274	396309	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
362	NC_013353	506432	506467	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
363	NC_013353	662371	662406	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
364	NC_013353	662468	662433	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* O103:H2 str. 12009, complete genome.
365	NC_013353	669878	669843	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
366	NC_013353	1219342	1219377	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
367	NC_013353	3285469	3285434	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
368	NC_013353	3554582	3554617	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O103:H2 str. 12009, complete genome.
369	NC_013353	3814220	3814255	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
370	NC_013353	3914792	3914827	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O103:H2 str. 12009, complete genome.
371	NC_013353	3914891	3914855	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
372	NC_013353	4174258	4174223	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
373	NC_013353	4343372	4343337	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* O103:H2 str. 12009, complete genome.
374	NC_013353	4343473	4343508	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
375	NC_013353	4389367	4389332	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
376	NC_013353	4503566	4503599	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* O103:H2 str. 12009, complete genome.
377	NC_013353	4610318	4610353	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
378	NC_013353	4754789	4754754	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O103:H2 str. 12009, complete genome.
379	NC_013353	5025374	5025409	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
380	NC_013353	5249621	5249654	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O103:H2 str. 12009, complete genome.
381	NC_013353	5253692	5253727	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
382	NC_013353	5304759	5304724	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O103:H2 str. 12009, complete genome.
383	NC_013353	5429614	5429581	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O103:H2 str. 12009, complete genome.
384	NC_013361	68495	68460	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
385	NC_013361	288479	288446	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
386	NC_013361	288520	288555	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
387	NC_013361	433953	433988	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
388	NC_013361	546651	546686	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
389	NC_013361	710708	710673	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
390	NC_013361	1423778	1423813	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
391	NC_013361	2875141	2875106	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
392	NC_013361	3151646	3151681	CCCTCTCCCCTTGCGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
393	NC_013361	3511609	3511574	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
394	NC_013361	3672574	3672541	CCCTCTCCCTTAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
395	NC_013361	3701151	3701116	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGGGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
396	NC_013361	3950381	3950416	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
397	NC_013361	4124972	4125007	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
398	NC_013361	4160945	4160980	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
399	NC_013361	4161044	4161008	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
400	NC_013361	4422993	4422958	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
401	NC_013361	4430692	4430657	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
402	NC_013361	4529160	4529195	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
403	NC_013361	4596931	4596896	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
404	NC_013361	4597032	4597067	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
405	NC_013361	4708379	4708344	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
406	NC_013361	4824010	4824043	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
407	NC_013361	4934980	4935015	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
408	NC_013361	5019544	5019577	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
409	NC_013361	5281085	5281120	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
410	NC_013361	5521062	5521097	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
411	NC_013361	5598398	5598363	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
412	NC_013361	5677540	5677507	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O26:H11 str. 11368 chromosome, complete genome.
413	NC_013364	68480	68445	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
414	NC_013364	297319	297286	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* O111:H- str. 11128, complete genome.
415	NC_013364	297360	297395	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
416	NC_013364	430450	430485	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
417	NC_013364	542324	542359	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
418	NC_013364	671002	670967	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* O111:H- str. 11128, complete genome.
419	NC_013364	1386866	1386901	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
420	NC_013364	2930630	2930665	CCCTCTCCCCTTGCGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
421	NC_013364	3223253	3223218	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
422	NC_013364	3357748	3357715	CCCTCTCCCTTAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O111:H- str. 11128, complete genome.
423	NC_013364	3386324	3386289	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
424	NC_013364	3626033	3626068	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O111:H- str. 11128, complete genome.
425	NC_013364	3899758	3899793	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O111:H- str. 11128, complete genome.
426	NC_013364	3899857	3899821	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
427	NC_013364	4160635	4160600	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
428	NC_013364	4168335	4168300	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
429	NC_013364	4266801	4266836	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
430	NC_013364	4335651	4335616	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* O111:H- str. 11128, complete genome.
431	NC_013364	4335752	4335787	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
432	NC_013364	4451436	4451403	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
433	NC_013364	4521716	4521681	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
434	NC_013364	4635821	4635788	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* O111:H- str. 11128, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
435	NC_013364	4750154	4750189	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
436	NC_013364	5046894	5046929	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
437	NC_013364	5237955	5237988	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O111:H- str. 11128, complete genome.
438	NC_013364	5242025	5242060	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
439	NC_013364	5281351	5281316	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O111:H- str. 11128, complete genome.
440	NC_013364	5351377	5351344	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O111:H- str. 11128, complete genome.
441	NC_013654	2294215	2294250	CCCTCTCCCCTTGCGGGAGAGGGGACGGGTGAGGGG	*Escherichia coli* SE15, complete genome.
442	NC_013654	4489452	4489487	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* SE15, complete genome.
443	NC_013941	168817	168850	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
444	NC_013941	168901	168867	CCCTCTCCCTTAAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
445	NC_013941	802984	802949	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
446	NC_013941	1392176	1392211	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
447	NC_013941	2700196	2700162	CCCTCGCCCCTTCGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
448	NC_013941	3124121	3124086	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
449	NC_013941	3868985	3869020	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
450	NC_013941	3869084	3869048	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
451	NC_013941	4113536	4113501	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
452	NC_013941	4582701	4582736	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. CB9615 chromosome, complete genome.
453	NC_016902	30121	30156	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
454	NC_016902	234031	233996	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
455	NC_016902	234132	234167	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* KO11FL chromosome, complete genome.
456	NC_016902	403317	403352	CCCTCGCCCCTATGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
457	NC_016902	502585	502620	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
458	NC_016902	654109	654145	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
459	NC_016902	654208	654173	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* KO11FL chromosome, complete genome.
460	NC_016902	690184	690149	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
461	NC_016902	863241	863276	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
462	NC_016902	863339	863304	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* KO11FL chromosome, complete genome.
463	NC_016902	1134739	1134774	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
464	NC_016902	3408939	3408974	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
465	NC_016902	3416349	3416384	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* KO11FL chromosome, complete genome.
466	NC_016902	3416446	3416411	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
467	NC_016902	3535814	3535779	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
468	NC_016902	3645957	3645922	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
469	NC_016902	3772575	3772608	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* KO11FL chromosome, complete genome.
470	NC_016902	3779737	3779702	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
471	NC_016902	3779778	3779811	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* KO11FL chromosome, complete genome.
472	NC_016902	3945916	3945950	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* KO11FL chromosome, complete genome.
473	NC_016902	3946000	3945967	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* KO11FL chromosome, complete genome.
474	NC_016902	4043491	4043526	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
475	NC_016902	4131550	4131583	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* KO11FL chromosome, complete genome.
476	NC_016902	4266380	4266345	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
477	NC_016902	4270451	4270418	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* KO11FL chromosome, complete genome.
478	NC_016902	4446313	4446278	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
479	NC_016902	4706123	4706088	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* KO11FL chromosome, complete genome.
480	NC_016902	4706176	4706211	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
481	NC_016902	4731626	4731591	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* KO11FL chromosome, complete genome.
482	NC_016902	4842319	4842352	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* KO11FL chromosome, complete genome.
483	NC_017625	102642	102607	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
484	NC_017625	102702	102735	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
485	NC_017625	275079	275044	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
486	NC_017625	378872	378907	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
487	NC_017625	493235	493270	CCCTCGCCCCTTTGGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
488	NC_017625	644066	644102	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
489	NC_017625	644165	644130	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* DH1 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
490	NC_017625	837676	837711	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
491	NC_017625	837774	837739	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* DH1 chromosome, complete genome.
492	NC_017625	2528438	2528473	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
493	NC_017625	3256443	3256478	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* DH1 chromosome, complete genome.
494	NC_017625	3256540	3256505	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
495	NC_017625	3373112	3373077	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
496	NC_017625	3482291	3482256	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
497	NC_017625	3582908	3582875	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* DH1 chromosome, complete genome.
498	NC_017625	3590053	3590018	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
499	NC_017625	3706746	3706780	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* DH1 chromosome, complete genome.
500	NC_017625	3706830	3706797	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* DH1 chromosome, complete genome.
501	NC_017625	3963146	3963181	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
502	NC_017625	4200296	4200261	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
503	NC_017625	4444018	4443983	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGAGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
504	NC_017625	4554979	4555012	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* DH1 chromosome, complete genome.
505	NC_017625	4605974	4605939	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* DH1 chromosome, complete genome.
506	NC_017626	73657	73622	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* 042, complete genome.
507	NC_017626	300859	300826	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* 042, complete genome.
508	NC_017626	300900	300935	CCCTCGCCCCTTTGGGGAGAGGGTTGGGGTGAGGGG	*Escherichia coli* 042, complete genome.
509	NC_017626	308095	308062	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* 042, complete genome.
510	NC_017626	308137	308171	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 042, complete genome.
511	NC_017626	1486849	1486814	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 042, complete genome.
512	NC_017626	1487021	1487056	CCCTCGCCCTTTCAGGGAGAGGGTTGGGGTGAGGGT	*Escherichia coli* 042, complete genome.
513	NC_017626	2824724	2824689	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 042, complete genome.
514	NC_017626	4612757	4612792	CCCTCTCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* 042, complete genome.
515	NC_017626	4789194	4789159	CCCTCTCCCCAGAGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* 042, complete genome.
516	NC_017626	4789172	4789207	CCCTCTCCCCTCTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* 042, complete genome.
517	NC_017626	4794472	4794505	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* 042, complete genome.
518	NC_017626	5222279	5222246	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* 042, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
519	NC_017628	551330	551365	CCCTCGCCCCTGTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* IHE3034 chromosome, complete genome.
520	NC_017631	2823098	2823063	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ABU 83972 chromosome, complete genome.
521	NC_017631	5045295	5045260	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ABU 83972 chromosome, complete genome.
522	NC_017632	3053024	3052989	CCCTCGCCCCTGTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UM146 chromosome, complete genome.
523	NC_017633	167865	167898	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* ETEC H10407, complete genome.
524	NC_017633	167949	167915	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* ETEC H10407, complete genome.
525	NC_017633	449528	449563	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
526	NC_017633	558837	558872	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
527	NC_017633	678209	678244	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
528	NC_017633	678306	678271	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* ETEC H10407, complete genome.
529	NC_017633	1489759	1489724	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
530	NC_017633	3348787	3348822	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* ETEC H10407, complete genome.
531	NC_017633	3348885	3348850	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
532	NC_017633	3630286	3630321	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* ETEC H10407, complete genome.
533	NC_017633	3630385	3630349	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
534	NC_017633	3777005	3776970	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
535	NC_017633	3891366	3891331	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
536	NC_017633	3994376	3994411	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
537	NC_017633	4162350	4162317	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
538	NC_017633	4162410	4162445	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
539	NC_017633	4330666	4330701	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
540	NC_017633	4381664	4381631	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* ETEC H10407, complete genome.
541	NC_017633	4776314	4776349	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
542	NC_017633	5000256	5000291	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* ETEC H10407, complete genome.
543	NC_017634	600518	600483	CCCTCGCCCCACCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O83:H1 str. NRG 857C chromosome, complete genome.
544	NC_017634	2561759	2561724	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O83:H1 str. NRG 857C chromosome, complete genome.
545	NC_017641	167379	167412	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* UMNK88 chromosome, complete genome.
546	NC_017641	167463	167429	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* UMNK88 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
547	NC_017641	533844	533879	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
548	NC_017641	3479510	3479545	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* UMNK88 chromosome, complete genome.
549	NC_017641	3479608	3479573	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
550	NC_017641	3877905	3877870	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
551	NC_017641	3992287	3992252	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
552	NC_017641	4098507	4098542	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
553	NC_017641	4271010	4270977	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
554	NC_017641	4271070	4271105	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
555	NC_017641	4411420	4411455	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
556	NC_017641	4462418	4462385	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* UMNK88 chromosome, complete genome.
557	NC_017641	4755321	4755356	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNK88 chromosome, complete genome.
558	NC_017644	2363016	2363051	CCCTCTCCCCTTGCGGGAGAGGGGACGGGTGAGGGG	*Escherichia coli* NA114 chromosome, complete genome.
559	NC_017646	81221	81186	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O7:K1 str. CE10 chromosome, complete genome.
560	NC_017646	706015	706048	CCCTCGCCCCTCGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* O7:K1 str. CE10 chromosome, complete genome.
561	NC_017646	4917716	4917681	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O7:K1 str. CE10 chromosome, complete genome.
562	NC_017646	4917756	4917791	CCCTCTCCCCTCTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O7:K1 str. CE10 chromosome, complete genome.
563	NC_017646	5119026	5119061	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O7:K1 str. CE10 chromosome, complete genome.
564	NC_017646	5182855	5182820	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O7:K1 str. CE10 chromosome, complete genome.
565	NC_017651	2789532	2789497	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. 'clone D i2' chromosome, complete genome.
566	NC_017651	4952285	4952250	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. 'clone D i2' chromosome, complete genome.
567	NC_017652	2789532	2789497	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. 'clone D i14' chromosome, complete genome.
568	NC_017652	4952285	4952250	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. 'clone D i14' chromosome, complete genome.
569	NC_017656	168817	168850	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
570	NC_017656	168901	168867	CCCTCTCCCTTAAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
571	NC_017656	797219	797184	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
572	NC_017656	1385750	1385785	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
573	NC_017656	2710349	2710315	CCCTCGCCCCTTCGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
574	NC_017656	3134358	3134323	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
575	NC_017656	3777468	3777503	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
576	NC_017656	3777567	3777531	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
577	NC_017656	4022007	4021972	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
578	NC_017656	4491146	4491181	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O55:H7 str. RM12579 chromosome, complete genome.
579	NC_017663	158502	158535	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* P12b chromosome, complete genome.
580	NC_017663	158586	158552	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* P12b chromosome, complete genome.
581	NC_017663	263073	263040	CCCTCTCCCTGTGGGAGAGGGTTGGGGTGAGGGC	*Escherichia coli* P12b chromosome, complete genome.
582	NC_017663	263114	263149	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
583	NC_017663	408818	408853	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
584	NC_017663	517978	518013	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
585	NC_017663	622702	622737	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
586	NC_017663	622799	622764	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* P12b chromosome, complete genome.
587	NC_017663	1966903	1966938	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
588	NC_017663	3278814	3278849	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* P12b chromosome, complete genome.
589	NC_017663	3278912	3278877	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
590	NC_017663	3510807	3510842	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* P12b chromosome, complete genome.
591	NC_017663	3510906	3510870	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
592	NC_017663	3661980	3661945	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
593	NC_017663	3772547	3772512	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
594	NC_017663	4268881	4268848	CCCTCTCCCTGTGGGTGAGGGTCGGGGTGAGGGC	*Escherichia coli* P12b chromosome, complete genome.
595	NC_017663	4628612	4628647	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* P12b chromosome, complete genome.
596	NC_017663	4914647	4914614	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* P12b chromosome, complete genome.
597	NC_017664	68411	68376	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
598	NC_017664	165903	165936	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* W chromosome, complete genome.
599	NC_017664	165987	165953	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* W chromosome, complete genome.
600	NC_017664	332126	332093	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* W chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
601	NC_017664	332167	332202	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
602	NC_017664	339329	339296	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* W chromosome, complete genome.
603	NC_017664	465947	465982	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
604	NC_017664	576090	576125	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
605	NC_017664	695458	695493	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
606	NC_017664	695555	695520	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* W chromosome, complete genome.
607	NC_017664	702965	702930	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
608	NC_017664	2967695	2967660	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
609	NC_017664	3239317	3239352	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* W chromosome, complete genome.
610	NC_017664	3239415	3239380	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
611	NC_017664	3412473	3412508	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
612	NC_017664	3448449	3448484	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* W chromosome, complete genome.
613	NC_017664	3448548	3448512	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
614	NC_017664	3600072	3600037	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
615	NC_017664	3699249	3699214	CCCTCGCCCCTATGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
616	NC_017664	3868434	3868399	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* W chromosome, complete genome.
617	NC_017664	3868535	3868570	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
618	NC_017664	4072433	4072398	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
619	NC_017664	4180405	4180372	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* W chromosome, complete genome.
620	NC_017664	4291098	4291133	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
621	NC_017664	4316548	4316513	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
622	NC_017664	4316601	4316636	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* W chromosome, complete genome.
623	NC_017664	4569259	4569294	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
624	NC_017664	4738902	4738935	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* W chromosome, complete genome.
625	NC_017664	4742973	4743008	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* W chromosome, complete genome.
626	NC_017664	4877804	4877771	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* W chromosome, complete genome.
627	NC_017906	168869	168902	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* Xuzhou21 chromosome, complete genome.
628	NC_017906	168953	168919	CCCTCTCCCTTAAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* Xuzhou21 chromosome, complete genome.
629	NC_017906	714567	714532	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* Xuzhou21 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
630	NC_017906	1495678	1495713	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* Xuzhou21 chromosome, complete genome.
631	NC_017906	2887527	2887493	CCCTCGCCCCTTCGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* Xuzhou21 chromosome, complete genome.
632	NC_017906	3310693	3310658	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* Xuzhou21 chromosome, complete genome.
633	NC_017906	3906491	3906526	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* Xuzhou21 chromosome, complete genome.
634	NC_017906	3906590	3906554	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* Xuzhou21 chromosome, complete genome.
635	NC_017906	4053023	4052988	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* Xuzhou21 chromosome, complete genome.
636	NC_017906	4152068	4152033	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* Xuzhou21 chromosome, complete genome.
637	NC_017906	4622185	4622220	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* Xuzhou21 chromosome, complete genome.
638	NC_018650	30884	30919	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
639	NC_018650	264667	264632	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
640	NC_018650	264768	264803	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
641	NC_018650	460686	460721	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
642	NC_018650	468385	468420	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
643	NC_018650	719687	719723	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
644	NC_018650	719786	719751	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
645	NC_018650	755763	755728	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
646	NC_018650	1016568	1016533	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
647	NC_018650	1252976	1253011	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
648	NC_018650	1281819	1281852	CCCTCTCCCTTAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
649	NC_018650	1705043	1705008	CCCTCTCCCCTTGCGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
650	NC_018650	3793440	3793475	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
651	NC_018650	3793537	3793502	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
652	NC_018650	3909097	3909062	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
653	NC_018650	4019091	4019056	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
654	NC_018650	4263523	4263557	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
655	NC_018650	4263607	4263574	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
656	NC_018650	4361112	4361147	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
657	NC_018650	4447892	4447925	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
658	NC_018650	4585690	4585655	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
659	NC_018650	4589761	4589728	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
660	NC_018650	4624960	4624925	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
661	NC_018650	4809210	4809177	CCCTCTCCCTCCGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
662	NC_018650	4809251	4809286	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
663	NC_018650	4812941	4812906	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
664	NC_018650	5035634	5035599	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
665	NC_018650	5035687	5035722	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
666	NC_018650	5172275	5172308	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2009EL-2050 chromosome, complete genome.
667	NC_018658	30107	30142	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
668	NC_018658	262590	262555	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
669	NC_018658	262691	262726	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
670	NC_018658	456058	456093	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
671	NC_018658	463757	463792	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
672	NC_018658	715173	715209	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
673	NC_018658	715272	715237	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
674	NC_018658	751249	751214	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
675	NC_018658	1013121	1013086	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
676	NC_018658	1249515	1249550	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
677	NC_018658	1277808	1277841	CCCTCTCCCTTAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
678	NC_018658	1717779	1717744	CCCTCTCCCCTTGCGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
679	NC_018658	3801436	3801471	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
680	NC_018658	3801533	3801498	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
681	NC_018658	3918441	3918406	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
682	NC_018658	4028435	4028400	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
683	NC_018658	4285814	4285848	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
684	NC_018658	4285898	4285865	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
685	NC_018658	4383402	4383437	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
686	NC_018658	4470182	4470215	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
687	NC_018658	4607841	4607806	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
688	NC_018658	4611912	4611879	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
689	NC_018658	4646333	4646298	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
690	NC_018658	4828985	4828952	CCCTCTCCCTCCGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
691	NC_018658	4829026	4829061	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
692	NC_018658	4832716	4832681	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
693	NC_018658	5055499	5055464	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
694	NC_018658	5055552	5055587	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
695	NC_018658	5192141	5192174	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2011C-3493 chromosome, complete genome.
696	NC_018661	30864	30899	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
697	NC_018661	252563	252528	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
698	NC_018661	252664	252699	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
699	NC_018661	449359	449394	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
700	NC_018661	457058	457093	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
701	NC_018661	708697	708733	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
702	NC_018661	708796	708761	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
703	NC_018661	744773	744738	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
704	NC_018661	1006228	1006193	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
705	NC_018661	1242717	1242752	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
706	NC_018661	1271561	1271594	CCCTCTCCCTTAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
707	NC_018661	1712733	1712698	CCCTCTCCCCTTGCGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
708	NC_018661	3851063	3851098	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
709	NC_018661	3851160	3851125	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
710	NC_018661	3968069	3968034	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
711	NC_018661	4078063	4078028	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
712	NC_018661	4322494	4322528	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
713	NC_018661	4322578	4322545	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
714	NC_018661	4420083	4420118	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
715	NC_018661	4506863	4506896	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
716	NC_018661	4644662	4644627	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
717	NC_018661	4648733	4648700	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
718	NC_018661	4683931	4683896	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
719	NC_018661	4868735	4868702	CCCTCTCCCTCCGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
720	NC_018661	4868776	4868811	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
721	NC_018661	4872466	4872431	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
722	NC_018661	5095064	5095029	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
723	NC_018661	5095117	5095152	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
724	NC_018661	5231704	5231737	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* O104:H4 str. 2009EL-2071 chromosome, complete genome.
725	NC_020163	29573	29608	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
726	NC_020163	304531	304566	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
727	NC_020163	308221	308186	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
728	NC_020163	308262	308295	CCCTCTCCCTCCGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
729	NC_020163	449498	449533	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
730	NC_020163	681615	681582	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* APEC O78, complete genome.
731	NC_020163	866174	866207	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* APEC O78, complete genome.
732	NC_020163	866258	866224	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* APEC O78, complete genome.
733	NC_020163	1100399	1100434	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
734	NC_020163	1209593	1209628	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
735	NC_020163	1344373	1344408	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
736	NC_020163	1344470	1344435	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* APEC O78, complete genome.
737	NC_020163	3407207	3407172	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
738	NC_020163	3553129	3553094	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
739	NC_020163	3789945	3789980	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* APEC O78, complete genome.
740	NC_020163	3790043	3790008	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
741	NC_020163	3999158	3999193	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* APEC O78, complete genome.
742	NC_020163	3999259	3999223	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
743	NC_020163	4274039	4274004	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
744	NC_020163	4431206	4431171	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* APEC O78, complete genome.
745	NC_020163	4431307	4431342	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
746	NC_020163	4613485	4613450	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* APEC O78, complete genome.
747	NC_020163	4724490	4724457	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* APEC O78, complete genome.
748	NC_020518	159372	159405	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
749	NC_020518	159456	159422	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
750	NC_020518	306315	306350	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
751	NC_020518	405586	405621	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
752	NC_020518	498095	498130	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
753	NC_020518	498192	498157	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
754	NC_020518	1155468	1155433	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
755	NC_020518	2552647	2552682	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
756	NC_020518	2552745	2552710	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
757	NC_020518	2712932	2712967	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
758	NC_020518	2713031	2712995	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
759	NC_020518	2858381	2858346	CCCTCGCCCCTTTGGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
760	NC_020518	2956818	2956783	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
761	NC_020518	3056701	3056736	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
762	NC_020518	3210275	3210242	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
763	NC_020518	3210335	3210370	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
764	NC_020518	3337708	3337743	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
765	NC_020518	3388704	3388671	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
766	NC_020518	3499665	3499700	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGAGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
767	NC_020518	3743387	3743422	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
768	NC_020518	3925949	3925914	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* str. K-12 substr. MDS42 DNA, complete genome.
769	NC_022364	68411	68376	CCCTCGCCCCTTTGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
770	NC_022364	165903	165936	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* LY180, complete genome.
771	NC_022364	165987	165953	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* LY180, complete genome.
772	NC_022364	332126	332093	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* LY180, complete genome.
773	NC_022364	332167	332202	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
774	NC_022364	339329	339296	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* LY180, complete genome.
775	NC_022364	468623	468658	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
776	NC_022364	569672	569707	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
777	NC_022364	689040	689075	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
778	NC_022364	689137	689102	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* LY180, complete genome.
779	NC_022364	696547	696512	CCCTCGCCCCATCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
780	NC_022364	2875157	2875122	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.

TM No	Code	Start	End	Sequence (5'-3')	Strain
781	NC_022364	3148117	3148152	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* LY180, complete genome.
782	NC_022364	3148215	3148180	CCCTCACCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
783	NC_022364	3322050	3322085	CCCTCGCCCCTTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
784	NC_022364	3359363	3359398	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* LY180, complete genome.
785	NC_022364	3359462	3359426	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
786	NC_022364	3512325	3512290	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
787	NC_022364	3611502	3611467	CCCTCGCCCCTATGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
788	NC_022364	3782025	3781990	CCCTCTCCCTTTCAGGGAGAGGGCTGGGGTGAGGGT	*Escherichia coli* LY180, complete genome.
789	NC_022364	3782126	3782161	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
790	NC_022364	3992714	3992679	CCCTCGCCCCTTTGGGGAGAGGGCTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
791	NC_022364	4100686	4100653	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* LY180, complete genome.
792	NC_022364	4214054	4214089	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
793	NC_022364	4239504	4239469	CCCTCGCCCCTACGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
794	NC_022364	4239557	4239592	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* LY180, complete genome.
795	NC_022364	4505311	4505346	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
796	NC_022364	4677101	4677134	CCCTCTCCCTTAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* LY180, complete genome.
797	NC_022364	4681172	4681207	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* LY180, complete genome.
798	NC_022364	4816003	4815970	CCCTCTCCCTGAGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* LY180, complete genome.
799	NC_022370	551016	551051	CCCTCGCCCCTGTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* PMV-1 main chromosome, complete genome.
800	NC_022648	2544412	2544447	CCCTCTCCCCTTGCGGGAGAGGGGACGGGTGAGGGG	*Escherichia coli* JJ1886, complete genome.
801	NZ_AGTD01000001	167993	168026	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
802	NZ_AGTD01000001	168077	168043	CCCTCTCCCTTGAGGGAGAGGGTTAGGGTGAGGGT	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
803	NZ_AGTD01000001	368581	368616	CCCTCGCCCCCTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
804	NZ_AGTD01000001	476396	476431	CCCTCGCCCCCTCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
805	NZ_AGTD01000001	611467	611502	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
806	NZ_AGTD01000001	611564	611529	CCCTCTCCCTTCCAGGGTGAGGGCTGGGGTGAGGGT	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
807	NZ_AGTD01000001	1623396	1623361	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.

TM No	Code	Start	End	Sequence (5'-3')	Strain
808	NZ_AGTD01000001	3535527	3535562	CCCTCGCCCTTTCAGGGAGAGGGCCGGGGTGAGGGT	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
809	NZ_AGTD01000001	3535625	3535590	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
810	NZ_AGTD01000001	3786283	3786318	CCCTCTCCCTTCCAGGGAGAGGGTCGGGGTGAGGGT	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
811	NZ_AGTD01000001	3786382	3786346	CCCTCGCCCCGTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
812	NZ_AGTD01000001	3938207	3938172	CCCTCGCCCCTTTGGGGTGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
813	NZ_AGTD01000001	4052237	4052202	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
814	NZ_AGTD01000001	4155251	4155286	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
815	NZ_AGTD01000001	4326260	4326227	CCCTCTCCCTGAGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
816	NZ_AGTD01000001	4326320	4326355	CCCTCGCCCCTCCGGGGAGAGGGCCGGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
817	NZ_AGTD01000001	4448314	4448349	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
818	NZ_AGTD01000001	4502965	4502932	CCCTCTCCCTGTGGGAGAGGGTCGGGGTGAGGGC	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
819	NZ_AGTD01000001	4613834	4613869	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGAGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
820	NZ_AGTD01000001	4858979	4859014	CCCTCGCCCCTCTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
821	NZ_AGTD01000001	5075028	5075063	CCCTCGCCCCTTTGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
822	NZ_AGTD01000001	5149171	5149136	CCCTCGCCCCTCCGGGGAGAGGGTTAGGGTGAGGGG	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.
823	NZ_AGTD01000001	5219449	5219416	CCCTCTCCCTGTGGGAGAGGGCCGGGGTGAGGGC	*Escherichia coli* UMNF18 chromosome, whole genome shotgun sequence.

Table 13.5: Description of the different TM loci.

TM loci and LCB are defined for all TMs found in the 56 *E. coli* genomes.

TM No	Blocks	Strand	Start	End	Length
TM locus 1	2_56|2_56	-	1071	7254	6184
TM locus 2	3_56|3_56	-	28627	42351	13725
TM locus 3	3_56|3_56	-	28627	42351	
TM locus 4	4_56|4_56	-	3028	36830	33803
TM locus 5	277_56|5_56	+	62854	66351	3498
TM locus 6	6_56|6_56	-	12880	12923	44
TM locus 7	9_56|9_56	+	14395	14580	186
TM locus 8	9_56|9_56	-	14395	14580	
TM locus 9	9_56|9_56	-	21882	21951	70
TM locus 10	10_56|10_56	+	24842	36460	11619
TM locus 11	16_56|16_56	-	44077	44367	291
TM locus 12	18_56|41_56	-	39293	39631	339
TM locus 13	21_56|21_56	+	1	7063	7063
TM locus 14	21_56|21_56	-	11093	11257	165
TM locus 15	21_56|21_56	+	11093	11257	
TM locus 16	21_56|21_56	+	16496	16650	155
TM locus 17	21_56|21_56	-	16496	16650	
TM locus 18	113_56|28_56	+	17193	21456	4264
TM locus 19	41_56|41_56	+	1	443	443
TM locus 20	416_56|47_56	-	25196	26762	1567
TM locus 21	200_56|48_56	-	9710	11456	1747
TM locus 22	200_56|48_56	+	9710	11456	
TM locus 23	53_56|53_56	-	1448	1917	470
TM locus 24	53_56|53_56	-	15264	15346	83
TM locus 25	54_56|54_56	-	8698	8867	170
TM locus 26	62_56|62_56	-	6885	6907	23
TM locus 27	66_56|66_56	+	2654	3643	990
TM locus 28	66_56|66_56	-	2654	3643	
TM locus 29	67_56|67_56	+	13821	13928	108
TM locus 30	67_56|67_56	-	13821	13928	
TM locus 31	70_56|70_56	+	5024	5157	134
TM locus 32	81_56|81_56	+	17454	17743	290
TM locus 33	82_56|82_56	-	8456	8620	165
TM locus 34	82_56|82_56	+	8456	8620	
TM locus 35	91_56|91_56	-	10154	10429	276
TM locus 36	91_56|91_56	+	10154	10429	
TM locus 37	97_56|97_56	+	2421	2615	195
TM locus 38	99_56|99_56	+	3383	3707	325
TM locus 39	99_56|99_56	+	7732	7845	114
TM locus 40	103_56|210_56	-	13820	15654	1835
TM locus 41	103_56|210_56	+	13820	15654	
TM locus 42	125_56|125_56	+	7861	7990	130

TM No	Blocks	Strand	Start	End	Length
TM locus 43	132_56|132_56	+	1496	1707	212
TM locus 44	145_56|145_56	+	7092	7258	167
TM locus 45	154_56|154_56	-	5186	5469	284
TM locus 46	165_56|165_56	+	1	7023	7023
TM locus 47	202_56|202_56	+	4125	4211	87
TM locus 48	228_56|228_56	+	4499	16331	11833
TM locus 49	273_56|273_56	+	4337	4634	298
TM locus 50	326_56|326_56	+	1	147	147
TM locus 51	390_56|390_56	+	1	216	216
TM locus 52	78_55|78_55	-	2238	2321	84
TM locus 53	102_55|102_55	+	10988	23960	12973
TM locus 54	102_55|102_55	-	10988	23960	
TM locus 55	383_15|383_15	-	1	11411	11411
TM locus 56	383_15|383_15	+	1	11411	
TM locus 57	383_15|383_15	-	1	11411	
TM locus 58	463_11|463_11	+	1	105	105
TM locus 59	633_5|633_5	+	1236	8263	7028
TM locus 60	1241_7|1241_7	+	701	732	32
TM locus 61	1381_11|1381_11	-	0	0	0
TM locus 62	1381_11|1381_11	+	0	0	
TM locus 63	NO|NO				

Table 13.6: Sequence variability.

Sequence variability [nt] calculated for the TMs and the different random control groups. Bold line indicates the average sequence variability.

TM	Random1	Random2	Random3	Random4
6184	311	1611	0	0
13725	0	0	1910	0
33803	0	0	75	0
3498	4504	0	226	0
44	0	0	0	863
186	0	0	12973	0
70	4246	0	0	0
11619	4169	0	1330	1917
291	0	8779	16340	19128
339	0	0	0	0
7063	0	0	0	0
165	7891	0	33803	0
155	0	13444	10967	9832
4264	0	3459	0	11619
443	0	0	6674	0

TM	Random1	Random2	Random3	Random4
1567	5621	0	4289	0
1747	0	0	0	0
470	14747	0	0	0
83	0	3081	134	0
170	0	0	13725	0
23	0	0	0	0
990	0	0	0	0
108	0	0	0	0
134	0	0	0	31
290	0	0	0	0
165	0	0	0	0
276	6240	14050	6600	0
195	0	0	0	0
325	0	0	0	0
114	0	0	2707	0
1835	14278	0	0	0
130	0	10677	4685	0
212	0	0	4911	1344
167	0	33803	0	2341
284	0	0	16340	6497
7023	0	0	0	0
87	0	0	0	0
11833	3992	0	0	0
298	0	0	0	0
147	0	0	0	0
216	0	0	0	0
84	0	1453	0	0
12973	113	0	842	6145
11411	0	0	0	4911
105	0	0	0	0
7028	0	0	0	0
32	0	0	1629	0
0	0	0	0	0
2966,063	**1377,33333**	**1882,4375**	**2920**	**1346,41667**

Table 13.7: Control regions.

Control regions analyzed for genomic instability in *E. coli* MG1655. Genomic range of control regions is indicated.

Control group 1		Control group 2		Control group 3	
start	**end**	**start**	**end**	**start**	**end**
1892791	1893870	990061	991180	778538	779587
2651670	2652959	770451	771616	4588729	4589773
4190011	4191099	775855	776860	3552213	3553320
3005531	3006557	1084321	1085336	367903	369012
1690811	1691872	1957862	1958871	2231277	2232320
3586371	3587411	377759	378910	499642	500640
1388311	1389358	363084	364145	112891	113893
1290876	1291940	1196126	1197128	239791	240816
3321371	3322414	528712	529723	3605411	3606512
2559711	2560740	4560326	4561326	4521491	4522654
2628578	2629612	2316436	2317457	2917422	2918521
2494160	2495188	4487978	4488980	2122483	2123610
2620428	2621474	3771502	3772593	1100107	1101228
4228961	4230004	3278313	3279351	2733611	2734747
1108629	1109663	1613008	1614097	62736	63835
2733738	2734756	376221	377344	140820	141920
519221	520300	4312633	4313700	4541228	4542361
1338730	1339780	1664686	1665820	1123535	1124550
3498771	3500065	1491965	1493010	3015295	3016310

Table 13.8: Plasmid constructs.

Name of the construct, cloning procedure, backbone and primer sequences used for cloning are shown.

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
(1) pQE-J06-eGFP-coreG_3T	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGGGGTGGGTGGGTGGGGTATAATGCTAGCACT AGTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCCCACCCACCCACCCCCGTAAACTCGAGGTGAAG
(2) pQE-J06-eGFP-coreG_3A	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGGGGAGGGAGGGAGGGGTATAATGCTAGCACT AGTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCCCTCCCTCCCTCCCCCGTAAACTCGAGGTGAAG
(3) pQE-J06-eGFP-coreG_2T	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCTGGTGGTGGTGGTGGTATAATGCTAGCACTA GTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCCACCACCACCACCAGCCGTAAACTCGAGGTGAAG
(4) pQE-J06-eGFP-coreG_2CT	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGGGCTGGCTGGCTGGGGTATAATGCTAGCACT AGTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCCCAGCCAGCCAGCCCCGTAAACTCGAGGTGAAG G
(5) pQE-J06-eGFP-corectrl1	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGGGGTGTGTGTGTGTGGTATAATGCTAGCACTA GTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCACACACACACACCCCGTAAACTCGAGGTGAAG
(6) pQE-J06-eGFP-corectrl2	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGTGTGGCGTGGGCGGGGTATAATGCTAGCACT AGTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCCCGCCCACGCCACACCGTAAACTCGAGGTGAAG
(7) pQE-J06-eGFP-corectrl3	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCACTCACTCACTCCCGTATAATGCTAGCACTA GTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
				TTATACGGGAGTGAGTGAGTGCCGTAAACTCGAGGTGAAG
(8) pQE-J06-eGFP-stopG3T	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: GGGTGGGTGGGTGGGGCCCGAAAGGAAGCTGAGTTG Rev: ATCCTTACTTGTACAGCTCGTCCATGC
(9) pQE-J06-eGFP-stopG2CT	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: GGCTGGCTGGCTGGGGCCCGAAAGGAAGCTGAGTTG Rev: ATCCTTACTTGTACAGCTCGTCCATGC
(10) pQE-J06-eGFP-stopctrl1	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: GGGTGTGTGTGTGTGGCCCGAAAGGAAGCTGAGTTG Rev: ATCCTTACTTGTACAGCTCGTCCATGC
(11) pQE-J06-eGFP-stopctrl3	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CACTCACTCACTCCCGCCCGAAAGGAAGCTGAGTTG Rev: ATCCTTACTTGTACAGCTCGTCCATGC
(12) pQE-J06-eGFP-20G_3T	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: GGGTGGGTGGGTGGGTTTAAGAAGGAGATATACCATGGGC Rev: GTGCTAGCATTATACCTAGGACTG
(13) pQE-J06-eGFP-20G_2CT	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: GGCTGGCTGGCTGGGTTTAAGAAGGAGATATACCATGGGC Rev: GTGCTAGCATTATACCTAGGACTG
(14) pQE-J06-eGFP-20ctrl1	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: GGGTGTGTGTGTGTGTTTAAGAAGGAGATATACCATGGGC Rev: GTGCTAGCATTATACCTAGGACTG
(15) pQE-J06-eGFP-20ctrl3	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CACTCACTCACTCCCTTTAAGAAGGAGATATACCATGGGC Rev: GTGCTAGCATTATACCTAGGACTG
(16) pQE-J06-eGFP-UPG_3T	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGGCCCTTGGGTGGGTGGGTGGGGTTTACGGCTAGCTCA GTCCTAGGTATAATGCTAGCACTAGTGGACGTTTAGCTTTAAGAAGGAGAT ATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCTAGGACTGAGCTAGCCGTAAACCCCACCCACCCACCCAAGGGC CTCGAGGTGAAG
(17) pQE-J06-eGFP-UPctrl1	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGGCCCTTGGGTGTGTGTGTGTGGTTTACGGCTAGCTCAG TCCTAGGTATAATGCTAGCACTAGTGGACGTTTAGCTTTAAGAAGGAGATA TACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCTAGGACTGAGCTAGCCGTAAACCACACACACACACCCAAGGGC CTCGAGGTGAAG
(18) pQE-J06-eGFP--10/GQG$_3$T	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCTAGCTCAGTCCTAGGTATAATGGGGTGGGT GGGTGGGGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense:

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
				TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACCCCACCCACCCACCCCA TTATACCTAGGACTGAGCTAGCCGTAAACTCGAGGTGAAG
(19) pQE-J06-eGFP--10/GQctrl1	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCTAGCTCAGTCCTAGGTATAATGGGGTGTGTG TGTGTGGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACCACACACACACACCCCA TTATACCTAGGACTGAGCTAGCCGTAAACTCGAGGTGAAG
(20) pQE-J06-eGFP-coreG_3T-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCCCACCCACCCACCCGTATAATGCTAGCACTA GTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACGGGTGGGTGGGTGGGCCGTAAACTCGAGGTGAAG
(21) pQE-J06-eGFP-coreG_3A-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCCCTCCCTCCCTCCCGTATAATGCTAGCACTA GTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACGGGAGGGAGGGAGGGCCGTAAACTCGAGGTGAAG
(22) pQE-J06-eGFP-coreG_2T-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCTCCACCACCACCAGGTATAATGCTAGCACTA GTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCTGGTGGTGGTGGAGCCGTAAACTCGAGGTGAAG
(23) pQE-J06-eGFP-coreG_2CT-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGGCCAGCCAGCCAGCCGTATAATGCTAGCACT AGTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACGGCTGGCTGGCTGGCCCGTAAACTCGAGGTGAAG
(24) pQE-J06-eGFP-corectrl1-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCACACACACACACCCGTATAATGCTAGCACTA GTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACGGGTGTGTGTGTGTGCCGTAAACTCGAGGTGAAG
(25) pQE-J06-eGFP-corectrl2-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCCCGCCCACGCCACAGTATAATGCTAGCACTA GTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACTGTGGCGTGGGCGGGCCGTAAACTCGAGGTGAAG
(26) pQE-J06-eGFP-corectrl3-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGGGGAGTGAGTGAGTGGTATAATGCTAGCACT AGTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense:

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
				TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCACTCACTCACTCCCCCGTAAACTCGAGGTGAAG
(27) pQE-J06-eGFP-stopG3T-as	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CCCACCCACCCACCCGCCCGAAAGGAAGCTGAGTTG Rev: ATCCTTACTTGTACAGCTCGTCCATGC
(28) pQE-J06-eGFP-stopG2CT-as	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CCCAGCCAGCCAGCCGCCCGAAAGGAAGCTGAGTTG Rev: ATCCTTACTTGTACAGCTCGTCCATGC
(29) pQE-J06-eGFP-stopctrl1-as	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CACACACACACACCCGCCCGAAAGGAAGCTGAGTTG Rev: ATCCTTACTTGTACAGCTCGTCCATGC
(30) pQE-J06-eGFP-stopctrl3-as	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CACTCACTCACTCCCGCCCGAAAGGAAGCTGAGTTG Rev: ATCCTTACTTGTACAGCTCGTCCATGC
(31) pQE-J06-eGFP-20G_3T-as	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CCCACCCACCCACCCTTTAAGAAGGAGATATACCATGGGC Rev: GTGCTAGCATTATACCTAGGACTG
(32) pQE-J06-eGFP-20G_2CT-as	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CCCAGCCAGCCAGCCTTTAAGAAGGAGATATACCATGGGC Rev: GTGCTAGCATTATACCTAGGACTG
(33) pQE-J06-eGFP-20ctrl1-as	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CACACACACACACCCTTTAAGAAGGAGATATACCATGGGC Rev: GTGCTAGCATTATACCTAGGACTG
(34) pQE-J06-eGFP-20ctrl3-as	whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CACTCACTCACTCCCTTTAAGAAGGAGATATACCATGGGC Rev: GTGCTAGCATTATACCTAGGACTG
(35) pQE-J06-eGFP-UPG_3T-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGGCCCTTCCCACCCACCCACCCGTTTACGGCTAGCTCAG TCCTAGGTATAATGCTAGCACTAGTGGACGTTTAGCTTTAAGAAGGAGATA TACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCTAGGACTGAGCTAGCCGTAAACGGGTGGGTGGGTGGGAAGGG CCTCGAGGTGAAG
(36) pQE-J06-eGFP-UPctrl1-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGGCCCTTCACACACACACACCCGTTTACGGCTAGCTCAGT CCTAGGTATAATGCTAGCACTAGTGGACGTTTAGCTTTAAGAAGGAGATAT ACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGTCCACTAGTGCTAGCA TTATACCTAGGACTGAGCTAGCCGTAAACGGGTGTGTGTGTGTGAAGGGC CTCGAGGTGAAG
(37) pQE-J06-eGFP--10/GQG3T-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCTAGCTCAGTCCTAGGTATAATGCCCACCCAC CCACCCGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense:

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
				TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGGGTGGGTGGGTGGGC ATTATACCTAGGACTGAGCTAGCCGTAAACTCGAGGTGAAG
(38) pQE-J06-eGFP--10/GQctrl1-as	insert	*E. coli* XL10 gold	pQE-J06-eGFP	Sense: CTTCACCTCGAGTTTACGGCTAGCTCAGTCCTAGGTATAATGCACACACAC ACACCCGTTTAGCTTTAAGAAGGAGATATACCATGGGCCA Antisense: TGGCCCATGGTATATCTCCTTCTTAAAGCTAAACGGGTGTGTGTGTGTGCA TTATACCTAGGACTGAGCTAGCCGTAAACTCGAGGTGAAG
(39) pQE-J06-eGFP-oxyR	Whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: GGCGATGGAGGATGGATAATGGGCCATCATCATCATC Rev: AGCTAAACGTCCACTAGTG
(40) pQE-J06-eGFP-oxyRctrl1	Whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CGTCGATGGAGGATTGATAATGGGCCATCATCATCATC Rev: AGCTAAACGTCCACTAGTG
(41) pQE-J06-eGFP-oxyRctrl2	Whole plasmid PCR	*E. coli* XL10 gold	pQE-J06-eGFP	Fw: CTTCGATGGAGGATGGATAATGGGCCATCATCATCATC Rev: AGCTAAACGTCCACTAGTG
(43) pBAD-eGFP-coreG_3T	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGGGTGGGTGGGTCTACTGTTTCTCCATACCC Rev: GCGTCAGGTAGGATCCGCTAATCTTATG
(44) pBAD-eGFP-coreG_3Tctrl1	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGTGTGTGTGTGTCTACTGTTTCTCCATACCC Rev: GCGTCAGGTAGGATCCGCTAATCTTATG
(45) pBAD-eGFP-coreG_3Tctrl2	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGAGTGAGTGAGTCTACTGTTTCTCCATACCC Rev: GCGTCAGGTAGGATCCGCTAATCTTATG
(46) pBAD-eGFP-20G_3T	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGGGTGGGTGGGAGGAGGAATTAACCATGGGC Rev: GGGTATGGAGAAACAGTAGAGAG
(47) pBAD-eGFP-20ctrl1	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGTGTGTGTGTGAGGAGGAATTAACCATGGGC Rev: GGGTATGGAGAAACAGTAGAGAG
(48) pBAD-eGFP-20ctrl2	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGAGTGAGTGAGAGGAGGAATTAACCATGGGC Rev: GGGTATGGAGAAACAGTAGAGAG
(49) pBAD-eGFP-stopG_3T	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGGGTGGGTGGGCAAAGCCCGAAAGGAAGCTGAG Rev: TTACTTGTACAGCTCGTCCATGCC
(50) pBAD-eGFP-stopctrl1	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGTGTGTGTGTGCAAAGCCCGAAAGGAAGCTGAG Rev: TTACTTGTACAGCTCGTCCATGCC
(51) pBAD-eGFP-stopctrl2	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGAGTGAGTGAGCAAAGCCCGAAAGGAAGCTGAG Rev: TTACTTGTACAGCTCGTCCATGCC

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
(43) pBAD-eGFP-coreG_3T-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CCCACCCACCCACCCTCTACTGTTTCTCCATACCC Rev: GCGTCAGGTAGGATCCGCTAATCTTATG
(44) pBAD-eGFP-coreG_3Tctrl1-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CACACACACACACCCTCTACTGTTTCTCCATACCC Rev: GCGTCAGGTAGGATCCGCTAATCTTATG
(45) pBAD-eGFP-coreG_3Tctrl2-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CTCACTCACTCACCCTCTACTGTTTCTCCATACCC Rev: GCGTCAGGTAGGATCCGCTAATCTTATG
(46) pBAD-eGFP-20G_3T-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CCCACCCACCCACCCAGGAGGAATTAACCATGGGC Rev: GGGTATGGAGAAACAGTAGAGAG
(47) pBAD-eGFP-20ctrl1-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CACACACACACACCCAGGAGGAATTAACCATGGGC Rev: GGGTATGGAGAAACAGTAGAGAG
(48) pBAD-eGFP-20ctrl2-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CTCACTCACTCACCCAGGAGGAATTAACCATGGGC Rev: GGGTATGGAGAAACAGTAGAGAG
(49) pBAD-eGFP-stopG_3T-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CCCACCCACCCACCCCAAAGCCCGAAAGGAAGCTGAG Rev: TTACTTGTACAGCTCGTCCATGCC
(50) pBAD-eGFP-stopctrl1-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CACACACACACACCCCAAAGCCCGAAAGGAAGCTGAG Rev: TTACTTGTACAGCTCGTCCATGCC
(51) pBAD-eGFP-stopctrl2-as	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: CTCACTCACTCACCCCAAAGCCCGAAAGGAAGCTGAG Rev: TTACTTGTACAGCTCGTCCATGCC
(52) pBAD-18-lacZ-relA	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGAGAGGACGATGGTTGCGGTAACCATGATTACGGATT CACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(53) pBAD-18-lacZ-relAm1	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw: ACGCGTCGACGGAGAGGACGATGGTTGCTGTAACCATGA TTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(54) pBAD-18-lacZ-relAm2	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACTTAGAGGACGATGGTTGCGGTAACCATGATTACGGATT C ACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(55) pBAD-18-lacZ-rseA	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGATACTGGATAAGGGTATTAGGCATGACCATGATTAC GGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(56) pBAD-18-lacZ-rseAm1	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGTATACTGGATAAGGGTATTAGTCATGACCATGATTAC GGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(57) pBAD-18-lacZ-rseAm2	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGTATACTGGATAAGGGTATTAGGCATGACCATGATTAC GGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
(58) pBAD-18-lacZ-napH	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGCTGGCTGGAGGGGAACAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(59) pBAD-18-lacZ-napHm1	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGAcggctgTCTGGAGGGGAACAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(60) pBAD-18-lacZ-napHm2	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGAcTTCTGGCTGGAGGGGAACAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(61) pBAD-18-lacZ-yadI	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGCTAAGGAGGAAGGCGATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(62) pBAD-18-lacZ-yadIm1	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGAcTTCTAAGGAGGAAGGCGATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(63) pBAD-18-lacZ-yadIm2	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGCTAATGAGGAAGGCGATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(64) pBAD-18-lacZ-oxyR	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGCGATGGAGGATGGATAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(65) pBAD-18-lacZ-oxyRm1	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGTCGATGGAGGATTGATAATGAGAGGATCGCATCACCATCACCATCACAC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(66) pBAD-18-lacZ-oxyRm2	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACTTCGATGGAGGATGGATAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(67) pBAD-18-lacZ-oxyRm3	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGCGTTGGAGGATGGATAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(68) pBAD-18-lacZ-oxyRm4	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGCGATGGTGGATGGATAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(68) pBAD-18-lacZ-oxyRm5	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGCGAAGGAGGATGGATAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(69) pBAD-18-lacZ-oxyRm6	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-18-lacZ	Fw:ACGCGTCGACGGCGATGGAGGATGAATAATGACCATGATTACGGATTCACTGGCCGTC Rev: CCCAAGCTTTTATTTTTGACACCAGACCAACTGGTAATG
(70) pBAD-eGFP-G_3U	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGGTGGGTCCACCCAGGAGGAATTAACCATGGGCCATCATC Rev:ACCCACCCAGGAGGAAACGGGTATGGAGAAACAGTAGAGAGTTG
(71) pBAD-eGFP-G_3Uctrl	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: TGTGTGTCCACCCAGGAGGAATTAACCATGGGCCATCATC Rev:CACCCACCCAGGAGGAAACGGGTATGGAGAAACAGTAGAGAGTTG

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
(72) pBAD-eGFP-G_3Umm	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw:TGGGTGGGTTCAACCAGGAGGAATTAACCATGGGCCATCATC Rev:CCCACCCAGAAGACAACGGGTATGGAGAAACAGTAGAGAGTTG
(73) pBAD-eGFP-G_3Ummctrl	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GTGTGTGTTCAACCAGGAGGAATTAACCATGGGCCATCATC Rev:ACCCACCCAGAAGACAACGGGTATGGAGAAACAGTAGAGAGTTG
(74) pBAD-eGFP-G_2CU	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: GGTGGTTCAAC AGGAGGAATTAACCATGGGCCATCATC Rev:ACCACCAGAAGAAAACGGGTATGGAGAAACAGTAGAGAGTTG
(75) pBAD-eGFP-G_2CUctrl	Whole plasmid PCR	*E. coli* XL10 gold	pBAD-eGFP	Fw: TTGTTGTTCAACAGGAGGAATTAACCATGGGCCATCATC Rev:CCACCAGAAGACAACGGGTATGGAGAAACAGTAGAGAGTTG
(76) pAJ-10_kdpD (S)	insert	*E. coli* BL21 (DE3) gold	pBAD-18	Fw CGACGCGTTCACGGCATAAAGCGATAGC Rev: CCCAAGCTTGGTCTGGCGTGAGGTGTATGATGATCGG
(77) pAJ-12_kdpD	insert	*E. coli* BL21 (DE3) gold	pBAD-18	Fw: ACGCGTCGACTCACGGCATAAAGCGATAGC Rev: GTGCGAGCTCGAGCGAAATGCATCATCACCATC
(78) pAJ-13_kdpD M1 (S)	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-10	Fw:GGCGTTGGCCTCGGCCTCGCCATTTGCCATGCTATCGTAGAGGTACAC Rev:GGGGACGGCGGACTCTTTG
(79) pAJ-14_kdpD M1	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-12	Fw:GGCGTTGGCCTCGGCCTCGCCATTTGCCATGCTATCGTAGAGGTACAC Rev:GGGGACGGCGGACTCTTTG
(80) pAJ-15_kdpD *E. coli* (S)	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-10	Fw:GGGGTAGGGCTTGGACTGGCAATTTGCCATGCTATCGTAGAGGTACAC Rev: GGGGACGGCGGACTCTTTG
(81) pAJ-16_kdpD *E. coli*	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-12	Fw:GGGGTAGGGCTTGGACTGGCAATTTGCCATGCTATCGTAGAGGTACAC Rev: GGGGACGGCGGACTCTTTG
(82) pAJ-17_kdpD M2 (S)	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-10	Fw:GGGGTGGGGCTGGGGCTGGCGATTTGCCATGCTATCGTAGAGGTACA C Rev: GGGGACGGCGGACTCTTTG
(83) pAJ-18_kdpD M2	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-12	Fw:GGGGTGGGGCTGGGGCTGGCGATTTGCCATGCTATCGTAGAGGTACA C Rev: GGGGACGGCGGACTCTTTG
(84) pAJ-21_kdpD Gal	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-12	Fw: AATGGGGCTGGGGCTGGCGATTTGCCATGCTATCGTAGAG Rev: TCCAGCCCCAGCCCCACGCCGGGGACGGCGGACTC
(85) pAJ-22_kdpD Gal M1	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-14	Fw: GAAATGGGACTAGGACTGGCGATTTGCCATG Rev: CAGTCCTAGTCCTACGCCGGGGACG
(86) pAJ-23_kefC	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-11	Fw:GTGCGAGCTCATGCATCATCACCATCACCACGATAGCCATACTCTACT GCAGGCGCTGATCTATCTTGGTTCG Rev: ACGCGTCGACTTAGATTGACGGTTTGACCTCGGGTTC
(87) pAJ-24_kefC M1	Whole plasmid	*E. coli* BL21 (DE3) gold	pAJ-23	Fw: AGCCCCAGCCCCAGCGCCTCCAGCGCCTGTC Rev: GGGGCTGGGGCGTTATGAAGCC

No and name of construct	Cloning procedure	Bacterial strain	Backbone	Primerpair / Oligonucleotides (from 5' to 3')
	PCR			
(88) pAJ-25_kefC M2	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-23	Fw: TAGGACTAGGACGTTATGAAGCCCGCGAG Rev: GTCCTAGTCCTAGCGCCTCCAGCGCCTGTC
(89) pAJ-26_eutE	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-11	Fw: ACGCGTCGACTTAAACAATGCGAAACGCATCGACTAATAC Rev: ACGCGTCGACTTAAACAATGCGAAACGCATCGACTAATAC
(90) pAJ-27_eutE M1	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-11	Fw: GACTAGGCGGAGAAGGCTGGACC Rev: CTAGTCCGGCAATGCACGG
(91) pAJ-28_kdpD short	Whole plasmid PCR	*E. coli* BL21 (DE3) gold	pAJ-12	Fw: TGAGTCCCCGGCGTGGGGCTGGG Rev: GGACTCTTTGTTGCCGCGGGCGAAC

Sequences of plasmid backbones:

pQE-J06-eGFP (5'-3')

CTCGAGTTTACGGCTAGCTCAGTCCTAGGTATAATGCTAGCACTAGTGGACGTTTAGCTTTAAGAAGGAGATATACCATGGGCCATCATCATCATCA
TCATCATCATCATCACAGCAGCGGCCATATCGAAGGTCGTCATATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAG
CTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCA
CCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGA
CTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTC
GAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACA
GCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGC
CGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCC
AACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAAGGATCCGGCTGCT
AACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATACCGCGGCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCTGAT
ACAGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAA
GTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGAC
TGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCG
GAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGAAGGCCATCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCT
TTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATTCTAGAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGC
AGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGC
AGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTG
AAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCG
AGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCC
AGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAA
ACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGC
CTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCAC
GAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCA
CTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGG
TATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTT
TGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTT
AAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTA
AACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTG
TAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAA
ACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAG
TTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCC
CAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAG
TGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTC
ATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATC
ATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAG
CATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACT
CATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATA
GGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGC
CCTTTCGTCTTCAC

pBAD-eGFP(5'-3')

AAGAAACCAATTGTCCATATTGCATCAGACATTGCCGTCACTGCGTCTTTTACTGGCTCTTCTCGCTAACCAAACCGGTAACCCCGCTTATTAAAAG
CATTCTGTAACAAAGCGGGACCAAAGCCATGACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGC
GTCACACTTTGCTATGCCATAGCATTTTTATCCATAAGATTAGCGGATCCTACCTGACGCTTTTTATCGCAACTCTCTACTGTTTCTCCATACCCGT
TTTTTGGGCTAACAGGAGGAATTAACCATGGGCCATCATCATCATCATCATCATCATCATCACAGCAGCGGCCATATCGAAGGTCGTCATATGGTGA
GCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGA
GGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTAC
GGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCT
TCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAA
GGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTG
AACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGC
CCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGG
GATCACTCTCGGCATGGACGAGCTGTACAAGTAAGGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATA
ACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATATCCCGCAAGAGGCCCGGCAGTA
CCGGCATAACCAAGCCTATGCCTACAGCATCCAGGGTGACGGTGCCGAGGATGACGATGAGCGCATTGTTAGATTTCATACACGGTGCCTGACTGCG
TTAGCAATTTAACTGTGATAAACTACCGCATTAAAGCTTGGGCCCGAACAAAAACTCATCTCAGAAGAGGATCTGAATAGCGCCGTCGACCATCATC
ATCATCATCATTGAGTTTAAACGGTCTCCAGCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAAGC
GGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTA
GTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTT
TGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCGCC
ATAAACTGCCAGGCATCAAATTAAGCAGAAGGCCATCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTTGTTTATTTTTCTAAATACATTC

AAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTT
ATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGG
GTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATG
TGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACA
GAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGA
TCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACC
AAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAA
TTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTG
AGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGA
TGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTA
AAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGT
CAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGT
GGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCG
TAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGT
GTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGG
GTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGAT
TTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACAT
GTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGC
GAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATGGTGCACTCTCAGTA
CAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCT
GACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCAT
CACCGAAACGCGCGAGGCAGCAGATCAATTCGCGCGCGAAGGCGAAGCGGCATGCATAATGTGCCTGTCAAATGGACGAAGCAGGGATTCTGCAAAC
CCTATGCTACTCCGTCAAGCCGTCAATTGTCTGATTCGTTACCAATTATGACAACTTGACGGCTACATCATTCACTTTTTCTTCACAACCGGCACGG
AACTCGCTCGGGCTGGCCCCGGTGCATTTTTTAAATACCCGCGAGAAATAGAGTTGATCGTCAAAACCAACATTGCGACCGACGGTGGCGATAGGCA
TCCGGGTGGTGCTCAAAAGCAGCTTCGCCTGGCTGATACGTTGGTCCTCGCGCCAGCTTAAGACGCTAATCCCTAACTGCTGGCGGAAAAGATGTGA
CAGACGCGACGGCGACAAGCAAACATGCTGTGCGACGCTGGCGATATCAAAATTGCTGTCTGCCAGGTGATCGCTGATGTACTGACAAGCCTCGCGT
ACCCGATTATCCATCGGTGGATGGAGCGACTCGTTAATCGCTTCCATGCGCCGCAGTAACAATTGCTCAAGCAGATTTATCGCCAGCAGCTCCGAAT
AGCGCCCTTCCCCTTGCCCGGCGTTAATGATTTGCCCAAACAGGTCGCTGAAATGCGGCTGGTGCGCTTCATCCGGGCGAAAGAACCCCGTATTGGC
AAATATTGACGGCCAGTTAAGCCATTCATGCCAGTAGGCGCGCGGACGAAAGTAAACCCACTGGTGATACCATTCGCGAGCCTCCGGATGACGACCG
TAGTGATGAATCTCTCCTGGCGGGAACAGCAAAATATCACCCGGTCGGCAAACAAATTCTCGTCCCTGATTTTTCACCACCCCCTGACCGCGAATGG
TGAGATTGAGAATATAACCTTTCATTCCCAGCGGTCGGTCGATAAAAAAATCGAGATAACCGTTGGCCTCAATCGGCGTTAAACCCGCCACCAGATG
GGCATTAAACGAGTATCCCGGCAGCAGGGGATCATTTTGCGCTTCAGCCATACTTTTCATACTCCCGCCATTCAGAG

pBAD-18 (5’-3’)

ATCGATGCATAATGTGCCTGTCAAATGGACGAAGCAGGGATTCTGCAAACCCTATGCTACTCCGTCAAGCCGTCAATTGTCTGATTCGTTACCAATT
ATGACAACTTGACGGCTACATCATTCACTTTTTCTTCACAACCGGCACGGAACTCGCTCGGGCTGGCCCCGGTGCATTTTTTAAATACCCGCGAGAA
ATAGAGTTGATCGTCAAAACCAACATTGCGACCGACGGTGGCGATAGGCATCCGGGTGGTGCTCAAAAGCAGCTTCGCCTGGCTGATACGTTGGTCC
TCGCGCCAGCTTAAGACGCTAATCCCTAACTGCTGGCGGAAAAGATGTGACAGACGCGACGGCGACAAGCAAACATGCTGTGCGACGCTGGCGATAT
CAAAATTGCTGTCTGCCAGGTGATCGCTGATGTACTGACAAGCCTCGCGTACCCGATTATCCATCGGTGGATGGAGCGACTCGTTAATCGCTTCCAT
GCGCCGCAGTAACAATTGCTCAAGCAGATTTATCGCCAGCAGCTCCGAATAGCGCCCTTCCCCTTGCCCGGCGTTAATGATTTGCCCAAACAGGTCG
CTGAAATGCGGCTGGTGCGCTTCATCCGGGCGAAAGAACCCCGTATTGGCAAATATTGACGGCCAGTTAAGCCATTCATGCCAGTAGGCGCGCGGAC
GAAAGTAAACCCACTGGTGATACCATTCGCGAGCCTCCGGATGACGACCGTAGTGATGAATCTCTCCTGGCGGGAACAGCAAAATATCACCCGGTCG
GCAAACAAATTCTCGTCCCTGATTTTTCACCACCCCCTGACCGCGAATGGTGAGATTGAGAATATAACCTTTCATTCCCAGCGGTCGGTCGATAAAA
AAATCGAGATAACCGTTGGCCTCAATCGGCGTTAAACCCGCCACCAGATGGGCATTAAACGAGTATCCCGGCAGCAGGGGATCATTTTGCGCTTCAG
CCATACTTTTCATACTCCCGCCATTCAGAGAAGAAACCAATTGTCCATATTGCATCAGACATTGCCGTCACTGCGTCTTTTACTGGCTCTTCTCGCT
AACCAAACCGGTAACCCCGCTTATTAAAAGCATTCTGTAACAAAGCGGGACCAAAGCCATGACAAAAACGCGTAACAAAAGTGTCTATAATCACGGC
AGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATCCATAAGATTAGCGGATCCTACCTGACGCTTTTTAT
CGCAACTCTCTACTGTTTCTCCATACCCGTTTTTTTGGGCTAGCGAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAA
GCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGC
AGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGA
ACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAA
ATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGAA
GGCCATCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACC
CTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCT
GTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTA
AGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGG
GCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTA
AGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTT
TGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGC
AGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTT
GCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC
TGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGG
TGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTACGCGCCCTGTAGCGGCGCATTAAGCGCGGC
GGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCC

GGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTTGGGTGATG
GTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTTGAAC
AACACTCAACCCTATCTCGGGCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTT
AACGCGAATTTTAACAAAATATTAACGTTTACAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAG
TTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAA
AACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATAC
TGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCT
GCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACAC
AGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAG
GTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCAC
CTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTT
GCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAG
CCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGC
ATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCG
ACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGT
CAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCG
AAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGA
TGCCGGCCACGATGCGTCCGGCGTAGAGGATCTGCTCATGTTTGACAGCTTATC

14 DANKSAGUNG

An dieser Stelle möchte ich mich ganz herzlich bei allen bedanken, die es mir möglich gemacht haben, meine Dissertation in der Arbeitsgruppe Hartig an der Universität Konstanz abzuschließen.

Ich bedanke mich herzlich bei Prof. Dr. Jörg Hartig für die Ermöglichung dieser Arbeit, die Unterstützung bei der Bearbeitung von vielen interessanten Forschungsthemen, sowie die ausgezeichnete Betreuung und die guten Arbeitsbedingungen.
Prof. Dr. Andreas Marx danke ich für die Übernahme des Zweitgutachtens und die hilfreichen Diskussionen während der Thesiskomitees.
Prof Dr. Tancred Frickey danke ich für die Übernahme des Prüfungsvorsitzes und die sehr gute Zusammenarbeit.

Ein riesengroßes Dankeschön an die ganze AG Hartig für die Unterstützung während der Arbeit, das tolle Klima im Labor und die schöne Zeit. Vor allem bedanke ich mich bei Astrid, unserem Lab-Manager, ohne die vieles nicht so einfach funktioniert; bei Charlotte und Bene für die zahlreichen Diskussionen und alles drum herum; bei Michele für den guten Espresso und den italienischen Flair, bei Lena und Malte für die würdige Nachfolge.
Ich bedanke mich auch bei allen ehemaligen Gruppenmitgliedern, vor allem bei Steffi, Fil, Vijay und Kangkan.

Vielen Dank auch an Peiwen Xiong und Andi Groß für die sehr gute und produktive Zusammenarbeit bei diversen kooperativen Projekten. Zarko Kulic und David Witte danke ich für die Hilfe bei NMR Messungen.

Ich bedanke mich bei allen meinen Studenten für die fleißige Mithilfe und die Abwechslung.

Herzlichen Dank an Boris und Charlotte für die Korrektur dieser Arbeit.

Bei allen meinen Freunden möchte ich mich für die schöne Zeit und den Ausgleich neben der Forschung bedanken. Besonders danke ich Dani und Caro, die einfach immer da sind, egal was ist.

Mein ganz besonderer Dank gilt meiner wunderbaren Familie und meinem Mann Robert für die stetige Unterstützung, den Rückhalt und das Zusammenhalten in jeglicher Situation.