Deciphering triplex formation using a synthetic biology-inspired and deep-sequencing approach

Beate Kaufmann

Deciphering triplex formation using a synthetic biology-inspired and deep-sequencing approach

Research thesis

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Beate Kaufmann

Submitted to the Senate of the Technion - Israel Institute of Technology

Chesvan, 5779, Haifa, November, 2018

Acknowledgments

The research thesis was carried out under the supervision of Assistant Professor Dr. Roee Amit in the Department of Biotechnology and Food Engineering of the Technion - Israel Institute of Technology. The generous financial help of the Technion - Israel Institute of Technology is gratefully acknowledged.

I would like to express my very great appreciation to Roee Amit for giving me the opportunity to work on an idea that I had back in 2014. I am particularly grateful for his continuous support of my Ph.D. study, his patience, motivation, and drive to add more value to the philosophical part of the Ph.D. through valuable discussions that we had throughout the years. I could not have imagined having a more challenging and great mentor for my Ph.D. project.

Next, Zohar Yakhini is greatly acknowledged for his expertise and advice that he gave me with respect to the analysis of the next-generation data. I greatly appreciated his help and critical questions along the way. Special thanks go to Oz Solomon who helped me with practical tips and insights into acquiring the knowledge that was necessary to analyze the deep-sequencing part of this thesis.

Furthermore, I would like to thank my colleagues for providing such a laid-back and comfortable atmosphere in the lab, an absolutely fantastic Israel experience and every advice on life and science ever given. Thanks to Inbal C., Inbal V., Lior, Michal, Naor, Noa, Orna, Roni, Sarah and Slava for having my back in every moment. I would also like to offer special thanks to Lisa, Lisa, Noa and Sapir for supporting me at the bench and helping me become a better mentor.

Moreover, I wish to acknowledge the Technion Genome Center. I am particularly grateful to Tal, Anastasia and Sally for help with the planning and execution of the next-generation sequencing experiments. Moreover, Efrat and Shay from the Life Science and Engineering (LS&E) Infrastructure Center are acknowledged for help with the flow cytometry experiments.

I would also like to thank Prof. Ben-Zion Levi's lab in particular Aviva, Nitzan and Mamduh for help with the PAGE and cell culture experiments, the Avi Sphigelman lab in particular Zoya for help with the circular dichroism experiments as well as Prof. Arnon Henn for supplying materials needed for the circular dichroism spectroscopy experiments.

Last, but not least I would like to thank my family and friends. Thank you so much for much for supporting and encouraging me throughout this thesis as well as for providing fun times together in Israel and Europe. This thesis would never have been completed without you.

Contents

List of Figures

List of Tables

Abstract

Nature provides a tremendously rich toolbox of dynamic nucleic acid structures that are widespread in cells and affect multiple biological processes. With technological advances in deep-sequencing and DNA synthesis, non-canonical structures gained renewed scientific as well as biotechnological interest. One particularly intriguing form of such structures is the formation of triplexes. It involves three nucleic acid strands and a mix of Watson&Crick as well as Hoogsteen hydrogen bonds. By applying low-throughput approaches *in vitro*, progress in the field has been made, but until today the underlying rules for triplex formation remain debated and evidence for such triplexes *in vivo* (e.g. in form of RNA*DNA-DNA triplexes) is circumstantial. In this Ph.D. project, I applied a combined strategy of synthetic biology-inspired circuit designs in bacterial and mammalian cells and the development of multiple deep-sequencing and DNA synthesis-based platforms to systematically refine the triplex code. I started with the design of synthetic long non-coding RNAs (slncRNAs) from the bottom-up and tested them in an enhancer-based circuit in bacteria and a gene-activation platform in mammalian cells. In both systems a non-specific and inconsistent up-regulatory effect of a reporter gene in presence of slncRNA molecules was observed, but yielded overall inconclusive results. The challenges I faced using the synthetic biology designs, prompted me to build several next-generation sequencing platforms to study triplex formation *in vitro* and in cells. To do so, I designed large libraries of short, single-stranded oligos containing putative triplex-forming sequences. Following transfection of the libraries into cells, or incubation with double-stranded DNA *in vitro*, a subset of oligos binds the double-stranded DNA via triplex formation, is selectively enriched (Triplex-Seq), or ligated to genomic DNA in close proximity (Triloci-Seq), and subsequently analyzed using next-generation sequencing. By applying the Triplex-Seq approach *in vitro* and in cells, I identified that triplexes are preferably formed in neutral compared to acidic pH, and G-rich oligos as well as G-rich double-stranded DNA form stable and highly specific triplexes. Furthermore, a minimal length of 7-10 nucleotides was shown to be sufficient for stable triplex formation. To identify genomic target sites to which the oligos were bound, I employed Triloci-Seq. Using this approach, I found putative genomic, GAA_{rich} motifs that exhibited a 3 nt periodicity in the sequence reads and have been predicted to form triplexes with GAA_{rich} and TTC_{rich} oligos that were used in this approach. These results, together with the complementary Triplex-Seq data, refine the sequence context required for triplex formation thus establishing a powerful tool to further study unusual nucleic acid interactions. I believe that my results demonstrate the power of deep-sequencing and synthetic biology platforms to explore triplex formation and build upon a growing interest in using DNA structures for bio– and nanotechnological applications.

List of Abbreviations

1D - one-dimensional, 3D - three-dimensional
A - adenine
Amp - ampicillin
ap - anti-parallel (pH 7 condition)
APS - ammonium persulfate
ATP - adenosine triphosphate
BA - bioassay media
bp - base pair
C - cytosine
C_4-HSL - N-butanoyl-L-homoserine lactone
Cas9 - CRISPR associated protein 9
CD - circular dichroism
CDS - coding sequence
CHO/K1 - Chinese hamster ovary cell line K1
CISTR-ACT - cis- and trans-chromosomal chondrogenic regulator transcript
CLAMP - chromatin linked adaptor for MSL proteins
CMV - cytomegalovirus
CO_2 - carbon dioxide
CoREST/REST - cofactor/repressor element-1 silencing transcription factor
CRISPR - clustered regularly interspaced short palindromic repeats
CSB - Crush and Soak buffer
dhfr - dihydrofolate reductase
DIT - digital integration time
Dlx-2 - Distal-Less Homeobox 2
DMEM - Dulbecco Eagle's Minimum Essential Medium
DNMT3b - DNA methyltransferase 3 beta
DNA - deoxyribonucleic acid
DPBS - Dulbecco's phosphate buffered saline
DRBP - DNA-RNA-binding proteins
dsDNA - double-stranded DNA
EDTA - ethylenediaminetetraacetic acid
EMSA - electrophoretic mobility shift assay
ENCODE - Encyclopedia of DNA Elements
eRNAs - enhancer RNAs
eYFP - enhanced yellow fluorescent protein
FACS - flow cytometry activated cell sorter
FBS - fetal bovine serum
FIRRE - functional intergenic repeating RNA element
FL - fluorescence
FP - fluorescent protein
FSC - forward scatterG - guanine
GAL4/UAS - galactose-responsive transcription factor/upstream activating sequence
GFP - green fluorescent protein
GR - glucocorticoid receptor
gRNA - guide RNA

HAC - human artificial chromosome
HDAC - histone deacetylase
HEK-293 - human embryonic kidney cells, clone 293
HOTAIR - HOX transcript antisense RNA
HOTTIP - HOXA transcript at the distal tip
IUPAC - International Union of Pure and Applied Chemistry
Kan - kanamycin
kb - kbp/10^3 basepairs
Khspk1 - antisense RNA of the sphingosine kinase 1 (sphk1) gene
LB - lysogeny broth/Luria-Bertani
lincRNA - long intergenic non-coding RNA
lncRNA - long non-coding RNA
mat2 - methionine adenosyltransferase
MECP2 - methyl CpG binding protein 2
mHG - minimum hypergeometric
mL - milliliter
mM, M - millimolar, molar
mRNA - messenger RNA
MS - mass spectrometry
MSL - male specific lethal
ncRNA - non-coding RNA
NGS - next-generation sequencing
NLS - nuclear localization signal
nt/nts - nucleotide/nucleotides
o/n - overnight
OD - optical density
oligos - oligonucleotides
ORF - open reading frame
p - parallel (pH 5 condition)
PAGE - polyacrylamide gel electrophoresis
PARTICLE - promoter of MAT2A-antisense radiation-induced circulating lncRNA
PCR - polymerase chain reaction
PEG - polyethylene glycol
PEI - polyethylene imine
PFA - paraformaldehyde
PRC2 - polycomb repressive complex 2
pRNA - promoter-associated RNA
RAP - RNA antisense purification
RBP -RNA-binding protein
RT - room temperature
rDNA - ribosomal DNA
RNA - ribonucleic acid
roX - RNA on X
RPM - reads per million (bioinformatic analysis)/rpm - rounds per minute
SAF-A - scaffold attachment factor A
sbfp2 - strongly enhanced blue fluorescent protein 2
SDS - sodium dodecyl sulfate

SHARP - SMRT and HDAC associated repressor protein
slncRNA - synthetic long non-coding RNA
SMRT - silencing mediator for retinoid or thyroid-hormone receptors
Sox2 - sex determining region Y-box 2
SPR - surface particle resonance
SSC - saline-sodium citrate (buffer)/side scatter (flow cytometry)
ssDNA - single-stranded DNA
T - thymine
TBE - Tris/Borate/EDTA
TDB - triplex disfavoring buffer
tdPCP - tandem-dimer PP7 phage coat protein
TE - Trizma/Tris-EDTA
TEMED - N,N,N',N'-Tetramethylethylenediamine
TF - transcription factor
TFO - triplex-forming oligo
TLS - translocated in lipocarsoma
TMP - Trioxsalen/4,5,8-Trimethylpsoralen
TTS - triplex target site
U - Units
UTR - untranslated region
UV - ultraviolet
V - Volt
Xi - inactive X-chromosome
XIST - X-inactive specific transcripts)
µL- microliter/ µM - micromolar / µg - microgram

1 Introduction

Scientists love to solve puzzles and riddles and if a scientific mystery presents itself, researchers will definitely engage in deciphering the unresolved question. And it would not be science if there are not such big questions out there. One said enigma of recent years has been termed the biological 'dark matter' of genomics. Only about 2 % of the mammalian genomes encode for proteins, so does that mean that the remaining 98 % of the genome lack any obvious function? In recent years, technological advances in genome sequencing[1] and low-cost gene synthesis[2] allowed researchers to shed light on the genome's 'dark matter'. Since then discoveries such as the findings that ultraconserved elements across mammalian genomes impair neurological functions[3] or that mutational hotspots frequently occur in the non-coding regions in cancer[4] have been elucidated. Furthermore, the Encyclopedia of DNA Elements (ENCODE), an international research consortium initiated to build a list of functional DNA elements, has identified and annotated thousands of non-coding transcripts (ncRNAs)[5]. These ncRNAs lack any obvious protein-coding potential, but are nevertheless implicated in various biological processes[6]. Due to their length and their inability to be translated into proteins these transcripts are termed long non-coding RNAs (lncRNAs).

1.1 Long non-coding RNAs (lncRNAs)

lncRNAs are per definition longer than 200 nts and are predominantly localized in the nucleus[7]. This novel class of non-coding transcripts can be non-polyadenylated or polyadenylated, mono- or multi-exonic, are transcribed in sense or antisense orientation and are overall expressed at lower levels compared to mRNAs[7] (Figure 1a).

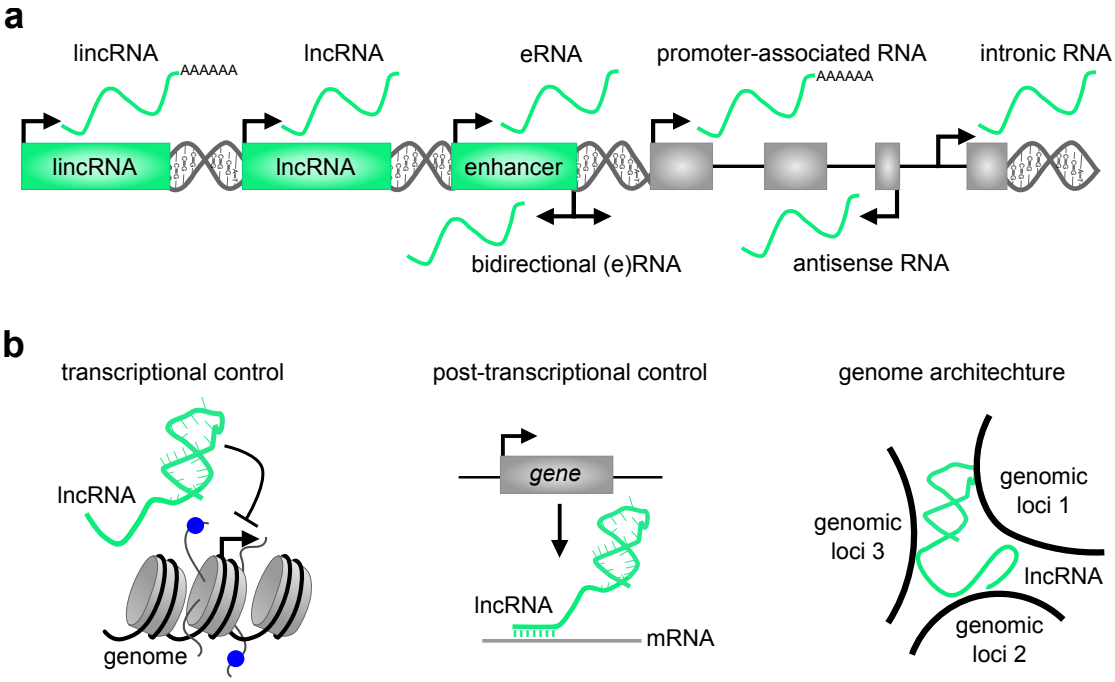


Figure 1| Characteristics and function of long non-coding RNAs (lncRNAs). **a**, lncRNAs are a diverse class of ncRNAs and are transcribed as intronic transcripts, in sense or antisense orientation, as promoter-associated transcripts (paRNAs), and as unidirectional or bidirectional enhancer RNAs (eRNAs). Additionally, one subset of RNAs are expressed as long intergenic ncRNAs (lincRNAs) that are spliced, polyadenylated and tissue specific. **b**, lncRNAs have been implicated in various biological processes such as regulation of gene expression at the level of transcription, as well as post-transcription, and have been proposed to guide the three-dimensional genome architecture via bridging genomic loci.

Thus, lncRNAs can be roughly classified into sub-groups depending on these characteristics such as long intergenic ncRNAs (lincRNAs), uni- or bidirectional enhancer RNAs (eRNAs) and promoter associated RNAs (pRNA)[8;7;9]. While many lncRNAs appear to be functional molecules and have been implicated in transcriptional regulation[9;10;11] and dosage compensation[12;13], genome architecture (personal correspondence, I. Farabella, 42nd FEBS congress), post-transcriptional processing[14] as well as translational regulation[15] (Figure 1b), some transcripts seem to be a mere by-product of pervasive transcription of the genome and lack apparent functions[16;17]. This diverse set of potential lncRNA functions and their impact on biological processes also results in emerging pathogenic patterns such as the involvement of lncRNAs in cancer, myopathies and genetically inherited disorders[18]. With a 10-fold increase of publications in the field of lncRNAs in a decade (ncbi: 2008: 222 publications/year, 2017: 2938 publications/year[19]), it represents an extensive amount of overall lncRNA studies. Hence, this work will be focusing on lncRNAs transcriptional regulators.

1.2 lncRNA transcriptional regulators

Given the higher-order structures of eukaryotic genomes, lncRNAs have developed multiple ways to control gene expression such as providing docking platforms for chromatin remodeling complexes and histone modifying proteins, modulating transcription factor (TF) activity and long-range gene regulation by bridging genomic loci that are not in close proximity to one another (Figure 2).

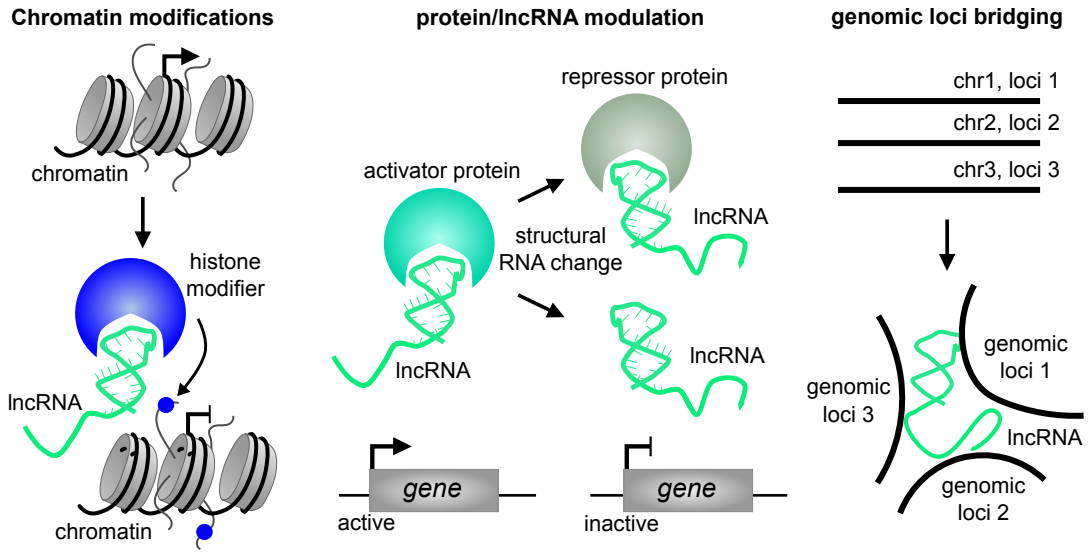


Figure 2| Schematic overview of transcriptional lncRNA regulators. lncRNAs mainly localize in the nucleus and have developed several ways to regulate gene expression. lncRNAs scaffold regulatory proteins such as histone modifiers and chromatin remodellers that change the sate of chromatin (left). Besides changing the chromatin state, lncRNAs can undergo conformational changes in their secondary structure, which lead to substitution of a transcriptional activator protein with a repressor complex or dissociation of transcription factors (center). Another strategy to regulate or orchestrate simultaneous gene expression is by bridging distant genomic loci of the same or different chromsomes using lncRNAs (right).

1.2.1 Chromatin modulation

One simple, yet intriguing way of how lncRNAs regulate gene expression relies on scaffolding large protein complexes that change the state of chromatin (Figure 2, left). One such example is Xist (X-inactive specific transcripts) which regulates dosage compensation at the early stage of the development of the female embryo of mammals. Dosage compensation describes the process

of silencing one X-chromosome to normalize expression of X-linked genes. The 17 kb long lncRNA is transcribed from the inactive X-chromosome (Xi) while being absent on the active chromosome[20]. Xist is found close to its site of transcription, coats the inactive chromosome[21] and induces heterochromatin formation[22] by scaffolding proteins such as SHARP (SMRT and HDAC associated repressor protein, also known as SPEN) which in turn recruits the histone deacetylase HDAC3[23].
Contrary to silencing gene expression of X-linked genes, roX (RNA on X) molecules activate gene expression in male *Drosophila* flies. roX compensates for the lack of the additional X-chromosome by up-regulating transcription of the X-linked genes. roX1 and roX2 contain conserved tandem stem loop structures that bind the male specific lethal (MSL) protein complex[24]. This ribonucleoprotein complex coats the chromosome and thereby activates gene expression through acetylation of specific histones[13;25].
While Xist and roX are examples of *cis*-regulatory molecules, HOTAIR (HOX transcript antisense RNA) was the first lncRNA to be described that acts *in trans* (at a distance)[26]. During development, HOTAIR is transcribed from the homeobox C (HoxC) locus and represses transcription of genes in the HoxD cluster located on another chromosome. Recently, Tsai and colleagues reported that HOTAIR assembles at least two protein complexes with histone-modifying activities by providing a modular secondary docking structure[27;11]. The lncRNA tethers the polycomb repressive complex 2 (PRC2) and the CoREST/REST repressor complex thereby organizing histone-modifying protein activity at specific genomic loci.

1.2.2 Modification of transcription factor activity

In addition to protein scaffolding to change epigenetic states, lncRNAs can also bind transcriptional regulators (Figure 2, center). For instance, the lncRNA Gas5 inhibits the glucocorticoid receptor (GR) and regulates gene expression by mimicking the GR-binding domain thereby reducing protein-DNA interactions[14]. Contrary to DNA mimicry, the lncRNA that regulates the *ccnd1* gene (cyclin D1) induces a conformational change in the TLS (translocated in liposarcoma) protein thus activating the protein for DNA binding[28]. Subsequently, histone acetyltransferases are repressed and *ccnd1* gene expression is silenced. While lncRNAs can induce conformational changes in TFs, they also bind both transcriptional activators and repressors. The polyadenylated lncRNA Evf2 for example recruits either the Dlx-2 (Distal-Less Homeobox 2) activator or MECP2 (methyl CpG binding protein 2) repressor to the intergenic enhancer elements Dlx-5/6 ei and Dlx-5/6 eii thus either mediating activation or repression, respectively[29].

1.2.3 Transcriptional regulation via spatial-compartment formation

Recent studies suggested that ncRNAs are transcribed in sense or antisense orientation from enhancer regions (enhancer RNAs, eRNAs) and enhance transcription. Recently, Li *et al.* published an article to support that eRNA transcripts have indeed functional roles in enhancing transcription[30]. The authors observed an eRNA-controlled increase of transcription at oestrogen-regulated enhancers and proposed an eRNA-mediated stabilization between promoter and enhancers. This hypothesis of lncRNA-enhancer bridging was supported by Melo and colleagues who reported a p53-dependent eRNA transcription[31]. Both studies tethered the transcribed eRNAs to a GAL4/UAS (galactose-responsive transcription factor/upsteam-activating sequence) reporter system via RNA-binding proteins (RBPs) such as λN and MS2, respectively. The tethering strategy significantly enhanced transcription in an eRNA-dependent manner at enhancer sites and promoters. In the looping-based transcriptional regulation, eRNAs assemble protein

complexes and bridge distal enhancer sites with specific promoter regions or provide a platform that can be easily accessed by the polymerase II transcriptional machinery. For instance, eRNAs stabilize the cohesin complex at enhancer sites by interacting with the Cohesin subunits thus enhancing transcription[30]. Another study showed that the homeodomain-containing TF Dlx-2 forms a complex with the Evf2 eRNA that is transcribed from the ultraconserved element *ei* of the Dlx-5/6 enhancer. Through Evf-2-mediated bridging of the TF Dlx-2, Dlx-5/6 enhancer activity is increased[28].

While enhancer-based gene regulation is one aspect of providing spatial proximity, other lncRNAs have been implicated to shape nuclear compartments (Figure 2, right) such as Xist and FIRRE (functional intergenic repeating RNA element). Xist has been described above as a lncRNA that scaffolds histone-modifying proteins thus silencing gene expression, but it also induces structural chromosomal changes by interacting with the lamin B receptor[32;23]. FIRRE, a lncRNA that is required for proper adipogenesis, has been shown to interact with the nuclear-matrix factor hnRNPU[33] and thereby co-localizes with distinct genomic loci such as DXZ4[34]. In a follow-up study on the mechanism on how such 3D-organizational characteristics can be established, the authors showed that while FIRRE plays an important role, it is not the only factor to form such topological-associated domains[35]. Whilst FIRRE and Xist shape the 3D-structure of the genome, NEAT (nuclear enriched abundant transcript 1) induces so called paraspeckles. These structures are sub-cellular compartments that contain proteins (mainly RBPs) as well as mRNAs[36] and have been proposed to act as transcription regulation centers[37] and form a nuclear RNA retention architecture[38].

1.3 lncRNA localization to genomic targets

Considering the implications of lncRNAs in regulation of gene expression and their role in recruiting regulatory proteins to specific genomic loci, it is imperative to ask how lncRNAs target genomic loci in a precise manner. Three strategies on how lncRNAs accomplish such interactions will be discussed below (Figure 3): (1) Structural genome architecture-guided interactions, (2) Protein-mediated genomic interactions, and (3) Direct RNA-DNA interactions via triple helix formation.

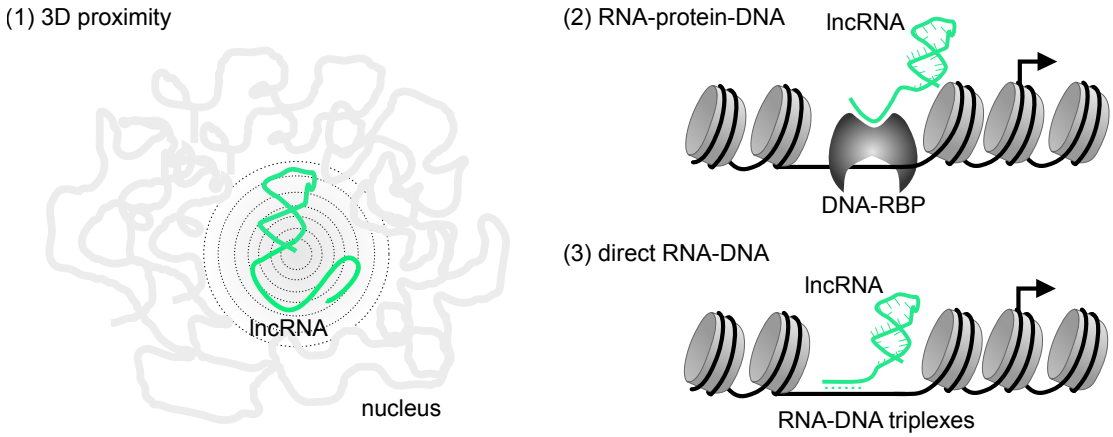


Figure 3| Strategies for targeting genomic loci by lncRNAs. Three main strategies of how lncRNAs target genomic loci are presented in this scheme. (1) The genome architecture provides a scaffold for lncRNAs to reach distal genes with respect to the linear location on the chromosome. (2) lncRNAs can bind to DNA-RNA-binding proteins (DRBPs) that recognize DNA motifs on the genome thereby recruiting the lncRNA to its site of regulation. (3) lncRNAs can target genomic loci by binding directly to DNA by forming triple helical structures (triplexes).

1.3.1 3D-proximity guides lncRNA interactions

To understand how lncRNAs can be restricted to specific genomic regions, I turn again to Xist, the lncRNA earlier described as a *cis*-regulatory acting lncRNA. Using the newly developed RNA antisense purification (RAP) technology, which purifies RNA and associated DNA regions via biotinylated oligos, it has been proposed that Xist exhibits a broad patterning with higher enrichment at gene-rich regions across the entire inactive chromosome[39]. Establishment of such a pattern was proposed to be guided by the proximity transfer model in which Xist initially (Xi initiation) localizes to DNA regions that are in close 3D-proximity to its own transcription site, but are distally located across the chromosome. Spreading of Xist was proposed to be achieved by modifying chromatin states and nuclear architecture through recruitment of the histone methyltransferase PRC2[39]. While there are numerous other lncRNAs that act in *cis*-regulatory manner such as FIRRE, HOTTIP[40] (HOXA transcript at the distal tip) and CISTR-ACT (Cis- And Trans-Chromosomal Chondrogenic Regulator Transcript)[41], the precise mechanism of how interchromosomal contacts contribute to *cis*-regulatory effects remains to be investigated.

1.3.2 Protein-mediated interactions with genomic loci

Contrary to the 3D-proximity model, protein-mediated interactions of lncRNAs with genomic sites require proteins such as RBPs, DNA-RNA-binding proteins (DRBPs) and DNA-binding proteins to be imported into the nucleus and target specific loci. Some lncRNAs such as roX interact with chromatin to target specific regions. The MSL complex, which consists of five proteins as well as the roX1 and roX2 lncRNAs, has been shown to interact with CLAMP (chromatin-linked adaptor for MSL proteins), a zinc finger protein that recognizes so called chromatin entry sites on the genome that are GA-rich[42;43]. Recently, McHugh *et al.* employed the RAP technology followed by quantitative mass spectrometry (RAP-MS) and showed that Xist interacts with 10 proteins[23] and at least three of them are required for silencing the X chromosome. Intriguingly, one of the proteins, a known chromatin-binding protein termed SAF-A (scaffold attachment factor A), has been proposed to bind to Xist and recruits the Xist/protein complex to sites on the X chromosome.

This protein-guided targeting of DNA and the similar mechanism of lncRNAs that bridge enhancer regions, as was discussed above, inspired me to hypothesize that RBPs may facilitate scanning of genomic target sites of lncRNAs in a similar fashion to how TFs efficiently find their binding sites on the genome. This hypothesis is based on the facilitated diffusion model that has been described mathematically over 30 years ago[44]. To control the rate of transcription, TFs need to recognize their specific DNA sequence in a fast and precise manner. Berg and colleagues described two limiting factors for the three-dimensional diffusion of protein-DNA target site search: (i) the small numbers of perfect DNA-binding sites/operators in the genome and (ii) structural similarities between perfect and non-target sites. Hence, over the years biophysical work suggested that the protein-DNA search is a combined mechanism of one-dimensional (1D) sliding of the protein along the DNA and three-dimensional (3D) diffusion in the cytoplasm. *In vitro* work using λ phage DNA-coated flow cells demonstrated that the C-terminal domain of the tumor suppressor protein p53 indeed slides along the DNA[45]. Furthermore, Hammar and colleagues described how TFs find their specific binding sites in living cells[46]. They demonstrated in a single-molecule tracking assay that the transcriptional repressor LacI slides along the DNA in a 1D-manner and 3D-diffuses in the bacterial cell. This facilitated diffusion allowed LacI to accurately scan for potential LacI-binding sites in the bacterial genome. Similarly, more recent

work by Chen *et al.* demonstrated that the enhancer-binding regulator Sox2 (sex determining region Y-box 2) diffuses in the nuclear space combined with DNA sliding to efficiently find its target site in the genome of embryonic stem cells[47]. Thus, even though research indicates a chromosomal structure guided approach for *cis*-regulatory effects in the case of Xist, it is possible that the affinity and scanning characteristics of proteins as described for Sox2 and LacI do facilitate and support other lncRNA regulatory interactions.

1.3.3 Direct lncRNA-DNA interaction

The third strategy highlights a slightly more unconventional mechanism in which lncRNAs directly interact with genomic sites via RNA*DNA-DNA triple helix formation[10;48;49;50;51;52;53;54]. These RNA-DNA interactions rely on single-stranded RNA molecules that associate with the targeted DNA duplex strand via Watson and Crick-independent hydrogen bonds and are termed triple helix structures (triplexes)[55]. Because of the unconventionality of such triplex structures, I will elaborate on non-canonical structures in general and triplexes in particular in the following paragraphs.

1.4 Non-canonical DNA structures

Triplex structures belong to the class of non-canonical nucleic acid structures (Figure 4). During biological processes such as transcription and replication, DNA unwinds and is able to form structures that differ from the Watson and Crick-proposed B-form of the DNA double helix (conformational polymorphism)[56]. These alternate structures fall into two categories: The first one adopts structures such as parallel duplexes[57;58], hairpins[59;60] and cruciforms[61;62] using Watson and Crick bonds, while the second group relies on Watson and Crick-independent hydrogen bonds termed Hoogsteen bonds. The most prominent members of this group are G-quadruplexes[63;64;65;66], the i-motif (C^+-C basepair)[67;68;69;70] and inter- as well as intramolecular triplexes which rely on both Watson and Crick as well as Hoogsteen interactions[71;72;73;74]. All non-canonical structures are implicated in biological processes such as transcriptional regulation[75;64], replication[76] or genome instability[77;78]. Despite two important publications that further shed light on the abundant occurrence of G-quadruplexes in the genome[79] and the existence of i-motifs in cells[80], there has been little experimental evidence supporting the existence of non-canonical structures and in particular triplexes *in vivo*.

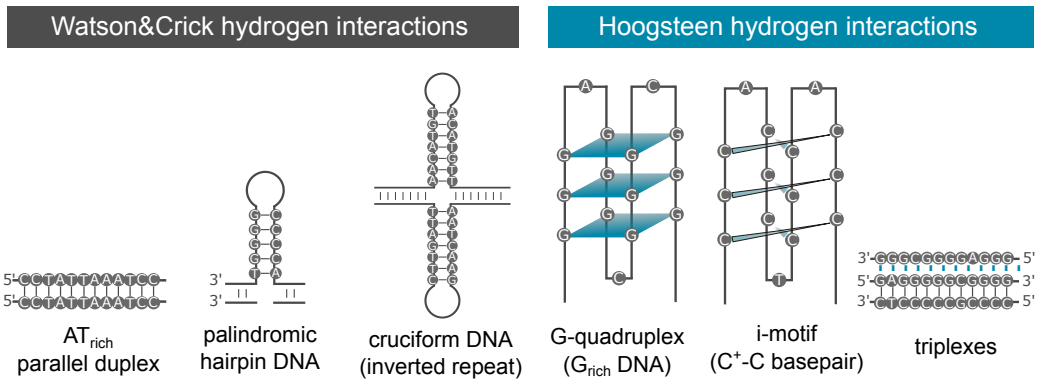


Figure 4| Schematic representation of non-canonical structures. Non-canonical DNA structures are found throughout the cell. These alternate structures can be divided into two groups. The first one uses the commonly known Watson and Crick hydrogen interactions and comprises parallel duplexes, hairpin and cruciform DNA, whereas the second group uses Hoogsteen interactions and is formed by G-quadruplexes, i-motifs and triplexes.

Triplexes are alternate structures that occur inter- and intramolecularly and are found in bacteria [81;82;72], yeast [73] and mammalian cells [83;78]. Triplex formation is the process in which a single-stranded nucleic acid molecule (RNA or DNA) binds predominantly to the purine-rich (guanine or adenine) major groove of double-stranded nucleic acid molecules (RNA or DNA) (Figure 5a, left). Such structures have been discovered shortly after Francis Crick and James Watson, with the help of X-ray diffraction images from Rosalind Franklin, published the DNA double-helix model. Most notably the existence of triple helical structures was first supported in a 1957 study [84]. Gary Felsenfeld put forward that, given the appropriate concentration of multivalent ions such as magnesium, mixtures of polyadenylic and polyuridylic acid form a double-stranded nucleic acid molecule and a single ribonucleic acid chain wraps around the double helix.
The underlying interactions between the third strand and the double helix are based on hydrogen interactions that differ from the known Watson and Crick bonds. Few years after the existence of triplexes was proposed, Karst Hoogsteen suggested that basepairs can adopt a different geometry through rotation of the purine base around the glycosidic bond thus providing different hydrogen acceptors (N7 position of purine base) and donors (C6 amino group of purine base) [55]. These hydrogen interactions have been termed Hoogsteen bonds and are the principal force in triplex structures (Figure 5a, right).

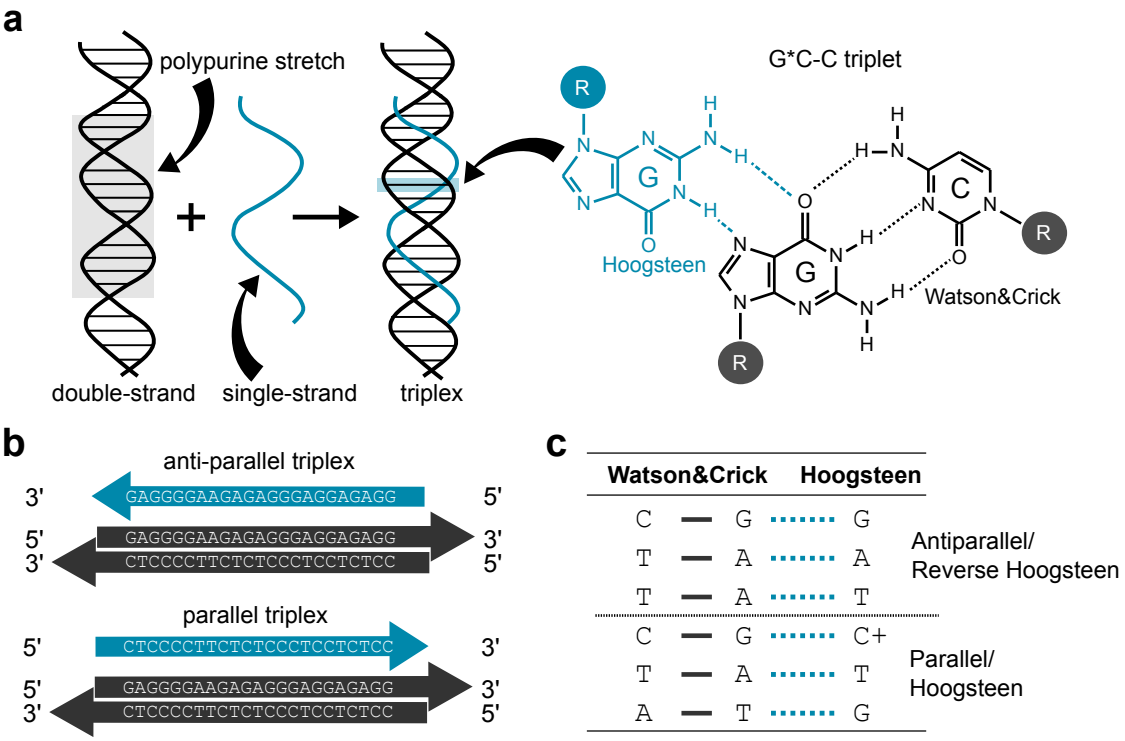


Figure 5| Triplex formation and the underlying 'triplex code'. **a**, Triple helices (triplexes) are formed when a single-stranded nucleic acid molecule interacts via Hoogsteen interactions with the purine strand of a Watson&Crick-based double helix. Hoogsteen interactions make use of free hydrogen acceptors and donors of the Watson & Crick basepair by rotation and adoption of a different geometry. **b**, Triplexes can adopt two different types of triplexes depending on salt concentration and pH. The first one is referred to as an anti-parallel triplex because the third strand is reversed with respect to the 5' –> 3' orientation of the purine sequence of the double helix. **c**, In literature, several nucleotide combinations were found that can occur in the same triplex. It has also been shown that anti-parallel triplexes use reverse Hoogsteen bonds, while parallel triplexes require Hoogsteen bonds. Some triplexes contain mixed motifs which can occur both in anti-parallel or parallel triplexes.

Single-stranded chains bind the polypurine stretch of the major groove of the duplex molecule via two possible Hoogsteen configurations: (i) Hoogsteen interactions promoting a parallel orientation of the third strand and (ii) reverse Hoogsteen bonds resulting in an anti-parallel orientation

of the third strand with respect to the duplex sequence (Figure 5b). Based on these restrictions in nucleotide geometry a 'triplex code' was proposed and is shown in Figure 5c.
Experiments performed *in vitro* indicated that parallel orientations are predominantly established by pyrimidine motifs (cytosine, thymine) and anti-parallel architectures are found in purine motifs (adenine, guanine). However, mixed motifs (guanine, thymine) of nucleotides adopt both parallel and anti-parallel triplexes. For triplex formation to occur certain conditions are required. Morgan and Wells showed in 1968 that double-stranded DNA forms triplex structures with cytosine-rich polyribonucleotides in acidic environments and random polymers did not yield triplexes, thus indicating that triplexes exhibit a clear specificity [85]. While pyrimidine-rich sequences require acidic conditions to obtain a N3-protonated cytosine in a C-G*C^+ triplex [86;87;88], it has been shown that triplex formation using guanine-rich single-stranded chains in physiological potassium concentrations was inhibited due to G-quadruplex formation [86;89;90]. Given this specificity in formation of triplexes [85;86;91;92], Moser and Dervan developed short triplex-forming oligonucleotides (TFOs) that can be targeted to double-stranded molecules and from triplexes by binding to the major groove of homopolypurine stretches in the triplex target site (TTS) via Hoogsteen bonds [93;86;94].
TFOs are generally 10-30 nucleotides long and due to their specificity in targeting double-stranded DNA, they were subsequently used as biotechnological tools both *in vitro* and *in vivo*. One of the first publications in the field demonstrated the successful induction of double-strand breaks within the TTS upon triple helix formation [93] (Figure 6a). In the following decades, TFOs have been used for site-specific transcriptional regulation such as transcription initiation [95] (Figure 6b) and elongation inhibition [96] (Figure 6c) as well as triplex-directed mutagenesis [97].

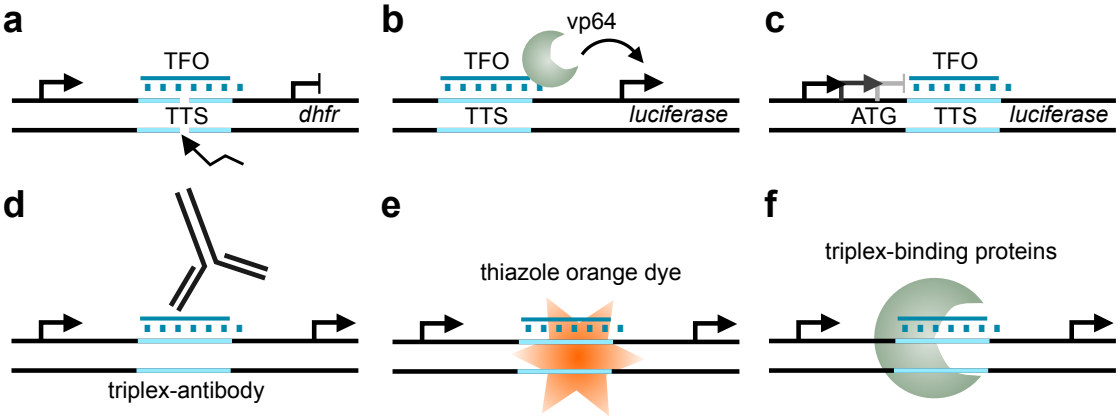


Figure 6| Applications of triplex-forming oligos and tools to detect triplexes. TFOs are short single-stranded oligos that can form triplexes with a double-stranded, purine-rich DNA sequence that is termed triplex target site (TTS). Several groups used TFOs as biotechnological tools to **(a)** induce double-strand breaks and use it for targeted mutagenesis, **(b)** activate gene expression by coupling TFOs to the activator peptide vp64 and **(c)** inhibit transcriptional elongation of a reporter gene. Furthermore, TFOs were used as tools to show endogenous triplex formation such as **(d)** generation of triplex-recognizing antibodies and **(e)** use of thiazole orange, a dye that was shown to possess a higher affinity towards triplexes compared to duplexes. **(f)** A vast list of naturally occurring triplex-binding proteins have been discovered in bacteria, yeast and mammalian cells.

Based on the initial study that showed triplex-mediated double-strand breaks, TFOs were further optimized to increase recombination events at sites of interest [98;99]. Despite extensive amount of work that focused on establishing TFOs as therapeutic and biotechnological agents, triplex formation under physiological conditions mostly was attenuated and the expected biological function (e.g. site-directed mutagenesis) was low. This constraint was highlighted recently, in which studies from Wang *et al.* [97;100] and Vasquez *et al.* [101] could not be reproduced by a different research group in 2017 [97;100;102]. In the original publication, the authors showed that both in

cells[97] and mice[101] significant triplex-induced mutagenesis occurred using non-modified TFOs. To overcome such limitations, a new field developed in which the backbone[92;103], base[94] and sugar component[104;105] of TFOs were intensively modified thus facilitating triplex formation of both guanine-rich (G-rich) or cytosine-containing TFOs. Additionally, chemical compounds such as psoralen were coupled to the TFOs to increase targeting rates of genomic DNA, mutagenesis frequency or inhibition elongation efficiency[106;98;107;99;97].

While exploiting triplexes upon introduction of exogenous TFOs mainly focuses on chemically modified TFOs, the existence of endogenously occurring triplexes in cells was long questioned due to the constraints of triplex formation in physiological conditions (pH of 7.0 disfavors parallel triplexes, physiological potassium concentrations of > 40 mM disfavor triplexes and favor G-quadruplexes). In recent years, there has been an increasing amount of experimental proof which supports the natural occurrence of triplexes *in vivo*. Additional experimental evidence supporting triplex formation *in vivo* has been generated using specific triplex-binding antibodies[108] (Figure 6d), triplex-intercalating dyes such as thiazole orange[109] (Figure 6e) and findings of endogenous triplex-binding proteins (Figure 6f) involved in DNA damage repair[110], recombination[111], transposition[112] and chromosome segregation[113]. While experimental support for endogenous triplexes accumulated, *in silico* analyses also suggested that polypurine regions, exhibiting characteristics of putative triplex target sites, are significantly over-represented in promoter regions and CpG islands in several genomes[114;115;116]. With increasing availability of sequenced genomes, bioinformatic tools such as the Triplexator[117], triplex domain finder[118] and the TTSMI database[119] were developed and confirmed that putative TTS sequences are highly enriched in gene regulatory elements. Taken together, the molecular understanding of triplex formation *in vitro* and the abundance of putative TTS in the genome, has prompted some researchers to suggest that RNA molecules might interact with the double-stranded genomic DNA and form RNA*DNA-DNA triplexes (Figure 7).

The first study about putative triplex-forming lncRNAs in eukaryotic cells demonstrated that a promoter-associated ncRNA silenced transcription of the *dihydrofolate reductase* (*dhfr*) gene likely through formation of a triple-helix structure between the ncRNA and the promoter region[48] (Figure 7a). The lncRNA that is transcribed from a minor promoter upstream of the major promoter of the *dhfr* gene interacts directly with TFIIB, as shown by RNA immunoprecipitation, leading to dissociation of the pre-initiation complex and thereby reducing DHFR expression. Specificity of targeting the non-coding RNA to the major promoter is proposed to be achieved trough triplex formation with a G-rich region containing an Sp1-binding site within the major promoter. This specific triplex interaction has been demonstrated by *in vitro* analysis of an H-form band-shift assay. While Martianov *et al.* proposed a RNA-mediated repressional switch in TFIIB, Schmitz and colleagues showed three years later that *ribosomal DNA (rDNA)* genes are silenced through triplex formation of a promoter-associated ncRNA (pRNA) with a sequence representing the binding site for the transcription factor TTF-1 termed T0 in the core rDNA promoter[10] (Figure 7b). The pRNAs induce *de novo* methylation by triplex formation with the putative TTS which is recognized by the DNA methyltransferase DNMT3b (DNA methyltransferase 3 beta). Furthermore, it was demonstrated that the pRNA is bound by the large subunit TIP-5 of the chromatin remodeling complex NoRC[120] which leads to silencing of the rDNA genes. lncRNAs are known to bind the repressive PRC2 complex and two prominent lncRNAs, Fendrr[50] (Figure 7c) and MEG3[52] (Figure 7d) have been proposed to target PRC2 to multiple genomic sites in *cis* (in close proximity) and *trans* (distal) via triplex formation. Contrary to promoter-associated RNAs, the lncRNAs Khspk1[51] (antisense RNA of the *sphingosine kinase 1* (*sphk1*) gene) and PARTICLE[53] (promoter of MAT2A-antisense radiation-induced

circulating lncRNA) are transcribed in antisense orientation of the genes they are regulating.

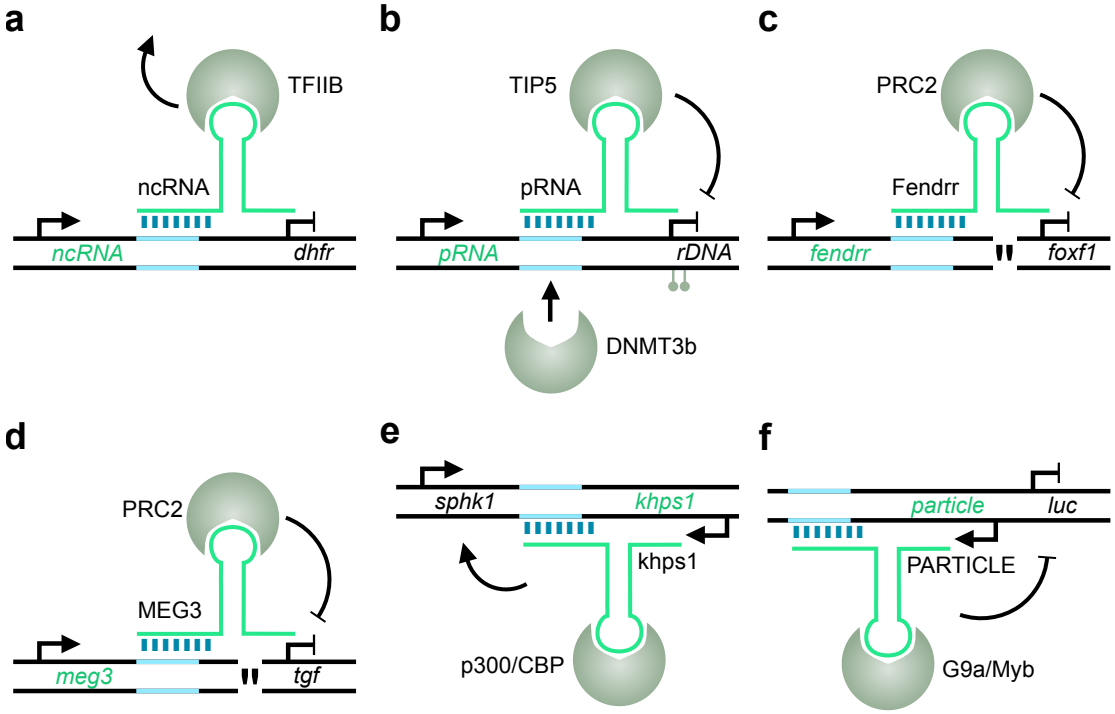


Figure 7| RNA*DNA-DNA triplex formation in cells. lncRNAs have been proposed to form triplexes with genomic loci in cells. Since 2007, several publications indicated the involvement of various lncRNAs in regulating expression via triplex formation such as **(a)** the ncRNA interacts with the Sp1-site located within the promoter for the *dihydrofolate reductase* (*dhfr*) gene, **(b)** the promoter-associated RNA (pRNA) interacts with a genomic sequence overlapping a TF-binding site of the core rDNA promoter, **(c + d)** Fendrr and MEG3 lncRNAs target genomic loci in *cis* and *trans*, **(e +f)** lncRNAs Khspk1 and PARTICLE are transcribed in antisense orientation and interact with promoter regions of the (e) *spkh1* gene or (f) *mat2a* gene.

While the lncRNAs that have been described so far promote silencing of gene expression, Khspk1 lncRNA has been shown to bind the histone acetyltransferase p300/CBP and recruits it to the promoter of the *spkh1* gene thereby enhancing gene transcription of *sphk1* which reduces E2F1-induced apoptosis (Figure 7e). Low-dose irradiation has been proposed to invoke biological processes which is in part regulated by PARTICLE. PARTICLE is transcribed upon low-dose radiation and silences the *mat2a* (methionine adenosyltransferase) gene via triplex formation within the CpG-island of the *mat2a* promoter[53] (Figure 7f). *mat2a* expression is silenced by bringing the lysine methyltransferase G9a into close proximity of the promoter thereby ensuring access to sufficient methyl groups required in methyl transfer reactions and polyamine biosynthesis. While all publications propose triplex formation as a potential mechanism of targeting lncRNAs to genomic loci, triplex formation was demonstrated only *in vitro* using RNA-based electrophoretic mobility shift assays (EMSAs) and surface particle resonance (SPR) assays.

1.5 Challenges in triplex formation

Despite accumulating evidence for in cell triplex formation, most researchers in the lncRNA field[121;122] doubt the involvement of triplex formation in lncRNA-chromatin interactions. Several critical points have been raised throughout the period of my Ph.D. project (including personal correspondence on ncRNA Keystone conferences in 2015 and 2017) to which I would like to comment on in the next paragraphs (Figure 8):

1. Inaccessibility of genomic loci through heterochromatin formation

2. Physiological environment disfavors triplex formation

3. Lack of understanding of triplex code *in vivo* and *in vitro*

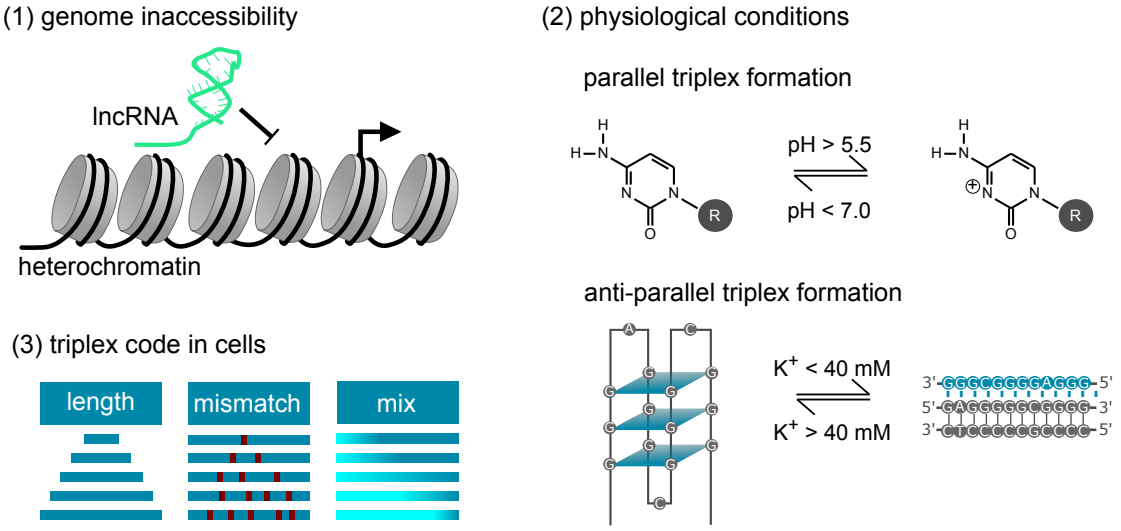


Figure 8| Challenging aspects in triplex formation. Triplex formation in cells is up until now slightly controversial due to three main challenges that need to be overcome and/or investigated: (1) The inaccessibility of triplex target sites due to densely packed chromatin. (2) Physiological environments disfavor triplex formation such as neutral pH disfavors parallel triplex formation and high monovalent ion concentrations such as potassium favor other non-canonical structures in G-rich stretches such as G-quadruplexes. (3) The underlying triplex code is still not completely understood and lacks answers to questions regarding the minimal length, how many mismatches can be tolerated and how triplexes can be formed with pyrimidine and purine mixes.

1.5.1 Inaccessibility of genomic loci through heterochromatin formation

Given the known compaction of eukaryotic genomes into tightly packed nucleosomes, investigators raised the question of how lncRNAs can penetrate such densely packed structures (Figure 8a). While studies have shown that short TFOs can target non-chromatinized (naked) genomic DNA *in vitro*[123], chromosomal DNA in permeabilized cells[124] or the genomic DNA in human fibroblast cells[107] and in mice[101], a conflicting study showed that targeting chromatinized DNA using TFOs was unsuccessful[125]. A follow-up study demonstrated that TTS located in actively transcribed genomic loci can be accessed by TFOs, while TFO access to the same TTS in silent loci seems to be inhibited[126]. However, given the recent finding that genomes are pervasively transcribed, one might argue that it depends on the dynamic interplay of various factors to determine when genomic loci are actually accessible[127].

1.5.2 Physiological environment disfavors triplex formation

In parallel triplex formation cytosines need to be protonated and, until recently, it was assumed that cytosine protonation only occurs in an acidic environment. Zeraati and colleagues however reported that the non-canonical i-motif structure does occur under physiological conditions[80]. Similar to the C-G*C^+ triplex, this structure requires a hemiprotonated cytosine to form a C^+:C basepair. The authors demonstrated that i-motif structures are found throughout human cells by developing an antibody against this structure and propose that molecular crowding, as well as cytosine modifications positively influence i-motif formation and stability.
Coming back to RNA as the third strand in triplexes, several studies suggested that N3 atoms in cytosines are thermodynamically favored for protonation compared to other residues in the cytosine base[128;129]. While most of the *in vitro* studies were carried out using a dilute polymeric

environment, physiological conditions are assumed to be characterized by intracellular macromolecular crowding effects. In accordance with Zeraati *et al.*, molecular crowding simulations using polyethylene glycol (PEG) have been shown to shift the apparent pKa of cytosine closer to pH neutrality thus favoring the formation of i-motifs[130] in physiological-like conditions. Similarly, PEG also increases the stability of the triplet T-A*T thereby favoring triplex formation and destabilizing duplex formation[130].
Interestingly, lncRNAs are known to be post-transcriptionally modified. Modifications such as conversion of cytosine to uracil and adenosine to inosine, known as RNA editing[131;132], or methylation of C5 in cytosine[133;134] contribute to overcoming the limitations thought to attenuate natural intermolecular triplex formation. And since it has been shown 30 years ago that inosine forms stable triplexes *in vitro*[86], this makes a strong argument for potential triplex formation of lncRNAs with genomic DNA. Consequently, the primary genomic sequence may not be the only determinant in the strength and specificity of triplex interactions.

1.5.3 Lack of understanding of triplex code *in vitro* and *in vivo*

One of the publications, which indicated that the promoter-associated pRNA is involved in triplex formation[10] (Figure 7b) and has been discussed earlier, is an interesting example that it is until now unclear what is required to form triplexes in cells. As mentioned above, triplexes can occur with either pyrimidine or purine motifs or mixed GT-motifs, but the putative triplex-forming sequence of the pRNA seems to be a unique pyrimidine/purine mix which does not correspond to the conservative motifs described earlier. The same research group recently published a study in which key questions regarding triplex structures, stability and the underlying triplex code *in vitro* were addressed[135]. Thus, both papers highlight the lack of knowledge regarding triplex structures *in vitro* and *in vivo* as well as the definition of high-affinity triplex target sites. Contributing to this lack is the fact that the publication from Wang *et al.*[100] using TFOs targeted to the genome in mammalian cells could not be reproduced in 2017[102] leaving the question of how reliable triplex formation occurs wide open.

1.6 Synthetic-biology approaches to study biological functions

Synthetic biology is an emerging field that combines ideas and concepts from multiple disciplines spanning biology, engineering, medicine and even philosophy. It centers around the design and construction of new artificial pathways, devices or chassis and the use of (non-)existent building blocks to replicate natural biological systems. In the past two decades, synthetic biology-inspired research drove innovations in agriculture, sustainable and renewable energy production, tissue engineering, targeted drug delivery and understanding biological design principles. Synthetic biology started off with the generation of genetic circuits such as the design of a synthetic oscillator in bacterial cells[136] and continued with the production of an anti-malaria drug precursor in engineered yeast cells[137] as well as the generation of the CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats-associated protein 9) genome editing toolbox for medical purposes[138;139;140]. With recent technological advances in next-generation sequencing, gene synthesis and genome assembly technologies[141], researchers aim to address question in understanding basic, underlying design principles in bacteria[142;143;144], yeast[145;146;147] and mammalian cells[148;79;149]. The concept of combining next-generation sequencing technologies with the rational synthetic biology design significantly expanded the understanding of mechanisms of regulatory gene functions on a plasmid as well as on a genome-wide scale. In this thesis I aimed to use those two powerful technologies to study triplex formation *in vitro* and in cells.

2 Research Objectives

For some researchers the mystery of the genomic dark matter seems to be solved: non-coding RNAs (ncRNAs) have been observed, and functions attributed. But this is where the questions start for me: how can ncRNAs exert their function and recognize genomic target sites? Can targeting DNA be as simple as using hydrogen interactions between three nucleic-acid strands (triplex formation)? And what might be the code between such triplex interactions that have potential far beyond merely understanding biological processes?

While the abundance of studies in the field of triplex formation is vast; direct evidence is circumstantial. Even after 60 years of research, there seems to be a lack of understanding of the underlying 'triplex code', the minimum length of purine/pyrimidine stretches that is required to from triplexes, and mixed purine/pyrimidine motif occurrences *in vitro* and *in vivo*. Moreover, the number of mismatches and locations within an oligo or lncRNA that can be tolerated has yet to be elucidated with more than only dozens of variants[94]. Finally, yet importantly the conflicting data and doubt in the lncRNA field of triplex-forming motives potentially being able to target genomic loci within the chromatin context puzzles researchers until today. Despite the intriguing hypotheses and preliminary results, detailed molecular mechanisms through which lncRNAs regulate transcriptional programs remain sparse. The postulated mechanisms of RNA*DNA-DNA triplex structures, 3D-proximity and protein-mediated screening for genomic target sites address intriguing questions, but lack systematic and high-throughput analysis to screen the vast sequence space.

To understand the triplex code that relies upon simple hydrogen interactions and decipher the minimal requirements for triplex formation could re-open the use of oligos in biotechnological applications and expand the existing toolbox and further strengthen the biological understanding of lncRNAs. Since lncRNAs are involved in multiple biological processes, and are associated in several diseases, dissecting the molecular lncRNA-mediated transcriptional mechanism would significantly contribute to the understanding of the complex eukaryotic gene expression and development of diseases.

In this work, I strove to understand triplex formation by using both a bottom-up and a top-down synthetic biology-based approach. For the bottom-up approach, I constructed synthetic biology-based massively parallel reporter assays using synthetic long-non-coding RNA as the desired regulatory subject that I wished to characterize. For the top-down approach, I developed high-throughput sequencing platforms for detection of triple-helix interactions both *in vitro* and *in vivo* to tackle basic questions that will shed light on how single-stranded nucleic acid molecules interact with double-stranded DNA. Furthermore, I aimed to dissect the proposed transcriptional mechanism of direct triplex formation in cells. Considering the complexity of transcriptional regulation in eukaryotic organisms, I applied the synthetic biology massively parallel reporter assay approaches in bacterial as well as eukaryotic cells to explore underlying molecular mechanisms. Thus, the thesis is divided into two main parts:

1. The synthetic-biology based circuits focus on building synthetic lncRNAs (slncRNAs) in bacterial and mammalian cells (starts on page 47).

2. The high-throughput deep-sequencing approach to study triplex formation *in vitro* and in cells (starts on page 65).

3 Material and Methods

3.1 Design of synthetic lncRNAs and triplex target sites (TTS)

3.1.1 Design and construction of synthetic lncRNAs

The synthetic long non-coding RNAs (slncRNAs) were designed using a rationale design and modular components. Each slncRNA consists of:

1. the DNA_{bind} motif that potentially forms triplexes with the respective triplex target site (TTS).
2. a short linker that connects the DNA_{bind} motif with the RBP_{bind} domain.
3. the RBP_{bind} motif that forms a hairpin structure with a protruding adenosine which can be recognized by the tandem-dimer PP7 phage coat protein (tdPCP).

Each module is mixed and matched with the two other modules resulting in a small library of slncRNAs. In Table 1, I listed all DNA sequences used to build the slncRNA molecules. The slncRNAs have been designed in such a way that they can be used in bacterial and mammalian systems.

Table 1| Sequences of each module in the designed slncRNA. Each slncRNA constitutes three parts (i) DNA_{bind} motif, (ii) linker and (iii) RBP_{bind} motif, which are mixed which each other and result in the library of slncRNAs shown in Table 2.

DNA_{bind} (5' –> 3')		**linker (5' –> 3')**		**RBP_{bind} (5' –> 3')**	
GAA	GAAGAAGAAG AAGAAGAAGA AGAA	**3 nt**	TTT	**$PP7_{x1}$**	TAAGGTACGTAATTGCCTAGAAAGGAGCAG ACGATATGGCGTCGCTCCCTGCAGCTCGAC
GGA	GGAGGAGGAG GAGGGGGAGG	**20 nt**	TCAATTGGAT TGTGCTATT	**$PP7_{x2}$**	TAAGGTACGTAATTGCCTAGAAAGGAGCAG ACGATATGGCGTCGCTCCCTGCAGCTCGAC ACTAGAAACCAGCAGAGCATATGGGCTCGC TGGCTGCAGTATTCCCGGCTTCATTAGATC C
pyr_{rich}	GCTCTTCTTT TCTTTCGG	**40 nt**	AAAACACCCA GGTCGAATAC ATATAAAATC TACACTACGT	**$PP7_{x3}$**	TAAGGTACGTAATTGCCTAGAAAGGAGCAG ACGATATGGCGTCGCTCCCTGCAGCTCGAC ACTAGAAACCAGCAGAGCATATGGGCTCGC TGGCTGCAGTATTCCCGGCTTCATTAGATC CTAAGGTACGTAATTGCCTAGAAAGGAGCA GACGATATGGCGTCGCTCCCTGCAGCTCGA C

Table 1| Sequences of each module in the designed slncRNA. Each slncRNA constitutes three parts (i) DNA_{bind} motif, (ii) linker and (iii) RBP_{bind} motif, which are mixed which each other and result in the library of slncRNAs shown in Table 2.

DNA_{bind} (5' –> 3')		linker (5' –> 3')		RBP_{bind} (5' –> 3')	
GAA	GAAGAAGAAG AAGAAGAAGA AGAA	**3 nt**	TTT	**$PP7_{x1}$**	TAAGGTACGTAATTGCCTAGAAAGGAGCAG ACGATATGGCGTCGCTCCCTGCAGCTCGAC
AA_{rich}	AAGGAAAGGA AAAAGAAAAG AGA			**$PP7_{x4}$**	TAAGGTACGTAATTGCCTAGAAAGGAGCAG ACGATATGGCGTCGCTCCCTGCAGCTCGAC ACTAGAAACCAGCAGAGCATATGGGCTCGC TGGCTGCAGTATTCCCGGCTTCATTAGATC CTAAGGTACGTAATTGCCTAGAAAGGAGCA GACGATATGGCGTCGCTCCCTGCAGCTCGA CACTAGAAACCAGCAGAGCATATGGGCTCG CTGGCTGCAGTATTCCCGGCTTCATTAGAT CC
T0	AGGTCGACCA GTTGTTCC			**$PP7_{x5}$**	TAAGGTACGTAATTGCCTAGAAAGGAGCAG ACGATATGGCGTCGCTCCCTGCAGCTCGAC ACTAGAAACCAGCAGAGCATATGGGCTCGC TGGCTGCAGTATTCCCGGCTTCATTAGATC CTAAGGTACGTAATTGCCTAGAAAGGAGCA GACGATATGGCGTCGCTCCCTGCAGCTCGA CACTAGAAACCAGCAGAGCATATGGGCTCG CTGGCTGCAGTATTCCCGGCTTCATTAGAT CTAAGGTACGTAATTGCCTAGAAAGGAGCA GACGATATGGCGTCGCTCCCTGCAGCTCGA C

The modular design described above allows for the construction of bacterial and mammalian vectors. Table 2 lists all plasmids that were generated for both the bacterial and mammalian systems.

Table 2| List of slncRNAs for both bacterial and mammalian plasmids. slncRNAs have been designed in such a way that they can be subcloned into both the bacterial recipient vector and the mammalian recipient vector. The abbreviations used to describe the slncRNA motifs are as follows: the first corresponds to the DNA_{bind} motif, the second to the length of the linker in nucleotides and the third to the number of PP7-binding sites. The x in the plasmid column refers to x = 1 or 3; in which 1 refers to the bacterial vector and 3 to the mammalian vector. While all bacterial plasmids were tested in this study, only some of the plasmids were tested in mammalian cells. These plasmids are highlighted in blue.

plasmid	motif	plasmid	motif	plasmid	motif
pRNAx001	1-GAA-3-PP7x1	pRNAx024	24-AA-40-PP7x2	pRNAx047	47-AA-20-PP7x4
pRNAx002	2-GAA-20-PP7x1	pRNAx025	25-GAA-3-PP7x3	pRNAx048	48-AA-40-PP7x4

Table 2| List of slncRNAs for both bacterial and mammalian plasmids. slncRNAs have been designed in such a way that they can be subcloned into both the bacterial recipient vector and the mammalian recipient vector. The abbreviations used to describe the slncRNA motifs are as follows: the first corresponds to the DNA_{bind} motif, the second to the length of the linker in nucleotides and the third to the number of PP7-binding sites. The x in the plasmid column refers to x = 1 or 3; in which 1 refers to the bacterial vector and 3 to the mammalian vector. While all bacterial plasmids were tested in this study, only some of the plasmids were tested in mammalian cells. These plasmids are highlighted in blue.

plasmid	motif	plasmid	motif	plasmid	motif
pRNAx003	3-GAA-40-PP7x1	pRNAx026	26-GAA-20-PP7x3	pRNAx049	49-pyr-3-PP7x4
pRNAx004	4-GGA-3-PP7x1	pRNAx027	27-GAA-40-PP7x3	pRNAx050	50-pyr-20-PP7x4
pRNAx005	5-GGA-20-PP7x1	pRNAx028	28-GGA-3-PP7x3	pRNAx051	51-pyr-40-PP7x4
pRNAx006	6-GGA-40-PP7x1	pRNAx029	29-GGA-20-PP7x3	pRNAx052	52-T0-3-PP7x4
pRNAx007	7-AA-3-PP7x1	pRNAx030	30-GGA-40-PP7x3	pRNAx053	53-T0-20-PP7x4
pRNAx008	8-AA-20-PP7x1	pRNAx031	31-AA-3-PP7x3	pRNAx054	54-T0-40-PP7x4
pRNAx009	9-AA-40-PP7x1	pRNAx032	32-AA-20-PP7x3	pRNAx055	55-GAA-3-PP7x5
pRNAx0x0	x0-pyr-3-PP7x1	pRNAx033	33-AA-40-PP7x3	pRNAx056	56-GAA-20-PP7x5
pRNAx011	11-pyr-20-PP7x1	pRNAx034	34-pyr-3-PP7x3	pRNAx057	57-GAA-40-PP7x5
pRNAx012	12-pyr-40-PP7x1	pRNAx035	35-pyr-20-PP7x3	pRNAx058	58-GGA-3-PP7x5
pRNAx013	13-T0-3-PP7x1	pRNAx036	36-pyr-40-PP7x3	pRNAx059	59-GGA-20-PP7x5
pRNAx014	14-T0-20-PP7x1	pRNAx037	37-T0-3-PP7x3	pRNAx060	60-GGA-40-PP7x5
pRNAx015	15-T0-40-PP7x1	pRNAx038	38-T0-20-PP7x3	pRNAx061	61-AA-3-PP7x5
pRNAx016	16-GAA-3-PP7x2	pRNAx039	39-T0-40-PP7x3	pRNAx062	62-AA-20-PP7x5
pRNAx017	17-GAA-20-PP7x2	pRNAx040	40-GAA-3-PP7x4	pRNAx063	63-AA-40-PP7x5
pRNAx018	18-GAA-40-PP7x2	pRNAx041	41-GAA-20-PP7x4	pRNAx064	64-pyr-3-PP7x5
pRNAx019	19-GGA-3-PP7x2	pRNAx042	42-GAA-40-PP7x4	pRNAx065	65-pyr-20-PP7x5
pRNAx020	20-GGA-20-PP7x2	pRNAx043	43-GGA-3-PP7x4	pRNAx066	66-pyr-40-PP7x5
pRNAx021	21-GGA-40-PP7x2	pRNAx044	44-GGA-20-PP7x4	pRNAx067	67-T0-3-PP7x5
pRNAx022	22-AA-3-PP7x2	pRNAx045	45-GGA-40-PP7x4	pRNAx068	68-T0-20-PP7x5
pRNAx023	23-AA-20-PP7x2	pRNAx046	46-AA-3-PP7x4	pRNAx069	69-T0-40-PP7x5

While all bacterial plasmids were tested and results are shown in this thesis, only a subset all mammalian plasmids has been tested and will be shown in the results section of this work. The samples that have been used in the mammalian system are highlighted in blue. The table shows

the plasmid name as well as a short description of the slncRNA starting with the DNA_{bind} motif, followed by the linker length and ending with the number of PP7-binding sites (RBP_{bind} motif).

3.1.2 Design and construction of TTS

The triplex target sites (TTS) are the double-stranded corresponding target sequences for the DNA_{bind} motif and have been described previously. For the purpose of gene activation, I constructed TTS that can be inserted in the template as well as non-template strand of the bacterial and mammalian reporter plasmids to maximize the potential gene regulatory effect. Furthermore, two different TTS insertion sites upstream of the mammalian minimal CMV (cytomegalovirus) promoter were chosen (110 bp and 145 bp upstream of the promoter) which generates double the amount of reporter plasmids for the mammalian system compared to the bacterial one. All reporter plasmids were tested in bacterial cells, while only a subset of reporters were used in the mammalian setup. The plasmids that have been tested are highlighted in blue.

Table 3| List of triplex target sites. The mammalian system contains all mentioned triplex target sites (TTS), whereas the bacterial system comprises only the even numbers (such as #70, 72 etc.). The nomenclature of the reporters is as follows: the first number identifies each reporter plasmid, the name target indicates that the plasmids are reporter plasmids and contain a TTS. The next name shows which TTS motif was used, the next letters indicate whether the TTS has been cloned into the template (temp.) or non-template strand (nontemp) and the last numbers indicate the TTS insertions site which is relevant for the mammalian system. The reporters that have been tested in the mammalian system are highlighted in blue.

plasmid	**sequence (5'–> 3')**
79-target AA-temp-145	AGAGAAAAGAAAAAGGAAAGGAA
70-target GAA-temp-110	CTTCTTCTTCTTCTTCTTCTTCTT
71-target GAA-temp-145	CTTCTTCTTCTTCTTCTTCTTCTT
72-target GAA-nontemp-110	AAGAAGAAGAAGAAGAAGAAGAAG
73-target GAA-nontemp-145	AAGAAGAAGAAGAAGAAGAAGAAG
74-target GGA-temp-110	CCTCCTCCTCCTCCCCCTCC
75-target GGA-temp-145	CCTCCTCCTCCTCCCCCTCC
76-target GGA-nontemp-110	GGAGGGGGAGGAGGAGGAGG
77-target GGA-nontemp-145	GGAGGGGGAGGAGGAGGAGG
78-target AA-temp-110	AGAGAAAAGAAAAAGGAAAGGAA
80-target AA-nontemp-110	TTCCTTTCCTTTTTCTTTTCTCT
81-target AA-nontemp-145	TTCCTTTCCTTTTTCTTTTCTCT
82-target pyr-temp-110	GCTCTTCTTTTCTTTCGG
83-target pyr-temp-145	GCTCTTCTTTTCTTTCGG
84-target pyr-nontemp-110	AGGAACAACTGGTCGACCT
85-target pyr-nontemp-145	AGGAACAACTGGTCGACCT
86-target T0-temp-110	AGGTCGACCAGTTGTTCCT

Table 3| List of triplex target sites. The mammalian system contains all mentioned triplex target sites (TTS), whereas the bacterial system comprises only the even numbers (such as #70, 72 etc.). The nomenclature of the reporters is as follows: the first number identifies each reporter plasmid, the name target indicates that the plasmids are reporter plasmids and contain a TTS. The next name shows which TTS motif was used, the next letters indicate whether the TTS has been cloned into the template (temp.) or non-template strand (nontemp) and the last numbers indicate the TTS insertions site which is relevant for the mammalian system. The reporters that have been tested in the mammalian system are highlighted in blue.

plasmid	**sequence (5'–> 3')**
79-target AA-temp-145	AGAGAAAAGAAAAAGGAAAGGAA
87-target T0-temp-145	AGGTCGACCAGTTGTTCCT
88-target T0-nontemp-110	AGGAACAACTGGTCGACCT
89-target T0-nontemp-145	AGGAACAACTGGTCGACCT
90-target T0-temppar-110	TCCTTGTTGACCAGCTGGA
91-target T0-temppar-145	TCCTTGTTGACCAGCTGGA

3.2 Cloning of bacterial and mammalian plasmids

All constructs have been cloned using standard molecular biology techniques such as polymerase chain reaction (PCR), restriction eznyme digestion, oligonucleotide annealing and enzymatic assembly of DNA molecules, also referred to as Gibson cloning[141], standard ligations and heat-shock transformation of Top10 cells (Invitrogen). Further details of libraries and constructs generated can be found in respective sections for design and construction of respective vectors.

3.3 Design and construction of bacterial vectors

To insert the slncRNA (Gen9) and TTS (annealed oligos, Sigma-Aldrich) libraries, two main bacterial recipient plasmids (plasmids that can be used to subclone the libraries) have been constructed. The description of the cloning process in shown in Table 4.

1. The pRNA plasmid comprises the DNA-encoded slncRNAs with DNA_{bind}, linker and PP7-binding sites which are placed under the C_4-HSL inducible promoter RhlR. All plasmids that were generated are listed in Table 2.

2. The pRep plasmid contains the enhancer-like cassette (modified from Amit *et al.*[142]) with or without putative TTS. All constructs that were cloned are shown in Table 3.

Table 4| Description of cloning of bacterial pRNA and pRep plasmid. The slncRNAs have been subcloned into a recipient vector that contains the C_4-HSL inducible promoter rhlR and the TTS were inserted into the synthetic enhancer plasmid designed by Amit and colleagues [142].

Plasmid	Description of cloning	Backbone/ Antibiotic rest.
pRNA1000 (recipient vector)	Ligation of the gBlock "promoter-mCherry-terminator" (KpnI, BglII, XbaI) into the bacterial A133 plasmid (KpnI, XbaI) thereby replacing tdPCP-FP and resulting in a C_4-HSL-inducible, bacterial vector which can be used for expression of slncRNAs.	A133-pRhlR/ AmpR
pRNA10xx	Ligation of digested, linear fragments obtained from Gen9 (XbaI, KpnI) into pRNA1000 (XbaI, KpnI) resulting in a vector with inducible transcription of different slncRNAs. The full slncRNA library has been constructed in this way. Sequences of individual slncRNA modules are described; x = #01-#69	A133-pRhlR/ AmpR
pRep0000	Ligation of annealed oligos into pLP-RbsK-RA51-mCh-correct2 [142] (NheI) resulting in theh reporter plasmid lacking any putative triplex target site (TTS) This plasmid can additionally be used to insert putative TTS via digestion with BsaI.	pLP-RbsK-RA51-mCh-correct2/ KanR
pRepx0xx	Ligation of annealed oligos into pRep0000 (BsaI). This generates the TTS reporter library.	pLP-RbsK-RA51-mCh-correct2/ KanR

3.4 Design and construction of mammalian vectors

Four different plasmids are required for the functionality of the slncRNA-dependent gene activation system and the description of the modular constructions of these plasmids has been listed in Table 5. The four plasmids that are required have the following features:

1. the pRNA plasmid transcribes the slncRNA and a *sbfp2* (strongly enhanced blue fluorescent protein 2*)* gene under a strong, constitutive CMV promoter. The slncRNA and the SBFP2 mRNAs are separated by Csy4 recognitions sites. The library of slncRNA plasmids that was generated can be seen in Table 2.

2. the pRBP plasmid encoding the fusion protein mKate2-vp64-tdPCP-NLS. mKate2 is the

fluorescent marker protein to confirm expression of the fusion protein, vp64 is the viral transactivator[150;151;152] which has been used to activate gene expression in synthetic system, tdPCP is the tandem dimer phage coat protein PP7[153] which binds with high-affinity to the RBP_{bind} hairpin structure on the slncRNAs and the nuclear localization signal (NLS) which localizes the fusion protein to the nucleus.

3. the reporter plasmid pRep carrying putative triplex target sites (TTS) upstream of a *eYFP* (enhanced yellow fluorescent protein) reporter gene. The TTS are inserted the template or non-template strand and 110 nt as well as 145 nt upstream of the minimal promoter CMV_{min}. All plasmids that were generated are listed in Table 3.

4. the plasmid pCsy4 encoding the Csy4 endonuclease which recognizes its cognate Csy4 binding sites on the slncRNA molecules and cleaves the slncRNA from the sbfp mRNA.

Table 5| List of plasmids for eukaryotic expression system. List of plasmids required for slncRNA-dependent mammalian expression system and plasmids used for microscopy/flow cytometry analysis.

Plasmid	**Description of cloning**	**Backbone/ Antibiotic rest.**
pRBP	pPolII_mKate2_vp64_tdPCP-nls: Gibson Assembly of PCR products of mKate2, vp64 and tdPCP-nls into pMS2-GFP (SpeI, ClaI) resulting in a constitutively expressing vector comprising the $polII_{subunit}$ promoter.	pEGFP/AmpR
pMS2-GFP (recipient vector)	Addgene, #27121, pMS2-GFP_dlFG_V29I	pEGFP/AmpR
pRNA3000 (recipient vector)	Addgene, #22880, pSBFP2-C1	pEGFP-C1/KanR
pRNA30xx	Ligation of digested Gen9 constructs (XbaI, NheI) into pRNA3000 (XbaI, NheI) resulting in a pRNA30xx plasmid encoding the slncRNA with varying DNA_{bind} and RBP_{bind} motif; x = #01-69	pEGFP-C1/KanR
pRep3000 (recipient vector)	Addgene, #55197, P1-EYFP-pA; plasmid has been kindly provided by Lior Nissim, MIT, Boston, USA[154]	pGL5-Luc/ AmpR

Table 5| List of plasmids for eukaryotic expression system. List of plasmids required for slncRNA-dependent mammalian expression system and plasmids used for microscopy/flow cytometry analysis.

Plasmid	**Description of cloning**	**Backbone/ Antibiotic rest.**
pRBP	pPolII_mKate2_vp64_tdPCP-nls: Gibson Assembly of PCR products of mKate2, vp64 and tdPCP-nls into pMS2-GFP (SpeI, ClaI) resulting in a constitutively expressing vector comprising the $polII_{subunit}$ promoter.	pEGFP/AmpR
pMS2-GFP (recipient vector)	Addgene, #27121, pMS2-GFP_dlFG_V29I	pEGFP/AmpR
pRep30xx	Gibson Assembly of Gen9 (#70-96) into pRep3000 (NotI, NheI) resulting in the reporter plasmids containing the putative triplex target site (TTS) upstream of the minimal promoter $CMV_{min}$and the *eYFP* gene.	pGL5-Luc/ AmpR
pPGK1-Csy4	Endonuclease Csy4; plasmid has been kindly provided by Lior Nissim, MIT, Boston, USA [154]	pGL2-Luc/AmpR
pCMV-eYFP	Gibson Assembly of digested CMV promoter (KpnI, SacI) and Reverse PCR from #55197 resulting in a strong constitutive expression vector of eYFP that can be used for compensation control in flow cytometry and microscopy.	pGL5-Luc/AmpR

Table 5| List of plasmids for eukaryotic expression system. List of plasmids required for slncRNA-dependent mammalian expression system and plasmids used for microscopy/flow cytometry analysis.

Plasmid	Description of cloning	Backbone/ Antibiotic rest.
pRBP	pPolII_mKate2_vp64_tdPCP-nls: Gibson Assembly of PCR products of mKate2, vp64 and tdPCP-nls into pMS2-GFP (SpeI, ClaI) resulting in a constitutively expressing vector comprising the $polII_{subunit}$ promoter.	pEGFP/AmpR
pMS2-GFP (recipient vector)	Addgene, #27121, pMS2-GFP_dlFG_V29I	pEGFP/AmpR
pCMV-mKate2 (#55200)	CMVp-dsRed2-Triplex-28-gRNA1-28: that can be used for fluorescence microscopy and flow cytometry. Plasmid has been kindly provided by Lior Nissim, MIT, Boston, USA [154]	pGL2-Luc/AmpR

3.5 Bacterial enhancer-slncRNA bioassay

To test the bacterial enhancer-assay, the automated liquid handling platform Freedom Evo (Tecan Group) was used. Briefly, co-transformed Top10 cells (i.e. Top10 cells containing an enhancer-based reporter plasmid pRep and a pRNA plasmid encoding the slncRNA) were grown overnight in 1.5 mL Luria Bertani (LB) medium complemented with appropriate antibiotics (100 µg/mL ampicillin (amp), 15 µg/mL kanamycin (kan)). Subsequently, the overnight culture was diluted 1:100 in low growth and low auto-fluorescence bioassay buffer (BA: 0.5 g/L tryptone, 0.3 mL/L glycerol, 86 mM NaCl, 0.05 $MgSO_4$, 1 mL/L 10xPBS pH7.4), supplemented with 5 % LB medium and appropriate antibiotics (kan/amp). Cells were dispensed into black 96-well plates (Greiner bio-one) that are compatible with fluorescent microplate reader measurements. Subsequently, 24 C_4-HSL (Cayman Chemical) inducer concentrations (0-300 µM) were added to the bacterial strains and incubated at 37 °C in humidified atmosphere while vigorously shaking. Starting 2 hours post induction, fluorescence measurements were taken every 30 minutes over a period of 8 hours to cover mid-log growth range. mCherry and Cerulean intensities were measured using the microplate reader Infinite Pro200 (Tecan Group) and excitation and emission filters are listed in Table 6. Fluorescence values (FL) were averaged, normalized by dividing FL by OD_{600}, and background auto-fluorescence was eliminated based on non-transformed Top10 cells. The normalized values were further processed by dividing FL/OD_{600} values by the average of the two minimal fluorescence levels in absence of C_4-HSL an at 0.018 µM C_4-HSL inducer concentrations. These values were termed "fold-change" and the distribution of fold-change values were plotted against increasing C_4-HSL concentrations.

Table 6| Excitation emission filters used for microplate reader assays.Two different filter sets have been used for the two fluorescent proteins (FP) Cerulean and mCherry.

FP	Excitation filter [nm]	Emission filter [nm]	Gain
Cerulean	420/10	485/10	55
mCherry	560/10	610/10	70

3.6 Mammalian activation-based slncRNA bioassay

3.6.1 Cell culture

The human embryonic kidney cell line (HEK-293, kindly provided by Arie Admon's lab, Technion) was incubated and maintained in 100x20 mm cell culture dishes (Nunclon cell culture treated, Thermo Scientific) under standard cell culture conditions at 37 °C in humidified atmosphere containing 5 % CO_2 and were passaged at 80-85 % confluence. Cells were washed once with 1x DPBS (Dulbecco's phosphate buffer saline, Biological Industries), and subsequently treated with 1 mL trypsin/EDTA (ethylenediaminetetraacetic acid, Biological Industries) followed by incubation at 37 °C for 1-2 minutes. $DMEM_{complete}$ (Dulbecco Eagle's Minimum Essential Medium, Biological Industries), complemented with 10 % FBS (fetal bovine serum, Biological Industries, Lot.no: 1418110) and final concentrations of 100 U penicillin plus 100 µg streptomycin (Biological Industries), was added and transferred into fresh $DMEM_{complete}$ in subcultivation ratios of 1:10.

3.6.2 Transient transfection

Either 1.2x10^4 HEK-293 cells were seeded in a 96-well tissue culture plate (Nunclon cell culture treated, Thermo Scientific) or 1.3x10^5 HEK-293 cells were seeded in 24-cell culture plates (Thermo Scientific) 24 hours prior transfection. At time of transfection, 0.1-0.5 µg DNA (depending on plate format used) was mixed with 0.3-1.5 µL of the transfection reagent TransIT-LT1 (Mirus Bio LLC), respectively. Volume was adjusted with OptiMEM (Gibco/Life Technologies) to 10-50 µL per well (depending on plate format used). DNA/TransIT-LT1 mix was incubated for 15 min at room temperature (RT) and subsequently added drop-wise to the cells.

3.6.3 Flow cytometry

48-72 h post-transfection, HEK-293 cells were washed once with 1xDPBS and incubated for 1-2 minutes at 37 °C with trypsin/EDTA. Trypsin was inactivated with 1xDPBS complemented with 1 % FBS and 3 mM ethylenediaminetetraacetic acid (EDTA, J.T. Baker (now available through Thermo Scientific). Data acquisition was performed on the LSRII Analyzer (BD Bioscience) using the 96-well mode of the FACS Diva Software (BD Bioscience) or on the MACSQuant (Miltenyi Biotec) analyzer using the proprietary MACSQuantify software . Usage of type of flow cytometer will be indicated in results. Laser-line, bandpass emission filters and detectors for respective fluorescent proteins (FPs) of both flow cytometry analyzers are listed in Table 7.

Data collected from the experiments were analyzed using FlowJo analysis software (FlowJo LLC). 1x10^4living cells were analyzed for each sample. Parameters during data acquisition were set as follows: Forward side scatter (FSC) was used as discriminator. Two-dimensional dot plots (FSC/SSC) were used to define population of living cells (gating tree), whereas histograms

Table 7| Laser and emission filters of flow cytometry analyzers used for data acquisition. Three different laser-lines were used in combination with three bandpass emission filters to measure fluorescence intensities of SBFP, eYFP and mKate2 for (A) LSRII Analyzer and (B) MACSQuant. FP, Fluorescent protein

(A) LSRII Analyzer

Laser (Excitation)	Bandpass filter (Emission)	Detector	FP
405 nm (25 mW)	450/50 nm	B	SBFP2
488 nm (22 mW)	530/30 nm	D	eYFP
633 nm (20 mW)	660/20 nm	B	mKate2

(B) MACSQuant

Laser (Excitation)	Bandpass filter (Emission)	Detector	FP
405 nm (40 mW)	450/50 nm	V1	SBFP2
488 nm (50 mW)	525/50 nm	B1	eYFP
561 nm (100 mW)	661/20 nm	Y3	mKate2

were adjusted by altering the voltage and gain to determine auto-fluorescence of non-transfected cells which was set to 0.5-1.0 % positive cells. Compensation was performed with the FlowJo analysis software after data acquisition to remove false-positive cells. The parameters used for compensation are shown in Table 8. The percentage of eYFP, mKate2 and SBFP2-positive cells as well as the median fluorescence intensities were exported and used to calculate eYFP fold-change and compare among data sets. To compute fold-change values for the eYFP expression, first the weighted median eYFP values were calculated as previously described [154] and described by:

$$weighted\,median\,eYFP = \%\,positive\,cells * median\,eYFP\,intensity \tag{1}$$

To compare among samples with and without slncRNAs, the weighted median eYFP values were normalized by dividing the weighted median eYFP levels with or without slncRNA respectively by the weighted median eYFP values obtained for reporter plasmids only:

$$normalized\,eYFP = \frac{weighted\,median\,eYFP\,(+/-\ slncRNA)}{weighted\,median\,eYFP\,(reporter\,only)} \tag{2}$$

Lastly, the ratio of these normalized eYFP values (+/- slncRNAs) were computed:

$$fold\,change = \frac{weighted\,median\,eYFP\,(+slncRNA)}{weighted\,median\,eYFP\,(-slcnRNA)} \tag{3}$$

3.7 Triplex-Seq

3.7.1 Design of oligonucleotides (oligos) and primers for Triplex-Seq

The developed Triplex-Seq assay requires oligos which are referred to as triplex-forming oligos (TFOs), triplex target sites (TTS), DNA adapters which are ligated to TFOs in the downstream

Table 8| Compensation matrix of flow cytometry analysis with spillover values Spillover values in compensation matrix were generated using the FlowJo compensation tool after data acquisition.

	SBFP2	**eYFP**	**mKate2**
SBFP2	100	1.58	2.92
eYFP	0.003	100	0.02
mKate2	0.29	8.52	100

protocol and primers for PCR amplification. In the next paragraphs, detailed descriptions of the design of all oligos, adapters and primers are given.

3.7.2 Design of triplex-forming oligonucleotides (TFOs)

The TFOs were designed with the following features and were ordered as single-stranded, desalted, non-modified DNA (ssDNA) oligonucleotides (oligos) from integrated DNA technologies (IDT). Each TFO consists of two parts: (i) a common capture sequence and (ii) a triplex-forming sequence:

1. The 19 nt long capture sequence consists of fixed bases and serves as a platform for PCR amplification (*in vitro* and in cell Triplex-Seq) as well as binding platform for a complementary biotinylated capture oligo (in cell Triplex-Seq) to enrich the TFOs. The capture sequence is the same for all TFOs that were tested in this study.

2. For the positive controls for triplex formation, the 20-30 nt long triplex-forming sequence consists of either adenine and guanine bases (anti-parallel TFO)[97;102] or cytosine and thymine bases (parallel TFO)[155]. The TFO libraries were synthesized using the mixed bases tool (standard) from IDT. Mixed base oligos are synthesized according to the International Union of Pure and Applied Chemistry (IUPAC) convention (Table 9). During chemical synthesis, each mixed base is integrated with percentages between 25 % (for 'N' mixed bases), 33 % (for 'D' and 'B' mixed bases) and 50 % (for 'R', 'S', 'Y', 'W', 'M', 'K' mixed bases) for every choice of base.

Table 9| IPUAC base code used in this study. All TFOs were synthesized according to the the International Union of Pure and Applied Chemistry.

Mixed bases used	**Mixed base code**
A, G	R
C, T	Y
A, C	M
G, T	K
G, C	S
A, T	W
G, C, T	B
A, G, T	D
A, C, G, T	N

Control TFOs were generated (i) without the triplex-forming sequence (adapter only) or (ii) without the adapter sequence. Full lists of all control TFOs (Table 10), TFO libraries (Table 11) and verification-TFOs (Table 12) are shown below. Each table includes the TFO name, sequence and number of variants per TFO library.
The TFOs that served as positive controls were used to verify that triplex formation is possible in our lab and two of these TFOs (TFO_AG30 and TFO_TC) were used in the *in vitro* Triplex-Seq protocol to confirm in every experiment that the conditions are optimal to induce triplex formation (Table 10).

Table 10| Literature TFOs and control TFOs for *in vitro* Triplex-Seq.All TFOs that were used to confirm triplex formation *in vitro* with positive controls from literature and designed TFOs in this study with the support of the prediction software "triplexator" [117].

name	**TFO sequence (5' –> 3')**	**length [nt]**	**reference**
TFO_AG30	AGGAAGGGGGGGGTGGTGGGGGAGGGGGAG	30	Wang *et al.* [97]
TFO_TC	CCTCTTCCTCCTCTTCCTTTCTC	23	Chiou *et al.* [155]
TFO_p _G20	CTTCAGCTTGGCGGTCTGGCTTTTCTCTTTTCTTCTTTT	39	this work
TFO_ap_G84	CTTCAGCTTGGCGGTCTGGGGAGGGGGAGGGGGGGGAGG	39	this work

The TFO libraries were constructed as described above and are listed in Table 11 including a short description highlighting the number of fixed bases, the mixed bases used, the sequence as it was ordered, the total length of the TFO and the number of variants for each library. Furthermore, single-variant TFOs were ordered based on the enrichment of sequences after the Triplex-Seq analysis. In Table 11 and Table 12 below, all ordered and tested TFO sequences are listed.

Table 11| List of TFO libraries for the Triplex-Seq approaches. Each TFO library was synthesized according to the IUPAC convention described on page 29. The name and the description indicate the nature of the TFO (e.g. which mixed base was used, how many fixed bases can be found within the TFO and whether a capture sequence was attached), the sequence, length and the number of variants for each library follow the description.

name	**description**	**TFO sequence (5' –> 3')**	**length [nt]**	**variants**
D-TFO	20 nt TFO with 7 fixed positions - G,A,T	CTTCAGCTTGGCGGT CTGGDDADTDDGAD DDDDTDDADG	39	$1.59x10^6$
D-TFO (PCR)	20 nt TFO with 7 fixed positions - G,A,T	CTTCAGCTTGGCGGT CTGGDDADTDDGAD DDDDTDDADGAGATCGGA AGAGCACACGTCTGAACT CCAGTCAC	73	$1.59x10^6$
B-TFO	20 nt TFO with 7 fixed positions - G,C,T	CTTCAGCTTGGCGGTCTG GCBGBBCBBBBBTCBBCB GBB	39	$1.59x10^6$

Table 11| List of TFO libraries for the Triplex-Seq approaches. Each TFO library was synthesized according to the IUPAC convention described on page 29. The name and the description indicate the nature of the TFO (e.g. which mixed base was used, how many fixed bases can be found within the TFO and whether a capture sequence was attached), the sequence, length and the number of variants for each library follow the description.

name	**description**	**TFO sequence (5' –> 3')**	**length [nt]**	**variants**
3D-TFO (stretch)	20 nt TFO with 3 bases stretch - G,A,T	CTTCAGCTTGGCGGTCTG GCATCCTGADDDCTGACA TGC	39	27
3B-TFO (stretch)	20 nt TFO with 3 bases stretch - G,C,T	CTTCAGCTTGGCGGTCTG GCATCGTAABBBAAGACA TGC	39	27
5D-TFO (stretch)	20 nt TFO with 5 bases stretch - G, A,T	CTTCAGCTTGGCGGTCTG GCATCCTGDDDDDTGACA TGC	39	243
5B-TFO (stretch)	20 nt TFO with 5 bases stretch - G,C,T	CTTCAGCTTGGCGGTCTG GCATCGTGBBBBBAGACA TGC	39	243
R-TFO	20 nt TFO with 6 fixed positions - G, A	CTTCAGCTTGGCGGTCTG GRRGRTRRRARRRRRGRR ART	39	$1.63x10^4$
M-TFO	20 nt TFO with 6 fixed positions - ctrl with capture and only C,A	CTTCAGCTTGGCGGTCTG GTMAMMGMMMMMCTMMMM AMM	39	$1.63x10^4$
K-TFO	20 nt TFO with 6 fixed positions - G,T	CTTCAGCTTGGCGGTCTG GCKGKKTKKKKKGTKKKK AKK	39	$1.63x10^4$
Y-TFO	20 nt TFO with 6 fixed positions - C,T	CTTCAGCTTGGCGGTCTG GGYCYYCYYYYYCTYYYY GYY	39	$1.63x10^4$
N-TFO	20 nt TFO with 6 fixed positions - G,A,T,C	CTTCAGCTTGGCGGTCTG GCNANNTNNNNNTGNNNN CNG	39	$2.68x10^7$
7D-TFO (stretch)	20 nt TFO with 7 bases stretch - G,A,T	CTTCAGCTTGGCGGTCTG GCATCCTDDDDDDDGACA TGC	39	2187
7B-TFO (stretch)	20 nt TFO with 7 bases stretch - G,C,T	CTTCAGCTTGGCGGTCTG GCATCGTBBBBBBBGACA TGC	39	2187

Table 11| List of TFO libraries for the Triplex-Seq approaches. Each TFO library was synthesized according to the IUPAC convention described on page 29. The name and the description indicate the nature of the TFO (e.g. which mixed base was used, how many fixed bases can be found within the TFO and whether a capture sequence was attached), the sequence, length and the number of variants for each library follow the description.

name	description	TFO sequence (5' –> 3')	length [nt]	variants
9D-TFO (stretch)	20 nt TFO with 9 bases stretch - G,A,T	CTTCAGCTTGGCGGTCTG GCATCCDDDDDDDDDACA TGC	39	$1.96x10^4$
9B-TFO (stretch)	20 nt TFO with 9 bases stretch - G,C,T	CTTCAGCTTGGCGGTCTG GCATCGBBBBBBBBBACA TGC	39	$1.96x10^4$
W-TFO	20 nt TFO with 6 fixed positions - T,A	CTTCAGCTTGGCGGTCTG GTWAWWWWWTWWWWAWW TWA	39	$1.63x10^4$

Table 12| List of TFOs used in verification experiments of Triplex-Seq. Following NGS analysis of the Triplex-Seq reads that were enriched in the downstream Triplex-Seq protocol, several single-variants of the most reactive ('positive TFOs') and least reactive hits ('negative TFOs') of the triplex band from the *in vitro* Triplex-Seq protocol were ordered and tested. The descriptions in the second column highlights in brief from which library that was tested and analyzed the TFO single-variants are derived from.

name	description	sequence (5'–> 3')	length [nt]
N-TFO pos_3	third hit (enriched in triplex band) of N-TFO library tested in pH 7 condition	CTAGTTGGGGGTGGGGGCGG	20
N-TFO neg_1	last hit (non-enriched in triplex band) of N-TFO library tested in pH 7 condition	CGAGGTTATGATGAAACCGG	20
G80_motif1	first hit (enriched in triplex band) of N-TFO with TTS (80 % guanine) in pH 7	CGAGGTGGGGGTGTTGCCGG	20
G80_motif2	second hit (enriched in triplex band) of N-TFO with TTS (80 % guanine) in pH 7	CTAGTTGGGGGTGGGAGCGG	20
G80_motif3	third hit (enriched in triplex band) of N-TFO with TTS (80 % guanine) in pH 7	CTAGGTGGGGGTGGGGGCTG	20

3.7.3 Design of triplex target sites (TTS)

The TTS were designed based on sequences found in literature[156;155;102] and were tested *in vitro*. I used the original TTS sequences as they have been described in the publications for the verification experiments and as an initial test to see whether triplex formation works in the lab. After verification that they worked in my hands, I chose to continue with two TTS in the Triplex-Seq process. For the purpose of the Triplex-Seq protocol, I expanded the sequence of the original TTS by approx. 20 nt on each side (5' and 3'). The TTS were generated by annealing single-stranded oligos (95 ºC for 2 minutes, cool-down to RT over a course of 45 minutes) and the sequence of each oligo is shown in Table 13. In addition to the TTS that were based on literature sequences (thus termed positive controls), I also designed new TTS with increasing frequency of guanines within the sequence, starting from 20 % guanines up to 84 % guanines. The design of these TTS as well as corresponding TFOs (see last two rows of Table 10 on page 30) were supported by the triplexator software[117]. This prediction software analyzes potential TFO/TTS pairs by matching the ssDNA to the dsDNA applying user-specific parameters. This powerful, computational framework also predicts putative TFOs within a single-stranded sequence, or potential TTS in dsDNA. The main parameters that were used to generate TFOs matching the TTS with increasing percentages of guanines are listed in below:

- maximum error-rate : 10%; maximum total error : 3; maximum number of tolerated consecutive pyrimidine interruptions in a target: 1
- minimum guanine content with respect to the target : 10% ; maximum guanine content with respect to the target : 100%
- minimum length : 15 nt- maximum length : 30 nt
- minimum guanine-percentage in anti-parallel mixed motif TFOs : 0% ; maximum guanine-percentage in parallel mixed motif TFOs : 100%

Table 13| List of triplex target sites used for the *in vitro* Triplex-Seq setup. Forward and reverse sequences of TTS oligos are shown. TTS were constructed by annealing the two respective oligos thereby generating the double-stranded TTS. The positive control TTS (modified from literature) were used for *in vitro* triplex formation experiments as well as for the Triplex-Seq protocol, while the TTS variants were generated based on the Triplexator[117] predictions with corresponding TFOs (see Table 10). The names of the TTS variants indicate the percentage of guanines in the TTS. ap, anti-parallel (pH 7 condition); p, parallel (pH 5 condition); fw, forward; rev, reverse

TTS (positive controls)

name	sequence (5'–> 3')	length [nt]	reference
TTS_1_37 (fw)	GTTCGAATCCTTCCCCCCCCACCACCCCCTCCCCCTC	37	Saleh *et al.*[102]
TTS_1_37 (rev)	GAGGGGGAGGGGGTGGTGGGGGGGGAAGGATTCGAAC	37	Saleh *et al.*[102]
TTS_2_45 (fw)	CATGCTACGTTGGAGAAGGAGGAGAAGGAAAGAGTCCTCTATACG	45	Chiou *et al.*[155]

TTS_2_45 (rev)	CGTATAGAGGACTCTTTCCTTCTCCTCCTT CTCCAACGTAGCATG	45	Chiou *et al.*[155]
TTS_1 (fw)	GTATCGTAATACGATGCGGTTCGAATCCTT CCCCCCCCACCACCCCCTCCCCCTCCAGAC TCAAGCTGACC	71	Adapted from Saleh *et al.*[102]
TTS_1 (rev)	GGTCAGCTTGAGTCTGGAGGGGGAGGGGGT GGTGGGGGGGGAAGGATTCGAACCGCATCG TATTACGATAC	71	Adapted from Saleh *et al.*[102]
TTS_2 (fw)	GTATCGTAATACGATGCGCATGCTACGTT GGAGAAGGAGGAGAAGGAAAGAGTCCT CTATACGCAGACTCAAGCTGACC	79	Adapted from Chiou *et al.*[155]
TTS_2 (rev)	GGTCAGCTTGAGTCTGCGTATAGAGGACTC TTTCCTTCTCCTCCTTCTCCAACGTAGCAT GCGCATCGTATTACGATAC	79	Adapted from Chiou *et al*[155]

TTS (variants)

name	**sequence (5'–> 3')**	**length [nt]**	**reference**
TTS_G20 (fw)	GGCCGCTTTTCTTTTCTCTTTTCTTCTTTT TTCTTTGACGT	41	this work
TTS_G20 (rev)	CAAAGAAAAAAGAAGAAAAGAGAAAAGAAA AGC	33	this work
TTS_G33 (fw)	GGCCGCTCTTCTTTTCTTCTTTCTTCCTTC TTCCTTGACGT	41	this work
TTS_G33 (rev)	CAAGGAAGAAGGAAGAAAGAAGAAAAGAAG AGC	33	this work
TTS_G53 (fw)	GGCCGCTCCTCCTCCCTTCTTTCTTCCTTC TTCCCTGACGT	41	this work
TTS_G53 (rev)	CAGGGAAGAAGGAAGAAAGAAGGGAGGAGGAGC	33	this work
TTS_G75 (fw)	GGCCGCTCCCCCTCCCTCCTTTCCTCCTCC CCCCCTGACGT	41	this work
TTS_G75 (rev)	CAGGGGGGGAGGAGGAAAGGAGGGAGGGGG AGC	33	this work

TTS_G84 (fw)	GGCCGCTCCCCCTCCCCCCCCTCCCCCTCC CCCCCTGACGT	41	this work
TTS_G84 (rev)	CAGGGGGGGAGGGGGAGGGGGGGGAGGGGG AGC	33	this work

3.7.4 Design of primers and other oligos

For the preparation of the TFO sequences for Illumina sequencing (see detailed experimental setup below), a single-stranded DNA (ssDNA) adapter was designed for ssDNA ligation as well as primers for PCR amplification of TFO sequences and simultaneous addition of Illumina sequences. A full list of primers and adapter sequences is shown in Table 14 and highlights the sequences as well as modifications of the primers.

Table 14| Primers and oligos for Triplex-Seq protocol. All primers that are used in PCR amplification step of Triplex-Seq protocol, as well as adapter and capture oligos are listed including modifications. All primers were ordered as desalted ssDNA oligos from IDt. Deviations of standard primers are mentioned in description.

oligo name	sequence (5' –> 3')	description
Common Illumina sequence	CAAGCAGAAGACGGCATACGAGAT NNNNNN GTGACTGGAGTTCAGACGTGTGCTC	Illumina reverse primer - sequence (N = barcode)
Illumina Index #1	CGTGAT	barcode sequence
Illumina Index #2	ACATCG	barcode sequence
Illumina Index #3	GCCTAA	barcode sequence
Illumina Index #4	TGGTCA	barcode sequence
Illumina Index #5	CACTGT	barcode sequence
Illumina Index #6	ATTGGC	barcode sequence
Illumina Index #7	GATCTG	barcode sequence
Illumina Index #8	TCAAGT	barcode sequence
Illumina Index #9	CTGATC	barcode sequence
Illumina Index #10	AAGCTA	barcode sequence
Illumina Index #11	GTAGCC	barcode sequence
Illumina Index #12	TACAAG	barcode sequence
Illumina Index #13	TTGACT	barcode sequence
Illumina Index #14	GGAACT	barcode sequence
Illumina Index #15	TGACAT	barcode sequence
Illumina Index #16	GGACGG	barcode sequence

Table 14| Primers and oligos for Triplex-Seq protocol. All primers that are used in PCR amplification step of Triplex-Seq protocol, as well as adapter and capture oligos are listed including modifications. All primers were ordered as desalted ssDNA oligos from IDt. Deviations of standard primers are mentioned in description.

oligo name	sequence (5' –> 3')	description
Illumina Index #17	CTCTAC	barcode sequence
Illumina Index #18	GCGGAC	barcode sequence
Illumina Index #19	TTTCAC	barcode sequence
Illumina Index #20	GGCCAC	barcode sequence
Illumina Index #21	CGAAAC	barcode sequence
Illumina Index #22	CGTACG	barcode sequence
Illumina Index #23	CCACTC	barcode sequence
Illumina Index #24	GCTACC	barcode sequence
Illumina Index #25	ATCAGT	barcode sequence
Illumina Index #26	GCTCAT	barcode sequence
Illumina Index #27	AGGAAT	barcode sequence
Illumina Index #28	CTTTTG	barcode sequence
Illumina Index #29	TAGTTG	barcode sequence
Illumina Index #30	CCGGTG	barcode sequence
Illumina Index #31	ATCGTG	barcode sequence
Illumina Index #32	TGAGTG	barcode sequence
Illumina Index #33	CGCCTG	barcode sequence
Illumina Index #34	GCCATG	barcode sequence
Illumina Index #35	AAAATG	barcode sequence
ssDNA adapter	/5Phos/ AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC /3SpC3/	HPLC purification, 5' Phosphorylation, 3' C3 spacer
Biot-TriSeqNGS-ddC	/5BiosG/ GAAGTCGAACCGCCAGACC /3ddC/	5' Biotin, 3' Dideocy-C (used as capture oligo)
PE_forward	AATGATACGGCGACCACCGAGATCTACACTCTT TCCCTACACGACGCTCTTCCGATCT	forward primer to add Illumina sequences

Table 14| Primers and oligos for Triplex-Seq protocol. All primers that are used in PCR amplification step of Triplex-Seq protocol, as well as adapter and capture oligos are listed including modifications. All primers were ordered as desalted ssDNA oligos from IDt. Deviations of standard primers are mentioned in description.

oligo name	sequence (5' –> 3')	description
TriSeqNGS001	CTTTCCCTACACGACGCTCTTCCGATCTCTTCA GCTTGGCGGTCTGG	forward primer to add part of Illumina sequence

3.7.5 Triplex formation *in vitro*

1000 pmole of Triplex-forming oligonucleotides (TFOs) and 50 pmole of respective triplex target site (TTS) were mixed at a molar ratio of 20:1 (TFO:TTS) and incubated in appropriate buffer conditions (see Table 15 for details) at 37 °C for 2 hours in a final volume of 25 µL. Samples were either subjected to the DNA ScreenTape assay (2200 Tapestation, Agilent) using 1 µL of each sample or mixed with 1x DNA loading dye (NEB) and loaded on a 10 % native polyacrylamide gel (PAGE) for separation of TFO, duplex and triplex fragments (see more details in description of electrophoretic mobility shift assay below).

Table 15| Triplex-forming buffer compositions. The 1x buffer compositions that were used to form triplexes *in vitro*.

triplex-forming conditions	composition of 1x triplex buffer
anti-parallel (pH 7)	10 mM Tris-HCl pH7.2, 10 mM $MgCl_2$
parallel (pH 5)	10 mM sodium acetate pH 5.0, 10 mM $MgCl_2$
triplex disfavoring	10 mM Tris-HCl pH7.5, 140 mM KCl

3.7.6 Electrophoretic mobility shift assay (EMSA)

To separate triplexes from duplex DNA and non-bound TFOs, a native polyacrylamide gel (PAGE) was used. The 10 % PAGE (15 % PAGE for verification experiments) was prepared by polymerizing the acrylamide/bis-acrylamide 40 % solution (Sigma) using N,N,N',N'- Tetramethylethylenediamine (TEMED, Alfa Aesar) and ammonium persulfate (APS, Sigma) in respective buffers (anti-parallel: 4 mM TBE,2.5 mM $MgCl_2$, parallel: 8 mM sodium acetate pH 5.0, 2.5 mM $MgCl_2$ and triplex disfavoring buffers: 4 mM TBE (Tris/Borate/EDTA), 140 mM KCl). Following PAGE preparation, samples were mixed with 1x purple loading dye (6x, NEB) and 7.5 µL of low molecular weight DNA ladder (NEB, #N3233) was loaded onto the gel. The 1x running buffer is the same that has been used for PAGE preparation. For sufficient band separation between triplexes and duplex DNA, electrophoresis was operated for 2 hours at a field strength of 7.5 V/cm^2. Subsequently, the gel was removed from the electrophoresis chamber and transferred to 1x running buffer containing 0.1 mg/mL of ethidium bromide (1 mg/mL, hylabs) to stain DNA for 20 minutes at RT while carefully shaking. Images of gels were acquired using a UV gel documentation system.

3.7.7 DNA fragment isolation from PAGE

Following triplex/duplex/TFO separation, DNA was isolated from PAGE using the Crush and Soak Method as it has been described by J. Sambrook and D. Rusell[157]. While UV illuminating the PAGE (305 nm), (putative triplex) bands in gel of triplex or TFO lanes were excised at the same height corresponding to the triplex DNA from the positive control in the triplex sample using a clean scalpel and transferred the gel slices into a 1.5 mL microcentrifuge tube. The weight of the slice was determined and 2 volumes of 1x Crush and Soak buffer (CSB, 200 mM NaCl, 10 mM Tris-HCl pH7.5, 1 mM EDTA pH8.0) was added. The gel was crushed into smaller fragments using a sterile pipette tip or inoculation loop and incubated overnight at 37 °C while slowly shaking. Following the overnight incubation, samples were centrifuged at maximum speed (16,000 *g*) for 2 minutes at 4 °C. The supernatant was transferred to a fresh microcentrifuge tube and an additional 2 volumes of CSB were added to the gel pellet, centrifuged (16,000 *g*, 2 minutes, 4 °C) and supernatants were pooled. Subsequently, DNA was ethanol-precipitated by addition of 3 volumes of ice-cold ethanol, 1/10 of volumes sodium acetate (pH5.0) and 1 µg GlycoBlue Coprecipitant (glycoblue, 15 mg/mL, Thermo Fischer Scientific). Samples were incubated for at least 1 hour at -80 °C followed by centrifugation (16,000 *g*, 30 minutes, 4 °C). Supernatant was carefully decanted, DNA was air-dried for 5 minutes at RT and dissolved in 15 µL of ultra-pure water (Ultra Pure Water, Biological Industries).

3.7.8 Heat-separation of duplex and TFO DNA (triplex disruption)

To ensure that TFOs are not bound to duplex DNA, which is required for the ssDNA adapter ligation in the next step, the DNA from the previous steps was mixed with 1x triplex disfavoring buffer (TDB: 10 mM Tris-HCl pH7.5, 140 mM KCl) and incubated at 95 °C for 5 minutes to separate duplex DNA from TFO. Subsequently, DNA was reannealed by decreasing temperature by 1 °C every 30 seconds until room temperature (RT) was reached. Following reannealing of duplex DNA and simultaneous prevention of triplex formation, DNA was ethanol-precipitated as has been described above (3 volumes of ice-cold ethanol, 1/10 of volumes sodium acetate (pH5.0) and 1 µg GlycoBlue) and resuspended in 23 µL ultra-pure water.

3.7.9 Single-stranded adapter ligation

After TFOs have been separated from duplex DNA, ssDNA adapter ligation was performed by using the CircLigase ssDNA ligase kit (#CL4115K, Epicentre). Briefly, samples were mixed in 1x CircLigase buffer, 2.5 mM $MgCl_2$, 50 µM adenosine-triphosphate (ATP), 100 U CircLigase and 50 pmole ssDNA adapter which contains a 5' phosphorylated terminus (to act as donor) and a 3' carbon spacer (for more details see Table 14). The reaction mix was incubated for 2 hours at 60 °C with subsequent deactivation of the enzyme for 10 minutes at 80 °C. The obtained TFO fragments were purified using Agencourt AMPure beads (Coulter Beckman), according to manufacturer's instructions. In brief, 1.8x volumes well-resuspended AMPure XP bead slurry was added to the PCR reaction mix, incubated for 5 minutes at RT and transferred to the DynaMag-96 Side Magnet (#12331D, Thermo Fisher Scientific). Following incubation of the sample on the magnet for 2 minutes (or until sample is clear), supernatant was removed and beads were washed twice with 200 µL freshly-prepared 70 % ethanol without removing samples from the magnet. Subsequently, samples were removed from plate, air-dried for 5 minutes to ensure no residual ethanol was left, resuspended in 25 µL ultra-pure water and incubated for 5 minutes at RT before transferring to magnet. Following a 2 minutes incubation, supernatant

was carefully transferred to a fresh tube.

3.7.10 Preparation of sequencing library

The final step of the protocol is the PCR amplification of the enriched TFOs and simultaneous addition of Illumina adapter sequences including indexes to multiplex samples. For a detailed list of Illumina oligonucleotides, index sequences and other PCR primers see Table 14. For the PCR mix, 0.01 µM of primer TriSeqNGS001 which binds the capture sequence of the TFO, 0.5 µM Illumina primer index #1-#34 that bind the ligated ssDNA adapter sequence and adds Illumina indexes, 200 µM dNTPs (each dNTP 100 mM Solution, Thermo Fisher Scientific), 1 U Q5 Hot Start High-Fidelity Polymerase (Q5, NEB), 3 % dimethylsulfoxide (DMSO) were mixed in 1x Q5 Reaction buffer and the following PCR program was executed: Initial denaturation for 2 minutes at 98 °C, followed by 15 cycles of 30 seconds at 98 °C, 30 seconds at 65 °C and 10 seconds at 72 °C which preceded the final elongation step for 2 minutes at 72 °C. PCR samples were purified using AMPure XP beads as has been described above. In a second PCR, 0.5 µM Illumina primer index #1-#34, 0.5 µM primer PE_forward, which adds the sequence that is complementary to the Illumina flow cell, were added to the reaction mix as has been described above. The same PCR program was used and after PCR completion, samples were cooled down to 4 °C, 5 U of Exonuclease I (ExoI, NEB) were added to the PCR mix and incubated at 37 °C for 30 minutes. Samples were subsequently purified using AMPure XP beads as described above. DNA Screen TapeAssay and Illumina sequencing 1 µL of prepared double-stranded DNA (dsDNA) libraries were analyzed by the DNA ScreeTape assay and size of DNA fragments was determined. To multiplex and prepare sequencing libraries, the molarity of the PCR-amplified dsDNA libraries was calculated by determining the average length based on the Tapestation results and the concentration of the dsDNA fragment which was measured by Qubit 4 Fluoremeter (Thermo Fisher Scientific) and samples were mixed to obtain a 10 nM pooled and multiplexed library.

3.7.11 Illumina sequencing

The multiplexed libraries were sequenced on an Illumina HiSeq 2500 (High Output Run Mode V4 or Rapid Run Mode) or MiSeq and were operated at the Technion Genome Center in Haifa. Depending on the number of library variants, different volumes of the TFO libraries were mixed (each TFO library was 2-10 nM) and the pooled library was run as a single-read 50 bp run. Due to the low diversity of sequences in the libraries, added 20 % PhiX Control v3 Library (Illumina, FC-110-3001) was added. The well-balanced GC/AT PhiX genome is derived from the small, well characterized bacteriophage PhiX and serves as an in-run control for cluster generation, sequencing and alignment. The overall read yield ranged between 150 - 300 Mio reads per HiSeq run, and 10-20 Mio reads per MiSeq run.

3.7.12 Bioinformatic analysis

First, Illumina sequencing read quality was validated, adapter sequences were trimmed using cutadapt[158] and aligned to the PhiX genome bowtie2[159] in local alignment mode (bowtie2 --local). Second, TFO sequences were extracted by (i) identifying the 19 nt long capture sequence '`CTTCAGCTTGGCGGTCTGG`', (ii) selecting sequences with exactly 39 nt and (iii) searching for identical matches to all possible combinations of the TFO sequence. Next, the number of reads were normalized by dividing each read count by the total number of reads for each sample and multiplied by 10^6(reads per million, RPM). For the *in vitro* Triplex-Seq analysis, the triplex reactivity was calculated. Triplex reactivity is defined as the ratio of the RPM of the triplex

band divided by the RPM of the TFO band and subtraction of 1. Thus, every value above 1 is defined as "triplex reactive".

$$triplex\ reactivity = \frac{RPM_{triplex}}{RPM_{TFO}} - 1$$

3.8 Circular dichroism spectroscopy

For circular dichroism (CD) spectroscopy, TFO (12.5 µM) and TTS (2.5 µM) were mixed in 1x triplex buffer (depending on condition either 10 mM Tris-HCl pH7.2, 10 mM $MgCl_2$ or 10 mM sodium acetate pH5.0, 10 mM $MgCl_2$), incubated for two hours at 37 °C and subsequently cooled down to 25 °C/RT. CD spectra were recorded on a J-1100 CD spectrophotometer (Jasco) using a 1 mm quartz cuvette (kindly provided by Arnon Henn's lab, Technion) with a total volume of 200 µL. The scanning speed was 100 nm/min, Digital Integration Time (D.I.T.) of 2 seconds and 2 accumulations (average of two consecutive recordings per sample) were recorded at 25 °C. The CD spectra were baseline-corrected using the respective buffers.

3.9 In cell Triplex-Seq

3.9.1 Cell culture

CHO-K1-MI-HAC (kindly provided by Y. Kazuki and M. Oshimura and hereby referred to as simply CHO cells) were grown in F-12 Nutrient Mixture (HAM's) medium (BI), supplemented with 10 % FBS (04-005-1A, LOT: 1630662, Biological Industries) and 1 % Penicillin-Streptomycin solution (Biological Industries) ($F12_{complete}$), and cultured at 37 °C and 5 % CO_2 in humidified atmosphere. CHO cells were subcultured at a 1:5 to 1:10 ratio and transfected by using the fast-forward transfection method (for more details see below).

3.9.2 Transfection of TFO libraries and cell harvest

13 µg of TFO was transfected by mixing with 65 µg of 1 mg/mL polyethylene imine (PEI Linear, #23966, Polysciences), incubated at room temperature for 15 minutes and subsequently added to $6x10^7$ CHO cells. Transfections were carried out either on 80 % confluent cells that have been seeded 24 hours prior transfection, or confluent CHO cells tissue culture plates were harvested on the day of transfection and cells were transfected in solution and subsequently transferred to the Nunclon (245 mm × 245 mm × 20 mm) tissue culture plates (fast-forward transfection). 24 hours post-transfection, supernatant was collected and transferred to a conical centrifuge tube, cells were harvested by adding enough trypsin (2-5 mL) to cover the tissue culture plates, resuspended in F-12 Nutrient Mixture (HAM's) medium and subsequently added to the supernatant. Cells were centrifuged at 300 *g* for 10 minutes at RT, washed once with 1x phosphate buffered saline (PBS) and genomic DNA was isolated.

3.9.3 Genomic DNA isolation and digestion

DNA was isolated using the Exgene Cell SV kit (GeneAll Biotechnology) according to the manufacturer's instructions. In brief, 400 µg Proteinase K and 400 µg RNAseA (#R6513, Sigma) were added to DNA sample and incubated for 2 minutes at RT, followed by addition of lysis buffer (Buffer BL) to tube, vortexing and incubation at 56 °C for 10 minutes. After addition of 1 volume ethanol, sample was applied on provided SV column and centrifuged (6000 *g*, RT, 1 minute). Column was washed with (Buffer BW and Buffer TW) and resuspended in ultrapure water. Following DNA/TFO isolation, gDNA was fragmented with xx U of EcoRI-HF (New

England Biolabs, NEB) for 4-6 hours at 37 °C in 1x Cutsmart buffer (NEB), purified using 1.8x AMPure beads as described above and resuspended in 1x triplex-disfavoring buffer (TDB) and incubated overnight at 37 °C.

3.9.4 Enrichment of TFOs

Following the incubation in 1xTDB, 100 pmole of biotinylated capture oligo (Biot-oTriSeqNGS-ddC, for details see Table 12) was added, incubated at 95 °C for 5 minutes and slowly cooled down to RT with 1 °C/30 seconds. The TFO/gDNA/oligo mix was precipitated using 3 volumes ice-cold ethanol, 1/10 volume sodium acetate (pH5.0) and 1 µg glycoblue. Samples were stored at -80 °C for at least 30 minutes followed by centrifugation at 4 °C for 30 minutes at 16,000 *g*. Precipitated DNA was resuspended in ultra-pure water and incubated with equal volumes of Dynabeads MyOne Streptavidin T1 (final conc.: 5 µg/µL) for 10 minutes at RT. Magnetic beads were placed on DynaBead magnetic plate and washed five times with wash buffer (10 mM TrisHCl pH7.2, 1 mM EDTA pH8.0, 4 M NaCl, 0.2 % Tween) by incubating the beads for 10 minutes at RT and subsequently removal of the wash buffer. To dissociate the non-biotinylated fragments, beads were washed two times with 1x saline-sodium citrate (SSC, Bio-Lab Chemicals) buffer and eluted in 50 µL 1xSSC buffer while being heated for 5 minutes at 95 °C. After this first release, beads were incubated for 10 minutes at 95 °C in 95 % Hi-Di formamide (Thermo Fisher Scientific) and 10 mM EDTA (pH8.0) which disrupts binding of biotin ad streptavidin. The samples from both the SSC and formamide fraction were ethanol precipitated by adding 3 volumes of ice-cold ethanol, 1/10 volume sodium acetate (pH5.0) and 1 µg glycoblue and protocol for ethanol precipitation was followed as described above.

3.9.5 Sequencing library preparation and Illumina sequencing

Heat-denaturation, ssDNA adapter ligation, preparation of sequencing library via PCR amplification and Illumina sequencing have been performed as has been described in *in vitro* Triplex-Seq.

3.9.6 In cell gDNA Triplex-Seq

For co-enrichment of genomic DNA (gDNA), cells were transfected and harvested as described above in the regular in cell Triplex-Seq protocol, washed with 1xPBS and incubated in 1 % paraformaldehyde (PFA) for 10 minutes at RT. Crosslinking reaction was stopped with 0.125 M glycine for 5 minutes at RT, followed by 15 minutes on ice. PFA was removed by centrifugation (400 *g,* 10 minutes) and washed once with 1x PBS. Cell membranes were lysed using a mild detergent-containing lysis buffer (10 mM Tris-HCl pH7.5, 10 mM NaCl, 0.2 % NP40) for 5 minutes and nuclei were harvested (500 *g*, 10 minutes) and resuspended in 1.5 % sodium dodecyl sulfate (SDS, Biological Industries) and 1x NEB buffer 2.1 (NEB). 300 pmole of capture oligo (see description above and in oligo list) was annealed by heating it first to 60 °C for 5 minutes followed by a cool-down of 1 °C/30 seconds. After annealing of capture oligo, SDS was sequestered by addition of 1.5 % Triton and 20 U MseI was added to this mix and incubated for 4-6 hours at 37 °C. To inactivate the enzyme, 1 mM EDTA (pH8.0) and 1 M NaCl were added and the mix were mixed with MyOne Streptavidin T1 beads (25 µL) and incubated for 15 minutes at RT. Following binding of TFOs/gDNA fragments to the beads, the supernatant containing non-bound gDNA was transferred to a new tube and stored at -20 °C until further processed. After removal of non-bound gDNA, the crosslink was reversed by incubating the beads-DNA reaction in reverse-crosslinking buffer (50 mM Tris-HCl pH7.2, 1 % SDS, 1 mM EDTA pH8.0, 100 mM NaCl, 500 µg Proteinase K (Sigma) overnight at 65 °C while carefully shaking. This step might

lead to dissociation of the capture oligo with its complementary capture sequence on the TFO, thus we performed reannealing of the capture oligo to the TFO in triplex-disfavoring buffer to prevent triplex formation. To do so, samples were mixed with 0.14 M KCl and 10 mM Tris-HCl (pH 7.5) and heated to 80 °C for 5 minutes after which the samples were cooled down at 1 °C/30 seconds. 1 M final concentration of NaCl was added to the reaction mix, and incubated for 10 minutes at RT. Subsequently, mix was transferred to magnetic stand and supernatant was transferred to a new tube.

3.10 In cell Triloci-Seq

3.10.1 Design of oligonucleotides (oligos) and primers

The developed in cell Triloci-Seq assay requires TFOs, a ss/ds-DNA adapter which is ligated to the TFOs in the downstream protocol, a cut_oligo that generates a double-stranded restriction site and primers for PCR amplification. The TFOs that were used in this study were adapted from literature and contain a triplex-forming motif as well as a region that corresponds to the Illumina sequencing primer. In Table 16, the TFOs and other oligos (adapter_oligo, cut_oligo and PCR primers) are shown.

Table 16| List of in cell Triloci-Seq primers. All primers that were used in this study to developed the in cell Triloci-Seq protocol are shown.

name	sequence (5-->3)	description	length [nt]
MEG3	CGGAGAGCAGAGAGGGAGCGAGATCGGAAGAG CGTCGT	Mondal *et al.*[52]	38
Fendrr	TCCCCTCCATCCTCTTCCTTCTCCTCCTCCTC TTCTTTAGATCGGAAGAGCGTCGT	Grote *et al.*[50]	56
HOTAIR-1	GAGAGAAGGGAGGAGAAGATCGGAAGAGCGTC GT	Kalwa *et al.*[54]	34
HOTAIR-2	GAGACCGAGAGAGAGAAGATCGGAAGAGCGTC GT	Kalwa *et al.*[54]	34
Khps1	CAGGGTCCCCCCTTTTTTTTTCCTCCTGGAGA TCGGAAGAGCGTCGT	Postepska *et al.*[51]	46
TTC	TTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT CTTCTTCTTCTTCAGATCGGAAGAGCGTCGT	Zheng *et al.*[49]	63
GAA	GAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGA AGAAGAAGAAGAAAGATCGGAAGAGCGTCGT	Ohno *et al.*[160]	63
Particle	AAGGGGGGGGGAAAGATCGGAAGAGCGTCGT	O'Leary *et al.*[53]	31
DHFR	ACAAATGGGGACGAGGGGGGCGGGGCGGCCAG ATCGGAAGAGCGTCGT	Martianov *et al.*[48]	47
T0	GTCGACCAGTTGTTCCTTTGAGATCGGAAGAG CGTCGT	Schmitz *et al.*[10]	38

oligo_cut	TCGTGTAGGGAGGATCCGTTCAGACGTGTGCTCT	this work	34
adapter_fw	/5Phos/GTAGGGAGGATCCGTTCAGACGTGT GCTCTTCCGA/iBiodT/CTTTAAGTA	this work	45
adapter_rev	TACTTAAAGATCGGAAGAGCACAC	this work	24
Index#1	CAAGCAGAAGACGGCATACGAGATCGTGATGT GACTGGAGTTCAGACGTGTGCTCTTCCGATCT	NEBNext Index primer 1	25
Index#2	CAAGCAGAAGACGGCATACGAGATACATCGGT GACTGGAGTTCAGACGTGTGCTCTTCCGATCT	NEBNext Index 2 Primer for Illumina	64
Index#3	CAAGCAGAAGACGGCATACGAGATGCCTAAGT GACTGGAGTTCAGACGTGTGCTCTTCCGATCT	NEBNext Index 3 Primer for Illumina	64

3.10.2 Cell culture

HEK-293 cells were grown in Dulbecco's Modified Eagle's Medium (DMEM) high glucose (#D5796, Sigma), supplemented with 10 % FBS (04-005-1A, LOT: 1630662) and 1 % Penicillin-Streptomycin solution herewith referred to as $DMEM_{complete}$, and cultured at 37 °C and 5 % CO_2 in humidified atmosphere. HEK-293 cells were subcultured at a 1:10 ratio and transfected by using the fast-forward transfection method.

3.10.3 Transfection of TFO libraries and cell harvest

13 µg of the pool of mixed TFOs (each TFO was mixed at molar ratios and for a full list of TFOs see Table 16 on page 42) was transfected by mixing with 65 µg of 1 mg/mL PEI incubated at RT for 15 min and subsequently added to $6x10^7$ HEK-293 cells that have been collected in 50 mL conical tubes. Transfection was carried out by harvesting confluent HEK-293 cells on the day of transfection and cells were transfected in solution and subsequently transferred to the tissue culture plates (fast-forward transfection). 24 hours post-transfection, supernatant was collected and transferred to a conical centrifuge tube and cells were harvested by adding enough trypsin (5 mL) to cover the tissue culture plates, resuspended in $DMEM_{complete}$ and subsequently added to supernatant. Cells were centrifuged at 300 *g* for 10 minutes, washed once with 1x phosphate buffered saline (PBS) and resuspended in 1x PBS.

3.10.4 Crosslinking of cells

PFA crosslink: Cells were mixed with 16 % paraformaldehyde (PFA) aqueous solution (cryoEM grade, Electron Microscopy Sciences) resulting in a final concentration of 1 %, incubated for 10 minutes at RT and inactivated by addition of 1 M glycine (final concentration of 0.125 M glycine). The mix was incubated for 5 minutes at RT, followed by 15 minutes on ice. Subsequently, cells were washed once with ice-cold 1x PBS and cell pellets (300 *g*, 5 minutes, RT) were either flash-frozen in liquid nitrogen and stored at -80 °C or transferred to cell permeabilization and linker ligation.

TMP crosslink: 10 µg/mL final concentration of Trioxsalen (4,5,8-Trimethylpsoralen (TMP), 200 µg/mL dissolved in ethanol, #T6137, Sigma) was added to cells and incubated in the dark

at RT for 5 minutes prior to crosslinking. The plates were placed on ice 15 cm away from the light source of the UV crosslinker (Spectroline, Select Series, 365 nm, Long Wave). Samples were irradiated six times at 365 nm for 30 seconds with 300 µJ/cm^2. After crosslinking cells were centrifuged (400 *g*, 10 minutes, RT) and washed once with 10 mM Tris-HCl (pH8.0), transferred to ice and either flash frozen in liquid nitrogen (and stored at -80 °C) or samples were further processed.

3.10.5 Nucleus permeabilization, adapter ligation, gDNA fragmentation

Cell pellets were resuspended in permeabilization buffer (10 mM TriS-HCl pH8.0, 10 mM NaCl, 0.2 % Nonidet-P40 (kindly provided by the Dganit Danino lab) and incubated for 30 minutes on ice, with occasional agitation, followed by nuclei centrifugation at 1100 g for 12 minutes at RT and washed once with ice-cold 1x PBS. Nuclei were resuspended in 1x T4 RNA ligase I reaction buffer (NEB). Samples were split so that each sample contains 20-25 Mio cells per reaction and 10 % SDS to a final concentration of 1.5 % was added and incubated for 15 minutes at 65 °C while occasional inverting. To sequester the SDS, 10 % Triton-X-100 (BioLab Chemicals) was added to a final concentration of 1.5 % and carefully mixed while avoiding air-bubble formation followed by addition of 0.5 µg RNAse A (10 mg/mL, R6513, Sigma) and incubation for 30 minutes at 37 °C while carefully shaking (70 rpm). Nuclei were mixed with 1 mM ATP (NEB), 20 % PEG8000 (NEB) and 20 U of T4 RNA ligase I (ssRNA ligase, NEB) and supplemented with 100 pmole ss/dsDNA adapter chimera. The adapter contains a 5' phosphorylated end and an internal biotinylated nucleotide (see more details in Table 16 on page 42). Sample were incubated for 4-6 hours at 22 °C followed by 12 hours at 16 °C while shaking at 100 rpm. The reaction was stopped by addition of final concentrations of 6.7 mM Trizma (pH8.0) (Sigma) as well as 1.6 mM EDTA (pH8.0) and incubated for 10 minutes at RT. Inactivated ligation mix was supplemented with 1x Cutsmart buffer (NEB) as well as 50 U of MseI (4-base butter, NEB) and incubated at 37 °C for at least 8 hours or overnight while slowly shaking (100 rpm). The enzyme was heat-inactivated by incubation in 1.5 % final concentration of SDS for 2 minutes at 65 °C while shaking followed by 1 % Triton-X100 addition.

3.10.6 Proximity-based ligation

After ssDNA ligation of ss/dsDNA adapter and digestion of gDNA, the fragmented gDNA pieces will be ligated to the TFOs containing the ligated adapter sequence. Here, we will use the approach of proximity-ligation was has been described before [161;162]. Briefly, 1000 U of T4 DNA ligase (NEB) were mixed with T4 DNA ligase buffer and the digested DNA sample in a volume that is 25-fold larger than the initial volume after digestion to reduce non-specific ligation and increase probability of fragments being ligated that are in close-proximity to one another. The ligation mix was incubated overnight at 16 °C while shaking (70 rpm).

3.10.7 Crosslink reversal

To reverse the PFA crosslink, 500 µg of proteinase K (#P2308, Sigma) was added to each sample and incubated for 8-12 hours at 56 °C while shaking. To reverse TMP crosslink, samples were placed on ice at a distance of 15 cm to the light source and irradiated with 254 nm at 1200 µJ/cm^2 for 2 seconds (Spectroline, Select Series, 365 nm, Long Wave) and repeated three times.

3.10.8 Phenol-chloroform based DNA isolation and DNA-fill in reaction

One volume of ice-cold UltraPure Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v, saturated with 10 mM Trizma pH8.0, 1 mM EDTA, Sigma-Aldrich) solution is added to the samples, vortexed for 120 seconds and centrifuged at RT for 10 minutes at 4500 *g*. The upper, aqueous phase was carefully transferred to a fresh tube and one volume of chloroform was added, briefly vortexed and again centrifuged (10 minutes, 4500 g, at RT). This step was repeated two times. The supernatant was transferred to a fresh tube, the volumes of all samples were filled up to the same volume with 1x TE (10 mM Trizma pH8.0, 1 mM EDTA) and 1/10 volume of sodium acetate was mixed, followed by 2.5 volumes of ice-cold 100 % ethanol and 1 µg glycoblue. The mix was stored at -80 °C for at least 30 minutes and subsequently centrifuged (4 °C, 60 minutes, 4500 *g*). The pelleted DNA was washed once with 70 % ethanol, ant the split samples were pooled and resuspended in 1x NEB2.1 buffer (NEB). To the resuspended DNA, 10 U of T4 DNA polymerase (NEB) was added, supplemented with 200 µM dNTPs and incubated for 2 hours at 11 °C. When using T4 DNA polymerase, make sure to keep samples below 12 °C to inhibit 3´ → 5´ exonuclease activity of the enzyme.

3.10.9 Streptavidin-coupled magnetic bead purification

250 µg MyOne Strepatvidin C1 Dynabeads (Thermo Fisher Scientific) were washed three times with 1x binding buffer (10 mM Tris-HCl pH7.2, 1 mM EDTA, 1 M NaCl) and resuspended in 2x binding buffer prior mixing with DNA. Following the addition of the DNA, samples were incubated for 15 minutes at RT using gentle rotation and transferred to the magnetic 96-well plate to separate the DNA-bound beads from the binding buffer. Subsequently, the beads were washed three times with a wash buffer (10 mM TrisHCl pH7.2, 1 mM EDTA, 4 M NaCl, 0.2 % Tween) which included a 10 minutes incubation step in wash buffer followed by an additional washing step with 10 mM Tris-HCl (pH7.2) which was repeated twice. The DNA was then eluted in a denaturing solution (100 mM NaOH, 0.1 mM EDTA pH8.0) by incubation at RT for 5 minutes while gently rotating. The supernatant was neutralized by adding 100 mM HCl as well as 0.1 mM Tris-HCl pH 7.2 final concentrations. The enriched DNA was then purified using 1.8x AMPure XP beads as has been described above and the DNA was eluted in ultra-pure water.

3.10.10 Circularization of ssDNA, oligo annealing and dsDNA digestion

To bring TFOs and ligated gDNA next to one another on a linear fragment, I use the principal of single-stranded DNA circularization, oligo annealing and digestion of DNA as has been described previously[162]. In brief, the eluted DNA is mixed with 1x Circligase buffer, 2.5 mM $MnCl_2$, 0.05 mM ATP and 1 U CircLigase ssDNA Ligase (#CL4111, epicentre) and incubated for 240 minutes at 60 °C followed by a heat-inactivation step for 10 minutes at 80 °C. The circularized DNA is mixed with 0.5x Cutsmart buffer, cut_oligo (see Table 16) which generates a double-stranded DNA recognition site for BamHI and the mix is heated to 95 °C for 2 minutes and the oligo is annealed by decreasing the temperature 1 °C/20 seconds. Subsequently, 60 U of BamHI-HF (NEB) is added, incubated for 60 minutes at 37 °C followed by AMPure XP beads purification and resuspended in ultra-pure water. This DNA can be used to be PCR amplified and prepared for Illumina sequencing.

3.10.11 Illumina sequencing library preparation

The final step of the protocol is the amplification of the enriched gDNA/TFO fragments and simultaneously addition of the Illumina adapter sequences. To do so, 0.2 µM of primer 1 (see list for details of primer sequences) and 0.2 µM of primer 2 (that adds the Illumina index sequence which is required for multiplexing libraries), 300 µM dNTPs and 1 U Q5 Hot Start Polymerase were mixed in 1x Q5 Reaction buffer with the total volume of eluted DNA and run with the following PCR program: Initial denaturation for 2 minutes at 98 °C, followed by 18 cycles of 30 seconds at 98 °C, 30 seconds at 65 °C and 30 seconds at 72 °C which preceded the final elongation step for 2 minutes at 72 °C. Samples were cooled down to 4 °C, 5 U of ExoI was added to the PCR mix and incubated for 37 °C for 30 minutes. Samples were subsequently purified using AMPure XP beads as described above.

3.10.12 Illumina sequencing

The multiplexed libraries were sequenced on an Illumina HiSeq 2500 (High Output Run Mode V4 or Rapid Run Mode) (operated at the Technion Genome Center, Haifa) of the pooled library as a 100 bp paired-end run. Due to the low diversity of sequences in the libraries, 10-20 % PhiX Control v3 Library (Illumina, FC-110-3001) was added to each run. The overall read yield ranged between 150 - 300 Mio reads per HiSeq run.

3.10.13 Bioinformatic analysis

As described above, Illumina sequencing read quality was validated, adapter sequences were trimmed using cutadapt and aligned to the PhiX genome as well as to the human genome (Human genome assembly, Genome Reference Consortium Human Build 38, GRCh38, GCA_000001405.15) as well as the hamster genome (Cricetulus griseus unplaced genomic scaffold, CriGri_1.0 scaffold984, NW_003614642.1) bowtie2 in local alignment mode (bowtie2 --local). After the alignment, TFO sequences were extracted by searching for identical matches to all possible combinations of the TFO sequence.

4 Results

4.1 Synthetic long non-coding RNAs (slncRNAs)

To test triplex formation using synthetic lncRNAs (slncRNAs) in cells, I devised two strategies:

1. a triplex-mediated repression/enhancer system in bacterial E.*coli* cells (Figure 9a, top)
2. a slncRNA-dependent gene activation approach in human cells (Figure 9a, bottom)

In both strategies, I generated the same library of slncRNAs and TTS on a reporter plasmid to screen for optimal gene regulation results. As shown in Figure 9b, the modular design of the slncRNAs comprises a (i) DNA targeting motif (DNA_{bind}) that forms triplexes with corresponding triplex target sites (TTS) on a reporter plasmid and (ii) a RNA-binding protein domain (RBP_{bind}). These two modules are connected via a flexible linker and is highlighted in gray in Figure 9b.

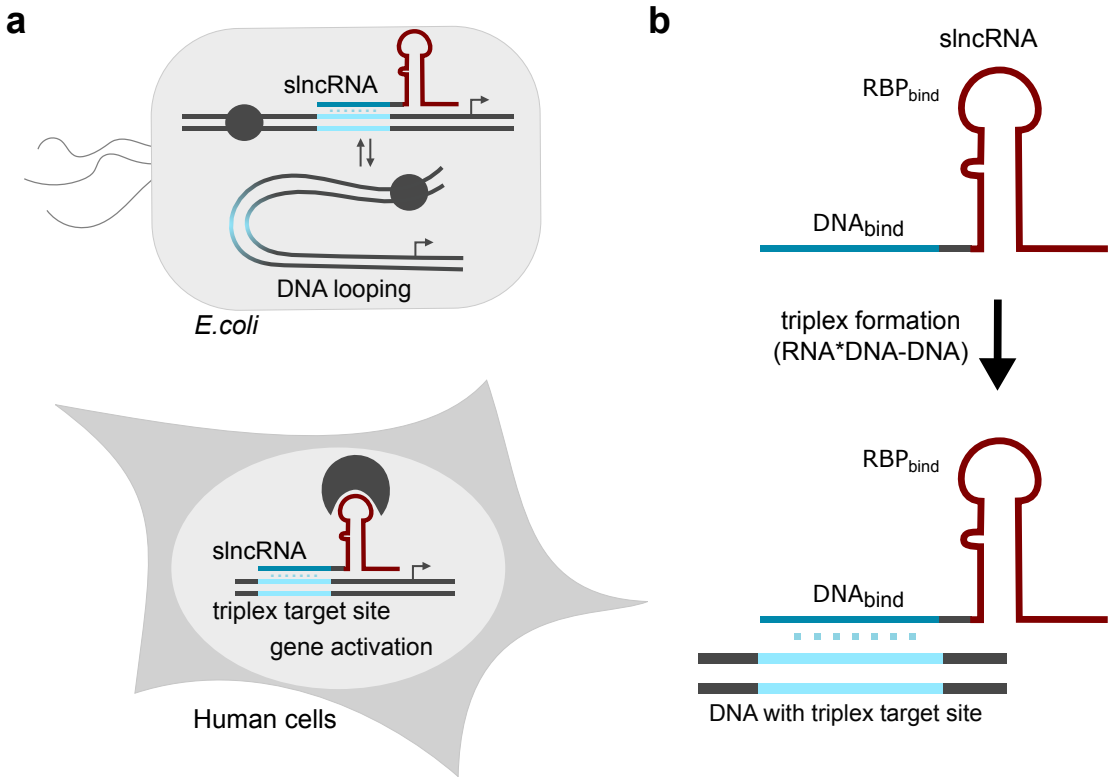


Figure 9| Schematic overview of slncRNA strategies and design. a, In bacteria, an enhancer-based genetic circuit was designed in which slncRNA binding to a triplex target site (TTS) sequence within the looping region is expected to alter its possibility to loop thus changing reporter gene levels. The strategy in mammalian cells relies on the recruitment of transcriptional activators to minimal promoters in a slncRNA-guided manner. **b,** The designed slncRNAs consist of two modules: Module I is the DNA_{bind} motif, which induces triplex formation by interaction with the TTS on the reporter plasmid. Module II consist of the RNA-binding protein sites (RBP_{bind}) and is connected to module I via a flexible linker (highlighted in dark gray). The RBP_{bind} domain can be recognized and bound by RNA-binding proteins and fusion proteins thereof.

To generate the putative DNA_{bind} motif, I selected five triplex-forming domains from literature that have been shown to induce triplex formation under *in vitro* conditions[10;49;163;164;165] and it has been proposed that some of these triplex-forming domains might induce triplex formation *in vivo*/in cells[10;49]. They consist of either purine (guanine or adenine) or pyrimidine (cytosine or thymine)-rich sequences and are 18 to 24 nts long (more information of the slncRNA library and sequences of DNA_{bind} motifs can be found on page 19 in the Materials and Methods section).

The DNA_{bind} motif is connected by a generic linker (5-40 nts) to the RBP_{bind} domain. The linker's main function is to preserve the secondary structure of the RBP_{bind}domain and keep the DNA_{bind} motif in a linear, non-basepaired state. As RBP_{bind} motif I chose the sequence of bacteriophage PP7 coat protein (PCP). In the PP7 phage, the coat protein assembles into the mature viral capsid and regulates translation of its proteins by binding to RNA molecules. The RNA molecules form a hairpin structure with a highly conserved protrusion of an adenosine nucleotide that is recognized and bound with high affinities by the PP7 phage coat protein [153]. The viral PCP sequence is well characterized and has been used for imaging RNA molecules in bacterial and eukaryotic cells by fusing PP7 to fluorescent proteins such as the green fluorescent protein (GFP) [166;167;168]. For imaging purposes, researchers have constructed cassettes of 24 repeats of PCP (24xPCP) or the tandem version of PCP (tdPCP) which has been shown to be more suitable for quantitative experiments [167]. As I did not intend to image RNA molecules, but use the tdPCP for activation or repression purposes, I generated slncRNA molecules with one to five RBP_{bind} motifs (one RBP_{bind} motif = one PP7 hairpin).
After designing the slncRNA variants, which are combinations of the DNA_{bind} motif, the linker sequence and the RBP_{bind} domain, I evaluated the 100 to 350 nts long slncRNA sequences (excluding 3' and 5' UTRs (untranslated regions)) *in silico* using the RNA prediction and design software NuPACK [169] to check whether the secondary structure of the PP7 binding site is preserved as this is crucial for binding of the tdPCP (Figure 10).

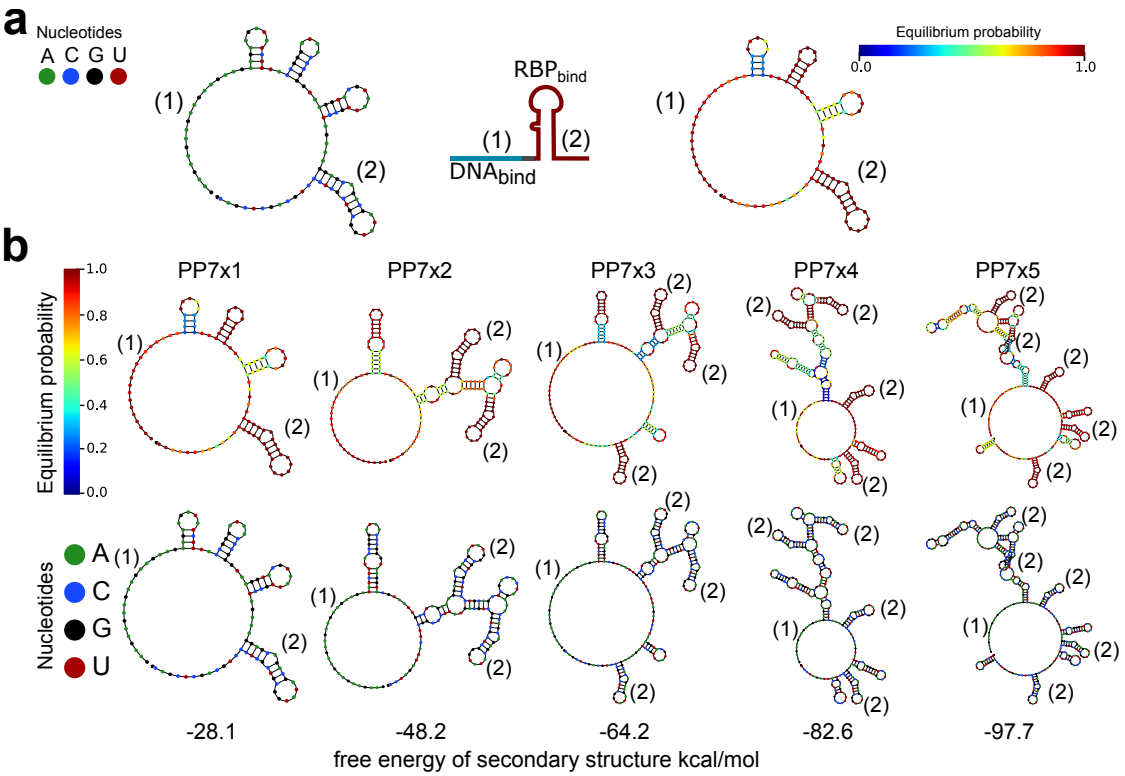


Figure 10| Prediction of slncRNA secondary structures using NuPACK. The secondary structures of the slncRNAs with the DNA_{bind} motif and the RBP_{bind} domain were analyzed using NuPACK [169]. **a,** The slncRNA with the DNA_{bind} motif GAA_{rich} and $PP7_{x1}$ RBP_{bind} domain are shown. The left image displays the nucleotides with different coloring, and the right image shows a color-coded scale which depicts the base-pairing probabilities. The DNA_{bind} motif (1) as well as the RBP_{bind} domain (2) are numbered and highlighted in each scheme. **b**, The same two schemes (nucleotides and base pairing probabilities) are shown for increasing numbers of PP7-binding sites. The RBP_{bind} motifs are highlighted. It can be seen that the secondary structure is preserved for all slncRNAs and the DNA_{bind} motif remains linear.

In Figure 10a, one can see the predicted secondary structures for the slncRNA with the DNA_{bind}

motif GAA (further details about the exact sequences can be found on page 19 in the Materials and Methods section), 3 nt linker and one PP7 binding site (GAA-3-PP7$_{x1}$). The left panel shows the nucleotide sequence of the predicted secondary structure of the slncRNA molecule, whereas the right panel depicts the base-pairing probabilities indicated by the color-coded scheme. It can be seen that the DNA$_{bind}$motif is very likely to be single-stranded at 37 ^{o}C, whereas the RBP$_{bind}$domain appears to exhibit the expected hairpin structure with the protruding adenosine[167]. In Figure 10b the nucleotide composition as well as equilibrium probability of the slncRNA with the same DNA$_{bind}$ motif with one, two, three four and five PP7 binding sites is shown. It can be observed that the hairpin structure is well preserved for all PP7-binding sites and the DNA$_{bind}$ motif remains single-stranded.

Following the *in silico* analysis of the slncRNA molecules, I designed the TTS, which functions as the double-stranded, purine-rich stretch on a reporter plasmid that can be bound by slncRNAs via triplex formation. The TTS sequences were extracted from the same publications which I used to obtain the DNA$_{bind}$ motif. The library of slncRNAs and TTS was ordered as linear, dsDNA fragments from Gen9 and subsequently cloned into respective plasmids.

4.2 slncRNAs in a bacterial enhancer circuit

4.2.1 Bacterial design of slncRNA-based enhancer circuit

The bacterial pRNA plasmid contains the N-butyryl-L-Homoserine lactone (C_4-HSL) inducible pRhlR-promoter[170] driving expression of the slncRNAs (Figure 11a, bottom) . The TTS sequences were subcloned into the template and the non-template strand of the reporter plasmid (pRep) that is based on the synthetic enhancer circuit described in Amit *et al.*[142] (Figure 11a, top).

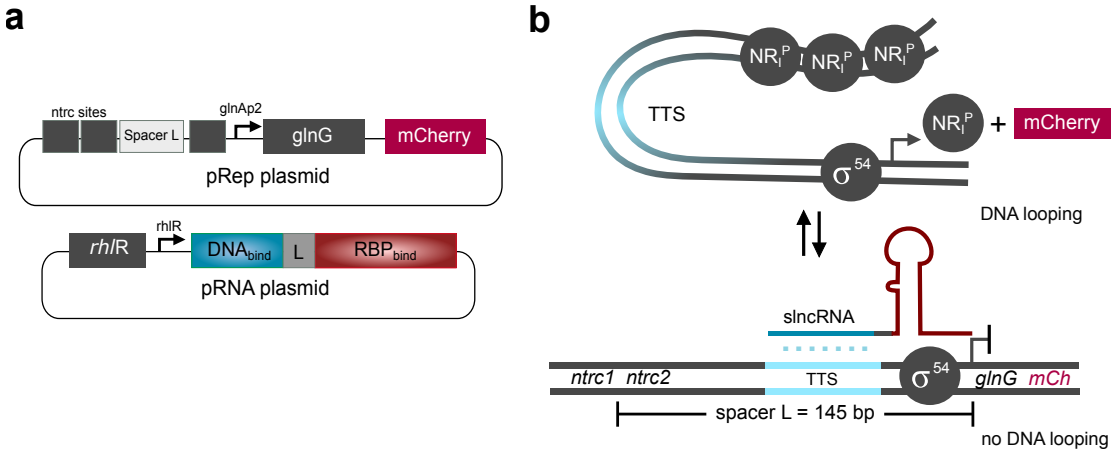


Figure 11| Bacterial enhancer-based circuit design. **a**, For the synthetic enhancer-based circuit, two plasmids are required. The pRNA plasmid harbors the slncRNA molecule and its transcription is controlled by the C_4-HSL inducible rhlR-promoter, and the reporter plasmid pRep that contains the σ^{54} promoter glnAP2 which can be activated by the enhancer NR$_I$ upon DNA looping. A spacer region (L) was introduced into the plasmid between the promoter glnAp2 and the NR$_I$-binding sites (*ntrC*). The spacer region will be used for insertion of triplex target sites (TTS) that are targeted by the DNA$_{bind}$ domains of the slncRNAs. **b**, The basic synthetic enhancer setup was modified[142]. DNA looping is now controlled by slncRNA binding to the TTS. In absence of slncRNAs, DNA looping occurs and the phosphorylated NR$_I$-dimers oligomerize at the enhancer binding sites and initiate DNA transcription by interaction with the poised σ^{54} polymerase. Upon induction of slncRNAs with C_4-HSL, triplex formation might occur and alters DNA looping probabilities thus changing transcriptional *mCherry* rates.

The bacterial enhancers regulate transcription in an enhancer-based mechanism by changing DNA looping characteristics. The enhancer cassette comprises the poised σ^{54} polymerase at the

glnAp2[171;172] promoter driving expression of the *glnG* (ntrC) gene followed by a strong ribosome binding site (RBS) and a mCherry gene as the reporter (Figure 11b). The center of the ntrC-binding sites are located 154 bp away from the center of the glnAp2 promoter referred to as the spacer sequence L that forms the loop. Amit and colleagues showed substantial regulation of looping-based transcription by strategically placing transcription factor (TF) binding sites inside the looping region of the synthetic enhancer plasmid. I hypothesized that triplex-forming slncRNA molecules may be capable of generating a similar regulatory effect by substituting the TF-binding sites with the 17-23 nt long TTS sequences within this looping sequence L. As a control sequence, the reporter plasmids contain a random sequence that should not form triplexes with DNA_{bind} motifs of the slncRNAs. I expected that, upon induction with C_4-HSL, the DNA_{bind} domain of the transcribed slncRNAs targets the TTS located centrally in the looping region of the reporter plasmid (Figure 11b, highlighted in light blue). The triplex interaction that may ensue will either induce a conformational change in the local 'looping' DNA, or alter the topology of the plasmid as a whole. This in turn will generate a shift in bending of the looping DNA or elastic constant that will alter the overall probability of looping of the bacterial enhancer. As a result, the steady state expression of the mCherry reporter will change as well, leading to an indirect detection of the triplex structure.
However, since both structural features of triplexes as well as the kinetics that are associated with such slncRNA-dsDNA interactions, are unknown, it was not clear that a simple design of slncRNA-DNA triplex formation would be sufficient to detect the desired regulatory effect within the context of the bacterial looping assay. As a result, I opted to add RBP binding domains of PP7 to the triplex-forming domain to simultaneously check for two additional regulatory effects, which may act synergistically by either amplifying the triplex effect on looping, or substantially increase the probability of triplex formation in the first place. In the first effect, I hypothesized, based on previous observations with TFs, that binding of a co-expressed tandem-dimer PP7-Cerulean (tdPCP-Cerulean) fusion protein to its cognate PP7-binding sites on a triplex-bound slncRNA molecule, may amplify the regulatory effect generated by the triplex alone. The effect here should be similar to the excluded volume regulatory mechanism that was recently uncovered, when they are bound inside the loop[143]. For the second strategy, we reasoned that a possible pitfall for the slncRNA approach is based on the search mechanism of the triplex site by the slncRNA. Here, unlike for TFs, it is not likely that a facilitated 1D-3D diffusion process is taking place. As a result, the chance that a slncRNA will find its binding site within a context of $>10^6$ potential binding sites is very small given only a 3D diffusion-based search process. Thus, in order to observe a specific triplex binding effect, a very large steady state concentration of slncRNA molecules within the cell may be needed, which may lead to adverse or toxic biological behavior.

4.2.2 Experimental setup of the bacterial enhancer assay

The general procedure of the enhancer-dependent reporter assay was as follows: The E.*coli* K12 strains containing the plasmids pRep and pRNA were grown in low-autofluorescence bioassay (BA) buffer, induced with increasing C_4-HSL concentration and fluorescence intensities of the cell populations were measured over time. Figure 12 shows a typical dose response with biological duplicates of E.*coli* strains containing the slncRNA GGA-3-$PP7_{x2}$ in which GGA represents the DNA_{bind} motif, the number corresponds to linker length followed by the number of PP7 binding site repeats. The optical density (OD) of bacterial growth rises as a function of time (Figure 12a), while the mCherry fluorescence (FL) increases as a function of time and increasing C_4-

HSL concentrations (Figure 12b). FL values were divided by OD, normalized by eliminating background auto-fluorescence levels of the BA buffer and bacterial cells, averaged over a range of 8 hours and plotted against increasing C_4-HSL inducer concentrations (Figure 12c). In this experiment, the pRNA plasmid has been co-transformed with the reporter plasmid (pRep76) containing the corresponding, putative TTS sequence for the respective slncRNAs. For this particular strain, a 2-fold increase in mCherry fluorescence is observed.

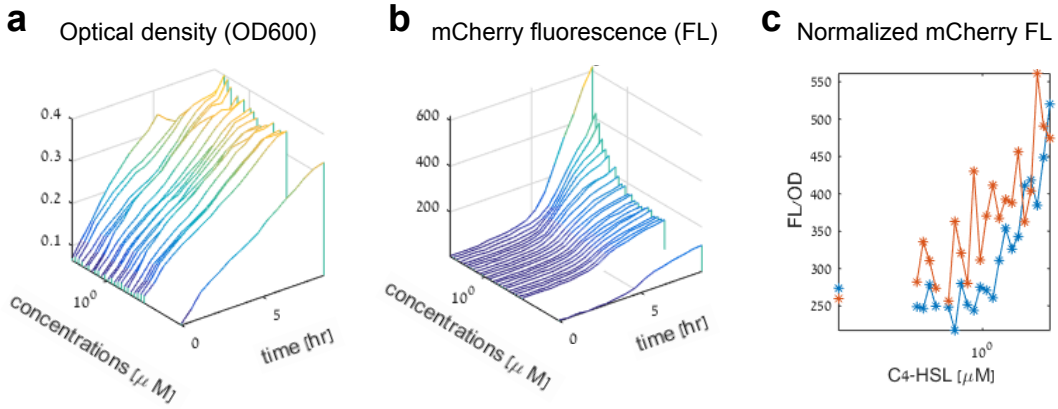


Figure 12| Representative data set of bacterial enhancer-based circuit. The sample set of an E.*coli* strain harboring a slncRNA with the DNA_{bind} motif GGA_{rich}, a 3 nt long linker and 2xPP7-binding sites (GGA_{rich}-3-$PP7_{x2}$) with a reporter plasmid harboring the respective TTS for the DNA_{bind} motif is shown. Top10 cells were induced with increasing concentrations of C_4-HSL and fluorescence/OD measurements were taken every 30 minutes starting 2 hours post-induction over a period of 8 hours during which the **a**, optical density (OD) increases over time and **b**, the mCherry levels rise as a function of time and C_4-HSL concentration. **c**, Normalized mCherry values (FL/OD) of biological duplicates (orange and blue lines) have been plotted against increasing C_4-HSL concentrations. It is clear that biological duplicates behave similarly, but overall the data is noisy. Nevertheless, a 2-fold increase in mCherry levels can be observed for these strains.

4.2.3 TTS-independent, enhancer-based upregulation of reporter gene

In total 556 E.*coli* strains have been tested, including biological duplicates, different slncRNAs and reporter plasmids. To analyze the large volume of collected data and compare across slncRNA adn reporter strains, fluorescence values recorded at low inducer concentrations for individual data sets were normalized to one. To do this, the FL/OD values of every measurement in a particular data set were divided by the average of the two minimal fluorescence levels in absence of the inducer C_4-HSL and at 0.018 µM. These normalized data sets, termed 'fold-change', were then used for a composite analysis of multiple configurations. To properly analyze the composite 'fold-change' data, I first computed a suitably weighted (due to different numbers in measured strains per data set or inducer concentration) distribution of fold-change values at a particular C_4-HSL concentration. In each graph below, I therefore use a gray-scale heat-map plot to depict the underlying distribution with dark and light gray corresponding to large and small bin occupancy values, respectively. In addition, I overlay the concentration-specific heat-maps with a data point corresponding to the value of the first moment (e.g. mean) for each distribution.

In Figure 13a, I compared data sets of the normalized fold-change of E.*coli* strains containing pRNA plasmids co-transformed with either the putative TTS (green marker) inserted in the looping region or reporter plasmids containing random sequences (blue marker). In this case, the fold-change levels calculated for slncRNAs with four different DNA_{bind} motifs (GAA_{rich}, $(GAA)_{x8}$, AA_{rich} and pyr_{rich}) were averaged (for further information about DNA_{bind} motifs check Table 1 on page 19 in the Materials and Methods section) as no difference in up-regulatory behavior has been observed (data not shown). Putative TTS sequences correspond to each slncRNA

with respective triplex-forming DNA_{bind} motif, whereas control sequences remain the same for each co-transformed slncRNA. Contrary to my expectations, the data shows that first moment values of both the TTS (blue markers) and non-TTS containing target sequences are indistinguishable across all inducer concentrations for the four different DNA_{bind} motifs. Interestingly, however both data sets also exhibit a consistent upward trend in fold-change. I observe a first slight increase (20 %) starting at 1 µM C_4-HSL followed by a second up-regulatory response (50 %) between 40 µM and 218 µM C_4-HSL. In addition, the distributions (gray-scale boxes) at higher inducer concentrations are much wider as compared with lower inducer concentrations. Together, these results indicate that while a difference due to the TTS is not observed, some form of interaction is apparently taking place leading in some cases to a 2-fold increase in expression.

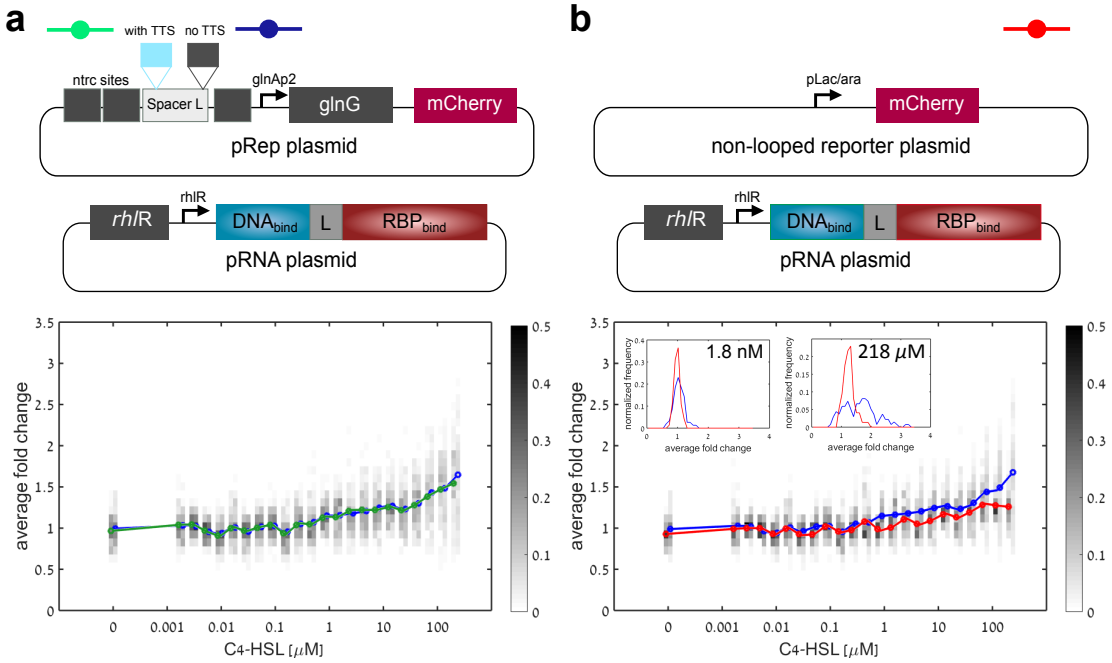


Figure 13| Up-regulatory effect mediated by enhancer-based reporter plasmids. **a**, Schematic representation of experimental setup and comparison of data sets of looped reporter. The green markers represent slncRNAs with four different DNA_{bind} motifs that have been co-transformed with reporter plasmids with respective TTS, whereas the blue marker contains strains with the same slncRNAs, but the control reporter plasmid lacking putative TTS sequences. The first moment values of fold-changes were plotted against increasing C_4-HSL concentrations. The gray-scale boxes represent distributions corresponding to bin occupancy values at given concentrations. Both data sets display similar up-regulatory trends towards higher concentrations starting at 1 µM C_4-HSL followed by a second, up-regulatory behavior between 40 µM and 218 µM C_4-HSL. **b**, The pLac/ara-mCherry plasmid containing a σ^{70} promoter is a non-looping based plasmid and is co-transformed with pRNA plasmids and compared to the control sequences from the left figure. The data set containing the non-looped reporter plasmid (red markers) exhibits the similar, slightly up-regulatory response in presence of increasing concentrations of C_4-HSL compared to the set of looping-based reporter plasmid (blue markers). Examining higher C_4-HSL concentrations, the looping-based reporter displays a higher up-regulatory increase than the σ^{70} containing data set that reaches a plateau. Even more intriguingly is the variability and shape of the distributions. While the non-looped reporter displays a unimodal distribution of fold-changes (right inset), several peaks are formed for the looping-based reporter at concentrations between 70 µM and 218 µM, whereas at low C_4-HSL concentrations distributions overlap (left inset).

As I could not distinguish between strains with the putative TTS compared to strains lacking the TTS, I next asked whether the up-regulatory effect can be observed in reporter plasmids containing a non-looping dependent promoter (σ^{70}) as well (Figure 13b). For this, I co-transformed pLac/ara-mCherry, containing the constitutive σ^{70} pLac/ara promoter, which does not require looping to regulate gene expression, with various pRNA plasmids and compared the distributions and first moment values of fold-changes for each inducer concentration with the results obtained for the control sequences shown. For concentrations until 1 µM, no up-regulatory response can

be noted for both enhancer-based (blue marker) and non-looped reporters (pLac/ara-mCherry, red marker). At concentrations ranging between 1 µM - 40 µM C_4-HSL, I observe a similar, slightly up-regulatory response for both data sets. When examining the three highest concentrations of the looping-based reporter, I recognize a slight up-regulation trend for the looping-based reporter, whereas a plateau is reached in the σ^{70}-containing reporter, but it is questionable whether the effect is significant given the relative distributions of the values (gray-scale boxes). Intriguingly however, both the variability and the overall shape of the distributions diverge markedly between the strains carrying the looping-based σ^{54} promoter and the σ^{70} promoter as the C_4-HSL concentration increases. In particular, at the highest concentration (Figure 13b, 218 µM – right inset), the distribution for the looping case is characterized by multiple peaks that are spread over a large range of fold-ratios, while the non-looping peak forms a unimodal distribution of fold ratios. In addition, the distributions do not overlap leading to the shift in first moment values plotted in Figure 13b. This is contrasted by the distributions shown at lower concentrations (see left inset: 1.8 nM), where the data sets for both promoter types overlap and are nearly indistinguishable. Even though the data shown in Figure 13a+b provides support for the hypothesis that some sort of interaction is taking place between the DNA loop and the slncRNA, I cannot draw definitive conclusions.

4.2.4 Control strains strengthen slncRNA involvement in up-regulation

I further wanted to further verify that the slight up-regulatory response in both the non-looped reporter and the two-step effect in the enhancer-based reporter is not an artifact of high C_4-HSL concentrations.

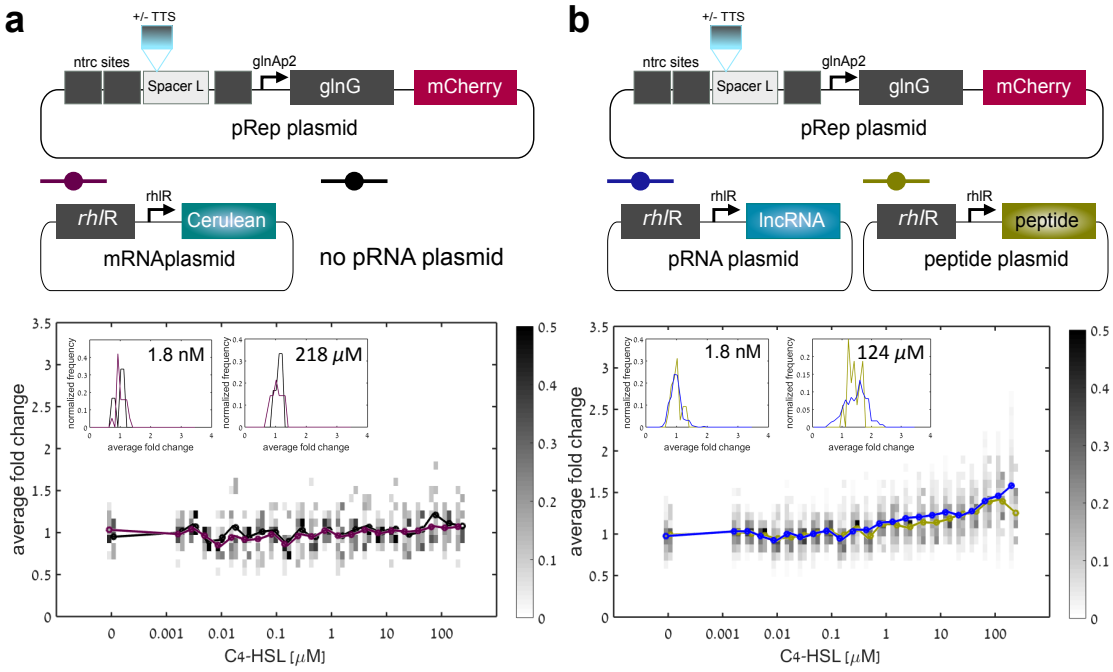


Figure 14| mCherry expression changes in absence of slncRNAs and presence of mRNA. a, E.*coli* strains have been transformed either with the looped reporter plasmids and a mRNA plasmid that encodes tdPP7-Cerulean, or with the looped reporter plasmid only. No significant up-regulation of mCherry can be observed for neither of the two data sets (purple markers, co-expression of mRNA, black markers, reporter plasmid only). Even at high concentrations, the mCherry values stay at basal levels. Hence, transcription of mRNA alone does not enhance mCherry transcription. **b**, E.*coli* strains containing the slncRNA sequences with an additional start codon (ATG) thereby encoding a short peptide termed peptide plasmid (yellow markers) were compared to the E.*coli* strains lacking the ATG (blue markers). The data sets of strains with and without ATG exhibit a similar up-regulatory increase up until 124 µM C_4-HSL, but the peptide plasmid strains display a unimodal distribution of fold-changes (right inset) at high C_4-HSL concentrations.

To do so, I examined control sets expressing no slncRNAs, mRNAs and slncRNAs that contain an open-reading frame (ORF) termed slncRNA peptide (Figure 14). In Figure 14a, the black marker corresponds to E.*coli* strains that lack the slncRNA-expressing plasmid and contain only the mCherry reporter plasmid. Thus, reporter distribution levels should be basal, independent of inducer concentrations. The second data set (purple marker) corresponds to normalized mCherry values for Top10 strains that contain the reporter plasmids and pRNA plasmids that instead of slncRNA transcribe mRNA. I chose the open reading frame encoding *pp7-cerulean* as mRNA for visualization of its expression (data not shown). I observed the same average fold-change and no significant up-regulation of mCherry for both control sets indicating that in absence of slncRNAs no regulatory effect on transcription occurs. Furthermore, insertion of a start codon, thus encoding slncRNA peptides (Figure 14b), reduced the observed up-regulatory effect thus indicating that the slncRNAs are indeed involved in enhancing gene expression.
Initially, I designed and constructed various slncRNAs with five different DNA_{bind} motifs. As shown in Figure 13, four DNA_{bind} domains resulted in a similar, up-regulatory response independently of the TTS in the loop of the reporter plasmid. Only the fifth DNA_{bind} sequence termed T0 shows a slightly different behavior dependent on the TTS of the reporter plasmid (Figure 15). The DNA_{bind} motif T0 derives from the core promoter of rDNA genes and has been reported to be the responsible sequence for triplex formation, DNA methylation and subsequent gene silencing of eukaryotic rDNA genes[10]. Interestingly, this sequence does not follow the predicted triplex-forming code that states that either polypurine or polypyrimidine stretches are required for triplex formation.

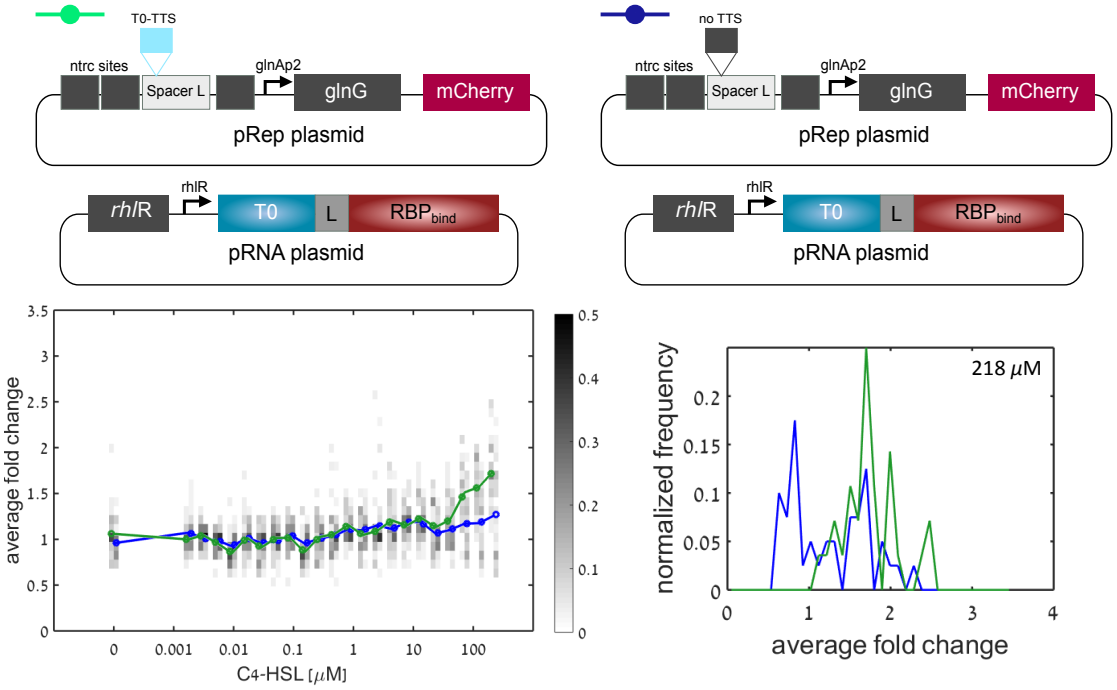


Figure 15| Specific up-regulatory response using slncRNAs with DNA_{bind} motif T0. The triplex-forming sequence T0 has been shown to form triplexes at eukaryotic promoters of rDNA genes *in vitro*[10]. T0/slncRNA-containing plasmids were co-transformed with the respective reporter plasmids comprising the putative TTS and were compared to strains carrying the control reporter plasmids. Contrary to the other four DNA_{bind} motifs that did not differ from the control set (blue markers), the up-regulatory response starting from 40 µM differs between the two data sets of T0-slncRNAs varies (green markers). While the set with control sequences exgibits the 20 % up-regulatory effect, the increase of mCherry for T0 with its putative TTS is (green marker) considerably stronger (70 %) and the distributions (right panel) are similar, but shifted to the right.

The data illustrates the differential response in the two sets (green marker, T0-containing re-

porter; blue marker, control sequence in looping sequence) starting at 40 µM C_4-HSL. While the control strains respond to increasing inducer concentrations from 1 µM-218 µM in the same way as the other control sequences in Figure 13 on page 52 (blue marker with line), the strains that contain the reporter plasmids with putative T0-TTS insertions in the looping sequence display a stronger up-regulatory effect. This indicates a potential sequence dependent up-regulation, which may be due to triplex formation.

4.2.5 AT-rich spacer sequence reduces non-specific up-regulatory effect

One possible explanation of the non-specific up-regulatory effect of mCherry might be the promiscuity of potential triplex target sites within the spacer sequence L flanking the actual TTS insertion site (Figure 16a, upper image). Therefore, the spacer sequence was altered by removing longer stretches of purine/pyrimidine stretches. To achieve this, most of the sequence was replaced with adenines/thymines (AT) repeats (Figure 16a, lower image). This AT-rich spacer was cloned into the enhancer-circuit plasmid and transformed to E.*coli* cells harboring the pRNA plasmids and mCherry expression levels were compared to strains containing the original spacer (Figure 16b). All E.*coli* strains shown in these data sets lacked specific TTS within the spacer. As has been shown before, E.*coli* strains with the original TTS-promiscuous spacer (blue markers) displays non-specific mCherry up-regulation and a wide distribution of fold-change values at high C_4-HSL concentrations (inset, blue line), the E.*coli* strains with the AT-rich spacer sequence exhibit a reduced up-regulatory effect and only at high C_4-HSL concentrations (> 70 µM) shows a minimal up-regulation of mCherry. Intriguingly, the distribution of these E.*coli* strain lack the wide distribution (Figure 16b, insets) and are comparable to the non-looped reporter plasmids (Figure 14a).

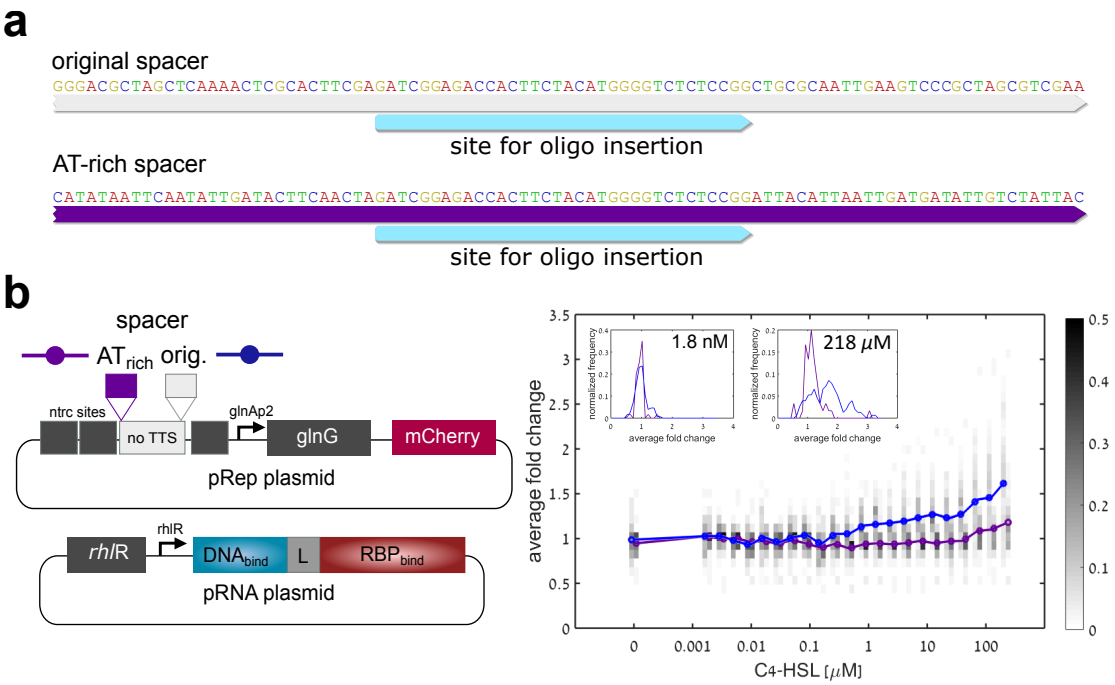


Figure 16| Modified spacer sequence reduces non-specific up-regulation. a, The original spacer sequence contains promiscuous purine/pyrimidine stretches. The purine-rich stretches were replaced with AT-repeats. **b**, Insertion of the AT-rich spacer (purple markers) reduces the non-specific up-regulatory effect observed in E.*coli* strains with the original spacer (blue markers) and a slight up-regulation can be observed for high C_4-HSL concentrations while the distributions (inset, purple line) remain unimodal.

4.2.6 Co-expression of RNA-binding proteins and slncRNAs

Lastly, I asked whether co-expression of a RBP-fusion protein influences the probability of looping in the enhancer circuit (see Figure 11c). Therefore, I introduced the tandem-dimer *PP7 phage coat protein* (tdPCP) gene into the pRNA plasmids upstream of the rhlR gene that is constitutively expressed. I co-transformed the pRNA-RBP plasmid with respective reporter plasmids and induced slncRNA transcription with increasing C_4-HSL concentrations. I compared the data sets with the same reporter and pRNA plasmids in absence and presence of the td-PCPs. Contrary to my expectations, mCherry expression as well as distributions display no differential behavior in presence (Figure 17, orange markers) or absence (Figure 17, gray markers) of the tdPCP proteins with reporter plasmids containing putative TTS sites. This indicates that the slncRNAs interact with the reporter plasmid in a protein-independent manner and the tdPCP-fusion protein does not influence local looping probabilities.

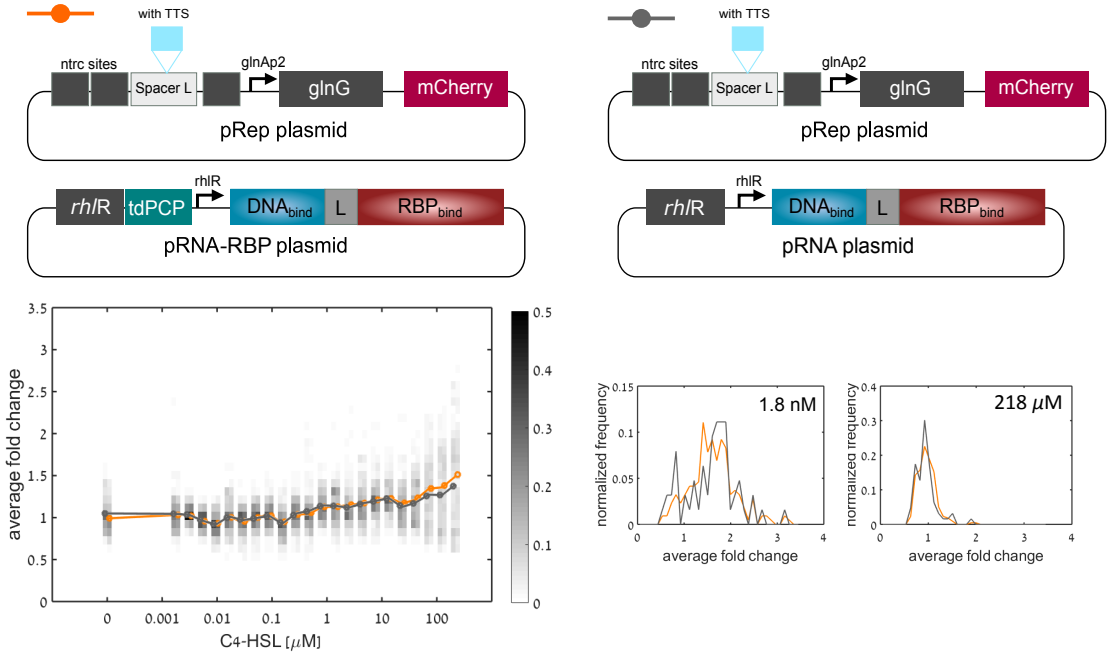


Figure 17| Comparison of mCherry fold-changes in absence or presence of tdPCP. Exemplified representation of experimentally tested E.*coli* strains. mCherry expression was measured in presence and absence of tandem-dimer phage coat protein (tdPCP) fused to Cerulean. No significant up-regulation of mCherry can be observed for neither of the two data sets when comparing the reporter plasmids with putative TTS in presence (orange marker) and absence (gray marker) of tdPCP-Cerulean (left panel) as well as when comparing distributions (right panel).

4.2.7 Summary bacterial enhancer-based assay

The enhancer-based circuit was designed to be responsive to slncRNA binding in the DNA looping region via triplex formation thereby inducing changes in mCherry expression levels. I based the idea of slncRNA-mediated gene regulation on work using synthetic enhancers that responded to transcription factors which were placed inside or outside the looping region and induced significant up- or down-regulation [142;143]. Unfortunately, this TF-dependent regulatory effect could not be exhibited in my experimental setup using slncRNAs. However, I did observe a non-specific up-regulatory effect independent of (i) slncRNAs the TTS within the spacer sequence L (Figure 13). Furthermore, insertion of a start codon into the slncRNA open reading frame (Figure 14b) reduced the non-specific up-regulation trend observed previously. While the up-regulatory effect of mCherry seems to be derived from DNA looping (Figure 14a) and exchange

of the purine-rich spacer with AT-repeats reduced this non-specific regulatory effect (Figure 16), I believe that pursuing this line of experimental setup is sub-optimal at this stage and yields inconclusive results. Thus, I decided to move on to the mammalian chassis and test the ability of slncRNAs to activate gene expression in a CRISPR/Cas9-like circuit.

4.3 Mammalian slncRNAs in gene-activation circuit

To study slncRNA-mediated gene activation in mammalian cells, I designed a CRISPR/Cas9-like gene activation circuit by applying the RNA-guided Cas9 activation of a reporter gene to a solely RNA-mediated gene activation circuit (Figure 18). To do so, I exploited the modular design of the slncRNAs (Figure 9b) and cloned them into a mammalian vector. The setup in eukaryotic cells for analyzing and evaluating each module of the slncRNAs is as follows: the slncRNA consists of a single DNA_{bind} motif and a RBP_{bind} domain connected by a flexible linker. I hypothesize that these DNA_{bind} motifs of the slncRNAs will interact with a putative TTS sequence located upstream of a minimal CMV promoter ($pCMV_{min}$) on the reporter plasmid. Triplex formation between the slncRNA and the putative TTS activates expression of the yellow fluorescent protein (eYFP) located downstream of the $pCMV_{min}$. Monitoring the fluorescent reporter protein levels allows for the analysis of the functionality of the slncRNA components.

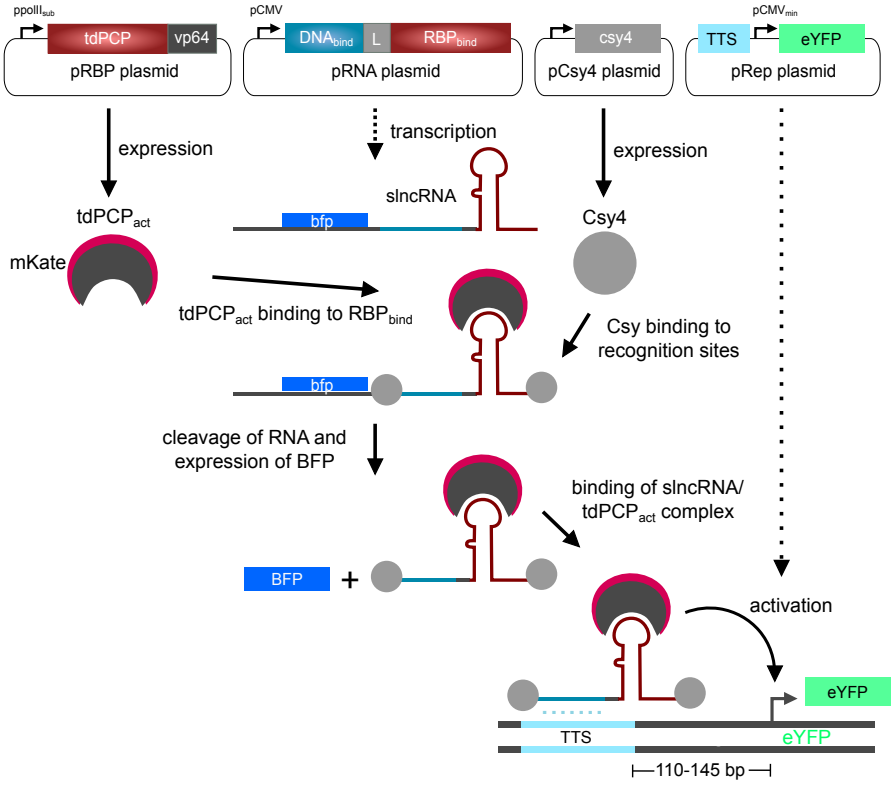


Figure 18| Design of mammalian slncRNA gene-activation circuit. For the functionality of the system, four constructs are required. Plasmid pCsy4 encodes the endonuclease Csy4 that recognizes and cleaves a 28 nt long RNA stretch that flanks the slncRNA seqeunce. Plasmid pRNA carries the *strongly enhanced blue fluorescent protein (sbfp2)* gene as a reporter for efficient expression and the slncRNA, which consists of two modules: (1) DNA_{bind} motif and (2) RBP_{bind}. Plasmid pRBP encodes a fusion protein ($RBP_{activator}$) comprising the fluorescent reporter mKate2, a RNA-binding protein (RBP) and a viral transactivator (vp64). Plasmid pRep, the reporter plasmid, contains a *yellow fluorescent protein* (*yfp*) gene under control of a minimal cytomegalovirus (CMV) promoter. Cleavage by the Csy4 endonuclease releases the slncRNA and the transcript *sbfp2*. The slncRNAs will be bound by fusion proteins ($RBP_{activator}$). The RNA-protein complex is relocated into the nucleus and the DNA_{bind} motif of the lncRNA interacts with the TTS on the reporter plasmid pRep. Triplex formation brings the transactivator in close proximity of $pCMV_{min}$ thus initiating transcription and fluorescent levels of SBFP2, eYFP and mKate2 are measured.

While most CRISPR/Cas9 approaches use polymerase III to transcribe the short guide RNAs (gRNAs), Nissim *et al.* published in 2014 the usage of polymerase II to regulate gRNA transcription [154]. Various approaches to achieve gRNA transcription from a polymerase III promoter were tested in the publication, and I chose to use the endonuclease Csy4 platform. Csy4 is known to recognize a specific RNA-hairpin and cleaves immediately downstream of the secondary structure

and has been successfully applied in the publication by Nissim and colleagues. In my system which is schematically represented in Figure 18, the slncRNA will be DNA-encoded (plasmid pRNA) and flanked by recognition sites for the endonuclease Csy4 that allows for the usage of a polymerase II apparatus and prevents rapid degradation of the slncRNAs due to Csy4 binding. As mentioned above, the slncRNA modules consist of a DNA_{bind} domain and the the RBP_{bind} module and are inserted into the pRNA plasmid downstream of the blue fluorescent protein (*sbfp2*) gene. The reporter plasmid (plasmid pRep) contains a minimal CMV promoter unit with the putative TTS sequences placed 110-145 bp upstream of the promoter. The $pCMV_{min}$ can be activated by the viral transactivator vp64[151]. The transactivator in turn is fused to the tandem RNA-binding protein PP7[153] ($RBP_{activator}$). Co-transfection of the four plasmids is expected to lead to transcription of the slncRNAs and SBFP2 mRNA with subsequent slncRNA cleavage by Csy4, protein expression of SBFP2, binding of the $RBP_{activator}$ to the slncRNAs via tdPP7 and activation of eYFP gene expression on the reporter plasmid via triplex formation of the slncRNA with the TTS upstream of the $pCMV_{min}$. Fluorescent proteins (SBFP2, eYFP and mKate2) have been carefully chosen using the fluorescence spectrum viewer multicolor tool provided by BD Biosciences to reduce spillover of fluorescence (FL) into the other FL channels and to establish a system for analyzing each component of the slncRNA-dependent gene activation. The slncRNA library contains the sequences for the slncRNAs (see Table 1 for variations in DNA_{bind} motif, linker length and number of RBP_{bind} modules) and the TTS sequences that are inserted into the reporter plasmids (see Table 3). The reporter plasmids contain putative TTS sequences and varying TTS positions upstream of the promoter region (-110 bp and -145 bp). Positions of the TTS have been chosen in accordance with unpublished data about multiplexing networks using CRISPR/Cas9 (personal correspondence with L. Nissim). The construction of the rationally designed slncRNA library provides the basis for a first broad round of screening and allows for preliminary conclusions of the experimental setup such as determination of optimal parameters for positioning of TTS sequences in reporter plasmids, as well as linker length and number of RBP-binding sites in the slncRNAs.

4.3.1 Characterization of expression and localization of reporter proteins

To characterize successful expression of fluorescent proteins (SBFP2 and mKate2), the RNA and pRBP plasmids were transfected into human embryonic kidney cells (HEK-293), respectively, fixed with paraformaldehyde 48 hours post-transfection, and visualized using the inverted microscope Nikon Eclipse Ti-E (Figure 19a). While the transfection efficiency for the pRNA plasmid ranges between 70-80 % (Figure 19a, bottom), transfection efficiencies for the RBP plasmid are lower (10-15 %) and a broader variation in mKate2 intensities between cells can be observed (Figure 19a, top).

To assess simultaneous mKate2 and SBFP2 expression and localization, all four plasmids (pRNA, pRBP, pCsy4, pRep) that are required for the functionality of the slncRNA-mediated gene activation circuit (Figure 19b) were co-transfected and expression of mKate2 and SBPF2 was imaged 48 h post-transfection. Examining the microscopy pictures, SBFP2 is localized throughout the cytoplasm as well as the nucleus, while mKate2 ($RBP_{activator}$) is predominantly located in the nucleus as expected due to the fusion of a nuclear localization signal (NLS) to the $RBP_{activator}$. Furthermore, in some cells, brighter mKate2 sports can be detected. This might be due to the presence of the slnRNA molecule which provides a docking platform to which the $RBP_{activator}$ can bind. As has been shown previously, imaging of single RNA molecules in mammalian cells using the RBP-mRNA labeling technique resulted in distinct spots in the cytoplasm[173] or nucleus[174].

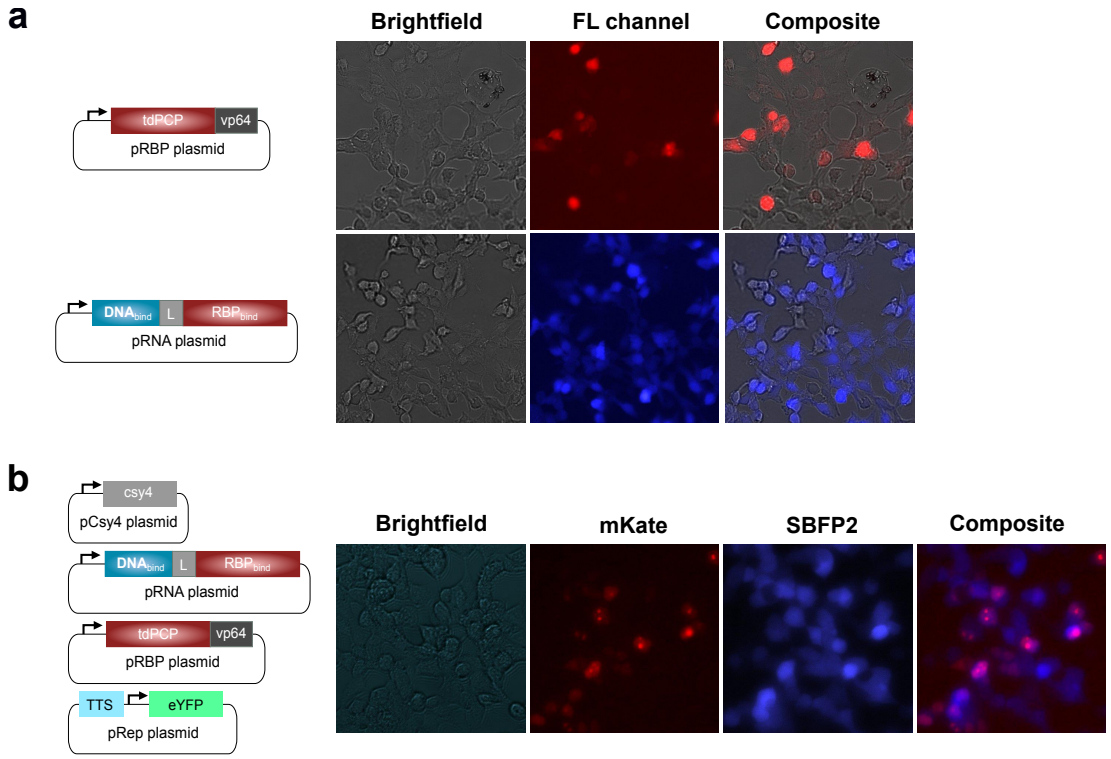


Figure 19| Microscopy analysis of transfection efficiencies and protein localization. **a**, The pRBP and pRNAs plasmid have been transfected into HEK-293 cells, respectively and fluorescence was imaged using fluorescence microscopy. 48 h post-transfection, HEK-293 cells were fixed on coverslips with paraformaldehyde. 70 %-80 % positive cells for the plasmid pRNA was observed, whereas 5 %-10 % transfection efficiencies for the mKate2-NLS (pRBP) plasmid can be seen. **b**, To confirm that the $RBP_{activator}$ is located in the nucleus, I co-transfected all four plasmids (pCsy4, pRNA, pRBP and pRep) into HEK-293 cells and detected fluorescence 48 post-transfection. In the composite picture it can be seen that while SBFP2 is located both cytoplasmatically and nuclear, mKate2 ($RBP_{activator}$) is located mostly in the nucleus and forms distinct spots.

4.3.2 Flow cytometry analysis of protein expression in human cells

To validate the microscopy results, I compared the transfection efficiencies of all four plasmids (pRNA, pRBP, pCsy4 and pRep) with transfection efficiencies when one of the four plasmids was omitted using flow cytometry analysis. A representative dotplot analysis of the flow cytometry data is shown in Figure 20. Transfection efficiencies of the pRNA plasmid (SBFP2 fluorescence) are high (70-80 %), while mKate2 expression levels (representing transfection of $RBP_{activator}$) are lower and range between 20-30 % (Figure 20). Interestingly, mKate2 levels seems to increase when the endonuclease Csy4 is present. If Csy4 is absent mKate2 positive cells are lower than 10 % (Figure 20a, second from the right), while the percentage increases to 19 % when all four plasmids are present (Figure 20a, left) and is highest when Csy4 is present and the slncRNA is absent (40 %, Figure 20a, right). The significant difference of mKate2 positive cells with and without slncRNA might be due to the strong, constitutive promoter CMV that is placed upstream of the slncRNAs thereby occupying large fractions of the transcription and translation machinery. The lower mKate2 expression levels in absence of Csy4 cannot be explained at this time. Based on the observation of SBFP2 and mKate2 expression, I calculated the weighted median fluorescence (FL) of eYFP, SBFP2 and mKate2 by multiplying the percentage of positive cells with the median FL intensity of those positive cells. This calculation has been previously described to represent a valid method to quantify FL levels [154].

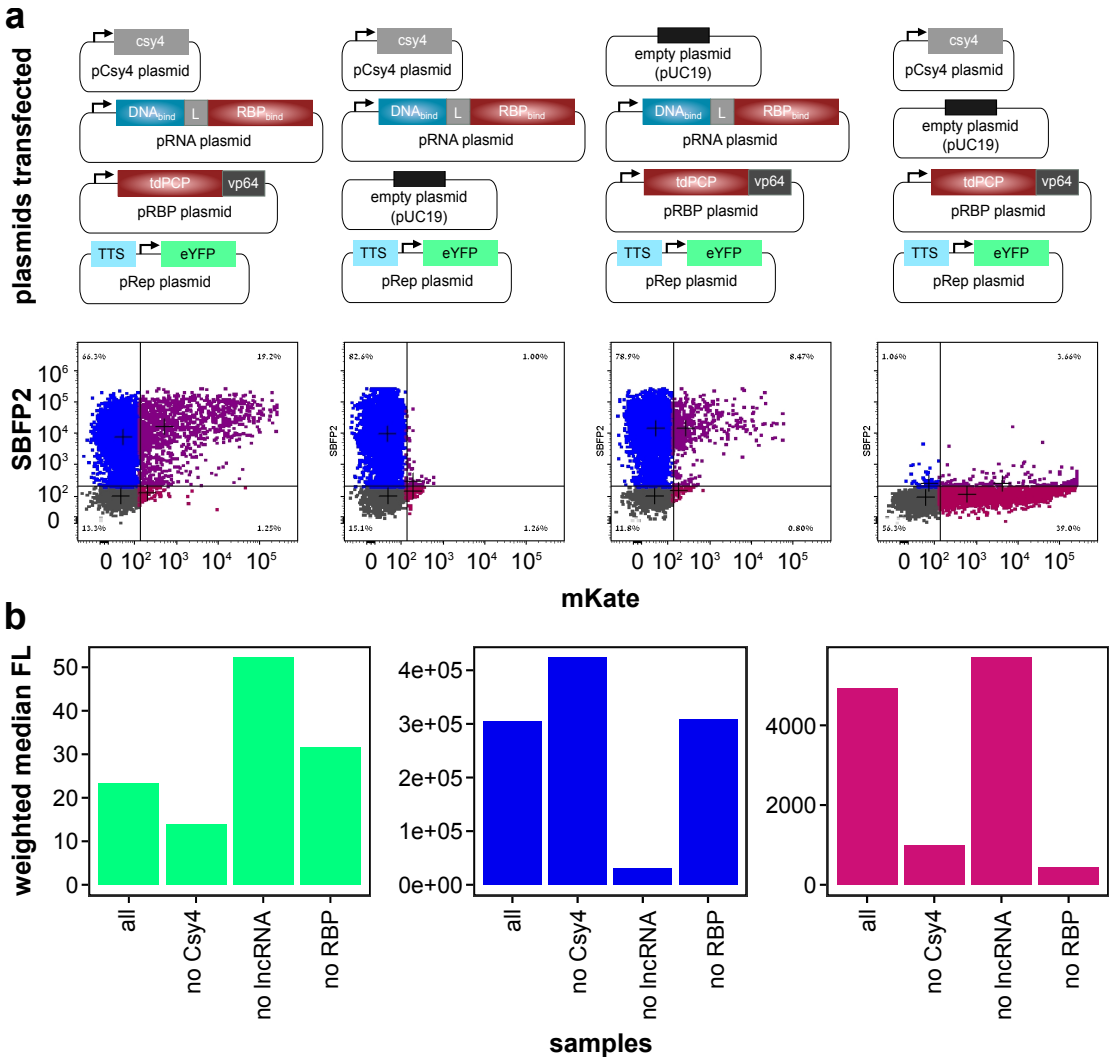


Figure 20| Flow cytometry analysis of transfection efficiencies. HEK-293 cells were transfected and flow cytometry analysis was performed on the MACSQuant analyzer 48 hours post transfection. **a**, Dot-plot analysis of SBFP2 (pRNA plasmid) positive HEK-293 cells and mKate2 (RBP plasmid) positive cells is shown. Each dot represents one exemplary sample with either all four plasmids (left), lacking pRBP plasmid (second from the left), lacking pCsy4 (second from the right) and lacking the pRNA plasmid (right). Little spillover is observed of the compensated fluorescence channels and confirms frequency of positive cells exhibited using the microscope. **b**, Bar plots of weighted median fluorescence (median fluorescence intensities multiplied by % of positive cells) of all three channels (eYFP for reporter plasmid, SBFP2 for pRNA plasmid and mKate2 (for pRBP plasmid).

The weighted median FL intensity of each three FL channels is shown in Figure 20b. While eYFP levels are low for all samples and range close to background noise, strong FL intensities are observed for SBFP2 expect in the sample where no slncRNA plasmid was transfected. Median FL intensities can be observed for mKate2 and confirm the lower mKate2 expression levels in the sample where Csy4 has not been transfected. Background mKate2 expression can be seen in the sample where the $RBP_{activator}$ was omitted.

4.3.3 Screen of slncRNA/TTS mix reveals little to no gene activation

A total of 22 different slncRNAs with 13 different reporter plasmids were tested (Figure 21). Each slncRNA contains a DNA_{bind} motif, a linker as well as 1-5 PP7-binding sites. The reporters contain different TTS motifs that have been inserted upstream (110 to 145 bp) of the $pCMV_{min}$ promoter. For detailed information on the slncRNAs and reporters tested, see Table 2 on page 20 (for slncRNAs) as well as Table 3 on page 22 (for TTS/reporters). To compare the samples within

a data set, I normalized the weighted median eYFP values of the samples with and without slncRNAs (Figure 21a, left) by dividing the weighted median eYFP values of the samples with and without slncRNAs by the eYFP values of samples with the reporter only.

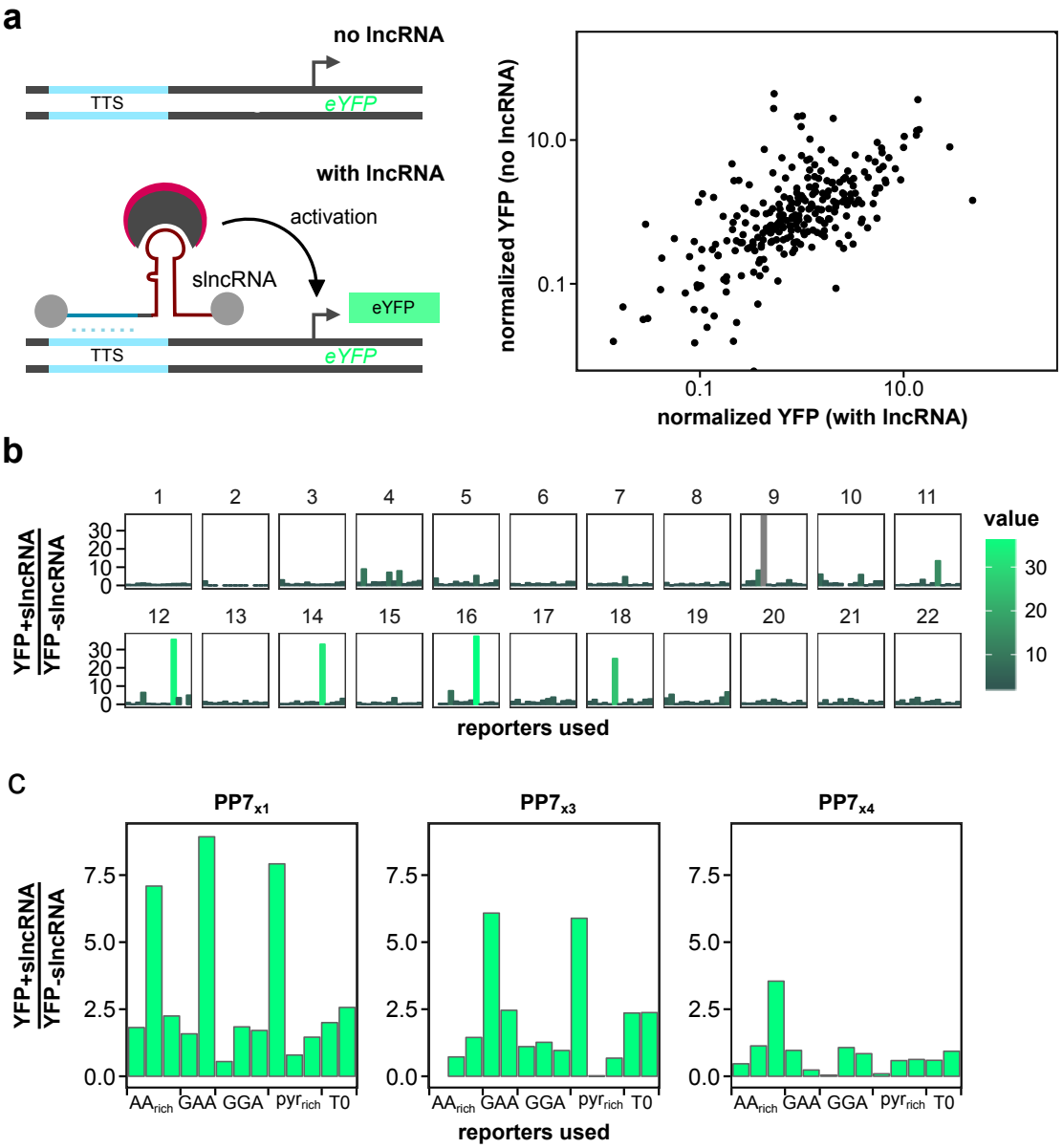


Figure 21| Analysis of slncRNA-mediated gene activation. 22 constructs were transfected along with 13 different reporter plasmids into HEK-293 cells and eYFP expression levels were analyzed 48 hours post-transfection. To compare expression levels of samples with and without slncRNAs, the weighted median eYFP values were normalized by dividing them by the weighted median eYFP values of the basal expression activity of the reporter plasmids only. **a**, Normalized eYFP values of samples without slncRNA were plotted against samples with slncRNAs. Each dot represents the normalized eYFP values for one reporter plasmid. As can be seen most samples show no activation. **b**, To check whether the samples exhibit a trend for activation for a particular slncRNA (#1-22), I plotted the ratio of samples with and without slcnRNA as a barplot in which each bar represents a reporter plasmid. Some slncRNAs display higher eYFP values. **c**, To detect a pattern in the samples that showed higher eYFP ratios, I selected sample #4 which contains the slncRNA GAA-40-PP7$_{x1}$that activates gene expression for reporters that comprise the TTS AA$_{rich}$, GAA$_{rich}$ and pyr$_{rich}$ motifs (left). This trend decreases when looking at slncRNAs with the same DNA$_{bind}$ motif, but an increase of PP7-binding sites (middle and right panel).

In Figure 21a (right panel), the normalized eYFP values of all tested samples without the slncRNAs were plotted as a function of the normalized eYFP values of all samples with the slncRNAs. In the scatter-plot, no difference was made between the reporter plasmid that was

co-transfected. Contrary to the expectations, in which I would expect higher eYFP values (indicating an increase in eYFP expression) for the data sets containing the slncRNAs (x-axis) compared to the samples where the slncRNA is absent (y-axis). While some samples seem to exhibit 10-fold higher eYFP values, most eYFP levels cluster at around 1 which indicates no change of eYFP reporter levels in presence or absence of the slncRNAs.

As this representation does not distinguish between any difference in reporter plasmids and slcnRNAs, I plotted the ratio of the normalized eYFP values shown in Figure 21a from samples with slncRNAs (+ slncRNAs) and without slncRNAs (- slncRNAs) in Figure 21b. Each slncRNA that was tested (indicated by the number on top of each plot) is plotted against the 13 different reporters. While most slncRNAs do not activate gene expression, some slncRNAs potentially activate gene expression (e.g. slncRNA #4, slncRNA #12, slncRNA #14, slncRNA #16 and slncRNA #18).

Having a closer look at data set #4, which corresponds to the slncRNA GAA-40-PP7$_{x1}$(Figure 21c, left panel), I would expect that the higher eYFP ratios correspond to the reporters with corresponding TTS which would be in this case the GAA-TTS. I do observe higher eYFP ratios for the GAA-TTS placed 145 bp upstream of pCMV$_{min}$ on the reporter plasmid, but similar eYFP ratios can also be observed for AA-TTS as well as pyrimidine-rich (CT) TTS indicating that the eYFP increase might not be triplex mediated. Furthermore, the trend that was observed for the slncRNA with one PP7-binding site (PP7$_{x1}$) should be similar if we increase the number of PP7-binding sites (see middle and right panel in Figure 21c.), but there seems to be an overall decrease in eYFP levels with increasing numbers of PP7-binding sites and no induction for the reporters harboring the GAA-TTS can be observed.

4.3.4 Lack of TTS position-related up-regulatory effect

While plotting the distribution of the eYFP ratios for each reporter construct as has been done in Figure 21a, I noticed that there might be a difference in eYFP levels depending on the position (-110 bp vs -145 bp) of the TTS with respect to the minimal promoter (Figure 22).

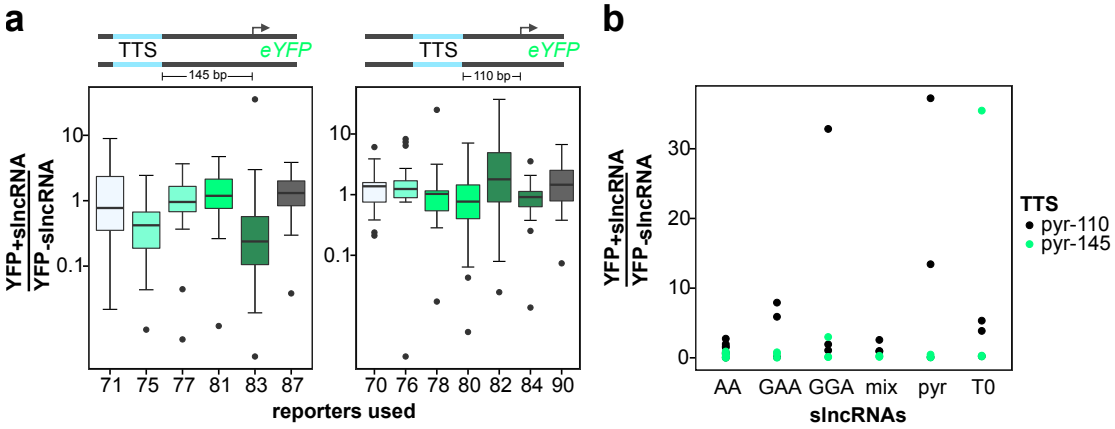


Figure 22| slncRNA-guided eYFP activation as a function of distance. a, Distributions of eYFP ratios for samples that were co-transfected with reporter plasmids in which the TTS was inserted 145 bp (left boxplot) and 110 bp (right boxplot) upstream of the pCMV$_{min}$ promoter. Overall eYFP ratios are slightly higher if they are closer to the minimal promoter (110 bp) compared to 145 bp. Two reporters exhibit opposite behaviors. While the reporter #83 (pyr$_{rich}$ TTS-motif) shows slight down-regulation in presence of slncRNAs, reporter #82 (pyr$_{rich}$ TTS-motif) slightly increases eYFP ratios. **b**, To have a closer look at these two reporters, I plotted the eYFP ratio for both reporters (#83 = pyr-145, #82 = pyr-110) against the slncRNAs with a specific DNA$_{bind}$ motif. While most slncRNAs did not increase eYFP ratios for the pyr-145 reporters, slncRNAs with the pyr$_{rich}$ and GAA$_{rich}$ DNA$_{bind}$ motif increase eYFP expression. The mix of slncRNAs corresponds to samples in which the eYFP ratios were sampled over slncRNAs with AA, GAA, GGA, pyr and T0 motifs.

In Figure 22a, the distributions of the eYFP ratios were plotted as a function of reporters (numbers indicate different reporters, details on page 22) and sorted according to their distance. The median of almost all samples ranges at around 1 indicating and confirming that no slncRNA-mediated gene activation occurs. Two reporters (#82 and #83) that differ only in their position while maintaining the same TTS (pyrimidine-rich TTS) and orientation (TTS has been inserted in the template strand) display opposite behaviors. While the reporters with the TTS being inserted 145 bp upstream of the minimal promoter exhibit a slightly down-regulatory effect of eYFP levels in presence of slncRNAs, the reporter with the TTS inserted 110 bp upstream of the promoter show a tendency for up-regulation. To further evaluate these reporters, I plotted the eYFP ratios of each reporter against the slncRNA DNA_{bind} motifs in a separate scatter plot in Figure 22b. As can be seen, most slncRNAs do not increase eYFP levels, only one slncRNA with GGA as DNA_{bind} motif as well as two slncRNAs with a pyrimidine motif has a slightly up-regulatory effect on eYFP expression levels.

4.3.5 Summary mammalian gene-activation system

The slncRNA-mediated gene activation circuit was designed to activate gene expression of a reporter gene upon binding of a slncRNA to its putative triplex target site upstream of a minimal promoter. While various groups showed the successful gene activation using the RNA-guided CRISPR/Cas9-activation[175;176;177;178], the here presented gene circuit using slncRNAs requires further optimization to successfully activate gene expression. Testing of (i) various triplex-forming motifs, (ii) increasing the number of PP7-binding sites within the slncRNAs, and (iii) changes in the positioning of TTS sequences in reporter plasmids did not result in a clear trend of activation of the reporter gene. Given the complexity of the circuit including the (i) co-transfection of four plasmids, (ii) the choice of activator proteins and their low expression levels in cells, (iii) the rapid degradation of RNA molecules and (iv) the small number of triplex-forming motifs that were tested, I decided to develop a simpler, high-throughput technology using short-single stranded DNA molecules to screen for triplex formation *in vitro* and in cells.

4.4 Deep-sequencing platforms to detect triplex formation

Despite various efforts to elucidate the underlying code for triplex formation *in vitro* using dozens of single-stranded, triplex-forming oligos (TFOs) [179;180;135], and the establishment of platforms to further investigate triplex formation in cells [97;100;95;101;102], little has been done to evaluate triplex formation using high-throughput technologies *in vivo*. In this chapter I will describe the development of two deep-sequencing platforms to decipher triplex formation *in vitro* and in cells:

1. Triplex-Seq platform

   The Triplex-Seq approach is a technology that was developed for *in vitro* as well as for *in cell* use. The technology tackles questions of the underlying triplex code, preference of nucleotides in triplexes and minimal length requirements of TFOs via a specific enrichment of single-stranded oligos bound to double-stranded DNA.

2. Triloci-Seq platform

   While the Triplex-Seq approach characterizes the single-stranded oligos in triplexes, but neglects the double-stranded counterpart, the Triloci-Seq platform was developed to identify both the single-stranded and double-stranded sequences found in a triplexes. This protocol is an important addition to screen for triplex target sites within the genome.

4.5 *In vitro* Triplex-Seq

4.5.1 Design of the *in vitro* Triplex-Seq platform

I developed a next-generation sequencing and DNA synthesis-based platform to study triplex formation *in vitro*. To do this, I combined an electrophoretic mobility shift assay (EMSA) (Figure 23a) with Illumina sequencing and DNA synthesis technologies (Figure 23b+c). Briefly, the Triplex-Seq platform comprises (i) a single variant of a triplex target site (TTS) and (ii) a library of triplex-forming oligos (TFOs). The double-stranded TTS (between 30-80 bp long) harbors the purine-rich segment that can accommodate a third strand. The short, single-stranded TFOs (up to 30 nts) contain the putative DNA stretches that form triplexes with the TTS in a parallel or anti-parallel orientation. To screen for TFO sequences that were found in a triplex, TTS and TFO libraries were mixed (1:20 molar ratio) and incubated in triplex-forming buffers favoring parallel or anti-parallel triplex formation. After a 2 hour incubation at 37 °C in ab anti-parallel favoring condition (10 mM Tris-HCl pH7.2, 10 mM $MgCl_2$, henceforth referred to as pH 7) or a parallel environment (10 mM sodium acetate pH 5.0, 10 mM $MgCl_2$, henceforth referred to as pH 5), the products were separated on a native polyacrylamide gel (PAGE) in which a DNA duplex migrates faster through the gel compared to a DNA triplex (Figure 23a). Subsequently, the presumed triplex bands were cut from gel and two samples for each data set were obtained: (i) DNA from the TFO only lane and (ii) DNA from the triplex lane. Following DNA extraction, TFO sequences were enriched, a ssDNA adapter was ligated to the 3' end of the TFOs, the double-stranded TTS was discarded and samples were prepared for next-generation sequencing via PCR amplification. Following Illumina sequencing, the samples were further processed and the enrichment was bioinformatically analyzed. To screen for single-stranded TFO sequences that form triplexes, I designed large TFO libraries (Figure 23c) using the randomized or mixed-base tool from integrated DNA technologies (IDT). The mixed-base TFOs consist of 6 fixed bases (either G, A, T or C) at given positions that serve as an internal barcode, as well as 14 mixed bases. Mixed bases are represented by the IUPAC (International Union of Pure and Applied Chemistry) single-letter codes as a nomenclature to specify incomplete nucleic

acids. During DNA synthesis, degenerated bases are incorporated into the TFO sequences at the wobble/mixed-base positions (e.g. for the N-TFO, G/A/T/C will be incorporated with a 25/25/25/25 ratio) creating a diverse TFO library with a mixture of different TFO sequences (Figure 23c). Using this randomized TFO library generation, 16 small (27 variants) to large (270,000,000 variants) TFO libraries were designed (Figure 23d) based on the assumed underlying triplex formation rules (Figure 5c).

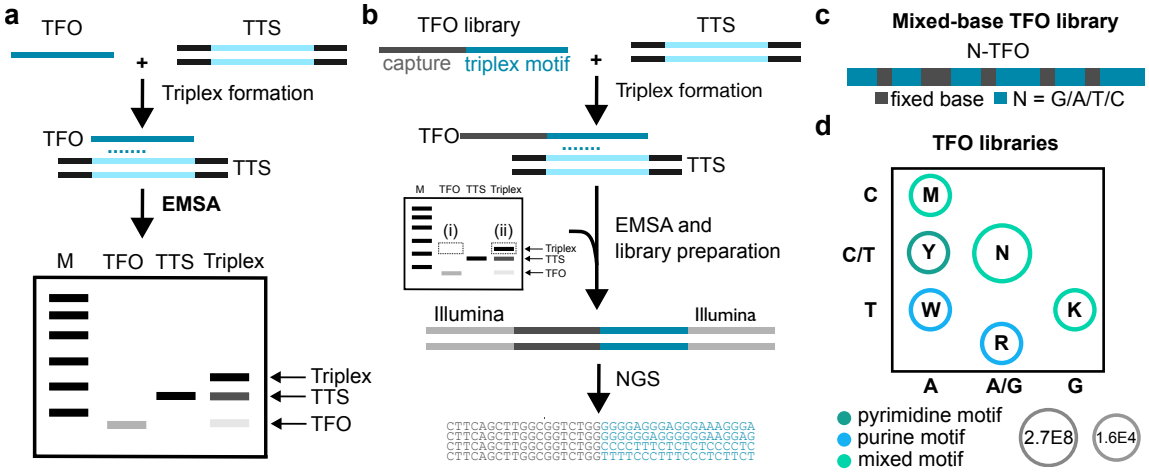


Figure 23| Design of the *in vitro* Triplex-Seq platform. a, Electrophoretic mobility shift assay (EMSA) is classically used for triplex formation experiments. Single-stranded triplex-forming oligos (TFOs) are mixed with a double-stranded triplex target site (TTS) in triplex-favoring conditions, the products are separated on a native polyacrylamide gel (PAGE) and migration of TFO, TTS and triplex is visualized. A shift between the faster-migrating duplex and slower-migrating triplex is expected as schematically described. **b,** The Triplex-Seq platform comprises (i) a single variant of a TTS and (ii) a library of mixed-base TFOs. To screen for TFO sequences that were found in a triplex, TTS and TFO libraries were incubated in triplex-favoring buffers. After incubation, the products were separated on a PAGE and visualized. Next, the presumed triplex bands were cut from the gel and two samples for each data set were obtained: (i) DNA from the TFO only lane and (ii) DNA from the triplex lane. Following DNA extraction, the double-stranded TTS was discarded, TFO sequences were prepared for next-generation sequencing via PCR amplification and subsequently bioinformatically analyzed. **c,** The mixed-base TFOs (20 nts) comprise a 19 nt long capture sequence that serves as a PCR primer docking site, 6 fixed bases (internal barcode), as well as 14 mixed bases which are incorporated at ratios depending on mixed/randomized bases represented by the IUPAC (International Union of Pure and Applied Chemistry) single letter code. **d,** A variety of TFO libraries were generated and are shown as a function of nucleotide content and size.

4.5.2 Example analysis of the *in vitro* Triplex-Seq approach

Following the implementation of the protocol that was described above on page 65, and subsequent sequencing of the TFO library, the DNA sequence reads obtained after the Illumina sequencing, were adapter trimmed and the quality of each read was assessed. Here, an example analysis for the R-TFO library (R = A/G, library size: 16,384 variants) in pH 7 is shown (Figure 24a). In the *in vitro* Triplex-Seq, each data set contains two samples as shown in the example PAGE (Figure 24b): (i) ssDNA from the TFO lane only (DNA that was extracted from gel in which only the TFO was applied) and (ii) ssDNA from the triplex lane (DNA that was extracted from gel which contained both the TFO and the TTS and potentially formed triplexes). To compare among data sets, I computed the normalized read counts by dividing each read count by the total number of reads and multiply the value by 1,000,000 (reads per million, RPM). The frequency of the RPMs of the TFO (grey line) as well as the triplex lane (blue line) were plotted in Figure 24c. I observe the expected exponential decay for the TFO lane in which most TFO sequences that were identified appear either once or twice (grey line), while in the triplex lane in addition to the initial exponential distribution, a heavier tail distribution is observed (blue line). In Figure 24d, I plotted the RPMs of the triplex and TFO per variant. Here, I divided

the groups of variants into hexagons defined by the relevant range of RPMs per variant for the TFOs and triplexes, respectively. All bins above a threshold of approx. 2.3 represent sequences that were found in the triplex lane only (highly-enriched sequences).

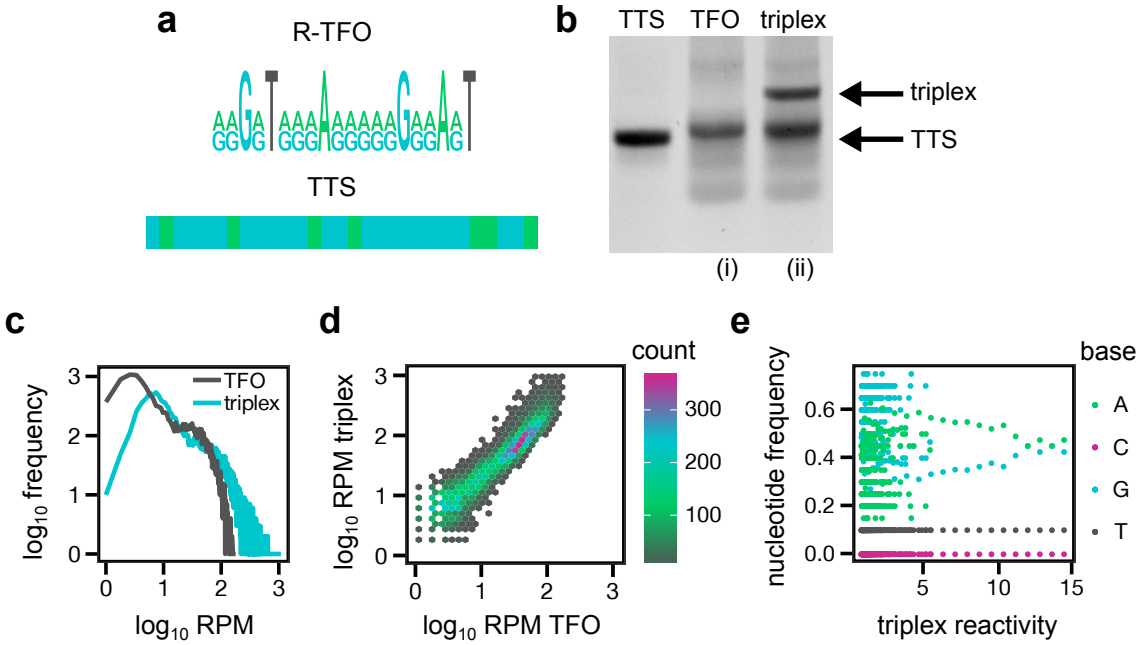


Figure 24| Example analysis of *in vitro* Triplex-Seq platform. **a,** An example analysis of the R-TFO library is shown and the sequence logo of R-TFO and TTS sequence is displayed. **b,** Products of three samples were separated on a 10 % native polyacrylamide gel (PAGE) for 2 hours. The triplex target sites (TTS) displays a strong band, while the triplex-forming oligo (TFO) displays several, but slightly less intense bands indicating the formation of secondary or double-stranded structures. The triplex lane shows the same bands as in the TFO and TTS lane, but displays an additionally band that migrated slower through the gel indicating that this band corresponds to triplexes. All deep-sequencing data was analyzed according to the scheme presented here. **c,** Example analysis of R-TFO library in which frequency of normalized read counts (reads per million, RPM) are plotted against the RPMs for both TFO (grey) and triplex lane (blue). An exponential distribution is observed for the TFO lane, and a heavier-tail distribution for higher RPMs is observed for the triplex lane. **d,** RPM values for TFO and triplex lane, respectively, were grouped into hexagons according to their range per variant and color-coded (RPM counts). Above a threshold of approx. 2.3 only bins in the triplex lane are detected (triplex-specific TFO sequences). **e,** A new measure to determine TFO sequences found in triplexes was introduced ('triplex reactivity') and is defined by the ratio of $\mathrm{RPM}_{\mathrm{triplex}}$ and $\mathrm{RPM}_{\mathrm{TFO}}$ and subtraction of one. A positive triplex reactivity score identifies triplexes and is plotted as a function of the average nucleotide frequency of each nucleotide within the TFO sequences at given triplex reactivity values.

To further characterize the sequences, I introduced a new measure termed 'triplex reactivity' which is defined as the ratio of $\mathrm{RPM}_{\mathrm{triplex}}$ and $\mathrm{RPM}_{\mathrm{TFO}}$ and subtraction of one ($\textit{triplex reactivity} = \frac{RPM_{triplex}}{RPM_{TFO}} - 1$). A positive reactivity score indicates an enrichment of a particular variant above the TFO only control. In Figure 24e, the triplex reactivity scores were plotted as a function of the mean nucleotide frequency of each nucleotide (G/A/T/C) within the TFO sequences at a given triplex reactivity value. For the R-TFO library, I observe that cytosines (C) and thymines (T) are absent or stay constant throughout each triplex reactivity score and adenines (A) decrease while guanines (G) increase with increasing triplex reactivities.

4.5.3 2-mixed base TFO libraries bind TTS in a pH-dependent manner

I first tested TFO libraries that contained a mixture of two mixed-bases (K, M, R, W) at 14 positions (for details see Table 11 on page 32) in pH 5 and pH 7 conditions (Figure 25a). The distribution of triplex reactivities of both conditions and each of these TFO libraries is shown in Figure 25b. The overall triplex reactivities in pH 5 are low, except for the M-TFO (adenine/cytosine). Contrary, triplex reactivity values for all TFO libraries tested in pH 7 are

higher than in pH 5 suggesting that the pH influences formation of stable triplexes and a neutral pH is preferred.

Next, to search for enriched motifs that can be found in the most reactive TFO sequences, I used DRIMust [181], a tool that identifies enriched k-mers and motifs based on a ranked list of sequences. In this work, I applied DRIMust on sorted triplex reactivity lists to detect k-mers that are significantly over-represented at variants with high-reactivity scores. From the k-mers the algorithm also computes an enriched consensus motif. DRIMust was applied for each of the 4 TFO libraries. In Figure 25c, the obtained consensus motifs are shown for pH 5 (left panel) and pH 7 (right panel) and all motifs range between 5-10 nts. For the pH 5 condition, a consensus motif was only detected for the M-TFO whereas consensus motifs for all TFO libraries were found in pH 7. In cases where no motif was identified, the algorithm was not able to detect enriched k-mers in the sorted list at a minimum hypergeometric (mHG) p-value of 10^{-6}. These results imply that triplex formation is pH-dependent for both parallel and anti-parallel triplex formation, contrasting the current view that anti-parallel triplex formation is pH-independent [91]. The motif sequences that were found contain stretches of thymines (K-TFO and W-TFO), stretches of adenines (R-TFO and W-TFO) and a mix of cytosines and adenines (M-TFO in pH 5 and pH 7).

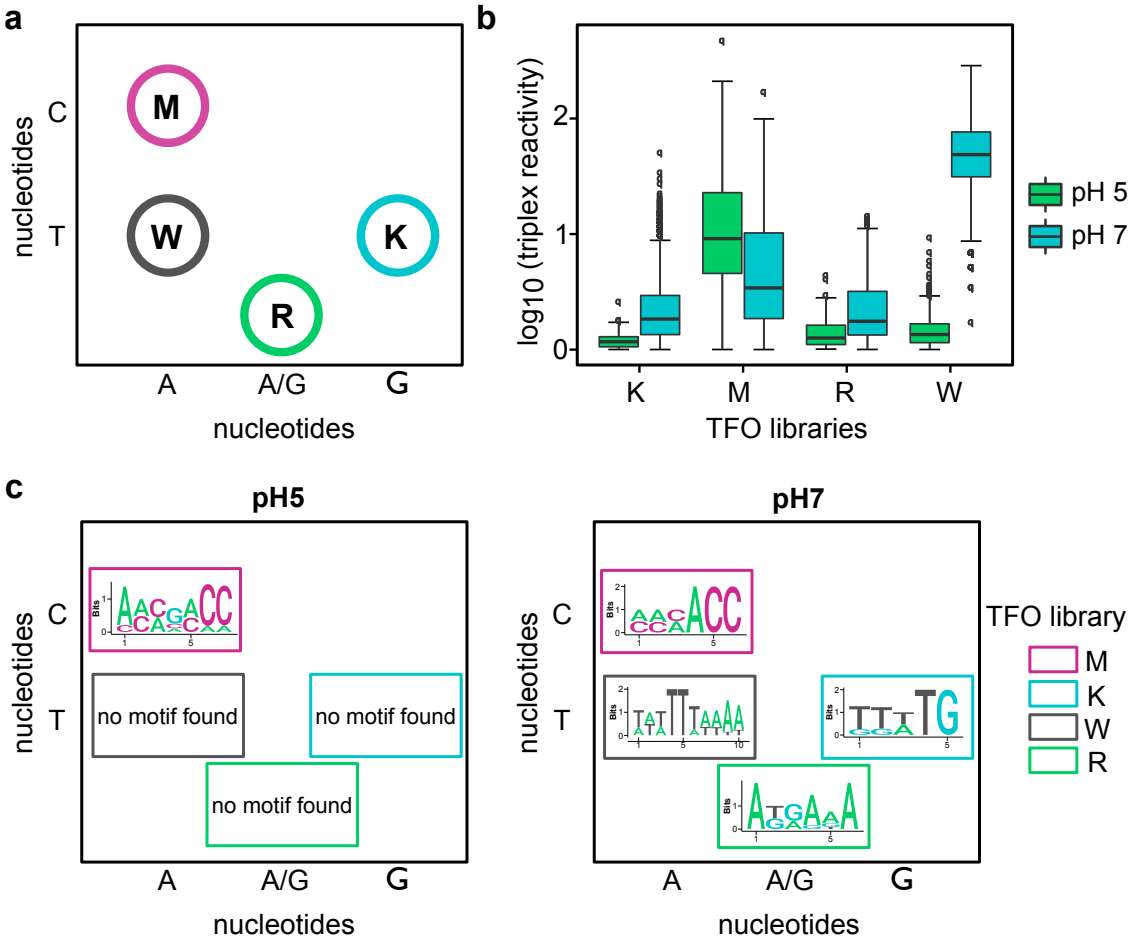


Figure 25| pH-dependent triplex formation using small TFO libraries. a, The enrichment of sequences of TFO libraries with 2-mixed bases (K, M, R, W) were tested using the Triplex-Seq platform. **b,** The distributions of triplex reactivities of the four TFO libraries that were tested in pH 5 and pH 7 indicate lower triplex reactivity values in pH 5 compared to pH 7, except for the M-TFO in which triplex reactivity values are high for both conditions. **c,** DRIMust motifs were computed based on ranked triplex reactivities and indicate consensus motifs of 5-10 nts for all libraries in pH 7 (with high triplex reactivity values) and only for the M-TFO library in pH 5. While for the M-TFO a mix of adenines and cytosines can be found, the other motifs exhibit stretches of thymines (K-TFO and W-TFO) and adenines (R-TFO and W-TFO).

4.5.4 G-rich TFO sequences preferred in triplex formation

I next sought to evaluate which TFO sequences will be enriched when testing the ~ 270,000,000 large N-TFO library (Figure 26a, top). In the basic Triplex-Seq experiment, a known TFO/TTS pair (pH 5: TTS2[155] in Figure 26a, middle; pH 7: TTS1[102] in Figure 26a, bottom) was used alongside the TFO libraries to determine the migration pattern of duplex and triplex. In Figure 26b, I plotted the mean nucleotide frequency against the triplex reactivities for pH 5 (Figure 26b, left) and pH 7 (Figure 26b, right) and observe a G-increase with higher triplex reactivities for both conditions.

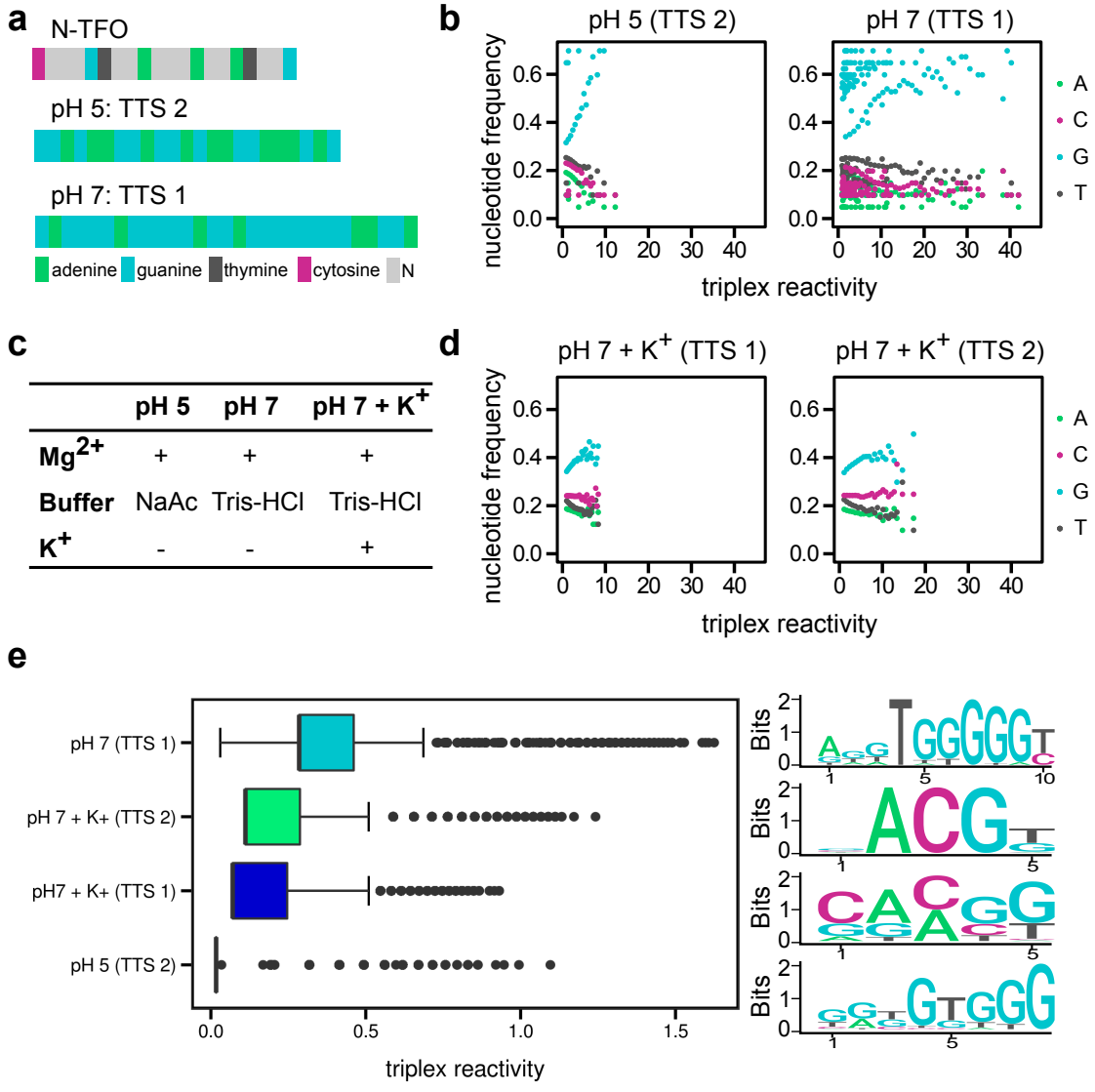


Figure 26| G-rich motif in TFO sequences forms stable and specific triplexes. a, The 270,000,000 large N-TFO library was tested with two triplex target sites (TTS 2 in pH 7, TTS 1 in pH 5). **b,** The mean nucleotide frequency of the two triplex-favoring conditions (pH 5 and pH 7) was plotted as a function of the triplex reactivity values and a G-increase is observed. **c,** The N-TFO library was also tested in triplex-disfavoring conditions (pH 7+K^+) in which high concentrations of potassium (140 mM) were added. **d,** The mean nucleotide of the N-TFO library in triplex-disfavoring conditions is shown and indicate a trend to G-rich TFO sequences. **e,** The distributions of triplex reactivity scores in all conditions with both TTS sequences are shown (left panel) and indicate the highest triplex reactivities for pH 7 (without potassium). In contrast to the overall G-enrichment in the TFO sequences for all conditions, the DRIMust consensus motifs differ (right panel). For the triplex-favoring conditions (pH 5: bottom logo, p-value < 10^{-82} and pH 7: top logo, p-value <$2x10^{-298}$), long stretches of thymine are observed (up to 5 nts), the motifs for the triplex-disfavoring condition are shorter (middle logos) and lack these G-stretches (p-values <$2x10^{-53}$).

To further characterize the behavior of the N-TFO, I additionally applied the Triplex-Seq platform in a buffer (Figure 26c) which has been described in literature to negatively affect triplex formation due to the presence of monovalent ions[88;102]. In Figure 26d the triplex reactivities of both TTS sequences in the triplex-disfavoring buffer are plotted and again, we observe a trend toward G-rich sequences with higher triplex reactivities, but this effect is weaker as compared with the conditions lacking potassium (Figure 26b). While it is commonly assumed that physiological concentrations of potassium abolish triplex formation, more sensitive assays showed that triplex formation is reduced to 10-20 % triplexes compared to triplex favoring conditions[182;183;184]. Comparing the triplex reactivities of all conditions shown in Figure 26e (left panel), a similar decrease of triplex reactivity values was observed (31 % for pH 5, 37 % for pH 7+K^+), suggesting and confirming the formation of more stable triplexes in neutral pH compared to acidic or neutral pH with high potassium concentrations. In contrast to the overall trend of G-rich TFOs in all conditions, the DRIMust consensus motifs display a clear pattern of G-stretches (up to five guanines for pH 7) within a 7-10 nt long consensus motif for acidic and neutral pH without potassium (Figure 26e, bottom and top logos, respectively). A shorter motif (5 nts long) for both TTS in the conditions containing potassium is observed and the significant G-enrichment is shifted to a motif that contains all four nucleotides (Figure 26e, middle logos).
The 7-10 nt long DRIMust consensus motif that was found with the N-TFO library prompted me to design TFO libraries (B- and D-TFOs) with a continuous stretch (3-9 nt) of mixed bases that are flanked by fixed bases (Figure 27a+b, top) to characterize the minimal sequence length required for triplex formation.

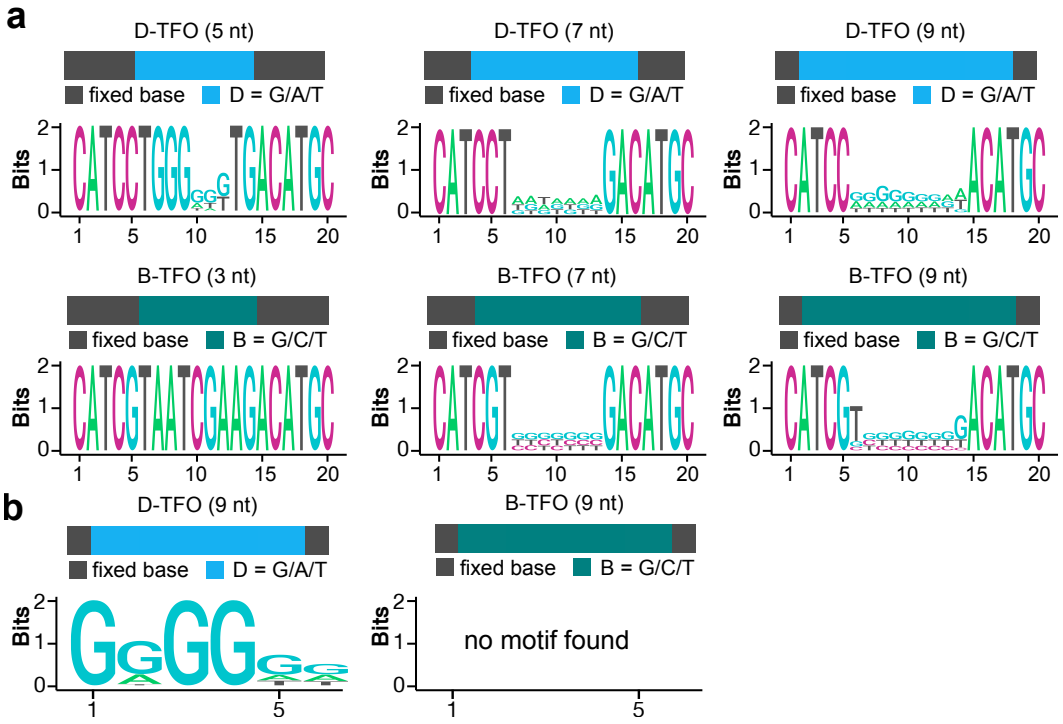


Figure 27| Mixed-base stretches *in vitro* identify minimal TFO length. To characterize a potential minimal length of a TFO that is required to form stable and specific triplexes, TFO libraries with mixed-base stretches (B- and D-TFO) of varying length (3-9 nt) flanked by fixed bases were designed and tested in pH 5 (B-TFO) and pH 7 (D-TFO). **a,** The sequence logos of the D-TFO with 5 nt (left), 7 nt (middle) and 9 nt (right) are shown. The 3 nt D-TFO library is not shown as no sequence reads were obtained. Only for the 5 nt and the 9 nt long D-TFO stretch, a trend towards G-rich stretches can be seen. The sequence logos for 3 nt (left), 7 nt (middle) and 9 nt (right) for the B-TFO are displayed. While only one TFO variant dominated in the 3 nt TFO library, a trend to G-rich sequences can be seen for the 7 nt and 9 nt long B-TFO stretches. **b,** To identify a consensus motif, DRIMust was applied and only in the D-TFO library a G-rich stretch of 5 nt was found, while for the other libraries no DRIMust motif was identified.

The B-TFO library was tested in pH 5 and the D-TFO library was tested in pH 7 with the respective TTS sequences shown in Figure 26a. The sequence logos of the TFO sequences with triplex reactivities > 1 are shown in Figure 27a. While no sequence reads were obtained for the 3 nt D-TFO and 5 nt B-TFO, the TFOs with shorter (3-5 nt) stretches contain no (Figure 27a, bottom) to little (Figure 27a, top) information content. The TFO sequence logos for longer stretches display a weak trend towards G-rich sequences in particular for the 9 nt stretch for the D-TFO. Applying DRIMust of the ranked triplex reactivity lists of all stretches yielded a G-rich DRIMust consensus motif (7 nt long) for only the 9 nt D-TFO library confirming the (i) pH dependence of triplex formation and (ii) suggesting that a minimal length of approx. 9 nt is needed for stable triplex formation.

To characterize single variants with high triplex reactivities obtained in the N-TFO library for their potential to form triplexes, I applied two classical techniques on (i) a variant with a high triplex reactivity score and (i) a variant with a negative triplex reactivity score (Figure 28a, bottom).

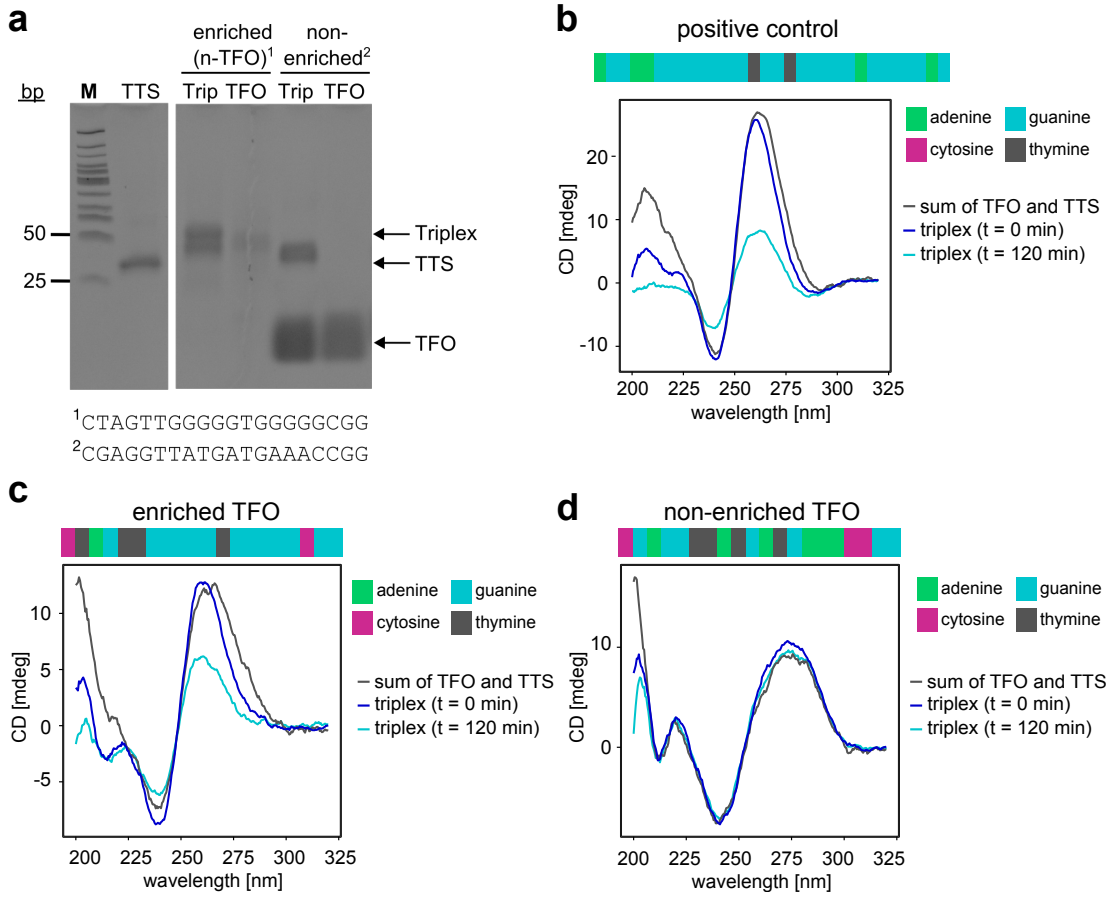


Figure 28| Classical techniques to characterize enriched N-TFO variants. **a,** An electrophoretic mobility shift assay (EMSA) was performed on a highly-enriched TFO variant (highest triplex reactivity score) and compared to a non-enriched variant (negative triplex reactivity score) and a clear shift for the enriched N-TFO variant is observed, contrary to the non-enriched N-TFO variant where no shift is noted. **b-d,** Circular dichroism (CD) spectroscopy of three different samples. CD spectroscopy is a technique to discriminate between DNA structures by the differential absorption of left- and right-handed circularly polarized light and the difference between the absorption is plotted as CD in millidegrees [mdeg]. **b,** The AG30 positive control shows a difference in absorption between the sample that was measured immediately after mixing (t=0 min) and after incubation in which triplexes had time to form (triplex sample). A reduced, positive CD peak at 260 nm and a flat, slightly negative absorption is observed for the triplex sample at 205 nm. No difference between the CD spectra directly after mixing and the CD spectra for the sum of the individual spectra for the TFO and the TTS is observed. **c,** Similar CD spectra are shown where the enriched N-TFO variant was tested and the most notable change is based on the slightly skewed CD curve with a peak at 260 nm which is only observed for the triplex sample (t=120 min). **d,** No differential absorption of all three samples is observed for the non-enriched TFO variant.

To confirm that the enriched N-TFO variant forms triplexes, I performed an EMSA (Figure 28a) with a 10 % PAGE. As expected, the enriched N-TFO variant (high triplex reactivities) displays a shift from the duplex to a triplex band (lane 2), whereas no shift is observed in the PAGE for the non-enriched N-TFO variant (lane 4). Comparing the lanes of the enriched TFO (lane 3) with the non-enriched TFO (lane 5), I note that the enriched TFO variant migrates at a height slightly below the TTS (lane 1), while the non-enriched TFO migrates faster through the gel.
To confirm the EMSA result, I also subjected the samples to circular dichroism (CD) spectroscopy. CD spectroscopy is a technique where the absorption of right- and left handed circularly polarized lights of optically active molecules can be used to characterize DNA structures. The difference in absorption is called CD, is measured in millidegree [mdeg] and plotted as a function of wavelength. I compared the CD spectra of the triplex sample directly after mixing (t = 0 min), after 2 hours incubation at 37 °C in pH 7 condition (t = 120 min) and the sum of the individual spectra of TFO and TTS controls. In Figure 28b, the positive control from literature[97;102] (TFO AG30) is shown and highlights the time-dependent change in the CD spectrum after 120 minutes of incubation compared to the sum of the individual TFO and TTS spectra and directly after mixing. A positive CD peak is observed at 260 nm for all samples, but the CD peak for the sample after 2 hours incubation is 2-fold lower, and no CD peak is observed at 205 nm for the sample after 2 hours incubation, whereas a positive CD peak is shown for the control samples (t=0 min and the sum of the individual spectra). A similar observation is made for the enriched N-TFO variant (Figure 28c). Here the difference in absorption is most visible for the slightly skewed CD spectrum with the peak at 260 nm for the sample after 2 hours incubation compared to nearly symmetrical CD spectra for the samples that were tested directly after mixing and for the sum of the individual TFO and TTS spectra. In contrast to the difference in CD spectra for the positive control and the enriched N-TFO variant, the three CD spectra for the non-enriched N-TFO variant (Figure 28d) look identical except at lower wavelength (200 nm). The combination of the EMSA and the CD spectroscopy further confirms that the enriched TFO variants with high triplex reactivity scores form triplexes while TFO variants with negative triplex reactivity scores are unable to do so.

4.5.5 Guanine increase in TTS stabilizes triplex formation

With the identification of G-rich TFO sequences, I next studied the relationship between TFO enrichment and variation of guanine/adenine ratios in the TTS. Five TTS sequences were *de novo* designed in which the ratio of guanine and adenine was systematically changed (20/80 to 80/20) (Figure 30a). To identify a positive control (e.g. a TFO sequence that forms triplexes with respective TTS) for the newly designed TTS sequences, I applied the Triplexator software[117] to detect triplexes (TFO/TTS pairs). Triplexator (triple-helix locator) is a computational tool that utilizes approximate pattern matching based on known triplex formation rules. To detect TFO/TTS pairs, the sequences (putative TFO and TTS sequences) are subjected to a set of user-defined parameters such as the minimal length, maximum error and error rate. Based on this set of constraints, the algorithm computes the optimal/maximal TFO/TTS pairs that were found in the given sequences (details of user-defined constraints can be found on page 33). The optimal TFO/TTS pairs were then experimentally verified via EMSA for pH 5 (Figure 29a) and pH 7 (Figure 29b). As shown in the 10 % PAGE for both conditions, only one out of five of the predicted TFO/TTS pairs worked. In the PAGE of the pH 5 condition (Figure 29a), a shift from duplex to triplex band is observed for the TTS with 20 % guanine (lane 4) as opposed to the other lanes where no shift is observed. In Figure 29b, a duplex-to-triplex transition is observed

for the TTS with 80 % guanine (lane 14) and no other shift can be seen in the other lanes. Intriguingly, all TFO/TTS pairs that were tested in the shift assays, were optimal TFO/TTS pairs according to the Triplexator tool, but only a success rate of 20 % (of triplex formation) was obtained experimentally. Hence, the TTS with 20 % guanine in pH 5 and the TTS with 80 % guanine and their respective TFOs were used as positive controls in the subsequent experiments which are described below.

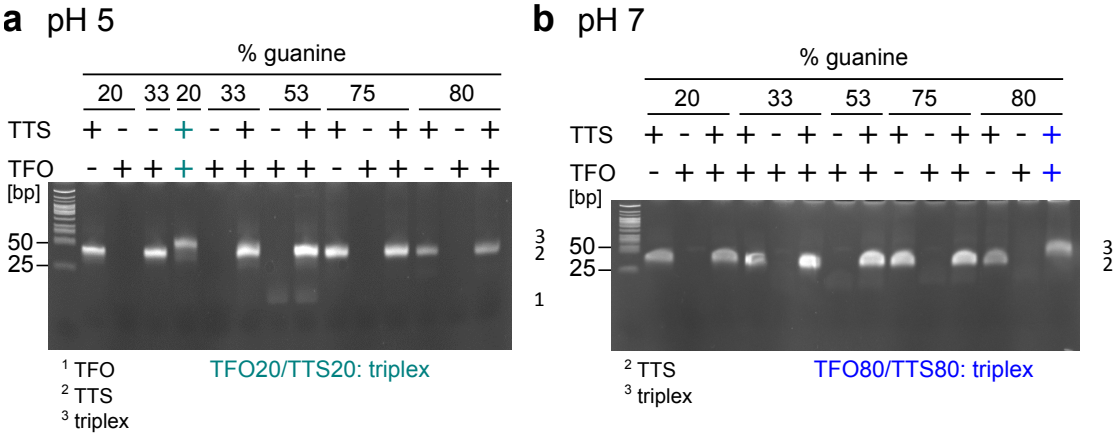


Figure 29| Mobility shift assay to identify *de novo* designed TFO/TTS pairs. An electrophoretic mobility shift assay (EMSA) was performed on the five *de novo* designed TTS with increasing guanine content. TFOs for every TTS were generated and triplex formation potential was predicted using the Triplexator software [117]. Each TFO/TTS pair (e.g. the TFO20 was tested with the TTS with 20 % guanine) was tested in pH 5 and pH 7 condition. **a,** The PAGE in the pH 5 condition identifies one working TFO/TTS pair that exhibits a shift from duplex to triplex (lane 4, highlighted in green) for the TTS with 20 % guanine. **b,** In the PAGE that was tested in the pH 7 condition, also one TFO/TTS pair was identified that forms triplexes (lane 14), which corresponds to the TTS with 80 % guanine. These TFO/TTS pairs were used as positive controls in the subsequent experiment where the five different TTS variants were used.

The five TTS variants (Figure 30a, bottom) were tested with the N-TFO library (Figure 30a, top) in pH 5 and pH 7 and the frequency of RPMs (Figure 30b) as well as the distribution of triplex reactivity scores for each TTS was plotted and compared (Figure 30c). I note an increase in enriched TFO sequences (Figure 30b) as well as a slight increase in triplex reactivities (Figure 30c), the higher the percentage of guanine within the TTS (mainly for 75 % and 80 %) suggesting the importance of G-rich TTS sequences for triplex formation in pH 5 and pH 7.

To evaluate the enriched TFO sequences for each TTS variant, the mean nucleotide frequency of the four bases is shown in Figure 30d. In the heat-map, each line corresponds to one TTS that was tested and each column represents the triplex reactivity that was divided into three ranges (low, medium, high). The heat-map is color-coded and displays the mean nucleotide frequency of TFO sequences in the respective triplex reactivity range.

While there is a clear increase for G-rich TFO sequences with (i) increasing percentage of guanines in the TTS and (ii) the higher the triplex reactivity, thymines were also found in the enriched sequences, but slightly decreased with higher triplex reactivities in pH 7 (Figure 30d). This is in accordance with findings from another study which identified G-rich patterns of over 80 % in the TTS indicating the importance of guanines in the TTS to form stable triplexes [179]. As the TTS variant were designed *de novo*, I wanted to confirm that the newly identified TFO sequences were able to bind the guanine-rich TTS sequences (75 % and 80 %).

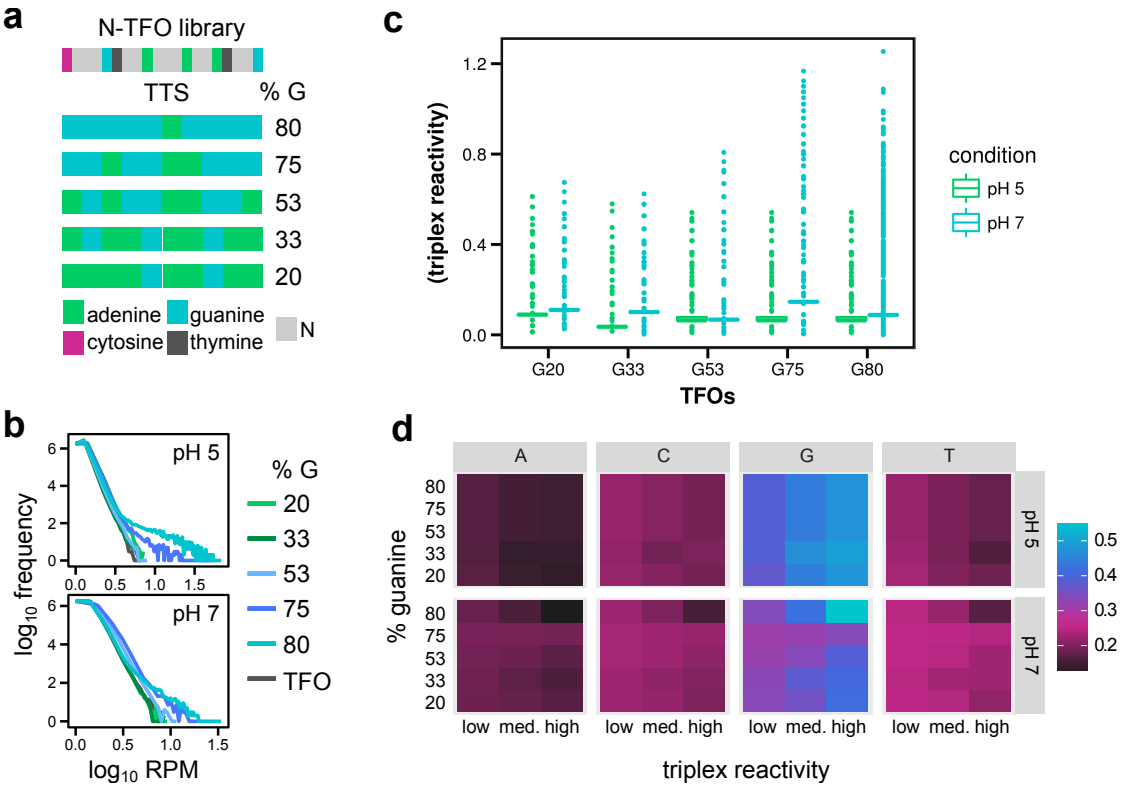


Figure 30| Increase in guanine content in TTS enhances triplex formation. a, To test the influence of guanine/adenine ratio, five triplex target sites (TTS) were designed, in which the guanine/adenine ratio was systematically altered. **b,** The N-TFO library was subsequently tested in pH 5 and pH 7 and the frequency of the RPMs was plotted against the RPMs. An increase in enriched sequences for higher guanine contents (75 % and 80 %) can be observed. **c,** In the heat-map, each column represents a range of triplex reactivity values (low, medium, high) and each line corresponds to a TTS variant. The nucleotide frequency of each nucleotide of each TFO sequence at a given RPM range is color-coded. A strong trend to G-rich TFOs is observed. This trend increases with (i) higher guanine-contents of the TTS and (ii) higher triplex reactivities. Additionally, thymines are found, albeit to a lesser extent, in the TFO sequences of pH 7, but decrease with higher triplex reactivities.

To do so, I ordered the three top hits of TFO sequences with high triplex reactivity scores from pH 7 as single variants (Figure 31, left panel) and performed a classical mobility shit assay. In Figure 31 (right panel), the 15 % PAGE of the three motifs and the AG-rich positive control $AG30^{97;102}$ is shown.

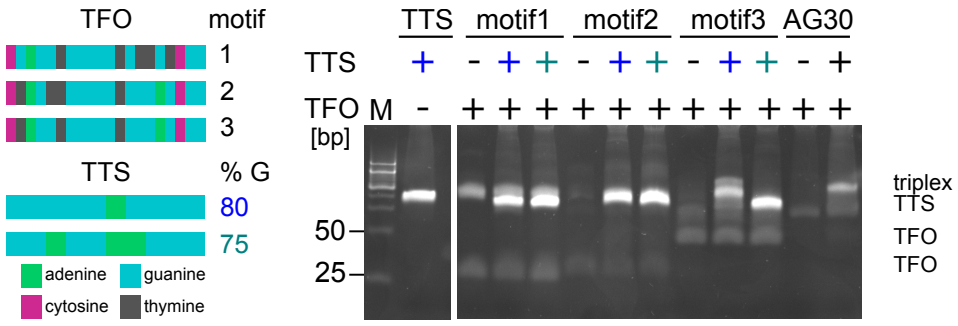


Figure 31| EMSA confirms stable triplex formation using G-rich TFO/TTS pairs. Three single TFO variants of the top of the triplex reactivity list (pH 7 condition) were ordered and a shift assay was performed on a 15 % PAGE with two guanine-rich TTS variants (TTS with 75 % and 80 %). Lane 1 shows a bright band corresponding to the TTS with 80 % guanine. It can be noted that motif 1 (lane 3) migrates at the same height as the triplex band and no clear shift is visible for this motif. In contrast to motif 2 which shows a clear shift from duplex to triplex for both TTS (lane 6 and 7). Motif 3 displays a shift only for the TTS with 80 % guanine (lane 9). In lane 12 the positive control with the purine rich AG30 TFO is shown and exhibits a slightly higher band due to its larger size (30 nt long AG30 in contrast to 20 nt TFO variants).

Contrary to motif 1, where no shift is observed between the duplex and the triplex band (lane

3), a clear shift for motif 2 for both TTS (lane 6+7) and a shift from duplex to triplex for TTS with 80 % with motif 3 (lane 9) can be observed. The shift in the positive control AG30 (lane 12) is slightly higher due to the difference in length (AG30: 30 nt, TFO: 20 nt).
To identify emerging consensus motifs from the different TTS sequences, a DRIMust analysis was applied. For TTS sequences with low guanine percentage (20 %-53%, Figure 32), the enriched TFO logos contain little information and together with the low triplex reactivity values (Figure 30c) and weak enrichment of TFO sequences (Figure 30b) imply that many variants interact weakly with these TTS variants. Conversely, for TTS sequences with high guanine percentage (75-80 %, Figure 32), the information content of the consensus motif increases and even indicates that one (for pH 5) to two (for pH 7) particular TFO variants interact with these TTS in both a reactive and specific fashion. Therefore, these results suggest that G-rich TTS sequences lead to a stronger and more specific triplex interaction with the TFOs

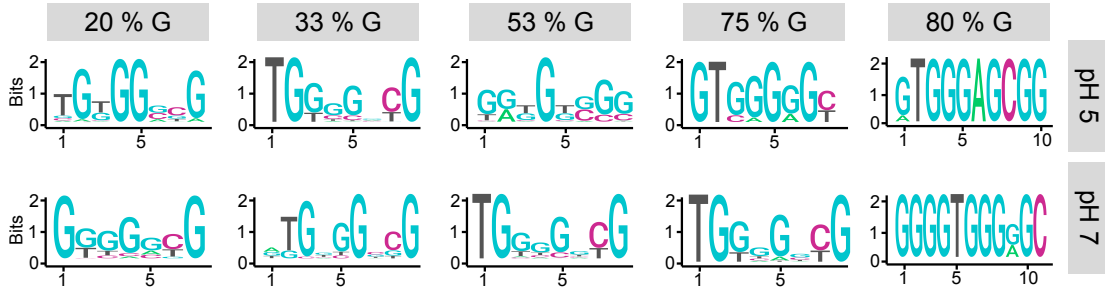


Figure 32| G-rich TFO consensus motifs identified with *de novo* designed TTS variants. DRIMust analysis was used to characterize consensus motifs for the N-TFO library that was tested with each TTS (identified by the percentage of G in each column) in pH 5 (upper row) and pH 7 (lower row). It can be noted that guanine is dominating in all consensus motifs, but little information can be retrieved for TFOs with low to medium guanine content in the TTS (20-50 % G) suggesting a weaker and less-specific triplex interaction of the TFOs with the TTS variants. Contrary to the TTS with high guanine content (in particular for 80 % G) only one or two motifs dominate in the TFO sequences implying a strong and specific interaction with TTS variants with high guanine content (p-values between $8x10^{-60}$ and $5x10^{-324}$).

4.5.6 Summary *in vitro* Triplex-Seq

In this part of the Ph.D. thesis, I focused on the development of a high-throughput sequencing technology termed *in vitro* Triplex-Seq. Through the use of the mixed-base tool for cost-effective oligo synthesis, and the fast turn-around time of the Triplex-Seq protocol, I believe that this platform adds a unique tool to the field of triplex formation and several insights were gained:

1. There is a pH-dependence for anti-parallel and parallel triplex formation using TFO libraries (Figure 25 + Figure 26). While lower triplex reactivities were detected in pH 5 and pH 7+K^+, high triplex reactivities were obtained in pH 7.

2. In addition, the combination of the triplex reactivity scores with the DRIMust analysis tool provides a powerful downstream process to compute consensus motifs of TFO libraries with high triplex reactivity values.

3. It was found that G-rich TFOs and TTS variants form stable and specific triplexes in pH 7 and to a slighter extent in pH 5 (Figure 26 + Figure 30).

4. The potential minimal length of a TFO ranges between 7-10 nt. This length was identified by the DRIMust motifs that vary between 5-10 nt (Figure 30 + Figure 26) as well as by the more direct characterization of utilizing mixed-base stretches of varying length where only clear G-rich patterns were found for the longer 7-9 nt long stretches (Figure 27).

4.6 In cell Triplex-Seq

4.6.1 Moving from *in vitro* to in cell sequencing platforms

Studying triplex formation in cells in a systematic fashion has been a challenge for many years. Unlike in the *in vitro* experiments where mobility shift assays, CD spectroscopy, microscale thermophoresis and absorbance melting curves can be used in a controlled environment, in cell approaches to study triplex formation has mostly been indirect [10;51]. To address this challenge, I devised a two pronged strategy: First, I developed an in cell Triplex-Seq approach to show enrichment of TFO sequences that are consistent with the *in vitro* Triplex-Seq data. Second, I developed Triloci-Seq which is designed to simultaneously detect TFOs and putative TTS motifs. In this chapter, I will describe results for the in cell Triplex-Seq data, and preliminary results for the Triloci-Seq approach.

4.6.2 Design of the in cell Triplex-Seq

The in cell Triplex-Seq approach is an adapted version of the *in vitro* Triplex-Seq platform where, instead of a single TTS variant, the putative TTS sequences in the mammalian genome are used (Figure 33). *In silico* analysis showed significant enrichment of putative TTS sequences in gene regulatory elements [117;119]. Here, the same TFO libraries that were used for the development of the *in vitro* Triplex-Seq protocol were utilized for the in cell approach. In brief, TFO libraries were transfected into mammalian cells (Chinese hamster ovary cells, CHO-K1-HAC cells) and a subset of TFOs were expected to bind to genomic sequences.

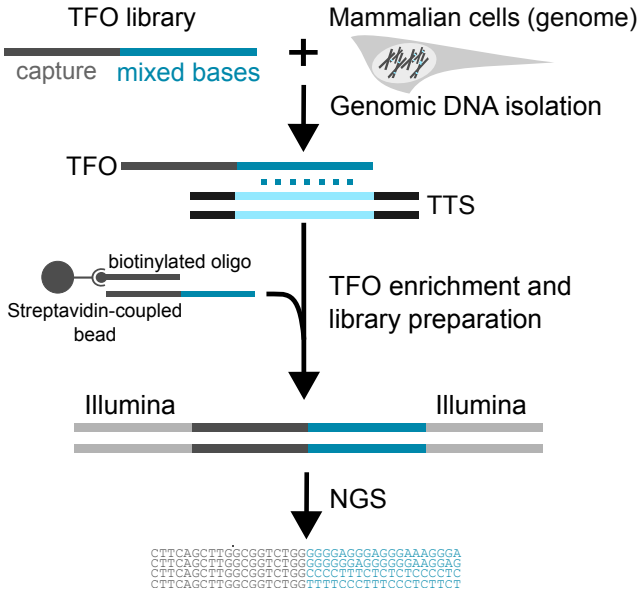


Figure 33| Design of the in cell Triplex-Seq platform. The scheme shows the in cell Triplex-Seq platform. Here, TFO libraries that were also used in the *in vitro* Triplex-Seq platform, were transfected into mammalian cells. 24 hours post-transfection, genomic DNA was isolated and TFOs that were bound to genomic DNA were co-isolated. To enrich for these TFO sequences, a biotinylated oligo that is complementary to the capture sequence of the TFO libraries was added and with the help of streptavidin-coupled magnetic beads separated from genomic DNA which was discarded. Subsequently, TFO libraries were prepared for next-generation sequencing (NGS) following the same protocol as it was described for the *in vitro* Triplex-Seq platform.

24 hours post-transfection, genomic DNA and the TFOs that were bound to genomic DNA were co-isolated. Next, genomic DNA was sheared using a restriction enzyme (6 bp-cutter) and the TFO sequences were enriched via the 19 nt long capture sequence that is shared by all TFOs. A complementary, biotinylated oligo was annealed to the capture sequence and enriched using

streptavidin-coupled magnetic beads and the genomic DNA was discarded. Finally, the TFO sequences were subsequently prepared for Illumina sequencing by PCR amplification as described in the *in vitro* Triplex-Seq platform, and TFO sequences were bioinformatically analyzed

4.6.3 In cell sequencing control data and TFO libraries

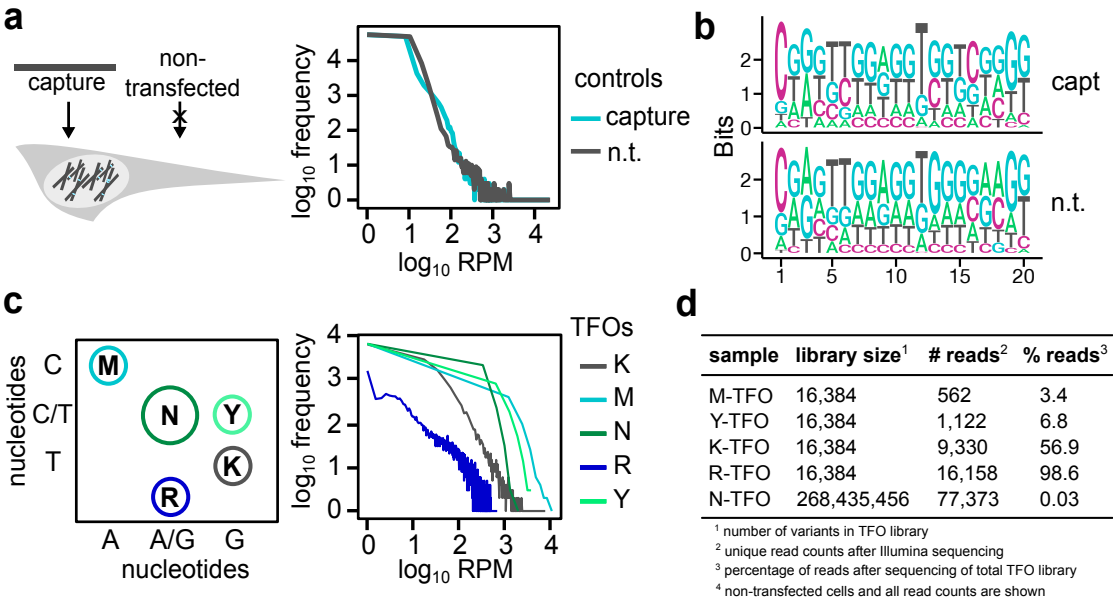
Contrary to the *in vitro* Triplex-Seq platform, where I had two samples for each data set (TFO only and triplex lane), the in cell Triplex-Seq lacks the 'TFO only' control thus no triplex reactivity can be computed. The lack of this control imposes a slightly different approach on the analysis and interpretation of the in cell Triplex-Seq data.

To determine the background noise of the in cell Triplex-Seq approach, I first looked at the data obtained from two control samples: (i) cells that have been transfected with the 19 nt long capture sequence only (capture only) and (ii) cells that were not transfected (n.t.) (Figure 34a, left panel). Approx. 80 % of all sequence reads of both samples corresponded to sequences that (i) contain the capture sequence and (ii) are 39 nt in length which corresponds to the full length of TFO plus capture sequence. To understand the underlying distribution of these 39 nt long sequence reads, the RPM (normalized read count) frequency was plotted as a function of RPM values. Following an initial flat decay for both samples (Figure 34a, right panel), the exponential behavior similar to the previously reported graph observed in the TFO only lane of the *in vitro* Triplex-Seq platform (Figure 24b) was seen. Having a closer look at the 20 nt long sequences downstream of the 19 nt long capture region of the sequence reads, I identified a bias towards longer stretches of adenines as well as thymines and GAA-rich sequences that appear in particular on top of the sequence reads list as shown in the presented sequence logos (Figure 34b). Furthermore, this background noise appears to be similar for both samples. To evaluate whether any sequences in the two control samples corresponded to TFO sequences, I searched for all TFO libraries that were transfected (Figure 34c, left and Figure 36, top) and found sequence reads of the R-TFO (4.8 % of all reads), K-TFO library (0.2 % of all reads) and 0.007 % of D9- and B9-TFO library (see scheme of these libraries in Figure 36).

Following the characterization of the background noise using the control samples, I next characterized actual TFO libraries (Figure 34c, left panel). Evaluating the file of all sequence reads of each TFO library revealed that a nearly identical background of sequences were found for all TFO libraries (data not shown) implying that these sequences are enriched through a common experimental step in the in cell Triplex-Seq protocol. To characterize the distribution of the TFO libraries, the sequences that corresponded to the TFO sequences of each library were extracted and these underlying distributions are shown in Figure 34c (right panel). The frequency was plotted as a function of RPMs and it can be seen that, depending on the TFO library, the distributions behave differently. The Y- and M-TFO libraries exhibit an initial flat decay of frequencies for lower RPMs and an exponential-like behavior was observed for all RPM values. The K-TFO library starts at similar RPM frequencies, but diverges into the exponential decay at lower RPMs and displays a similar behavior observed for the control samples of the in cell Triplex-Seq data (Figure 34a, right panel). In contrast to these four TFO libraries, the R-TFO library differs in the distribution of the frequencies. A step-like distribution is observed in which following a short, initial exponential decay, a flatter and noisier decay of frequencies with higher RPMs is observed. In contrast to the four smaller libraries (K, M, R, Y) which contain approx. 16,000 variants, the N-library is larger in size (over 260 Mio. variants), but exhibits a similar distribution as was seen for the M- and K- TFO libraries.

The difference in frequency distribution is also represented in the number of unique variants

after Illumina sequencing (Figure 34d). Here I compare percentages of sequence reads with respect to the total number of variants of each TFO library. While 3.4 - 6.8 % of the total variants for the Y- and M-TFO libraries were found, more than 50 % for the Y- and 95 % for the R-TFO libraries were detected. Given that 33 % of the R-TFO sequences are found in the background of the control samples, the underlying distribution needs to be regarded with caution. Furthermore, only 0.03 % of sequence reads was found in the N-TFO library suggesting that the read depth is likely to affect this percentage. Contrary to the *in vitro* Triplex-Seq platform, where I obtained enriched TFO sequences for these TFO libraries, in the in cell Triplex-Seq approach no apparent enrichment after sequencing was obtained. This difference either indicates an insufficient sequencing read depth or the lack of binding of TFOs to the genome.



sample	library size[1]	# reads[2]	% reads[3]
M-TFO	16,384	562	3.4
Y-TFO	16,384	1,122	6.8
K-TFO	16,384	9,330	56.9
R-TFO	16,384	16,158	98.6
N-TFO	268,435,456	77,373	0.03

[1] number of variants in TFO library
[2] unique read counts after Illumina sequencing
[3] percentage of reads after sequencing of total TFO library
[4] non-transfected cells and all read counts are shown

Figure 34| In cell Triplex-Seq data is noisier and differs from *in vitro* Triplex-Seq data. a, Two control samples were tested in which either the 19 nt long capture sequence was transfected (capture/capt) or cells were not transfected (n.t.) (left panel), compared with each other and the frequencies of RPMs was plotted as a function of RPMs (right panel). Following an initial flat distribution for lower RPMs, and exponential decay is observed and ends with noisier frequencies at high RPMs. **b,** To compare the sequence reads that were obtained for these two samples after the sequencing a sequence logo was generated, respectively and implies a preference for GGA_{rich} sequences as well as stretches of thymines for both samples. **c,** Several TFO libraries were transfected which are different in size (left panel) and the distributions of each TFO library are shown (right panel). No enrichment for these TFO libraries can be observed and while the M-, N- and Y-TFO show a long flat decay followed by an exponential decrease in frequencies, the K-TFO looks similar to the control samples and the R-TFO exhibits a two-step distribution. **d,** The difference in distribution is also reflected in the percentage of sequence reads obtained after sequencing with respect to the total number of reads.

4.6.4 Comparison of libraries before and after in cell Triplex-Seq

In the previous section, smaller (16,000) and larger libraries (269. Mio) were tested and no to little enrichment of TFO sequences was found. Hence, two libraries (D and B-TFO library) with each intermediate variant size (1,594,323 variants) were tested using the Triplex-Seq in cell approach (Figure 35a, left panel). The frequency distributions of the D-TFO library were compared before and after the in cell Triplex-Seq platform was applied. To do so, the D-TFO was ordered with the flanking Illumina adapter sequences (see Table 16) and PCR amplified (henceforth referred to as D-TFO (PCR)) to determine the distribution of sequence reads of the TFO library and to identify potential (PCR) biases. In Figure 35a (right panel), the distributions of the three samples are shown. The D-TFO (PCR) library displays the expected behavior, where after an initial steady distribution of frequencies, the exponential distribution is observed. In

contrast to the D- and B-TFO (in cells) libraries that were subjected to the in cell Triplex-Seq approach. These two libraries exhibit a similar behavior for lower RPMs as the D-TFO (PCR) sample, but display an enriched fraction of TFO sequences for higher RPMs. To determine the nucleotide preference for the TFO libraries, the mean nucleotide frequency of the three samples was plotted against the RPMs (Figure 35b). While a trend towards G-rich sequences (> 0.4) is observed for the B- and D-TFO (in cell) libraries, in particular for the enriched TFO sequences (high RPMs), the mean nucleotide frequency of the D-TFO (PCR) centers around 0.33 for the full range of RPMs with a slight tendency towards higher thymine frequencies at higher RPMs. The lack of any nucleotide preferences of the D-TFO library that was only *in vitro* PCR amplified suggests the absence of PCR biases. Thus, the G-enrichment observed in the in cell B- and D-TFO libraries cannot be attributed to biases in the downstream protocol, but likely is a direct result of an interaction of the TFOs with genomic DNA.

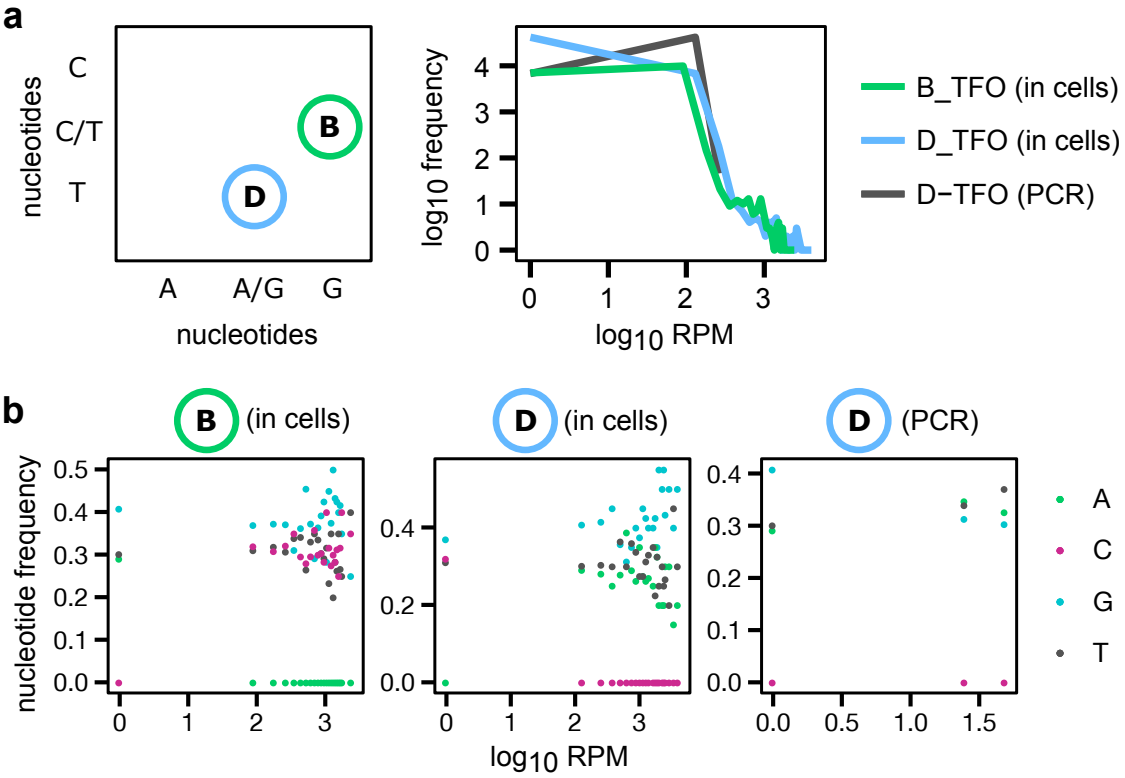


Figure 35| Comparison of in cell Triplex-Seq data and PCR-amplified TFO libraries. a, Two TFO libraries (1,594,323 variants each, left panel) were transfected into mammalian cells and the D-TFO library (D-TFO (in cells)) was compared to the D-TFO library that was only PCR amplified (PCR) and not subjected to the in cell Triplex-Seq protocol. The underlying distributions are plotted for the three libraries and while the D-TFO (PCR) exhibits the expected exponential distribution with an initial weak increase for lower RPMs, the B- and D-TFO libraries (in cells) are identical for lower RPMS, but show a heavier-tail distribution for higher RPMs. **b,** The mean nucleotide frequency of the three samples are plotted as a function of RPMs. A G-rich trend for higher RPMs is observed for the two B- and D-TFO libraries that were transfected in contrast to the D-TFO library (PCR) which displays a steady frequency of approx. 0.33 for all three nucleotides that are represented by the single letter code 'D' (G/A/T).

4.6.5 Short and G-rich TFO sequences interact with genome

Given the observed enrichment of G-rich sequences for the B- and D-TFO, I wanted to identify the nucleotide preferences of triplex-forming motifs using a continuous stretch of mixed bases required for triplex formation. Hence, the B- and D-TFO libraries with a range of 3-9 nt long mixed-base stretches flanked by fixed bases were designed and tested (Figure 36, top). In the heat-map, each column represents a range of RPMs (low, medium, high) and each line

corresponds to the different TFO libraries (Figure 36, bottom). The nucleotide frequency of the mixed-base stretch is color-coded where the black color indicates the absence of the nucleotides within the TFO sequence, and light gray corresponds to a lack of RPM values for the respective range. For the two different mixed-bases (B- and D-TFO) it can be seen that neither adenines (for the B-TFO library) nor cytosines (for the D-TFO library) are present, respectively. Moreover, an increase in G-rich sequences and a decrease for all other nucleotides to a similar degree is observed (i) the higher the RPMs (more enriched sequences) and (ii) the longer the stretch. The smaller TFO libraries (3nt-TFO: 27 variants) lack the strong preference for G-rich sequences suggesting that these libraries do not interact in a strong and specific fashion with the genome.

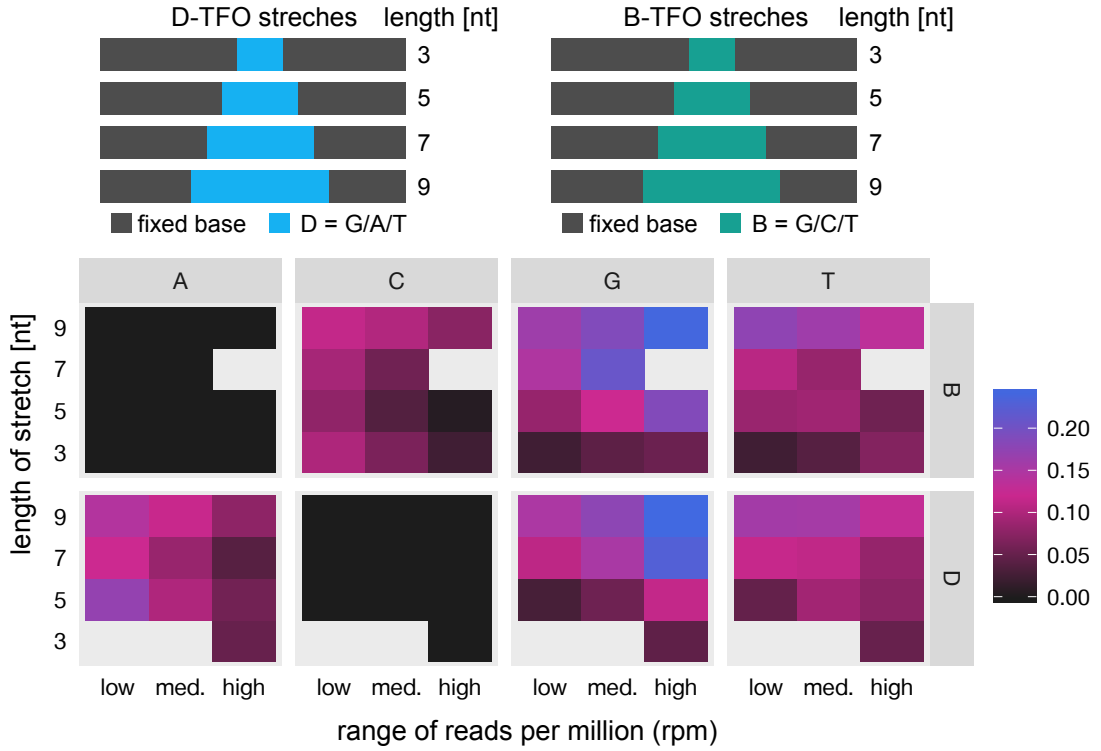


Figure 36| G-rich TFO sequences identified in in cell-Triplex-Seq data. Several TFO libraries were generated with mixed-base stretches of varying length (3-9 nt) using the B- and D-TFO libraries (top) and the nucleotide preference after sequencing is shown as a heat-map (bottom). Each row represents a TFO library with a mixed-base stretch of given length and each column corresponds to a range of RPMs (low, medium, high). The nucleotide frequency is color-coded (black over red to blue) and a decrease in cytosines and thymines (for the B-TFO libraries) and a decrease in adenines and thymines (for the D-TFO libraries) is observed. Conversely, for both mixed-bases (B- and D-TFOs) an increase in guanine is shown which is enhanced (i) the longer the stretch and (i) the more enriched the sequences are (higher RPM values).

Another representation of nucleotide preferences in TFO sequences can be done by generating a sequence logo of the highly enriched TFO sequences. In Figure 37a (left panel), the sequence logo of the D-TFO library with a 9 nt stretch is shown and the strong preference for guanines is confirmed in this representation as well. To evaluate the consensus motif of the enriched TFO sequences, I employed the DRIMust analysis on the ranked list of RPMs and obtained the motif shown in Figure 37a (right panel) where a a 7 nt long G-rich motif is observed. Similarly, the sequence logo of the D-TFO library with a 7 nt long stretch is shown (Figure 37b, left panel) together with a 5 nt long G-rich DRIMust logo that was extracted from the ordered list (Figure 37b, right panel). Likewise, a G-enriched motifs is shown for the 9 nt strech of the B-TFO library in the sequence logo (Figure 37c, left panel), whereas the DRIMust consensus motif is shorter (5 nt long G-rich motif) (Figure 37c, right panel) when compared to the one obtained for the 9-nt stretch D-TFO library. This trend is further enhanced with the 7-nt stretch for

the B-TFO, where a sequence logo with little information content is observed (Figure 37d, left panel), while DRIMust was unable to detect enriched k-mers of any form (Figure 37d, right panel). It seems therefore that the lack of cytosines in the D-TFO contributes to the increased enrichment observed for that TFO library as compared to the B-TFO library where a mix of guanine, cytosines and thymines is present in the TFO sequences.

This strong G-enrichment in the TFO libraries with longer stretches (7 and 9 nt) resembles a striking similarity to the DRIMust motif detected in the *in vitro* Triplex-Seq platform of the pH 7 condition where a G-rich motif of approx. 7 nt guanines with one thymine interruption is shown (Figure 37e, left panel) and a shorter motif of 5 nt was found for pH 5 (Figure 37e, right panel). Given this similarity of *in vitro* Triplex-Seq and in cell Triplex-Seq DRIMust motifs for two different TFO libraries (B- and D-TFO) highlights the robustness of the Triplex-Seq platform to study triplex formation *in vitro* and in cells.

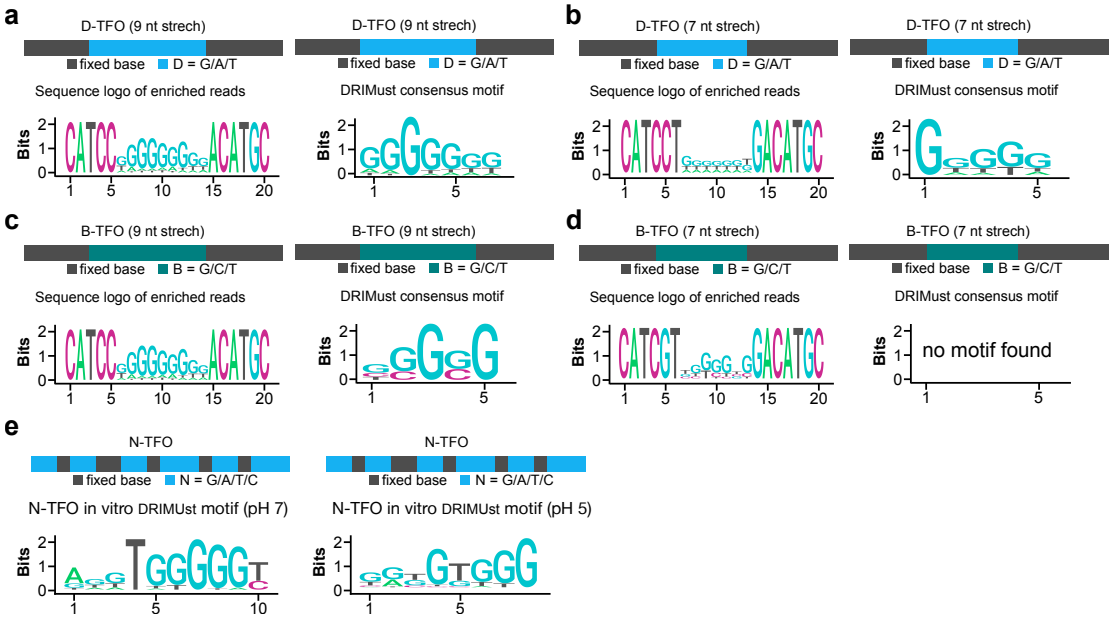


Figure 37| Minimal G-rich TFO motif in TFO sequences confirmed in in cell Triplex-Seq data. The sequence logos (left panels) and DRIMust motifs (right panels) are shown for a-d. **a, b,** The D-TFO library with 9 nt stretch and DRIMust motif (7 nt long) display a G-rich preference. A similar observation was made for the G-rich sequence logo for the 7 nt D-TFO library as well as for the DRIMust motif (5 nt long). **c, d,** Sequence logos for B-TFOs with 9 nt and 7 nt long stretches show G-rich pattern, but only 9 nt stretch results in DRIMust motif. **e,** The *in vitro* Triplex-Seq DRIMust motifs are shown for the N-TFO library in neutral (left panel) and acidic pH (right panel).

4.6.6 Summary of in cell Triplex-Seq data

The in cell Triplex-Seq platform was aimed at studying triplex formation in cells using a modified Triplex-Seq platform. The results imply that the in cell Triplex-Seq approach works, but requires further optimization.

1. The overall in cell Triplex-Seq data is noisier compared to the *in vitro* Triplex-Seq results and lack the 'TFO only' control to compute triplex reactivity scores.

2. The in cell Triplex-Seq data set, combined with the *in vitro* Triplex-Seq results, supports the hypothesis that G-rich TFOs (Figure 35 + Figure 36) lead to a stronger and more specific interaction with double-stranded DNA.

3. The identification of a minimal length of a TFO (7-10 nt) that is required to form triplexes has been confirmed with the in cell Triplex-Seq approach (Figure 36+ 37).

4.7 In cell Triloci-Seq

4.7.1 In cell Triloci-Seq approach

While the in cell Triplex-Seq platform displays advantages in (i) fast-turnaround times and (ii) a way to screen millions of putative TFO sequences, no knowledge of the corresponding genomic DNA to which the TFO was bound is gained. Here, I developed a slightly more complex high-throughput platform that relies on simultaneous sequencing of the TFO and TTS termed Triloci-Seq (Figure 38). The platform uses the same TFO library generation tool from IDT and has been described in the Triplex-Seq protocol. It is based on a ligation of the TFO to genomic DNA that was in close proximity in cells[162].

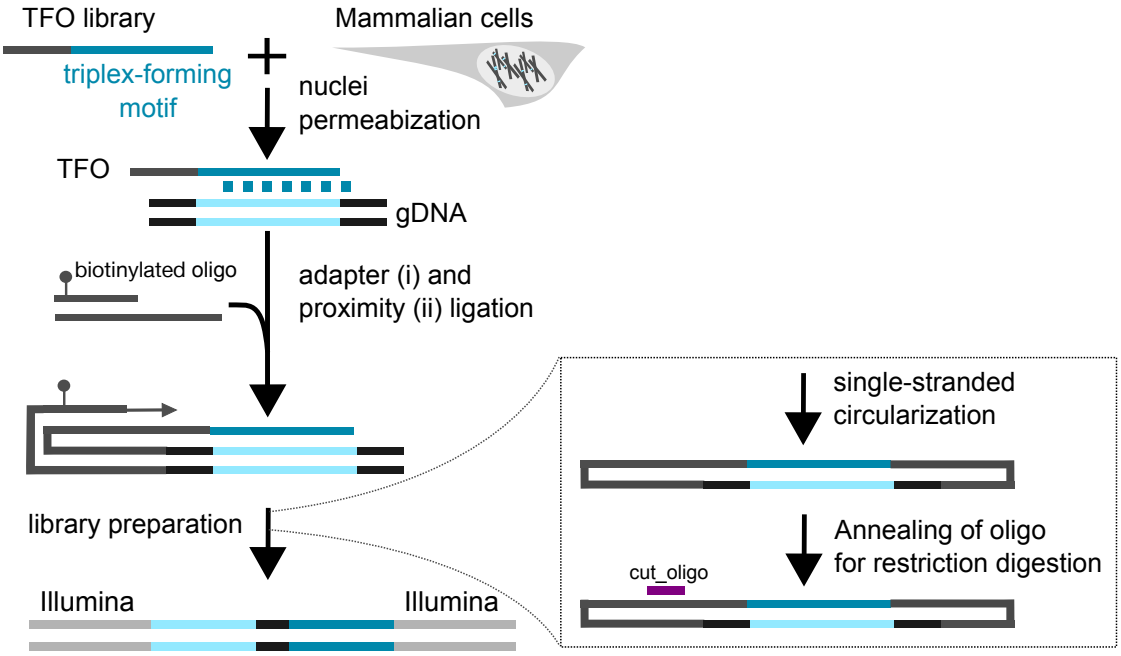


Figure 38| In cell Triloci-Seq design. The Triloci-Seq platform was develop to obtain information about both the TFO sequences, but also about the corresponding genomic DNA to which the TFOs were bound to. To do so, TFOs were generated using the mixed-base tool from IDT and transfected into mammalian cells. 24 hours post-transfection, cells were harvested, fixed and nuclei permeabilized. A biotinylated ss/dsDNA adapter was ligated to the single-stranded region of the TFOs and DNA (genomic DNA and TFO) was sheared using a restriction enzyme. Samples were diluted and TFO sequences that were bound to genomic DNA (thus were in close proximity) were ligated to the DNA fragments (proximity-based ligation). The ligated DNA fragments were isolated using streptavidin-coupled magnetic beads, and the non-biotinylated strand was circularized. An oligo was annealed that generated a restriction site, digested and subsequently PCR amplified to construct the Illumina sequencing libraries.

In briefs, TFO libraries were transfected into mammalian cells and a subset of the TFOs were expected to bind to genomic DNA. 24 hours post-transfection, cells were fixed and nuclei permeabilized. Following the single-stranded ligation of a biotinylated ssDNA/dsDNA chimeric oligo to the TFO libraries, genomic DNA was sheared using restriction enzymes (4 bp cutter), samples were diluted, and the TFO molecules that were in close proximity to the sheared DNA fragments were ligated to one-another (proximity-based ligation). The ligated DNA fragments were isolated, TFO/gDNA fragment enriched using streptavidin-coupled magnetic beads that bind to the biotinylated oligo of the TFO sequence, and the non-biotinylated double-strand sequence is further processed. Subsequently, the single-stranded fragments are circularized, a restriction site is created via annealing of a small oligo, DNA is digested and prepared for Illumina sequencing via PCR amplification.

4.7.2 In cell Triloci-Seq platform development

To validate and test the Triloci-Seq platform, ten single TFO variants were ordered (for details see list of Triloci-Seq primers on page 42). The triplex-forming motifs for the TFO variants were based on publications that suggested that lncRNAs regulate transcription via triplex formation. Thus the putative TTS sequences for RNA*DNA-DNA triplexes in the genome are supposedly known (Figure 39a). I mixed each TFO variant at molar ratios and the pool of ten TFOs was transfected into human embryonic kidney cells (HEK-293). Following the analysis of the sequence reads after sequencing, two TFO variants were found (GAA_{x15} and TTC_{x15}), while the remaining eight TFO variants could not be detected in the sequencing reads. In Figure 39b, the sequence logo of the reads containing the two TFO sequence is shown. Since the two TFOs that were found in the sequence reads, are identical in length with complementary sequence, I cannot distinguish whether these TFOs are indeed the two versions or in fact only one TFO whose sequence is shown in both directions. I will however talk about two TFOs in this section. In the sequence logo, it can be observed that the two TFO variants (the first 45 nucleotides) correspond to GAA and TTC triplex-forming motifs. The remaining nucleotides of the sequence reads represent a mix of all four nucleotides, it can however be noted that there is a preference for certain nucleotides (e.g. GAAGAA) at given positions indicating a recurring pattern. To identify the sequences that matched genomic DNA, the sequence reads were aligned to the human reference genome assembly GRCh38 (July 2018, GCF_000001405.38).

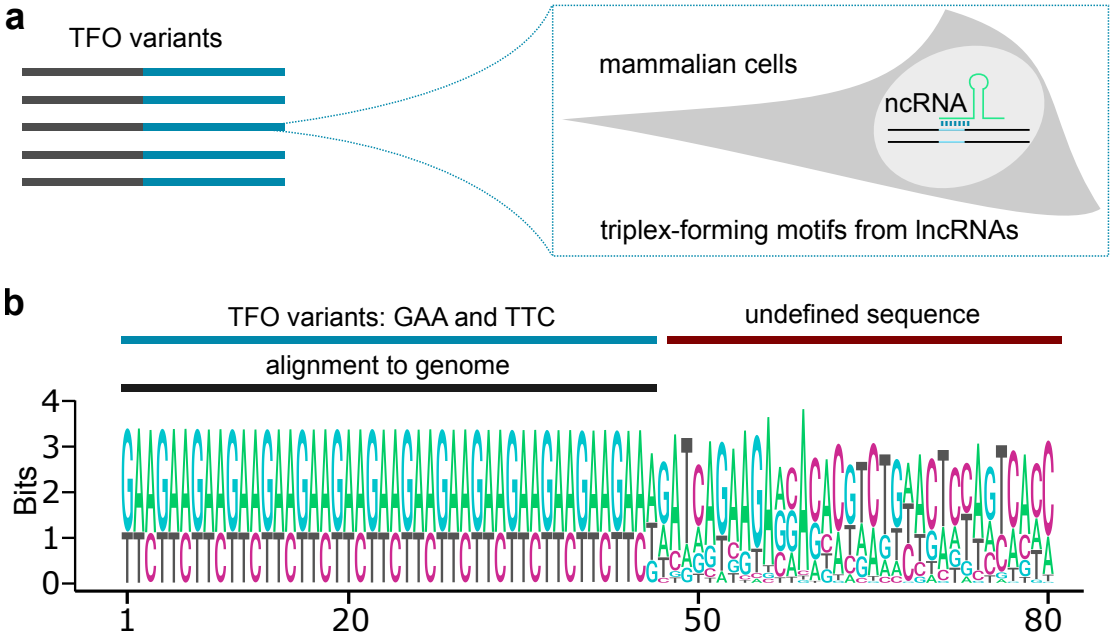
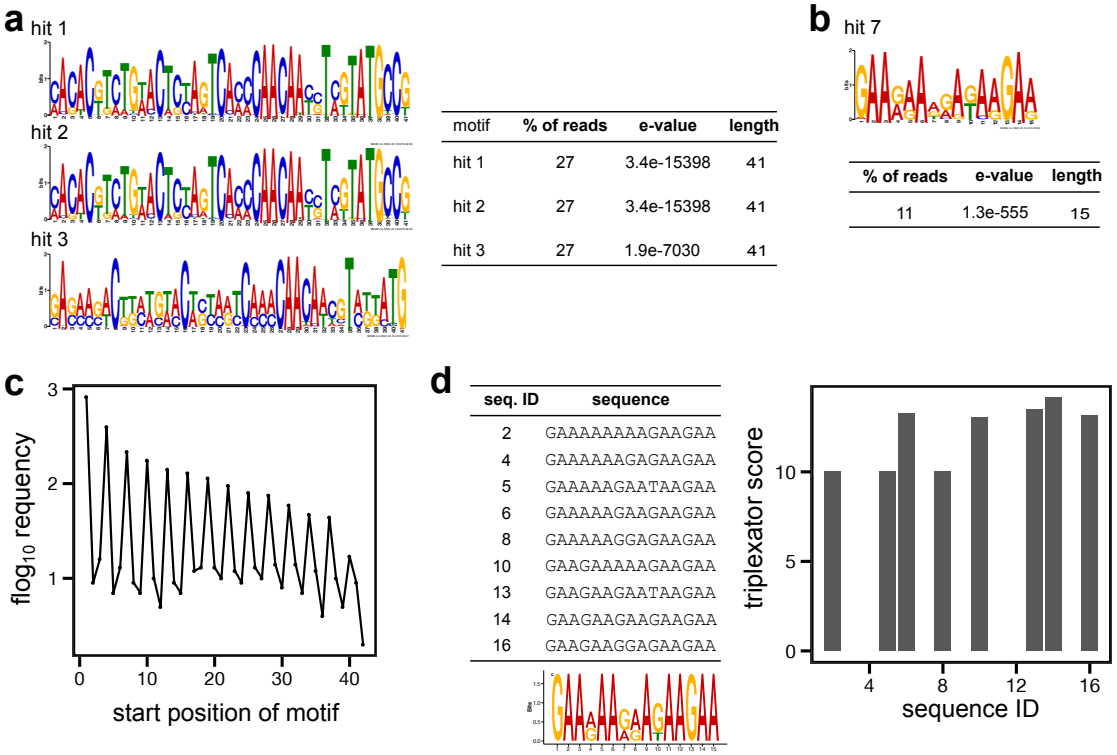


Figure 39| Validation of in cell Triloci-Seq approach using known triplex-forming motifs. a, Ten single TFO variants were designed to test and develop the in cell Triloci-Seq platform. These ten TFO sequences are based on motifs of lncRNAs that supposedly forming triplexes in cells. They were mixed at molar ratios and transfected as pool into mammalian cells. **b,** Out of the ten TFO variants, only two variants (TTC and GAA) were found in the sequence reads after sequencing the samples. The sequence logo of these sequence reads highlight the slight preference of GAA-sequences over TTC-sequences, and also imply a nucleotide preference in the sequence downstream of the TFO sequences (e.g. GAAGAA). To identify genomic DNA, the sequence reads were aligned to the genome. Most of the sequences that aligned to the genome correspond to the TFO sequences, there are however also some sequences downstream of the TFO sequence that aligned with the genome.

Surprisingly, the part of the sequence reads that correspond to the TFO variants were mostly the sequences that also aligned with the genome. If the TFO sequence (GAA or TTC repeats) aligned to the genome the full 45 nt of the TFO aligned with the genome. If the sequence downstream

of the TFO could be aligned with the genome, significantly shorter sequences aligned (15-20 nt) and overall corresponded to the GAA repeats that are observed in the sequence logo. In the next paragraph, I further characterized the sequences that are shown as the 'undefined sequences' in Figure 39b.

To identify recurring motifs in the sequences downstream of the TFO sequence, the MEME Suite (Motif-based sequence analysis tool) was used[185]. The TFO sequences were removed and the remaining part of the sequence reads (3739 unique sequences) was uploaded to the MEME (Multiple Em for Motif Elicitation) tool. Several motifs with different length patterns with high e-values were discovered. In Figure 40a, the first three motifs are shown as a sequence logo. Each motif appeared 1000 times (27 % of all sequence reads) and resemble the NEB Next primers that were used for preparing the sequencing libraries (see primer sequences on page 37), but were not detected during adapter trimming due to the occurrence of several mutations.



motif	% of reads	e-value	length
hit 1	27	3.4e-15398	41
hit 2	27	3.4e-15398	41
hit 3	27	1.9e-7030	41

% of reads	e-value	length
11	1.3e-555	15

seq. ID	sequence
2	GAAAAAAAAGAAGAA
4	GAAAAAAGAGAAGAA
5	GAAAAAGAATAAGAA
6	GAAAAAGAAGAAGAA
8	GAAAAAGGAGAAGAA
10	GAAGAAAAAGAAGAA
13	GAAGAAGAATAAGAA
14	GAAGAAGAAGAAGAA
16	GAAGAAGGAGAAGAA

Figure 40| Motif-based analysis of in cell Triloci-Seq data. a, MEME (Multiple Em for Motif Elicitation) analysis[185] of TFO-trimmed sequences identified three motifs as top hits that were 41 nt long and resemble the NEB Next primers that were used to PCR amplify the samples. The sequences were not removed during adapter trimming as they contain several point mutations (data not sown). **b,** A significantly enriched GAA-motif that does not align to the NEB Next primer. **c,** FIMO (Find Individual Motif Occurrences) analysis[186] was performed on GAA-motif (motif 7) against the sequence reads that were uploaded for the MEME analysis. The frequency of start positions is plotted as a function of the start positions within the sequence reads, and most sequences start directly downstream of the TFO sequence and decrease with higher start positions. Furthermore, some start positions occur at a frequency of 10 which results in a 3 nt periodicity. **d,** The 16 obtained motifs (derived from the sequence logo) were used to predict triplex formation with the two TFO sequences TTC and GAA using the triplexator software[117] and 9 out of 16 putative TTS sequences (represented by the sequence IDs in the table) form putative triplexes with one mismatch. The sequence logo of the motifs that form triplexes was generated and has a stronger preference to GAA sequences compared to the original sequence logo.

The first motif that differs from the primer sequence and was found 438 times in the sequence reads, is shown in Figure 40b. To further characterize this motif, another MEME Suite tool termed FIMO (Find Individual Motif Occurrences) was applied[186]. Here, the position of the discovered motif within the sequence read is detected and the frequency of motif positions was plotted as a function of the sequence read position (Figure 40c). It can be observed that most

motifs start at the beginning of the sequence read which is located immediately downstream of the TFO sequence (Figure 39b) and the number of motifs decrease with later start positions. Furthermore, a 3 nt periodicity is observed with high frequencies for start positions (S) that are divisible by 3 (except for the start position 1) and lower frequencies (~ 10) of motifs starting at nucleotides in between the $\frac{S}{3}$ start positions. This 3 nt periodicity is expected for a TFO motif with GAA/TTC repeats assuming that they interact with these motifs on the genome via triplex formation. The frequencies for these start positions which are significantly lower and are relatively constant might be attributed to mutations or sequencing errors. The high number of sequence motifs starting at the same position suggests a preference for ligation of GAA-rich sequences that are abundant in the genome.
To predict whether the discovered motif potentially forms triplexes with the two TFO sequences, I used the Triplexator software and tested the 16 discovered motifs (derived from the sequence logo) with the two TFO sequences. In Figure 40d, the Triplexator score (length of triplex) is plotted against the sequence ID (which is shown in Figure 40d in the table). 56 % of the discovered sequence motifs are forming triplexes according to the Triplexator prediction software and the sequence logo of these motifs (Figure 40d, bottom) is similar to the originally discovered motif (Figure 40b) with a slight trend towards GAA_{rich} sequences implying that these sequences might have been forming triplexes with either TTC- or GAA-TFO sequences.

4.7.3 Summary Triloci-Seq

The development of the Triloci-Seq approach serves the purpose to simultaneously sequence TFO and TTS sequences. In contrast to the Triplex-Seq platform, which is less complex, the Triloci-Seq approach has the potential to unravel genomic binding sites for triplex formation. Several insights based on the preliminary results were gained during development of the in cell Triloci-Seq protocol:

1. triplex-forming motifs that were described in triplex-forming lncRNAs were used as positive controls to test the in cell Triloci-Seq approach. 20 % of these TFOs that were transfected as a pool were found in the sequence reads and corresponded to the longest TFOs with 15 repeats of GAA and TTC, respectively. The abundance of GAA-repeats in the genome might explain the detection of only these two TFOs, and increasing the sequencing depth might increase the chance to find the other TFO motifs.

2. Motif-enriched analysis (MEME suite tools) identified mutated NEB Next primer sequences (noise) and a 15 nt long GAA-rich motif as a putative genomic triplex target site. The occurrence of the GAA-rich genomic motif in the sequence reads exhibited a 3 nt periodicity which is what I would expect for a TFO with a 3 nt repeat (GAA/TTC). Additionally, this motif was predicted to form triplexes with TTC and GAA-TFOs using the Triplexator software. These results imply that the genomic sites might have been bound by the TFO sequences.

5 Discussion

Studying triplexes has been and still is an intriguing research field to study. Understanding the underlying triplex formation rules *in vitro* using a controlled environment or cells has multiple impacts in both applied as well as basic research fields. Despite the extensive efforts to elucidate triplex formation, little has been done using high-throughput approaches. In this Ph.D. project, I aimed to develop a two-pronged approach: Building genetic circuits using synthetic biology-inspired logic and developing deep-sequencing platforms to tackle and systematically identify triplex formation rules. Initially, I designed two synthetic biology-based gene circuits to study triplex formation in bacterial and mammalian cells by constructing synthetic long non-coding RNAs (slncRNAs). In both designs, binding of slncRNAs to a triplex target site on a reporter plasmid via triplex formation was expected to change reporter gene transcription. Given the complexity of both designs only inconclusive results were obtained. Several factors for the weak or inconsistent changes in reporter gene levels have been identified and will be discussed in further detail below.

Based on these initial results from in cell experiments, I decided to simplify the experimental setup to understand triplex formation rules. To to so, I started with the design of a deep-sequencing platform in a controlled environment (*in vitro* Triplex-Seq). The successful implementation of the *in vitro* Triplex-Seq platform allowed me to identify a strong preference for G-rich triplex-forming sequences in the single-strands, as well as in the double-strands that participate in triplexes. Furthermore, stable triplexes are preferably formed in neutral pH and the minimal motif for TFO sequences that participate in triplex formation was pinpointed to approx. 7-10 nt. These results and their contribution to the research field of triplex formation will be discussed below.

The next logical step was then to adopt the Triplex-Seq platform to an in cell Triplex-Seq approach. Here, I could confirm the enrichment of G-rich TFO sequences that presumably were bound to the genome and identify again a minimal TFO motif of approx. 7-10 nt. These results will be discussed in the in cell Triplex-Seq paragraph. However, I would also like to comment on the background noise that was obtained in the in cell Triplex-Seq approach.

Despite these promising results, one limitation of the in cell Triplex-Seq platform is the lack of knowledge of putative triplex target sites to which the TFO was bound. Thus, the last part of the discussion will elaborate on the preliminary results and challenges during the development of the in cell Triloci-Seq approach for simultaneous sequencing of TFO and TTS.

5.1 Synthetic long non-coding RNAs

In the first part of this Ph.D. project, I used a synthetic biology-inspired approach to understand how lncRNAs interact with dsDNA. Based on the hypothesis of triplex formation between the single-stranded triplex-forming motif of the slncRNAs with double-stranded triplex target sites, slncRNAs were designed from the bottom-up and tested in bacterial as well as mammalian cells.

5.1.1 Enhancer-based circuit in bacterial cells

The rationale behind the bacterial enhancer-like circuit was to build a simple, fast and inexpensive platform to screen for slncRNAs in a medium-throughput manner. To achieve this, I designed a two-plasmid system comprising the enhancer-like reporter and a library of plasmids encoding slncRNAs under the C_4-HSL inducible promoter pRhlR (Figure 11). The slncRNAs carried different targeting sites (DNA_{bind}) and the enhancer-based reporters comprised either

respective, putative TTS or control sequences lacking the putative TTS sequences. I hypothesized that the slncRNA molecules would interact directly with the double-stranded TTS on the reporter plasmid by forming triplex structures and this slncRNA-TTS interaction presumably influences DNA looping capabilities and alter transcription (Figure 11).

Over 600 bacterial strains have been analyzed in an automated liquid handling platform and first moment mCherry values as well as distributions have been computed. For all data sets that have been compared, no up-regulatory effect until approximately 40 µM C_4-HSL was observed whereas higher C_4-HSL concentrations induced an up-regulatory response for bacterial strains with and without putative TTS in the reporter plasmids (Figure 13a). This up-regulatory effect observed at higher C_4-HSL concentrations is comparable to two publications that have shown a similar 2-2.5x-fold up-regulation of a luciferase reporter gene in presence of naturally occurring eRNA transcripts[28;31].

Furthermore, the results show a differential two step-behavior of slncRNAs in an enhancer-based reporter system using a σ^{54} promoter compared to reporter plasmids containing a σ^{70} promoter (Figure 13b). A significant difference for both the variability and the overall shape of the first moment value distributions was observed for concentrations higher than 40 µM, whereas first moment values and distributions did not differ for lower concentrations. This might indicate a non-specific interaction of the slncRNAs with the plasmid independently of the nature of the reporter (σ^{54} enhancer plasmid vs. σ^{70} non-enhancer plasmid) resulting in the first up-regulatory response of mCherry. Since changes in DNA supercoiling in bacteria play an underlying role for regulating gene expression[187], it is possible that the interaction of the slncRNAs with the plasmid might also change the DNA topology from DNA supercoiled plectonemes to a less tightly packed structure thus facilitating gene expression[188]. This however should not only influence the reporter plasmid (regulation in *trans*), but also the expression of the pRNA plasmid itself (*cis* regulation), which was not observed (data not shown).

Whereas the overall 20 % up-regulatory effect observed for all reporter plasmids might be a non-specific interaction with the plasmid, the second up-regulatory response observed only in the enhancer-like circuits may be due to a specific interaction of slncRNAs with promiscuous DNA sequences found in the looping region or on the reporter plasmid. To verify this hypothesis, these purine-rich stretches in the reporter plasmid were replaced by AT-rich sequences. In absence of any specific TTS, the strong non-specific up-regulatory response could not be detected; contrary to strains with the purine-rich spacer sequence (Figure 16) suggesting that four of the five designed DNA_{bind} motifs (GGA_{rich}, $(GAA)_{x8}$, pyr_{rich}, $AA_{rich,}$) bound non-specifically to the purine-rich repeats within the original spacer sequence. The fifth DNA_{bind} motif termed T0 contains a non-repetitive sequence, which stands in contrast to the other repetitive, purine/pyrimidine-rich DNA_{bind} sequences. The difference in behavior compared to the purine/pyrimidine-rich DNA_{bind} motifs (Figure 15) supports the hypothesis that this arbitrary sequence might not bind to promiscuous sequences in the looping region of the plasmid, but to its cognate TTS sequence in the loop and significantly enhances transcription.

The variability of mCherry distributions might indicate that a cell population or even a single cell has multiple states of bound slncRNAs. For instance, in one cell the non-specific interaction with the plasmid may dominate, whereas in the neighboring cell slncRNAs might interact via triplex formation with promiscuous sequences in the looping region that enhance transcription and a third cell might be bound by slncRNAs at different sequences that do not positively influence transcription. Given the wide range of distributions for bacterial strains that contain slncRNAs, the microplate reader assay might not be the ideal method to measure fluorescence intensities of single cells. Whether this interaction is protein-mediated, directly via triplex formation, or

due to alternative mechanisms could not be determined at this stage, and I decided to test the modular slncRNAs in a mammalian setup.

5.1.2 Triplex-mediated activation in mammalian cells

Since the bacterial enhancer-like circuit did not yield conclusive results, I decided to test a different synthetic circuit in mammalian cells because most naturally occurring lncRNAs have been described in mammalian/human cells[8;18] and the environment in the nucleus might favor triplex formation[189]. The CRISPR/Cas9-inspired gene activation system for mammalian cells is slightly more complex than the above described bacterial enhancer-like system. In the triplex-mediated gene activation circuit, slncRNAs were constitutively transcribed from a plasmid and provided a docking site for a $RBP_{activator}$ protein complex. Upon triplex formation of the slncRNA with a putative TTS on a reporter plasmid, the slncRNA/protein complex activates gene expression from a minimal promoter and fluorescent protein levels could be measured (Figure 18) .

I first tested the transfection efficiencies of the plasmids in presence and absence of the endonuclease Csy4 as well as the localization of the fluorescent protein reporters. The pRNA plasmid that carries the slncRNA, Csy4 regonition sites and the *sbfp2* gene exhibited over 80 % transfected cells in presence (Figure 19 and Figure 20) and absence (Figure 19a) of Csy4. Nissim *et al.* previously described the development of a multiplexable-gRNA cassette using the Csy4 strategy[154] and utilized mKate2 as a reporter gene for successful Csy4 cleavage. In this publication the authors described that in absence of Csy4 only low mKate2 levels were detected, whereas in presence of Csy4 a strong increase in mKate2 levels was observed. This stands in contrast to the here developed pRNA plasmid which displays similar levels of SBFP2 (which is the equivalent to the mKate2 reporter used by Nissim and colleagues) in presence and absence of Csy4. The reason for the difference in the behavior might stem from a polyA signal that was placed downstream of the *sbfp2*/slncRNA cassette which was absent in the design of Nissim and colleagues.

In contrast to the high transfection efficiencies of the pRNA plasmid, the pRBP plasmid encoding the PP7-mKate2-vp64 ($RBP_{activator}$) fusion protein displayed significantly lower transfection efficiencies of around 10 % in absence (Figure 19a) and 20 % in presence of Csy4 (Figure 20). The overall low expression levels of the PP7-mKate2 fusion could neither be traced back to the plasmid backbone nor promoter (data not shown) and cannot be explained at this stage, in particular based on the observation that mKate2 expression itself was strong, but limited to only 10 % of the cells. Contrary to the pRNA plasmid the pRBP plasmid does not harbor Csy4 sites thus should not be affected by the presence of Csy4. However, the increase in fluorescence levels of mKate2 that was observed in presence of Csy4 could be based on the formation of imperfect Csy4-recognition sites. It was previously shown that mutations in the RNA stem loop sequence identity affect binding of Csy4 to the RNA molecule to various degrees. While the hairpin structure is important for binding, cleavage is strongly impaired in absence of a guanine directly upstream of a scissile phosphate (the site of cleavage) or when secondary structures are formed below the stem structure[190]. In the 5' end of the polII promoter that was used to drive expression of the RBP-mKate2 fusion protein, I found a structure that could potentially resemble a weak Csy4-recognition site (data not shown), but formation of additional secondary structures downstream of the five-basepair hairpin might impair RNA cleavage. Thus, I propose that Csy4 may function as an RNA-binding protein with little to no cleavage activity and binding of the Csy4 protein to the 5'UTR of the RPB-mKate2 mRNA leads to the observed up-regulatory effect.

A similar up-regulatory observation was suggested recently by us where we placed an RBP in the 5'UTR of a reporter mRNA and were able to measure an increase in reporter expression levels (Katz *et al.*, submitted).
An additional observation that was made for the mKate2-fusion proteins is that they localized to the nucleus due to a nuclear localization signal (NLS), and form distinct spots when slncRNA molecules are present (Figure 19b). It has been described previously that PP7-FP fusion proteins can bind to RNA-hairpin structures [191] and were used for imaging mRNA molecules in bacteria [192] as well as mammalian cells [173;174] due to the formation of distinct spots, similar to what I have observed in presence of co-transfected slncRNA molecules. While this might indicate that the slncRNAs are in complex with the $RBP_{activator}$, I also observed similar spots in absence of slncRNA molecules, but presence of Csy4 (data not shown) suggesting that there might be another explanation for the formation of such spots within the nucleus.
Given the low protein expression levels of the $RBP_{activator}$ and the lack of responsiveness of the pRNA plasmid for Csy4 cleavage, it is not surprising that only a weak and inconsistent up-regulatory effect of the reporter gene is observed. The insufficient cleavage of the slncRNAs by the Csy4 protein translocates the slncRNAs into the cytoplasm and translates them into slncRNA/SBFP2 fusion proteins. Moreover, only 10-20 % of the cells contain the $RBP_{activator}$ that binds the slncRNA molecules in the cytoplasm and transports them via the NLS of the $RBP_{activator}$ protein back into the nucleus.
Given the discussed issues of few transfected cells and slncRNA localization, the overall lack or inconsistent up-regulation can be traced back to these factors. I would like however to comment on one trend observed for the slncRNA with the GAA-motif, where with one PP7-binding ($PP7_{x1}$) site an up to 7-fold higher expression was observed (Figure 21c), compared to slncRNA molecules containing more PP7-binding sites ($PP7_{x3}$ and $PP7_{x4}$). This overall decrease in reporter gene expression with higher numbers of PP7-binding sites is contrary to the expectations. It might be explained by interference of the longer slncRNA molecules with the RNA transcription machinery. It has been described that anchored, nascent RNA could destabilize the transcription complex by wrapping around the DNA template and the formation of R-loops [193]. The longer the RNA molecule, in this case the slncRNA that is bound to the TTS, the higher the probability that these RNA molecules can form R-loops in the transcribing reporter gene sequence thereby reducing gene expression.

5.2 Deep-sequencing platforms

Based on the inconclusive results from the synthetic biology-inspired gene circuits using slncRNAs, I decided to go one step back and use DNA-based TFOs to develop high-throughput platforms to study triplex formation.

5.2.1 *In vitro* Triplex-Seq

I designed and developed the *in vitro* Triplex-Seq platform to provide a unique high-throughput tool to study the underlying triplex formation rules in a controlled environment. The power of the platform lays in its combination of the IDT mixed-base tool to generate millions of different TFO variants with the classical shift assay and next-generation sequencing technologies (Figure 23). This makes the *in vitro* Triplex-Seq approach cost-effective and with its fast-turnaround time an ideal platform to systematically study triplex formation. Here, I would like to integrate the findings of the *in vitro* Triplex-Seq platform into the known literature context.

To evaluate successful triplex formation, I introduced a new measure termed triplex reactivity. After next-generation sequencing, I obtained a list of sequence reads for the sample where (i) the TTS and the TFO library were incubated and triplexes were formed (triplex sample) and where (ii) the TFO library alone was incubated (TFO sample). The TFO only control serves as the background noise. Hence, triplex reactivity is defined as the ratio of the normalized read counts of the triplex and TFO sample and subtraction of one. All positive triplex reactivity scores indicate formation of triplexes between the TFO variants and the TTS molecule.

I started by first characterizing the TFO libraries with 2-mixed bases. I observed a pH-dependent triplex formation where low triplex reactivity scores were found for acidic pH, and higher triplex reactivities were detected in neutral pH (Figure 25). By comparing the enriched motifs of these TFO libraries which were obtained using the DRIMust tool to past literature findings, I obtain support for the sensitivity and validity of the Triplex-Seq platform. Specifically, the R-TFO library (G/A) displays a short GARA-motif which has been associated in the Friedreich's Ataxia disease[83] and was found to down-regulate expression of the *dhfr* gene[194;48]. The K-TFO library (G/T) features a short consensus motif with 80 % thymines. Formation of triplexes with mixed GT-motifs have been described to preferably form anti-parallel triplexes[195] thus supporting my finding of higher triplex reactivities in pH 7 for the K-TFO library. The W-TFO library (A/T) exhibits stretches of adenines as well as thymines. Both nucleotides have been shown to form triplets with an A-T basepair, while adenines also interact with a G-C basepair[195]. It is however surprising that stretches of both nucleotides were found to form stable triplexes as this has not been reported before. The M-TFO library (adenine/cytosine) reveals an A/C mixed-consensus motif in both conditions. When looking at the top hits of the ranked triplex reactivity lists (data not shown), a preference for adenines can be observed (60 – 70 %) which can be explained by adenine's ability to bind to both G-C and A-T basepairs. The A/C-mixed motifs however might result from formation of non-canonical triplexes by binding to the C-T and T-A basepairs as was proposed previously[195] or the potential formation of parallel triplexes due to cytosine protonation at neutral pH[196;197].

In contrast to mixed motifs when testing smaller TFO libraries (16,000 variants), G-rich DRIMust consensus motifs were found when evaluating sequences from the N-TFO library (Figure 26). Stable G-rich triplexes have been described *in vitro*[198;199;200;184] and *in vivo*[97;201], but were also found to form G-quadruplexes in physiological environments[198;79]. The overall trend of G-rich TFOs combined with the short ACGT-DRIMust motifs in a triplex-disfavoring buffer (pH 7 + K^+) may indicate a triplex/G-quadruplex mix. G-rich TFO variants potentially assemble to G-quadruplexes in presence of potassium[202] and migrate through the gel at similar heights as the triplexes, while a weak interaction (low reactive) of TFOs with the TTS results in the ACGT-consensus motif.

Next, based on the observation of G-rich DRIMust consensus motifs (5-9 nt long), I designed TFO libraries with stretches of continuous mixed-bases (B- and D-TFO stretches) of varying length (3-9 nt) that are flanked by fixed bases (Figure 27). The TFO libraries with shorter mixed-base stretches (3-7 nt) lack a nucleotide preference and no DRIMust consensus motif could be computed. I hypothesize that these short TFO sequences bind weakly and in a non-specific fashion to the TTS, but are too short to strongly and specifically interact with the TTS. Conversely, only the TFO library with a 9 nt long D-stretch that was tested in neutral pH exhibited a clear G-rich DRIMust motif (5 nt long). These results confirm (i) the pH dependence proposed earlier because no DRIMust motif was found for the TFO with a 9 nt long B-stretch that was tested in pH 5 and (ii) identify approx. 9 nt as the minimal length at which TFO sequences form specific and strong triplexes. This is a shorter motif than what was described in

literature where the TFO motifs mostly ranged between 15-30 nt [189].
Lastly, the relationship between TFO enrichment and variation in guanine/adenine ratios in the TTS for triplex formation was evaluated and indicated a clear increase for G-rich TFO sequences which was correlated with both increasing guanine content in the TTS as well as higher triplex reactivity scores (Figure 30). This is in accordance with findings from another study which identified G-rich patterns of over 80 % in the TTS supporting the importance of guanines in the TTS to form stable triplexes [179]. The DRIMust motifs imply that with low guanine percentage little information is obtained and suggest that many variants interact weakly with these TTS variants (Figure 32). In contrast for TTS sequences with high guanine percentage, the consensus motif suggests that G-rich TTS sequences lead to a stronger and more specific triplex interaction with the TFOs.

5.2.2 In cell Triplex-Seq

After the successful development of the *in vitro* Triplex-Seq platform, I adopted it to an in cell approach to elucidate triplex formation in cells (Figure 33). In contrast to the *in vitro* approach, no 'TFO only' control was included in the initial implementation of the in cell Triplex-Seq protocol. Hence, no triplex reactivity score could be computed and used to determine the formation of triplexes above the background noise. Instead, the normalized read counts (RPMs) which indicate the level of enriched sequences were used. The lack of a 'TFO only' control obviously raises questions in the validity and scope of my interpretations regarding triplex formation in cells, and I will address this issue in the conclusion and outlook chapter on page 95 where I will propose modifications of the basic Triplex-Seq protocol to include a 'TFO only' control.
Several other controls were however tested in the in cell approach and included (i) non-transfected cells, (ii) cells that were transfected with the capture sequence only and (iii) a TFO library that was not transfected, but only PCR amplified. For the control samples (i) n.t. and (ii) capture almost 80 % of the sequence reads that were found corresponded to 39 nt long sequences and consisted of the 19 nt long capture sequence and a 20 nt downstream sequence (Figure 34b). These downstream sequences consisted of mainly long stretches of A-rich and GAA-rich sequences and were also found in all samples that were transfected with TFO libraries. Furthermore, over 30 % of R-TFO and 1.5 % of K-TFO variants were identified in the sequence reads and might point to contamination with these TFO libraries in the downstream processing of the Triplex-Seq protocol. The control (iii) that corresponded to the PCR amplification of the D-TFO library without transfection confirmed that during PCR amplification of the TFO libraries no PCR biases in form of mutations or preferences of certain nucleotide patters are introduced. The average nucleotide frequency for the TFO sequences with a given RPM value was equally distributed between the three nucleotides (33/33/33 ratio of G/A/T). This result strongly confirms that any nucleotide enrichment observed for the TFO libraries after transfection are likely caused by interaction with the genome (Figure 35).
After characterizing the background noise and PCR biases, several smaller libraries with 16,000 variants and a large library with over 260 Mio. variants were tested. The frequency distributions that were generated using the normalized read counts indicated multiple distributions depending on the TFO library. While for the M- (A/C) and Y-TFO (C/T) libraries less than 10 % of the variants of the total library were found and no enrichment was observed, the K-(G/T) and R-TFO (G/A) libraries exhibited either a similar distribution that was observed in the control samples (K-TFO) or a two step-distribution with a potential enrichment of certain sequences

(R-TFO). Given the fact that one-third of all R-TFO variants were found in the control samples and over 95 % of all variants were identified in the sequence reads of the R-TFO library itself, the question regarding the validity of the enrichment of these sequences arises and interpretations need to be regarded with caution. Lastly, no enrichment was observed for the N-TFO library which might be based on the lack of sequencing depth as only 0.03 % of all variants were detected, and an enrichment-like behavior as observed in the *in vitro* Triplex-Seq might be achieved with higher sequencing coverage (Figure 34c).

Since neither the smaller libraries nor the large N-TFO library resulted in an enrichment of TFO sequences, I tested intermediate-sized TFO libraries (B- and D-TFO). Contrary to the other TFO libraries, I obtained enriched TFO sequences for both B- and D-TFO libraries (Figure 35) and an increase of guanines in the enriched TFO sequences is found for both libraries suggesting that G-rich sequences are important for interactions of the TFOs with genomic DNA. This is in accordance with literature where purine-rich TFOs have been described to form stable triplexes in physiological conditions[182;203;179]. In the particular case of G-rich triplex-forming motifs, ambivalent results have been presented. While several lncRNAs contain purine/G-rich triplex-forming motifs[48;53;54] and TFOS have been used in cells[97;100] and even in mice[101] to successfully introduce site-specific mutations or regulate transcription[95], other studies showed a decreased bioactivity of G-rich TFOs due to formation of G-quadruplexes[200;184], or self-assembly of the G-rich TFOs[202;102]. Based on this literature context, I suggest to perform additional experiments using these enriched G-rich TFOs in cells to further shed light on this ambivalence (see on page 95).

Given the enriched TFO sequences with a trend towards G-rich sequences using the B- and D-TFO libraries, I wanted to determine the minimal length of TFO sequences (Figure 36). To do so, a set of TFO libraries with mixed-base B- and D stretches of varying length were designed and deep-sequencing revealed that the shorter mixed-base stretches (3-5 nt, to a certain extent also 7 nt long) did not yield any strong preference for any nucleotides contrary to the 9 nt long B- and D-TFO stretches where a clear G-rich pattern was observed. This suggests that TFOs shorter than 9 nt do not bind in a strong and specific manner to genome. Additionally, a 5 nt long G-rich DRIMust consensus motif was observed for the D-TFO (9 nt) and no motif was detected for TFOs with the same length but B-mixed bases (Figure 37). The difference in the mixed-base stretches of the B- and D-TFO libraries is based on the difference of containing either a cytosine (B-TFO library contains G/C/T nucleotides) or an adenine (D-TFO library is comprised of G/A/T). This is in accordance with literature where triplexes were proposed to form more stably with purine motifs (G/A)[97;48] while pyrimidine motifs (C/T) require the protonation of the cytosine residue to form triplexes[86;88]. Since the B-TFO library contains cytosines it it less likely to form stable triplexes thus implying and confirming that purine motifs are more likely to form triplexes in cells and require no more than 9 nt to be a specific and strong interaction partner with double-stranded DNA.

The experimental setup of the in cell Triplex-Seq approach is founded on the assumption that TFO molecules bind to genomic DNA sequences via triplex formation. Given the recurring pattern of G-rich TFO sequences, one might hypothesize that other interactions might play a role in TFO enrichment. It is well established that the genome contains an abundance of G-rich sequences such as in the end of chromosomes (telomeres[204]) and these G-rich stretches are often associated with the formation of G-quadruplex formation[79]. In case of telomeres, formation of these non-canonical DNA structures prevents the linear ends of the chromosome to be degraded[63]. Given the preference of G-quadruplex formation of G-rich sequence in physiological conditions[86;89], it suggests that TFO sequences might interact with the genome via

G-quadruplex formation. Other interactions within the cells (such as with proteins that prevent the degradation of G-rich TFO molecules[205]) or an experimental bias (e.g. via an increased transfection efficiency of G-rich TFOs) might be possibilities for the enrichment of G-rich TFOs that need to be ruled out. Although I am aware of these potential pitfalls in the experimental setup, I believe that the combined and almost identical results from the *in vitro* Triplex-Seq platform which was performed in a controlled environment, suggest that triplex formation is the underlying interaction with the genome and leads to the enriched G-rich TFO sequences.

5.2.3 In cell Triloci-Seq

Given the limitation of the in cell Triplex-Seq approach with respect to (i) only be able to sequence the TFO and (ii) the assumed, but not proven, triplex interaction of TFOs with genomic triplex target sites, the Triloci-Seq platform was developed to integrate TFO and TTS sequencing and obtain more information regarding the two main limitations of the in cell Triplex-Seq approach (Figure 38).

To develop the in cell Triloci-Seq protocol, a set of known triplex-forming motifs derived from lncRNAs were used (Figure 39). These sequences ranged from 13 to 45 nt in length and were transfected as a pool of sequences. Following the execution of the Triloci-Seq protocol, two out of ten TFO sequences were found in the sequence reads. The TFO sequences that were detected are the two longest TFOs from the set and contained 15 repeats of GAA and TTC, respectively. Long purine/pyrimidine stretches, and in particular GAA-repeats have been shown to be abundant within cells[160;49] and are significantly enriched in gene regulatory elements (5' and 3' regulatory regions)[49]. Furthermore, expansion of such repeats have been implicated in the Friederich's ataxia disease[83;206]. The abundance of these putative triplex target sites might be one explanation why only these two TFOs were detected in the sequence reads. This also highlights the importance of sequencing depth which needs to be increased significantly to find TFO sequences whose TTS might be less abundant throughout the genome.

Furthermore, most TFOs that were used to test the Triloci-Seq protocol are shorter than the two TFOs that were found. Another potential reason for the enrichment of the longer TFOs can be explained by polymer physics. In biophysics, nucleic acid molecules are viewed as polymers whose behavior can be effectively described by the worm-like chain (WLC) model. DNA is considered to be a rigid molecule and the parameter 'persistence length' is used as a measure for the structural rigidity of DNA molecules and the energy that is required for the DNA to bend. While the persistence length for double-stranded DNA ranges between 35-100 nm (naked vs chromatin DNA)[207;208], the persistence length for single-stranded DNA has been described to be significantly smaller (4 nm)[209;210]. Based on the Triplex-Seq results that identified a minimal stretch of approx. 9 nt that is required for triplex formation, the longer TFOs (45 nt) might not bind entirely to the genome. Given that the genome was sheared into 100-300 bp fragments (using a restriction enzyme recognizing a 4 bp DNA site), a longer, single-stranded (and more flexible) stretch of the TFO molecules increases the probability that the rigid double-stranded genomic region finds the ssDNA fragments. This is in contrast to shorter TFO sequences which might be bound entirely to the genome and thus cannot compensate for the stiffness of the double-stranded molecules.

To characterize the part of the sequence reads that do not correspond to the TFO sequence, I used the MEME suite tool and identified a 15 nt long GAA_{rich} motif which exhibited a 3 nt periodicity with respect to the motif occurrence (start position of the motif) in the sequence reads (Figure 40). This 3 nt periodicity in the sequence reads is expected, if we assume that the GAA_{x15} or

TTC_{x15} TFOs are bound to this motif via triplex formation. This observation suggests that the TFO/GAA-motifs have been identified in the sequence reads because the TFO was bound to the genome (or to these particular motifs) via triplex formation and, as part of the Triloci-Seq protocol, were subsequently ligated to one another. Lastly, I applied the Triplexator software on the detected GAA_{rich} motif and predicted triplex formation with the two TFOs (GAA_{x15} and TTC_{x15}). I generated a list of sequences based on the sequence logo of the GAA_{rich} motif (16 TTS variants) and used the two TFO sequences as interaction partners for triplex formation. Nearly 60 % of all sequences were found to form triplexes suggesting that triplex formation was the underlying interaction of the TFOs with the genome.
These preliminary results imply that the in cell Triloci-Seq approach works and can be used to further expand our knowledge in triplex interactions in cells. However, I would like to point out that these results are preliminary and no controls (such as non-transfected cells or a scrambled version of the long TFO sequences) were tested. Hence, improving the in cell Triloci-Seq protocol is the next step to obtain more significant results.

6 Conclusion and Outlook

In this Ph.D. project, I strove to decipher triplex formation and its underlying rules *in vitro* and in cells. To do so, I started out by designing and testing modular synthetic long non-coding RNAs (slncRNAs) to characterize triplex formation via synthetic gene circuits in bacterial and mammalian cells in a medium-throughput approach. Given the complexity of the systems and a lack of understanding of triplex formation in cells, this approach yielded overall inconclusive results. Thus through the course of time, I decided to design simpler high-throughput platforms which could both be used *in vitro* and in cells. By applying these approaches, I demonstrated the power of such high-throughput sequencing technologies to systematically analyze triplex formation in a sensitive, sequence-varied and cost-effective manner. My *in vitro* Triplex-Seq results suggest that triplex formation is pH-dependent, the minimal length of a TFO motif ranges between 7-10 nts, and G-rich TFO and TTS sequences lead to a stronger and more specific triplex interaction. The in cell Triplex-Seq data support the *in vitro* findings that G-rich TFOs are preferred dsDNA binding partners and the minimal length of TFOs ranges indeed between 7-10 nts. Given that no knowledge of the sequence content of the dsDNA is gained by the in cell Triplex-Seq approach, I developed in cell Triloci-Seq which allows for the simultaneous sequencing of TFO and genomic DNA. The preliminary results of the Triloci-Seq platform using ten different TFO variants of known triplex-forming motifs suggest that in cell Triloci-Seq requires a significant larger sequencing depth than was used in these initial experiments to detect TFO/TTS interactions that are less abundant. The two TFOs that were found to be ligated to genomic DNA have abundant putative TTS sequences within the genome. Given the GAA-repeats of the TFOs, the finding of a 3 nt periodicity in a putative TTS motif (15 nts long) implies that the in cell Triloci-Seq approach might have worked.

With the growing interest in dynamic non-canonical structures and their applications, understanding the underlying code of triplex formation *in vitro* is at the forefront of this endeavor. With the development of the *in vitro* Triplex-Seq platform, I contributed a unique tool and valuable insight to the field of triplex formation. This has direct implications for various research fields including material sciences through developing triplex-responsive hydrogels[211], synthetic biology-inspired diagnostics[212] and biosensors[213], nanotechnology-based switches[214;215] and ideas to expand the DNA origami toolbox[216].

Contrary to the *in vitro* Triplex-Seq results, the in cell approaches (Triplex-Seq and Triloci-Seq) have potential to tackle the challenges of studying triplex formation in cells, to expand the use of TFOs for therapeutic approaches and to understand how lncRNAs interact precisely with the genome. While I believe that I have made the first steps to fulfill these potentials, both platforms require further characterization using synthetic biology approaches and a general technology development.

As mentioned earlier, the in cell Triloci-Seq approach is still in the beginning of its development, in particular with respect to the use of better crosslinking techniques, novel ligation reactions, comparison of controls, increasing sequencing coverage and more elaborate bioinformatic analysis. As this requires mostly technical improvements and would expand beyond the scope of this section, I would like to focus on several approaches to improve the in cell Triplex-Seq platform. Here, I envision multiple strategies to further expand the toolbox of the developed in cell Triplex-seq approach. One challenge in studying triplex formation in cells is to directly identify that the enriched TFO molecules are bound to dsDNA (via triplex formation). To confirm that the enriched TFO variants from the in cell Triplex-Seq platform were bound to genomic DNA, I designed a synthetic-biology like gene activation circuit similar to what has been described

in literature[95] and is similar to what I have tried with the slncRNA-mediated gene activation circuit. A simple synthetic gene circuit will be applied to characterize selected variants obtained from the in cell Triplex-Seq platform by designing a reporter plasmid containing a minimal CMV promoter ($pCMV_{min}$) with 5-10 G-rich variants of putative triplex target sites that will be placed upstream of the $pCMV_{min}$ driving a reporter gene. Furthermore, a conjugation of small activator peptides (such as vp16) to 2-3 selected TFO variants ($TFO_{activator}$) will be performed. Upon triplex formation, the $TFO_{activator}$ is brought into close proximity of the minimal promoter thereby activating transcription. Once the system is successfully implemented two strategies can be followed: (i) use the enriched TFO variants and perform RNA-seq to determine up-regulation of genes and identify putative TTS in gene-regulatory elements or (ii) replace the DNA-based TFOs with RNA-TFOs to study RNA*DNA-DNA triplex formation.

Furthermore, the discussed challenge of lacking a 'TFO only' control in the in cell Triplex-Seq protocol prompted me to think about another way to perform the in cell Triplex-Seq approach. Here, I propose different capture approaches by exploiting the common capture sequence shared by all TFO variants which is used to enrich the TFOs after genomic DNA isolation. In the new capture approaches, the biotinylated oligo which is complimentary to the common capture sequence on the TFO, will be hybridized either (i) before transfection and (ii) after cell lysis, but before DNA isolation. Following the enrichment of the TFOs using the streptavidin-coupled magnetic beads, the in-cell Triplex-Seq protocol can be followed as described previously. Samples that will be taken at various time points (e.g. directly after transfection) then represent the 'TFO only' controls.

Coming to the end of this Ph.D. thesis, I would like to take a look at the broader picture and the scope of this thesis. I began with the development of synthetic-biology inspired circuits to study triplex formation. Retrospectively, what I should have started with, was the development of the high-throughput platforms to thoroughly understand triplex formation and then use synthetic biology as a way to verify and highlight that triplex formation can also be used for multiple applications in cells. Nevertheless, I believe I did gain important insights into triplex formation and established tools that can be used to further expand the knowledge of the exciting world of non-canonical DNA and continue the quest for shedding light on the genome's dark matter.

References

[1] Shawn E. Levy and Richard M. Myers. Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.*, 17(1):95–115, aug 2016.

[2] Sriram Kosuri and George M Church. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods*, 11(5):499–507, apr 2014.

[3] Diane E. Dickel, Athena R. Ypsilanti, Ramón Pla, Yiwen Zhu, Iros Barozzi, Brandon J. Mannion, Yupar S. Khin, Yoko Fukuda-Yuzawa, Ingrid Plajzer-Frick, Catherine S. Pickle, Elizabeth A. Lee, Anne N. Harrington, Quan T. Pham, Tyler H. Garvin, Momoe Kato, Marco Osterwalder, Jennifer A. Akiyama, Veena Afzal, John L.R. Rubenstein, Len A. Pennacchio, and Axel Visel. Ultraconserved Enhancers Are Required for Normal Development. *Cell*, 172(3):491–499.e15, 2018.

[4] Nils Weinhold, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.*, 46(11):1160–1165, nov 2014.

[5] Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, and Michael Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, sep 2012.

[6] Sarah Geisler and Jeff Coller. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.*, 14(11):699–712, nov 2013.

[7] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian A Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Nadav S Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J Luo, Eddie Park, Kimberly Persaud, Jonathan B Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E Antonarakis, Gregory Hannon, Morgan C Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, sep 2012.

[8] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, Julien Lagarde, Lavanya Veeravalli, Xiaoan Ruan, Yijun Ruan, Timo Lassmann, Piero Carninci, James B Brown, Leonard Lipovich, Jose M Gonzalez, Mark Thomas, Carrie A Davis, Ramin Shiekhattar, Thomas R Gingeras, Tim J Hubbard, Cedric Notredame, Jennifer Harrow, and Roderic Guigó. The GENCODE v7 catalog of human long noncoding RNAs:

analysis of their gene structure, evolution, and expression. *Genome Res.*, 22(9):1775–1789, sep 2012.

[9] Tae-Kyung Kim, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F Worley, Gabriel Kreiman, and Michael E Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, may 2010.

[10] Kerstin-Maike Schmitz, Christine Mayer, Anna Postepska, and Ingrid Grummt. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.*, 24(20):2264–2269, oct 2010.

[11] Miao-Chih Tsai, Ohad Manor, Yue Wan, Nima Mosammaparast, Jordon K Wang, Fei Lan, Yang Shi, Eran Segal, and Howard Y Chang. Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329(5992):689–693, aug 2010.

[12] Anton Wutz, Theodore P Rasmussen, and Rudolf Jaenisch. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet*, 30(2):167–74., 2002.

[13] Ibrahim A Ilik and Asifa Akhtar. roX RNAs: non-coding regulators of the male X chromosome in flies. *RNA Biol.*, 6(2):113–121, apr 2009.

[14] Tomoshige Kino, Darrell E Hurt, Takamasa Ichijo, Nancy Nader, and George P Chrousos. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal.*, 3(107):ra8, feb 2010.

[15] Chenguang Gong and Lynne E Maquat. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, 470(7333):284–288, feb 2011.

[16] Jesse M. Engreitz, Jenna E. Haines, Elizabeth M. Perez, Glen Munson, Jenny Chen, Michael Kane, Patrick E. McDonel, Mitchell Guttman, and Eric S. Lander. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539(7629):452–455, 2016.

[17] Marta Melé and John L. Rinn. "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription. *Mol. Cell*, 62(5):657–664, jun 2016.

[18] Pedro J. Batista and Howard Y. Chang. Long noncoding RNAs: Cellular address codes in development and disease. *Cell*, 152(6):1298–1307, 2013.

[19] US National Library of Medicine National Institutes of Health. Search on ncbi: long non coding RNA, 2018.

[20] Graeme D. Penny, Graham F. Kay, Steven A. Sheardown, Sohaila Rastan, and Neil Brockdorff. Requirement for Xist in X chromosome inactivation. *Nature*, 379(6561):131–137, jan 1996.

[21] Christine Moulton Clemson, J A McNeil, H F Willard, and J B Lawrence. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J. Cell Biol.*, 132(3):259–275, feb 1996.

[22] Kathrin Plath, Jia Fang, Susanna K Mlynarczyk-Evans, Ru Cao, Kathleen A Worringer, Hengbin Wang, Cecile C de la Cruz, Arie P Otte, Barbara Panning, and Yi Zhang. Role of histone H3 lysine 27 methylation in X inactivation. *Science*, 300(5616):131–135, apr 2003.

[23] Colleen A. McHugh, Chun-Kan Chen, Amy Chow, Christine F. Surka, Christina Tran, Patrick McDonel, Amy Pandya-Jones, Mario Blanco, Christina Burghard, Annie Moradian, Michael J. Sweredoski, Alexander A. Shishkin, Julia Su, Eric S. Lander, Sonja Hess, Kathrin Plath, and Mitchell Guttman. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature*, 521(7551):232–236, may 2015.

[24] Jeffrey J Quinn, Ibrahim A Ilik, Kun Qu, Plamen Georgiev, Ci Chu, Asifa Akhtar, and Howard Y Chang. Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nat. Biotechnol.*, 32(9):933–940, sep 2014.

[25] Ibrahim A Ilik, Jeffrey J Quinn, Plamen Georgiev, Filipe Tavares-Cadete, Daniel Maticzka, Sarah Toscano, Yue Wan, RobertC Spitale, Nicholas Luscombe, Rolf Backofen, HowardY Chang, and Asifa Akhtar. Tandem Stem-Loops in roX RNAs Act Together to Mediate X Chromosome Dosage Compensation in Drosophila. *Mol. Cell*, 51(2):156–173, jul 2013.

[26] John L. Rinn, Michael Kertesz, Jordon K. Wang, Sharon L. Squazzo, Xiao Xu, Samantha a. Brugmann, L. Henry Goodnough, Jill a. Helms, Peggy J. Farnham, Eran Segal, and Howard Y. Chang. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323, jun 2007.

[27] Srinivas Somarowthu, Michal Legiewicz, Isabel Chillón, Marco Marcia, Fei Liu, and Anna Marie Pyle. HOTAIR forms an intricate and modular secondary structure. *Mol. Cell*, 58(2):353–361, apr 2015.

[28] Xiangting Wang, Shigeki Arai, Xiaoyuan Song, Donna Reichart, Kun Du, Gabriel Pascual, Paul Tempst, Michael G Rosenfeld, Christopher K Glass, and Riki Kurokawa. Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature*, 454(7200):126–130, jul 2008.

[29] Allison M Bond, Michael J W VanGompel, Evgeny A Sametsky, Mary F Clark, Julie C Savage, John F Disterhoft, and Jhumku D Kohtz. Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat. Neurosci.*, 12(8):1020–1027, jul 2009.

[30] Wenbo Li, Dimple Notani, Qi Ma, Bogdan Tanasa, Esperanza Nunez, Aaron Yun Chen, Daria Merkurjev, Jie Zhang, Kenneth Ohgi, Xiaoyuan Song, Soohwan Oh, Hong-Sook Kim, Christopher K Glass, and Michael G Rosenfeld. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–520, jun 2013.

[31] Carlos A Melo, Jarno Drost, Patrick J Wijchers, Harmen van de Werken, Elzo de Wit, Joachim A F Oude Vrielink, Ran Elkon, Sónia A Melo, Nicolas Léveillé, Raghu Kalluri, Wouter de Laat, and Reuven Agami. eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol. Cell*, 49(3):524–535, feb 2013.

[32] Erik Splinter, Elzo de Wit, Elphège P. Nora, Petra Klous, H. J. G. van de Werken, Yun Zhu, L. J. T. Kaaij, W. van IJcken, Joost Gribnau, Edith Heard, and Wouter de Laat. The

inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev.*, 25(13):1371–1383, jul 2011.

[33] Ezgi Hacisuleyman, Loyal A Goff, Cole Trapnell, Adam Williams, Jorge Henao-Mejia, Lei Sun, Patrick McClanahan, David G Hendrickson, Martin Sauvageau, David R Kelley, Michael Morse, Jesse Engreitz, Eric S Lander, Mitch Guttman, Harvey F Lodish, Richard Flavell, Arjun Raj, and John L Rinn. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.*, 21(2):198–206, feb 2014.

[34] Emily M. Darrow, Miriam H. Huntley, Olga Dudchenko, Elena K. Stamenova, Neva C. Durand, Zhuo Sun, Su-Chen Huang, Adrian L. Sanborn, Ido Machol, Muhammad Shamim, Andrew P. Seberg, Eric S. Lander, Brian P. Chadwick, and Erez Lieberman Aiden. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U. S. A.*, 113(31):4504–4512, aug 2016.

[35] Rasim A. Barutcu, Philipp G. Maass, Jordan P. Lewandowski, Catherine L. Weiner, and John L. Rinn. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nat. Commun.*, 9(1):1444, dec 2018.

[36] Jeremy E Wilusz. Long noncoding RNAs: Re-writing dogmas of RNA processing and stability. *Biochim. Biophys. Acta*, jun 2015.

[37] Tetsuro Hirose, Giorgio Virnicchi, Akie Tanigawa, Takao Naganuma, Ruohan Li, Hiroshi Kimura, Takahide Yokoi, Shinichi Nakagawa, Marianne Bénard, Archa H. Fox, and Gérard Pierron. NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol. Biol. Cell*, 25(1):169–83, jan 2014.

[38] Ling-Ling Chen, Joshua N. DeCerbo, and Gordon G. Carmichael. Alu element-mediated gene silencing. *EMBO J.*, 27(12):1694–1705, jun 2008.

[39] Jesse M Engreitz, Amy Pandya-Jones, Patrick McDonel, Alexander Shishkin, Klara Sirokman, Christine Surka, Sabah Kadri, Jeffrey Xing, Alon Goren, Eric S Lander, Kathrin Plath, and Mitchell Guttman. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*, 341(6147):1237973, aug 2013.

[40] Kevin C. Wang, Yul W. Yang, Bo Liu, Amartya Sanyal, Ryan Corces-Zimmerman, Yong Chen, Bryan R. Lajoie, Angeline Protacio, Ryan A. Flynn, Rajnish A. Gupta, Joanna Wysocka, Ming Lei, Job Dekker, Jill A. Helms, and Howard Y. Chang. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, 472(7341):120–124, apr 2011.

[41] Philipp G Maass, Andreas Rump, Herbert Schulz, Sigmar Stricker, Lisanne Schulze, Konrad Platzer, Atakan Aydin, Sigrid Tinschert, Mary B. Goldring, Friedrich C Luft, and Sylvia Bähring. A misplaced lncRNA causes brachydactyly in humans. *J. Clin. Invest.*, 122(11):3990–4002, nov 2012.

[42] Marcela M L Soruco, Jessica Chery, Eric P Bishop, Trevor Siggers, Michael Y Tolstorukov, Alexander R Leydon, Arthur U Sugden, Karen Goebel, Jessica Feng, Peng Xia, Anastasia Vedenko, Martha L Bulyk, Peter J Park, and Erica Larschan. The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. *Genes Dev.*, 27(14):1551–1556, jul 2013.

[43] Jennifer A. Urban, Caroline A. Doherty, William T. Jordan, Jacob E. Bliss, Jessica Feng, Marcela M. Soruco, Leila E. Rieder, Maria A. Tsiarli, and Erica N. Larschan. The essential Drosophila CLAMP protein differentially regulates non-coding roX RNAs in male and females. *Chromosome Res.*, 25(2):101–113, jun 2017.

[44] Otto G. Berg, Robert B. Winter, and P H von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20(24):6929–6948, nov 1981.

[45] Anahita Tafvizi, Fang Huang, Alan R Fersht, Leonid A Mirny, and Antoine M van Oijen. A single-molecule characterization of p53 search on DNA. *Proc. Natl. Acad. Sci. U. S. A.*, 108(2):563–568, jan 2011.

[46] Petter Hammar, Prune Leroy, Anel Mahmutovic, Erik G Marklund, Otto G Berg, and Johan Elf. The lac repressor displays facilitated diffusion in living cells. *Science*, 336(6088):1595–1598, jun 2012.

[47] Jiji Chen, Zhengjian Zhang, Li Li, Bi Chang Chen, Andrey Revyakin, Bassam Hajj, Wesley Legant, Maxime Dahan, Timothée Lionnet, Eric Betzig, Robert Tjian, and Zhe Liu. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell*, 156(6):1274–1285, mar 2014.

[48] Igor Martianov, Aroul Ramadass, Ana Serra Barros, Natalie Chow, and Alexandre Akoulitchev. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature*, 445(7128):666–670, feb 2007.

[49] Ruiping Zheng, Zhen Shen, Vidisha Tripathi, Zhenyu Xuan, Susan M Freier, C Frank Bennett, Supriya G Prasanth, and Kannanganattu V Prasanth. Polypurine-repeat-containing RNAs: a novel class of long non-coding RNA in mammalian cells. *J. Cell Sci.*, 123(Pt 21):3734–3744, nov 2010.

[50] Phillip Grote and Bernhard G Herrmann. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol.*, 10(10):1579–1585, oct 2013.

[51] Anna Postepska-Igielska, Alena Giwojna, Lital Gasri-Plotnitsky, Nina Schmitt, Annabelle Dold, Doron Ginsberg, and Ingrid Grummt. LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. *Mol. Cell*, 60(4):626–636, nov 2015.

[52] Tanmoy Mondal, Santhilal Subhash, Roshan Vaid, Stefan Enroth, Sireesha Uday, Björn Reinius, Sanhita Mitra, Arif Mohammed, Alva Rani James, Emily Hoberg, Aristidis Moustakas, Ulf Gyllensten, Steven J.M. Jones, Claes M Gustafsson, Andrew H Sims, Fredrik Westerlund, Eduardo Gorab, and Chandrasekhar Kanduri. MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. *Nat. Commun.*, 6(1):7743, dec 2015.

[53] Valerie B. O'Leary, Saak V. Ovsepian, Laura G. Carrascosa, Fabian A. Buske, Vanja Radulovic, Maximilian Niyazi, Simone Moertl, Matt Trau, Michael J. Atkinson, and Nataša Anastasov. PARTICLE, a Triplex-Forming Long ncRNA, Regulates Locus-Specific Methylation in Response to Low-Dose Irradiation. *Cell Rep.*, 11(3):474–485, apr 2015.

[54] Marie Kalwa, Sonja Hänzelmann, Sabrina Otto, Chao-Chung Kuo, Julia Franzen, Sylvia Joussen, Eduardo Fernandez-Rebollo, Björn Rath, Carmen Koch, Andrea Hofmann, Shih-Han Lee, Andrew E. Teschendorff, Bernd Denecke, Qiong Lin, Martin Widschwendter, Elmar Weinhold, Ivan G. Costa, and Wolfgang Wagner. The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Res.*, 44(22):10631–10643, dec 2016.

[55] Kaarst Hoogsteen. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.*, 16(9):907–916, sep 1963.

[56] Mahima Kaushik, Shikha Kaushik, Kapil Roy, Anju Singh, Swati Mahendru, Mohan Kumar, Swati Chaudhary, Saami Ahmed, and Shrikant Kukreti. A bouquet of DNA structures: Emerging diversity. *Biochem. Biophys. reports*, 5:388–395, mar 2016.

[57] Rani V Parvathy, Sukesh R Bhaumik, Kandala V Chary, Girjesh Govil, Keliang Liu, Frank B Howard, and H Todd Miles. NMR structure of a parallel-stranded DNA duplex at atomic resolution. *Nucleic Acids Res.*, 30(7):1500–1511, apr 2002.

[58] Cecilia Noguez and Francisco Hidalgo. Ab Initio Electronic Circular Dichroism of Fullerenes, Single-Walled Carbon Nanotubes, and Ligand-Protected Metal Nanoparticles. *Chirality*, 26(9):553–562, sep 2014.

[59] Mahima Kaushik, Ritushree Kukreti, Deepak Grover, Samir K Brahmachari, and Shrikant Kukreti. Hairpin-duplex equilibrium reflected in the A–>B transition in an undecamer quasi-palindrome present in the locus control region of the human beta-globin gene cluster. *Nucleic Acids Res.*, 31(23):6904–6915, dec 2003.

[60] Shrikant Kukreti, Harpreet Kaur, Mahima Kaushik, Aparna Bansal, Sarika Saxena, Shikha Kaushik, and Ritushree Kukreti. Structural polymorphism at LCR and its role in beta-globin gene regulation. *Biochimie*, 92(9):1199–1206, sep 2010.

[61] Václav Brázda, Rob C. Laister, Eva B. Jagelská, and Cheryl Arrowsmith. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.*, 12(1):33, aug 2011.

[62] Natalie Saini, Yu Zhang, Karen Usdin, and Kirill S. Lobachev. When secondary comes first–the importance of non-canonical DNA structures. *Biochimie*, 95(2):117–123, feb 2013.

[63] Gary N. Parkinson, Michael P. H. Lee, and Stephen Neidle. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, 417(6891):876–880, jun 2002.

[64] Hisae Tateishi-Karimata, Noburu Isono, and Naoki Sugimoto. New Insights into Transcription Fidelity: Thermal Stability of Non-Canonical Structures in Template DNA Regulates Transcriptional Arrest, Pause, and Slippage. *PLoS One*, 9(3):e90580, mar 2014.

[65] Hong-Xin Jiang, Yunxi Cui, Ting Zhao, Hai-Wei Fu, Deepak Koirala, Jibin Abraham Punnoose, De-Ming Kong, and Hanbin Mao. Divalent cations and molecular crowding buffers stabilize G-triplex at physiologically relevant temperatures. *Sci. Rep.*, 5(1):9255, mar 2015.

[66] Amanda C. Hall, Lauren A. Ostrowski, Violena Pietrobon, and Karim Mekhail. Repetitive DNA loci and their modulation by the non-canonical nucleic acid structures R-loops and G-quadruplexes. *Nucleus*, 8(2):162–181, mar 2017.

[67] Jean-Louis Leroy, Maurice Guéron, Jean-Louis Mergny, and Claude Hélène. Intramolecular folding of a fragment of the cytosine-rich strand of telomeric DNA into an i-motif. *Nucleic Acids Res.*, 22(9):1600–1606, may 1994.

[68] Anh Tuân Phan, Maurice Guéron, and J L Leroy. The solution structure and internal motions of a fragment of the cytidine-rich strand of the human telomere. *J. Mol. Biol.*, 299(1):123–144, may 2000.

[69] Tracy A. Brooks, Samantha Kendrick, and Laurence Hurley. Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS J.*, 277(17):3459–3469, sep 2010.

[70] Miguel Garavís, Núria Escaja, Valérie Gabelica, Alfredo Villasante, and Carlos González. Centromeric Alpha-Satellite DNA Adopts Dimeric i-Motif Structures Capped by AT Hoogsteen Base Pairs. *Chemistry*, 21(27):9816–9824, jun 2015.

[71] Nicole A. Becker and James L. Maher. Characterization of a polypurine/polypyrimidine sequence upstream of the mouse metallothionein-I gene. *Nucleic Acids Res.*, 26(8):1951–1958, apr 1998.

[72] Yoshinori Kohwi and Yurij Panchenko. Transcription-dependent recombination induced by triple-helix formation. *Genes Dev.*, 7(9):1766–1778, sep 1993.

[73] Samir K. Brahmachari, Partha S. Sarkar, Sowmya Raghavan, Malathy Narayan, and Amit K. Maiti. Polypurine/polypyrimidine sequences as cis-acting transcriptional regulators. *Gene*, 190(1):17–26, apr 1997.

[74] Sathees C. Raghavan, Paul Chastain, Jeremy S. Lee, Balachandra G. Hegde, Sabrina Houston, Ralf Langen, Chih-Lin Hsieh, Ian S. Haworth, and Michael R. Lieber. Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation. *J. Biol. Chem.*, 280(24):22749–22760, jun 2005.

[75] Julian L. Huppert. Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes. *Chem. Soc. Rev.*, 37(7):1375, jul 2008.

[76] Christopher E. Pearson, Haralabos Zorbas, Gerald B. Price, and Maria Zannis-Hadjopoulos. Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J. Cell. Biochem.*, 63(1):1–22, oct 1996.

[77] Junhua Zhao, Albino Bacolla, Guliang Wang, and Karen M. Vasquez. Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.*, 67(1):43–62, jan 2010.

[78] Guliang Wang, Steve Carbajal, Jan Vijg, John DiGiovanni, and Karen M. Vasquez. DNA structure-induced genomic instability in vivo. *J. Natl. Cancer Inst.*, 100(24):1815–1817, dec 2008.

[79] Vicki S Chambers, Giovanni Marsico, Jonathan M Boutell, Marco Di Antonio, Geoffrey P Smith, and Shankar Balasubramanian. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, 33(8):877–881, aug 2015.

[80] Mahdi Zeraati, David B Langley, Peter Schofield, Aaron L Moye, Romain Rouet, William E Hughes, Tracy M Bryan, Marcel E Dinger, and Daniel Christ. I-motif DNA structures are formed in the nuclei of human cells. *Nat. Chem.*, 10(6):631–637, jun 2018.

[81] P. Karlovsky, P. Pecinka, M. Vojtiskova, E. Makaturova, and E. Palecek. Protonated triplex DNA in E. coli cells as detected by chemical probing. *FEBS Lett.*, 274(1-2):39–42, 1990.

[82] Y Kohwi, S.R. Malkhosyan, and T. Kohwi-Shigematsu. Intramolecular dG · dG · dC triplex detected in Escherichia coli cells. *J. Mol. Biol.*, 223(4):817–822, feb 1992.

[83] Maria M Krasilnikova and Sergei M Mirkin. Replication Stalling at Friedreich ' s Ataxia (GAA) n Repeats In Vivo. *Mol. Cell. Biol.*, 24(6):2286–2295, 2004.

[84] Gary Felsenfeld, David R. Davies, and Alexander Rich. FORMATION OF A THREE-STRANDED POLYNUCLEOTIDE MOLECULE. *J. Am. Chem. Soc.*, 79(8):2023–2024, apr 1957.

[85] Richard A Morgan and R D Wells. Specificity of the three-stranded complex formation between double-stranded DNA and single-stranded RNA containing repeating nucleotide sequences. *J. Mol. Biol.*, 37(1):63–80, oct 1968.

[86] Anthony G. Letai, Michael A. Palladino, Ethel Fromm, Victor Rizzo, and Jacques R. Fresco. Specificity in formation of triple-stranded nucleic acid helical complexes: studies with agarose-linked polyribonucleotide affinity columns. *Biochemistry*, 27(26):9108–9112, dec 1988.

[87] Giorgio Manzini, Luigi E. Xodo, Daniela Gasparotto, Franco Quadrifoglio, Gijs A. van der Marel, and Jacques H. van Boom. Triple helix formation by oligopurine-oligopyrimidine DNA fragments. *J. Mol. Biol.*, 213(4):833–843, jun 1990.

[88] Naoki Sugimoto, Peng Wu, Hideyuki Hara, and Yasunori Kawamoto. pH and cation effects on the properties of parallel pyrimidine motif DNA triplexes. *Biochemistry*, 40(31):9396–9405, aug 2001.

[89] Jeffery T. Davis. G-Quartets 40 Years Later: From 5'-GMP to Molecular Biology and Supramolecular Chemistry. *Angew. Chemie - Int. Ed.*, 43(6):668–698, jan 2004.

[90] Boris P Belotserkovskii, Erandi De Silva, Silvia Tornaletti, Guliang Wang, Karen M Vasquez, and Philip C Hanawalt. A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J. Biol. Chem.*, 282(44):32433–32441, aug 2007.

[91] P A Beal and P B Dervan. Second structural motif for recognition of DNA by oligonucleotide-directed triple-helix formation. *Science*, 251(4999):1360–1363, mar 1991.

[92] Richard W. Roberts and Donald M. Crothers. Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. *Science*, 258(5087):1463–1466, nov 1992.

[93] H E. Moser and P B. Dervan. Sequence-specific cleavage of double helical DNA by triple helix formation. *Science*, 238(4827):645–650, oct 1987.

[94] Richard W. Roberts and D M. Crothers. Specificity and stringency in DNA triplex formation. *Proc. Natl. Acad. Sci.*, 88(21):9397–9401, nov 1991.

[95] Svetlana Kuznetsova, Silmane Ait-Si-Ali, Irina Nagibneva, Frederic Troalen, Jean-Pierre Le Villain, Annick Harel-Bellan, and Fedor Svinarchuk. Gene activation by triplex-forming oligonucleotide coupled to the activating domain of protein VP16. *Nucleic Acids Res.*, 27(20):3995–4000, oct 1999.

[96] M Faria, C D Wood, Loic Perrouault, J S Nelson, A Winter, M White, Claude Helene, and Carine Giovannangeli. Targeted inhibition of transcription elongation in cells mediated by triplex-forming oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.*, 97(8):3862–3667, apr 2000.

[97] Gan Wang, David D Levy, Michael M Seidman, and Peter M Glazer. Targeted mutagenesis in mammalian cells mediated by intracellular triple helix formation. *Mol. Cell. Biol.*, 15(3):1759–1768, mar 1995.

[98] Pamela Havre, David J Gunther, F Gasparro, and Peter M Glazer. Targeted mutagenesis of DNA using triple helix-forming oligonucleotides linked to psoralen. *Proc. Natl. Acad. Sci. U. S. A.*, 90(16):7879–7883, aug 1993.

[99] Karen Vasquez, Gan Wang, Pamela Havre, and Peter M Glazer. Chromosomal mutations induced by triplex-forming oligonucleotides in mammalian cells. *Nucleic Acids Res.*, 27(4):1176–1181, feb 1999.

[100] Gan Wang, Michael M. Seidman, and Peter M. Glazer. Mutagenesis in mammalian cells induced by triple helix formation and transcription-coupled repair. *Science*, 271(5250):802–805, feb 1996.

[101] Karen Vasquez, Latha Narayanan, and Peter M Glazer. Specific mutations induced by triplex-forming oligonucleotides in mice. *Science*, 290(5491):530–533, oct 2000.

[102] Amer F. Saleh, Mick D. Fellows, Liming Ying, Nigel J. Gooderham, and Catherine C. Priestley. The Lack of Mutagenic Potential of a Guanine-Rich Triplex Forming Oligonucleotide in Physiological Conditions. *Toxicol. Sci.*, 155(1):101–111, 2017.

[103] Sabrina Buchini and Christian J. Leumann. Recent improvements in antigene technology. *Curr. Opin. Chem. Biol.*, 7(6):717–726, dec 2003.

[104] Miho Shimizu, Aritomo Konishi, Yasuaki Shimada, Hideo Inoue, and Eiko Ohtsuka. Oligo(2'- O -methyl)ribonucleotides Effective probes for duplex DNA. *FEBS Lett.*, 302(2):155–158, may 1992.

[105] Erika Brunet, Patrizia Alberti, Loïc Perrouault, Kavindra Babu, Jesper Wengel, and Carine Giovannangeli. Exploring cellular activity of locked nucleic acid-modified triplex-forming oligonucleotides and defining its molecular basis. *J. Biol. Chem.*, 280(20):20076–20085, 2005.

[106] Robert Besch, Christoph Marschall, Theda Schuh, Carine Giovannangeli, Claudia Kammerbauer, and Klaus Degitz. Triple helix-mediated inhibition of gene expression is increased by PUVA. *J. Invest. Dermatol.*, 122(5):1114–1120, may 2004.

[107] Dennis H. Oh and Philip C. Hanawalt. Triple helix-forming oligonucleotides target psoralen adducts to specific chromosomal sequences in human cells. *Nucleic Acids Res.*, 27(24):4734–4742, dec 1999.

[108] Yehenew M. Agazie, Gary D. Burkholder, and Jeremy S. Lee. Triplex DNA in the nucleus: direct binding of triplex-specific antibodies and their effect on transcription, replication and cell growth. *Biochem. J.*, 316:461–466, jun 1996.

[109] Irit Lubitz, Dragoslav Zikich, and Alexander Kotlyar. Specific High-Affinity Binding of Thiazole Orange to Triplex and G-Quadruplex DNA. *Biochemistry*, 49(17):3567–3574, may 2010.

[110] Aklank Jain, Sahay Akanchha, and Moganty R. Rajeswari. Stabilization of purine motif DNA triplex by a tetrapeptide from the binding domain of HMGBI protein. *Biochimie*, 87(8):781–90, aug 2005.

[111] Aklank Jain, Albino Bacolla, Prasun Chakraborty, Frank Grosse, and Karen M. Vasquez. Human DHX9 Helicase Unwinds Triple-Helical DNA Structures. *Biochemistry*, 49(33):6992–6999, aug 2010.

[112] Jason E. Rao and Nancy L. Craig. Selective recognition of pyrimidine motif triplexes by a protein encoded by the bacterial transposon Tn7. *J. Mol. Biol.*, 307(5):1161–1170, apr 2001.

[113] Marco Musso, Giovanna Bianchi-Scarrà, and Michael W. Van Dyke. The yeast CDP1 gene encodes a triple-helical DNA-binding protein. *Nucleic Acids Res.*, 28(21):4090–4096, nov 2000.

[114] Ramon J. Goñi, Xavier de la Cruz, and Modesto Orozco. Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res.*, 32(1):354–360, jan 2004.

[115] Sha He, Hai Zhang, Haihua Liu, and Hao Zhu. LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. *Bioinformatics*, 31(2):178–186, jan 2015.

[116] Claude Pasquier, Sandra Agnel, and Alain Robichon. The Mapping of Predicted Triplex DNA:RNA in the Drosophila Genome Reveals a Prominent Location in Development- and Morphogenesis-Related Genes. *Genes|Genomes|Genetics*, 7(7):2295–2304, jul 2017.

[117] Fabian A. Buske, Denis C. Bauer, John S. Mattick, and Timothy L. Bailey. Triplexator: Detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res.*, 22(7):1372–1381, jul 2012.

[118] Sonja Hanzelmann, Chao-Chung Kuo, Marie Kalwa, Wolfgang Wagner, and Ivan G. Costa. Triplex Domain Finder: Detection of Triple Helix Binding Domains in Long Non-Coding RNAs. *bioRxiv*, page 020297, 2015.

[119] Piroon Jenjaroenpun, Chee Siang Chew, Tai Pang Yong, Kiattawee Choowongkomon, Wimada Thammasorn, and Vladimir a Kuznetsov. The TTSMI database: a catalog of triplex target DNA sites associated with genes and regulatory elements in the human genome. *Nucleic Acids Res.*, 43(D1):D110–D116, jan 2015.

[120] Christine Mayer, Kerstin-Maike Schmitz, Junwei Li, Ingrid Grummt, and Raffaella Santoro. Intergenic transcripts regulate the epigenetic state of rRNA genes. *Mol. Cell*, 22(3):351–61, may 2006.

[121] Jesse M. Engreitz, Noah Ollikainen, and Mitchell Guttman. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat. Rev. Mol. Cell Biol.*, 17(12):756–770, dec 2016.

[122] Ryan A. Flynn and Howard Y. Chang. Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell*, 14(6):752–61, jun 2014.

[123] Evgeniy S. Belousov, Irina A. Afonina, Mikhail A. Podyminogin, Howard B Gamper, Michael W. Reed, Robert M. Wydro, and Rich B. Meyer. Sequence-specific targeting and covalent modification of human genomic DNA. *Nucleic Acids Res.*, 25(17):3440–3444, sep 1997.

[124] Carine Giovannangeli, Silvia Diviacco, Valerie Labrousse, Sergei Gryaznov, Pierre Charneau, and Claude Helene. Accessibility of nuclear DNA to triplex-forming oligonucleotides: The integrated HIV-1 provirus as a target. *Proc. Natl. Acad. Sci.*, 94(1):79–84, jan 1997.

[125] Philip M. Brown and Keith R. Fox. Nucleosome core particles inhibit DNA triple helix formation. *Biochem. J.*, 319(17):607–611, oct 1996.

[126] Erika Brunet, Maddalena Corgnali, Fabio Cannata, Loïc Perrouault, and Carine Giovannangeli. Targeting chromosomal sites with locked nucleic acid-modified triplex-forming oligonucleotides: study of efficiency dependence on DNA nuclear environment. *Nucleic Acids Res.*, 34(16):4546–4553, sep 2006.

[127] Oliver Bell, Vijay K. Tiwari, Nicolas H. Thomä, and Dirk Schübeler. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.*, 12(8):554–564, aug 2011.

[128] Mohit Chawla, Purshotam Sharma, Sukanya Halder, Dhananjay Bhattacharyya, and Abhijit Mitra. Protonation of Base Pairs in RNA: Context Analysis and Quantum Chemical Investigations of Their Geometries and Stabilities. *J. Phys. Chem. B*, 115(6):1469–1484, feb 2011.

[129] Antarip Halder, Sukanya Halder, Dhananjay Bhattacharyya, and Abhijit Mitra. Feasibility of occurrence of different types of protonated base pairs in RNA: a quantum chemical study. *Phys. Chem. Chem. Phys.*, 16(34):18383–18396, sep 2014.

[130] Shu-ichi Nakano, Daisuke Miyoshi, and Naoki Sugimoto. Effects of Molecular Crowding on the Structures, Interactions, and Functions of Nucleic Acids. *Chem. Rev.*, 114(5):2733–2758, mar 2014.

[131] Ernesto Picardi, Anna Maria D'Erchia, Angela Gallo, Antonio Montalvo, and Graziano Pesole. Uncovering RNA Editing Sites in Long Non-Coding RNAs. *Front. Bioeng. Biotechnol.*, 2(December):64, dec 2014.

[132] Jing Gong, Chunjie Liu, Wei Liu, Yu Xiang, Lixia Diao, An-Yuan Guo, and Leng Han. LNCediting: a database for functional effects of RNA editing in lncRNAs. *Nucleic Acids Res.*, 45(D1):D79–D84, jan 2017.

[133] Matthias Schaefer, Tim Pollex, Katharina Hanna, and Frank Lyko. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res.*, 37(2):e12–e12, nov 2008.

[134] Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic RNA Modifications in Gene Expression Regulation. *Cell*, 169(7):1187–1200, jun 2017.

[135] Rodrigo Maldonado, Michael Filarsky, Ingrid Grummt, and Gernot Längst. Purine- and pyrimidine-triple-helix-forming oligonucleotides recognize qualitatively different target sites at the ribosomal DNA locus. *RNA*, 24(3):371–380, mar 2018.

[136] Michael B. Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, jan 2000.

[137] Dae-Kyun Ro, Eric M. Paradise, Mario Ouellet, Karl J. Fisher, Karyn L. Newman, John M. Ndungu, Kimberly A. Ho, Rachel A. Eachus, Timothy S. Ham, James Kirby, Michelle C. Y. Chang, Sydnor T. Withers, Yoichiro Shiba, Richmond Sarpong, and Jay D. Keasling. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086):940–943, apr 2006.

[138] Rachel E Haurwitz, Martin Jinek, Blake Wiedenheft, Kaihong Zhou, and Jennifer a Doudna. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, 329(5997):1355–1358, 2010.

[139] Patrick D. Hsu, Eric S. Lander, and Feng Zhang. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, 157(6):1262–1278, 2014.

[140] Jeffry D Sander and J Keith Joung. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, 32(4):347–55, 2014.

[141] Daniel G Gibson, Lei Young, Ray-Yuan Chuang, J Craig Venter, Clyde a Hutchison, Hamilton O Smith, Clyde A Hutchison Iii, and Nature America. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, 6(5):343–345, 2009.

[142] Roee Amit, Hernan G. Garcia, Rob Phillips, and Scott E. Fraser. Building enhancers from the ground up: A synthetic biology approach. *Cell*, 146(1):105–118, jul 2011.

[143] Michal Brunwasser-Meirom, Yaroslav Pollak, Sarah Goldberg, Lior Levy, Orna Atar, and Roee Amit. Using synthetic bacterial enhancers to reveal a looping-based mechanism for quenching-like repression. *Nat. Commun.*, 7:10407, feb 2016.

[144] Lior Levy, Leon Anavy, Oz Solomon, Roni Cohen, Shilo Ohayon, Orna Atar, Sarah Goldberg, Zohar Yakhini, and Roee Amit. Short CT-rich motifs can trigger context-specific silencing of gene expression in bacteria. *bioRxiv*, 2016.

[145] Tali Raveh-Sadka, Michal Levo, Uri Shabi, Boaz Shany, Leeat Keren, Maya Lotan-Pompan, Danny Zeevi, Eilon Sharon, Adina Weinberger, and Eran Segal. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.*, 44(7):743–750, 2012.

[146] Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, 30(6):521–530, 2012.

[147] Michal Levo and Eran Segal. In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.*, 15(7):453–68, 2014.

[148] Simon Ausländer, David Ausländer, Marius Müller, Markus Wieland, and Martin Fussenegger. Programmable single-cell mammalian biocomputers. *Nature*, pages 5–10, 2012.

[149] Shira Weingarten-Gabbay, Shani Elias-Kirma, Ronit Nir, Alexey A Gritsenko, Noam Stern-Ginossar, Zohar Yakhini, Adina Weinberger, and Eran Segal. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science (80-.).*, 351(6270):1–24, 2016.

[150] T J Wu, G Monokian, D F Mark, and C R Wobbe. Transcriptional activation by herpes simplex virus type 1 VP16 in vitro and its inhibition by oligopeptides. *Mol. Cell. Biol.*, 14(5):3484–3493, 1994.

[151] Rui Lu, Ping Yang, Sharmila Padmakumar, and Vikram Misra. The herpesvirus transactivator VP16 mimics a human basic domain leucine zipper protein, luman, in its interaction with HCF. *J. Virol.*, 72(8):6291–6297, 1998.

[152] Albert J. Keung, Caleb J. Bashor, Szilvia Kiriakov, James J. Collins, and Ahmad S. Khalil. Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. *Cell*, 158(1):110–120, 2014.

[153] Jeffrey A. Chao, Yury Patskovsky, Steven C. Almo, and Robert H. Singer. Structural basis for the coevolution of a viral RNA-protein complex. *Nat. Struct. Mol. Biol.*, 15(1):103–105, jan 2008.

[154] Lior Nissim, Samuel D. Perli, Alexandra Fridkin, Pablo Perez-Pinera, and Timothy K. Lu. Multiplexed and Programmable Regulation of Gene Networks with an Integrated RNA and CRISPR/Cas Toolkit in Human Cells. *Mol. Cell*, 54(4):698–710, may 2014.

[155] Chiuan-chian Chiou, Shiau-wen Chen, Ji-dung Luo, and Yu-tzu Chien. Monitoring triplex DNA formation with fluorescence resonance energy transfer between a fluorophore-labeled probe and intercalating dyes. *Anal. Biochem.*, 416(1):1–7, sep 2011.

[156] Sabine Reither and Albert Jeltsch. Specificity of DNA triple helix formation analyzed by a FRET assay. *BMC Biochem.*, 3:27, 2002.

[157] Joseph Sambrook and David W. Russell. Isolation of DNA fragments from polyacrylamide gels by the crush and soak method. *CSH Protoc.*, 2006(1), jun 2006.

[158] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, may 2011.

[159] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, apr 2012.

[160] Mizuki Ohno, Tatsuo Fukagawa, Jeremy S. Lee, and Toshimichi Ikemura. Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. *Chromosoma*, 111(3):201–213, sep 2002.

[161] Adrian L. Sanborn, Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, Kristopher P. Geeting, Andreas Gnirke, Alexandre Melnikov, Doug McKenna, Elena K. Stamenova, Eric S. Lander, and Erez Lieberman Aiden. Chromatin extrusion explains key

features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.*, page 201518552, oct 2015.

[162] Bharat Sridhar, Marcelo Rivas-Astroza, Tri C. Nguyen, Weizhong Chen, Zhangming Yan, Xiaoyi Cao, Lucie Hebert, and Sheng Zhong. Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Curr. Biol.*, 27(4):602–609, feb 2017.

[163] Fedor Svinarchuk, Irina Nagibneva, Dmitry Cherny, Silmane Ait-Si-Ali, Linda L. Pritchard, Phillipe Robin, Claude Malvy, and Annick Harel-Bellan. Recruitment of transcription factors to the target site by triplex-forming oligonucleotides. *Nucleic Acids Res.*, 25(17):3459–64, sep 1997.

[164] Hidetaka Torigoe, Osamu Nakagawa, Takeshi Imanishi, Satoshi Obika, and Kiyomi Sasaki. Chemical modification of triplex-forming oligonucleotide to promote pyrimidine motif triplex formation at physiological pH. *Biochimie*, 94(4):1032–1040, apr 2012.

[165] Mrinal Kanti Ghosh, Anju Katyal, Ramesh Chandra, and Vani Brahmachari. Targeted activation of transcription in vivo through hairpin-triplex forming oligonucleotide in Saccharomyces cerevisiae. *Mol. Cell. Biochem.*, 278(1-2):147–55, oct 2005.

[166] Sami Hocine, Pascal Raymond, Daniel Zenklusen, Jeffrey A. Chao, and Robert H. Singer. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nat. Methods*, 10(2):119–121, feb 2013.

[167] Bin Wu, Jeffrey A. Chao, and Robert H. Singer. Fluorescence fluctuation spectroscopy enables quantitative imaging of single mRNAs in living cells. *Biophys. J.*, 102(12):2936–2944, jun 2012.

[168] Bin Wu, Jiahao Chen, and Robert H. Singer. Background free imaging of single mRNAs in live cells using split fluorescent proteins. *Sci. Rep.*, 4(1):3615, jan 2014.

[169] Joseph N. Zadeh, Conrad D. Steenberg, Justin S. Bois, Brian R. Wolfe, Marshall B. Pierce, Asif R. Khan, Robert M. Dirks, and Niles A. Pierce. NUPACK: Analysis and design of nucleic acid systems. *J. Comput. Chem.*, 32(1):170–173, jan 2011.

[170] Gerardo Medina, K. Juarez, Brenda Valderrama, and G. Soberon-Chavez. Mechanism of Pseudomonas aeruginosa RhlR Transcriptional Regulation of the rhlAB Promoter. *J. Bacteriol.*, 185(20):5976–5983, oct 2003.

[171] Mariette R. Atkinson, Narinporn Pattaramanon, and Alexander J. Ninfa. Governor of the glnAp2 promoter of Escherichia coli. *Mol. Microbiol.*, 46(5):1247–1257, dec 2002.

[172] Yi-Xin Huo, Zhe-Xian Tian, Mathieu Rappas, Jin Wen, Yan-Cheng Chen, Cong-Hui You, Xiaodong Zhang, Martin Buck, Yi-Ping Wang, and Annie Kolb. Protein-induced DNA bending clarifies the architectural organization of the sigma54-dependent glnAp2 promoter. *Mol. Microbiol.*, 59(1):168–180, jan 2006.

[173] Dahlene Fusco, Nathalie Accornero, Brigitte Lavoie, Shailesh M. Shenoy, Jean-Marie Blanchard, Robert H. Singer, and Edouard Bertrand. Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. *Curr. Biol.*, 13(2):161–167, jan 2003.

[174] Yaron Shav-Tal, Xavier Darzacq, Shailesh M. Shenoy, Dahlene Fusco, Susan M. Janicki, David L. Spector, and Robert H. Singer. Dynamics of single mRNPs in nuclei of living cells. *Science*, 304(5678):1797–1800, jun 2004.

[175] David Bikard, Wenyan Jiang, Poulami Samai, Ann Hochschild, Feng Zhang, and Luciano a. Marraffini. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.*, 41(15):7429–7437, 2013.

[176] Albert W Cheng, Haoyi Wang, Hui Yang, Linyu Shi, Yarden Katz, Thorold W Theunissen, Sudharshan Rangarajan, Chikdu S Shivalila, Daniel B Dadon, and Rudolf Jaenisch. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res.*, 23(10):1163–71, 2013.

[177] Fahim Farzadfard, Samuel D. Perli, and Timothy K. Lu. Tunable and multifunctional eukaryotic transcription factors based on CRISPR/Cas. *ACS Synth. Biol.*, 2(10):604–613, 2013.

[178] Luke a Gilbert, Matthew H Larson, Leonardo Morsut, Zairan Liu, A Gloria, Sandra E Torres, Noam Stern-ginossar, Onn Brandman, H Whitehead, Jennifer a Doudna, Wendell a Lim, and S Jonathan. CRISPR-Mediated Modular RNA-Guided Regualtion of Transcription in Eukaryotes. *Cell*, 154(2):442–451, 2013.

[179] A. Debin, C. Laboulais, M. Ouali, C. Malvy, M Le Bret, and F. Svinarchuk. Stability of G,A triple helices. *Nucleic Acids Res.*, 27(13):2699–707, jul 1999.

[180] Pierre Vekhoff, Alexandre Ceccaldi, David Polverari, Jean Pylouster, Claudio Pisano, and Paola B. Arimondo. Triplex Formation on DNA Targets: How To Choose the Oligonucleotide. *Biochemistry*, 47(47):12277–12289, nov 2008.

[181] Limor Leibovich, Inbal Paz, Zohar Yakhini, and Yael Mandel-Gutfreund. DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic Acids Res.*, 41(Web Server issue):174–179, 2013.

[182] Ann-Joy Cheng and Michael W. Van Dyke. Monovalent cation effects on intermolecular purine-purine-pyrimidine triple-helix formation. *Nucleic Acids Res.*, 21(24):5630–5, dec 1993.

[183] Fedor Svinarchuk, Dmitry Cherny, Arnaud Debin, Etienne Delain, and Claude Malvy. A new approach to overcome potassium-mediated inhibition of triplex formation. *Nucleic Acids Res.*, 24(19):3858–65, oct 1996.

[184] Faye A. Rogers, Janice A. Lloyd, and Meetu Kaushik Tiwari. Improved bioactivity of G-rich triplex-forming oligonucleotides containing modified guanine bases. *Artif. DNA PNA XNA*, 5(1):e27792, jan 2014.

[185] Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.*, 2(3):28–36, 1994.

[186] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, apr 2011.

[187] G Wesley Hatfield and Craig J Benham. DNA topology-mediated control of global gene expression in Escherichia coli. *Annu. Rev. Genet.*, 36:175–203, jan 2002.

[188] A C Albert, F Spirito, N Figueroa-Bossi, L Bossi, and A R Rahmouni. Hyper-negative template DNA supercoiling during transcription of the tetracycline-resistance gene in topA mutants is largely constrained in vivo. *Nucleic Acids Res.*, 24(15):3093–9, aug 1996.

[189] Fabian A. Buske, John S. Mattick, and Timothy L. Bailey. Potential in vivo roles of nucleic acid triple-helices. *RNA Biol.*, 8(3):427–439, may 2011.

[190] Samuel H Sternberg, Rachel E Haurwitz, and Jennifer A Doudna. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA*, 18(4):661–672, apr 2012.

[191] Jeffrey a. Chao, Young J. Yoon, and Robert H. Singer. Imaging translation in single cells using fluorescent microscopy. *Cold Spring Harb. Perspect. Biol.*, 4(11):1–12, 2012.

[192] Ido Golding and Edward C Cox. RNA dynamics in live Escherichia coli cells. *Proc. Natl. Acad. Sci. U. S. A.*, 101(31):11310–11315, 2004.

[193] Boris P. Belotserkovskii and Philip C. Hanawalt. Anchoring nascent RNA to the DNA template could interfere with transcription. *Biophys. J.*, 100(3):675–684, feb 2011.

[194] M. Cristina de Almagro, Silvia Coma, Véronique Noé, and Carlos J. Ciudad. Polypurine hairpins directed against the template strand of DNA knock down the expression of mammalian genes. *J. Biol. Chem.*, 284(17):11579–11589, 2009.

[195] Simon P. Chandler and Keith R. Fox. Specificity of Antiparallel DNA Triple Helix Formation â . *Biochemistry*, 35(47):15038–15048, jan 1996.

[196] P.L. Husler and H.H. Klump. Prediction of pH-Dependent Properties of DNA Triple Helices. *Arch. Biochem. Biophys.*, 317(1):46–56, feb 1995.

[197] Laurence Lavelle and Jacques R. Fresco. UV spectroscopic identification and thermodynamic analysis of protonated third strand deoxycytidine residues at neutrality in the triplex d(C(+)-T)6:[d(A-G)6.d(C-T)6]; evidence for a proton switch. *Nucleic Acids Res.*, 23(14):2692–705, jul 1995.

[198] Luigi E. Xodo. Characterization of the DNA triplex formed by d(TGGGTGGGTGGTTGGGTGGG) and a critical R-Y sequence located in the promoter of the murine Ki-ras proto-oncogene. *FEBS Lett.*, 370(1-2):153–157, aug 1995.

[199] Luigi E. Xodo, Marianna Alunni-Fabbroni, and Giorgio Manzini. G-RICH TRIPLEX-FORMING OLIGODEOXYNUCLEOTIDES AS TRANSCRIPTION REPRESSORS. *NUCLEOSIDES &. NUCLEOTIDES. NUCLEOTIDES*, 16(7-9):1695–1698, 1997.

[200] Boris P Belotserkovskii, Richard Liu, Silvia Tornaletti, Maria M Krasilnikova, Sergei M Mirkin, and Philip C Hanawalt. Mechanisms and implications of transcription blockage by guanine-rich DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 107(29):12816–12821, 2010.

[201] Karen M Vasquez and Peter M Glazer. Triplex-forming oligonucleotides : principles and applications. *Q. Rev. Biophys.*, 1(35):89–107, 2002.

[202] M P Knauert and P M Glazer. Triplex forming oligonucleotides: sequence-specific tools for gene targeting. *Hum. Mol. Genet.*, 10(20):2243–2251, 2001.

[203] Wendy M Olivas and L James Maher. Binding of DNA oligonucleotides to sequences in the promoter of the human bcl-2 gene. *Oligonucleotides*, 24(9):1758–1764, 1996.

[204] Woodring E Wright, Valerie M Tesmer, Kenneth E Huffman, Stephen D Levene, and Jerry W Shay. Normal human chromosomes have long G-rich telomeric overhangs at one end. *Genes Dev.*, 11(21):2801–2809, nov 1997.

[205] Pablo Armas, Sofía Nasif, and Nora B. Calcaterra. Cellular nucleic acid binding protein binds G-rich single-stranded nucleic acids and may function as a nucleic acid chaperone. *J. Cell. Biochem.*, 103(3):1013–1036, feb 2008.

[206] Cindy Follonier, Judith Oehler, Raquel Herrador, and Massimo Lopes. Friedreich's ataxia-associated GAA repeats induce replication-fork reversal and unusual molecular junctions. *Nat. Struct. Mol. Biol.*, 20(4):486–94, 2013.

[207] Sanneke Brinkers, Heidelinde R C Dietrich, Frederik H. de Groote, Ian T. Young, and Bernd Rieger. The persistence length of double stranded DNA determined using dark field tethered particle motion. *J. Chem. Phys.*, 130(21):215105, jun 2009.

[208] Jean-Michel Arbona, Sébastien Herbert, Emmanuelle Fabre, and Christophe Zimmer. Inferring the physical properties of yeast chromatin through Bayesian analysis of whole nucleus simulations. *Genome Biol.*, 18(1):81, dec 2017.

[209] Bernard Tinland, Alain Pluen, Jean Sturm, and Gilbert Weill. Persistence Length of Single-Stranded DNA. *Macromolecules*, 30(19):5763–5765, 1997.

[210] Qingjia Chi, Guixue Wang, and Jiahuan Jiang. The persistence length and length per base of single-stranded DNA obtained from fluorescence correlation spectroscopy measurements using mean field theory. *Phys. A Stat. Mech. its Appl.*, 392(5):1072–1079, mar 2013.

[211] Jiangtao Ren, Yuwei Hu, Chun-Hua Lu, Weiwei Guo, Miguel Angel Aleman-Garcia, Francesco Ricci, and Itamar Willner. pH-responsive and switchable triplex-based DNA hydrogels. *Chem. Sci.*, 6(7):4190–4195, jul 2015.

[212] Alon Singer, Srinivas Rapireddy, Danith H. Ly, and Amit Meller. Electronic Barcoding of a Viral Gene at the Single-Molecule Level. *Nano Lett.*, 12(3):1722–1728, mar 2012.

[213] Yubin Li, Xiangmin Miao, and Liansheng Ling. Triplex DNA: A new platform for polymerase chain reaction â based biosensor. *Sci. Rep.*, 5(1):13010, oct 2015.

[214] Andrea Idili, Alexis Vallée-Bélisle, and Francesco Ricci. Programmable pH-triggered DNA nanoswitches. *J. Am. Chem. Soc.*, 136(16):5836–9, apr 2014.

[215] Gabriel K. A. Minero, Jeppe Fock, J.S. McCaskill, and Mikkel F. Hansen. Optomagnetic studies of pH-switchable nanoparticle agglutination via triplex DNA formation. *Proc. Microtas 2016*, 8:1210–1211, 2016.

[216] Arun R. Chandrasekaran and David A. Rusling. Triplex-forming oligonucleotides: a third strand for DNA nanotechnology. *Nucleic Acids Res.*, 46(3):1021–1037, feb 2018.

פענוח יצירת טריפלקסים באמצעות גישות של ביולוגיה סינטטית וריצוף עמוק

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

דוקטור לפילוסופיה

ביאטה קאופמן

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

חשוון תשע"ט חיפה נובמבר 2018

פענוח יצירת טריפלקסים באמצעות גישות של ביולוגיה סינטטית וריצוף עמוק

ביאטה קאופמן