Investigating spatial organization and base modification of mammalian genomic DNA

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Philipp Trnka

geboren am 22.11.1989 in Berlin

Mainz, 2022

Dekan: Prof. Dr. Eckhard Thines

1. Berichterstatter: Prof. Dr. Christof Niehrs

2. Berichterstatter: [redacted]

Tag der mündlichen Prüfung: 04.10.2022

Table of Contents

1. Summary

The spatial organization of the DNA and chemical modifications of individual DNA bases add additional layers of information to the linear nucleotide sequence. In this thesis, I studied both of these layers, by (I) live cell imaging of genomic loci and (II) genome wide mapping of a DNA modification.

(I) Long-range interactions between regulatory genomic regions affect the transcription of associated genes. To study the dynamics of such interactions, live cell imaging of genomic loci would be desirable, but this method constitutes technical challenges due to the required targeting of a signal with high intensity, stability and enrichment over background. I employed catalytically dead Cas9 (dCas9) combined with a signal amplification-tag, called SunTag to establish targeting and imaging of repetitive loci in HeLa and mouse embryonic stem cells. However, imaging of non-repetitive sequences through expression of 36 single guide RNAs (sgRNAs) failed due to sgRNA competition and high background. Moreover, a new amplification tag, the HttTag, was developed to facilitate simultaneous imaging of two loci. The HttTag was functional *in vitro*, however, it did not result in targeted signal amplification in cells. Alternative methods for signal amplification and targeting of non-repetitive regions remain to be established for future live cell imaging studies.

(II) deoxyinosine (dI) is considered as DNA damage in mammalian cells because it is mutagenic. Genomic dI arises from spontaneous deamination of deoxyadenosine (dA) or via integration of dITPs during replication and can produce A>G transition mutations. However, recent reports indicate that enzymatic deamination of dA occurs in R-loops via adenosine deaminases acting on RNA (ADARs). Importantly, the genomic distribution of dI remains to be investigated and hence, I attempted to establish a next-generation sequencing (NGS) approach to map dI genome-wide. I developed dI-Endonuclease V enrichment (dIVe), which employs the inactive dI repair enzyme Endonuclease V (EndoV) to enrich dI-containing sequences. In combination with a custom NGS library preparation, dIVe-sequencing (dIVe-seq) was applied in HEK293T and MEF cells. In both cell types dIVe-seq revealed an enrichment for $(TG)_n/(CA)_n$ simple repeats. dIVe-qPCR confirmed a number of peak sites. However, dI manipulation by either overexpression or knockdown of dI effectors or by *in vitro* removal of dI failed to change the dIVe-qPCR signals, raising the question whether dIVe-seq is truly specific for dI. Mutation analysis revealed increased A>G transitions in a subset of dIVe-seq peaks, supporting the presence of dI in these peak sequences. Although further work will be required to improve the specificity of dIVe-seq, the data raise the possibility that dI is enriched in R-loop forming $(TG)_n/(CA)_n$ simple repeats.

2. Zusammenfassung

Die räumliche Organisation der Desoxyribonukleinsäure (DNA) und chemische Modifikationen einzelner Basen erweitern die lineare Nukleotid Sequenz um zusätzliche Informationen. In dieser Arbeit untersuche ich diese zusätzlichen Informationsebenen mit Hilfe von (I) Bildgebung in lebenden Zellen und (II) genomweiter Sequenzierung einer DNA-Modifikation.

(i) Wechselwirkungen zwischen regulatorischen DNA Sequenzen, können die Transkription assoziierter Gene beeinflussen. Die Dynamik solcher Interaktionen kann durch Bildgebung einzelner Gene in lebenden Zellen untersucht werden. Derartige Verfahren erfordern jedoch ein starkes und stabiles Signal an der Zielsequenz, dass sich signifikant vom Hintergrundsignal abhebt. In dieser Arbeit verwende ich katalytisch inaktives Cas9 (dCas9) in Kombination mit dem SunTag, einem signalverstärkendem Protein-Tag, um repetitive DNA-Sequenzen in humanen und murinen Zellen zu visualisieren. Die Detektion nicht repetitiver Regionen durch die Expression von 36 Leit-RNAs (sgRNAs), war aufgrund des hohen Hintergrundsignals und gegenseitiger Inhibition der sgRNAs nicht erfolgreich. Zusätzlich wurde ein weiterer Protein-Tag entwickelt, der HttTag, um simultane Detektion von zwei Sequenzen zu ermöglichen. Der HttTag war *in vitro* funktionsfähig, führte jedoch im Zellsystem nicht zur Signalverstärkung. Für zukünftige Bildgebungsstudien in lebenden Zellen, müssen alternative Methoden zur Signalverstärkung und zum Markieren von nicht-repetitiven DNA-Abschnitten etabliert werden.

(II) Deoxyinosin (dI) wird wegen seiner mutagenen Eigenschaften als DNA-Schädigung betrachtet. Genomisches dI entsteht durch spontane Desaminierung von Desoxyadenosin oder durch Integration von dI Triphosphat während der Replikation und kann A>G-Transitionen erzeugen. Neuere Forschungen legen jedoch nahe, dass dI in DNA:RNA Hybriden auch enzymatisch, durch Doppelsträngige RNA-spezifische Adenosin-Desaminasen (ADARs) integriert werden kann. Da die genomische Verteilung von dI bisher nicht bekannt ist, habe ich einen Ansatz zur Anreicherung von dI durch inaktiver Endonuklease V entwickelt, der als “dIVe” bezeichnet wird. dIVe in Kombination mit einer dI-kompatiblen Hochdurchsatz-Sequenzierung (dIVe-seq) wurde in HEK293T- und MEF-Zellen angewendet. In beiden Zelltypen zeigte dIVe-seq eine Anreicherung von $(TG)_n/(CA)_n$ Mikrosatelliten. Einige dieser Regionen konnten durch qPCR detektiert werden, jedoch wurden die dIVe-qPCR Signale weder durch Manipulation von dI-Effektoren, noch durch *in vitro*-Entfernung von dI beeinflusst. Dies stellt die Spezifität von dIVe-seq für dI in Frage. Eine Mutationsanalyse ergab erhöhte A>G-Transitionen in einigen der durch dIVe-seq angereicherten Regionen, was auf dI in diesen Sequenzen hinweisen könnte. Obwohl weitere Experimente erforderlich sind, um die Spezifität von dIVe-seq zu verbessern, weisen die Daten darauf hin, dass dI in DNA:RNA Hybriden angereichert sein könnte, die an $(TG)_n/(CA)_n$ reichen Mikrosatelliten gebildet werden.

3. General Introduction

3.1. DNA organization and modifications regulate cellular function

The genetic information encoding the structures and functions of an organism is stored in the linear DNA sequence. Additional layers of information are added by various levels of DNA organization and modification: from the positioning of chromosomes in the nucleus, down to epigenetic modifications of single nucleotides. These layers affect many cellular processes, including gene expression, DNA replication and repair, while aberrant organization and modification of the DNA can cause and accelerate a variety of diseases[1,2].

3.1.1. From chromosome territories to topologically associated domains

Fluorescence *in situ* hybridization (FISH) experiments have shown that chromosomes occupy discrete territories in the nucleus[3]. The relative position of these territories is conserved during cell division, resulting in similar positioning between parent and daughter cells[4,5]. In addition, a radial distribution is observed where gene rich chromosomes are located to the nucleus interior and gene poor chromosomes are located close to the nuclear envelope[6,7] (Figure 3-1A/B). This effect is even observed intra-chromosomally, where gene poor regions of a chromosome are oriented towards the nuclear envelope and gene dense regions can loop out towards the center of the nucleus[8,9]. Nucleoli associated regions are also observed to be gene poor and transcriptionally repressed[10,11]. In contrast, nuclear speckles are associated with active chromatin regions[11,12].

The development of next generation sequencing (NGS) techniques allowed to identify even finer levels of chromatin organization. Liebermann-Aiden and colleagues developed Hi-C, an evolution of chromosome conformation capture (3C), which allows unbiased sequencing of interactions between genomic loci. They observed the spatial segregation of open and closed chromatin into two genome-wide compartments A & B. The A compartments harbor active euchromatin and the B compartments inactive heterochromatin[13] (Figure 3-1B).

Higher resolution datasets revealed smaller organization units, called topologically associated domains (TAD) (Figure 3-1C). These sub-megabase units contain domain wide features like laminar association or epigenetic histone modifications that result in transcriptional co-regulation of a TAD. The local chromatin interaction domains are maintained between different cell lines and in various species[14–16].

TAD boundaries overlap with binding sites of the insulator protein CCCTC-binding factor (CTCF)[16]. CTCF binds to directional sequence motifs that span a domain in a convergent orientation[17]. In addition, CTCF sites are often bound by cohesin[18], a protein complex involved in cohesion of sister chromatids during metaphase and DNA repair.

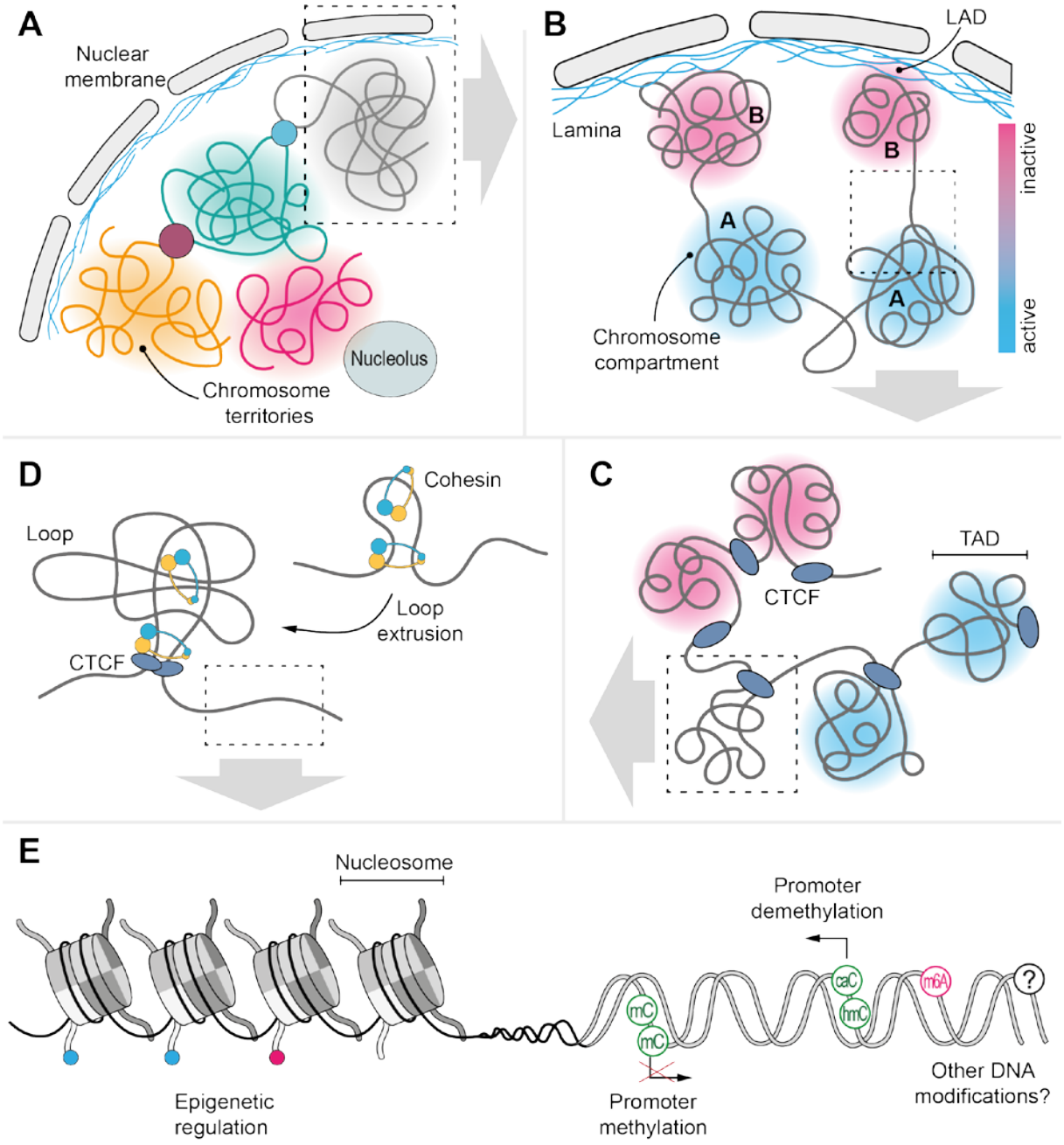


Figure 3-1: Chromatin organization regulates cellular function
A) Chromosomes occupy distinct regions called chromosome territories. Gene-rich chromosomes localize to the nuclear center, while gene poor chromosomes are located at the nuclear periphery or close to nucleoli **B)** Intrachromosomal radial distribution according to transcriptional activity. Inactive regions might localize to the nuclear laminar, forming laminar associated domains (LAD). Chromatin segregates into two genome wide compartments: active euchromatin (A) and inactive heterochromatin (B) **C)** Topologically associated domains (TAD) are self-interacting regions that show transcriptional co-regulation. TAD boundaries overlap with CTCF binding sites. **D)** Cohesin mediated loop extrusion: distant elements are brought into proximity by cohesin enclosing both regions and sliding along the DNA. CTCF sites stop cohesin progression. **E)** The linear DNA is wrapped around histone octamers forming nucleosomes. Histones can be modified by methylation and acetylation affecting transcription, DNA damage repair and replication. Similarly, DNA bases can be modified and affect cellular processes. For example, in DNA cytosine methylation is often associated with silenced promotors, while further oxidized 5mC derivatives 5hmC, 5fC and 5caC are correlated with gene activation. Figure modified from Wang *et al.*[19]

The discovery of cohesin at TAD boarders led to the loop extrusion model. In this model, cohesin is thought to clamp a DNA strand and loop it out, which brings distant genetic elements into proximity[20,21]. The loop extrusion is stopped by the insulating CTCF sites, thereby increasing the probability for these sites to interact (Figure 3-1D). Such a loop mediates the well-studied long distant interaction of the sonic hedgehog gene (*Shh*) with the zone of polarizing activity regulatory sequence. The two regulatory elements have a linear distance of 1 Mb but are brought into close proximity during *Shh* expression[22]. Deletion of the CTCF sites results in changes of the local TADs and abrogates the spatial proximity[23,24]. However, the effects on transcription are mild or not detectable. This is in line with other reports that describe a loss of loop domains upon cohesin removal, but only minor effects on transcription levels[25–27]. Moreover, perturbation of CTCF binding sites and cohesin does not directly affect deeper levels of chromatin organization, like nucleosome-interactions and histone modifications[25,28,29].

3.1.2. Nucleosomes and DNA modifications

Nucleosomes are formed by a histone core complex that is wrapped by a 147 bp long DNA segment. This unit is repeated throughout the genome, condensing the linear DNA strand into chromatin[30]. The histone core is an octamer composed of two H2A-H2B dimers and a H3/H4 tetramer. The positive charges of the outer histone tails facilitate wrapping of the negatively charged DNA helix[31]. Different posttranslational histone modifications are associated with various cellular processes, like transcription, replication and repair (Figure 3-1E). For example, active genes are associated with H3/H4 acetylation and H3K36- or H3K4 trimethylation (me3). H3K9me3 or H3K27me3, often mark repressed genes[32]. γH2A.X, the phosphorylated form of H2A.X, is an important factor in the repair of DNA double-strand breaks (DSB)[33]. Other known histone modifications include ubiquitination, sumoylation, and poly ADP-ribosylation[34,35].

The linear DNA sequence is the final and most fundamental layer of genome organization. On this level, dynamic chemical modifications can modulate cellular functions like gene expression, DNA repair and stability of sequence elements (Figure 3-1E).

One of the best characterized DNA modifications is 5-methylcytosine (5mC), introduced by DNA methyltransferases (DNMT)[36]. 5mC usually mediates transcriptional silencing and is generally low at active promotors[37,38]. However, it is also enriched over gene bodies of highly transcribed genes, to prevent spurious initiation of transcription[39]. Additionally, 5mC is involved in genomic imprinting, X-chromosome inactivation and suppression of retrotransposons[40]. Cytosine methylation is reversible via passive or active demethylation processes. Passive DNA demethylation results from replication without maintenance of the methylation pattern by DNMT1[41]. Active demethylation occurs by subsequent oxidation of 5mC into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) by

ten-eleven-translocation (Tet) enzymes[42,43]. 5fC and 5caC can be excised by base excision repair (BER) initiated by thymine DNA glycosylase (TDG)[44].

5hmC, 5fC and 5caC have all been suggested to be independent regulatory marks and not only byproducts of 5mC removal[45,46]. In neurons 5hmC is enriched over active genes and is specifically bound by methyl-CpG binding protein 2 (MeCP2)[47]. Similarly, 5fC and 5caC show distinct distribution patterns in mouse embryonic stem cells (mESC) and might convey different regulatory roles[48].

Another methylated base in prokaryotes and eukaryotes is N6-methyl-deoxyadenosine (m6dA). However, its origin and biological significance in mammals are controversial[49,50]. Other chemical DNA modifications like deoxyuracil (dU), 8-oxo-guanosine (8oxoG), or deoxyinosine (dI) are classically regarded as DNA damage because of their mutagenic properties[51]. Nevertheless, there is growing evidence that these modifications are not only random damages, but could modulate cellular processes in specific regions of the genome similar to other epigenetic modifications[52,53].

For dU, the deamination product of cytosine, the distribution is non-random[54,55] and was shown to affect the stability of repetitive DNA sequences[56] as well as gene expression by initiating BER[57–59]. Similarly, the distribution of 8oxoG, an oxidative change of guanosine is not random. In mouse cells specific gene loci are enriched for 8oxoG when compared to intergenic regions[60] and in human cells, 8oxoG is enriched in introns as well as transposable and repetitive elements[61]. Lastly, dI, the deamination product of adenosine, affects the stability of telomeric R-loops[62] and recruits the BER machinery to R-loops in micronuclei during pathologic chromosome fragmentation[63]. In line with these findings our lab has detected enrichment of dI in DNA:RNA hybrids, where it could modulate R-loop stability[64]. To this point, there is no genome wide data on the distribution of dI.

3.2. Structure of the thesis

In this thesis, I present two independent projects.

I: Live cell imaging of non-repetitive genomic regions

II: Mapping deoxyinosine in genomic DNA

I will present these projects in two separate parts, each containing an introduction, results and discussion. Materials and methods used during both projects are presented together at the end of the thesis.

Part-I

4. Live cell imaging of non-repetitive genomic loci

4.1. Introduction

4.1.1. Methods for live cell imaging of genomic regions

Genomic regions and nuclear architecture have classically been monitored by FISH or more recently by 3C based sequencing methods[65]. These methods require cell fixation or lysis and therefore only deliver a snapshot view of the current chromatin state. Moreover, 3C is usually performed on a cell population and the resulting data only gives a stochastic representation of possible interactions[65]. Both methods, FISH and 3C, cannot resolve spatio-temporal changes in genome organization on a single cell level. To circumvent the described issues, live cell imaging methods to capture chromatin dynamics have been established.

Imaging of genomic regions in living cells often relies on repetitive sequence elements like centromeres or telomeres[66–68]. By fusing a known binder of these repeats with a fluorophore, specific labeling of the repeat region is possible (e.g. CENPA to image centromeres). Similarly, integration of ectopic repeat arrays is used to image specific loci. By integrating 256 copies of the lac operator into the yeast genome, a lac repressor-GFP fusion protein has been employed for imaging[69]. Likewise, the interaction of the tetracycline operator and repressor have been used for imaging in living cells[70].

These live cell imaging approaches rely on a known binder for a specific repeat type. Moreover, the integration of an ectopic sequence into a genomic locus might perturb the local chromatin organization and gene expression. Therefore, more flexible and less invasive live cell imaging approaches were established using zinc finger proteins (ZFPs)[71] or transcription activator–like effectors (TALE)[72,73]. Both require the presence of repetitive regions, but they can be engineered to bind and target fluorophores to any sequence of interest. Single ZFPs interact with three base pairs and can be combined to recognize a selected target[74]. Similarly, TALE domains interact with a single base pair and can be engineered to bind a sequence of interest[75].

Both methods depend on protein-DNA interactions and require costly protein engineering. Hence, they became less popular when a programmable bacterial nuclease was characterized that allows specific targeting via a simple RNA sequence: CRISPR/Cas[76].

4.1.2. Imaging utilizing the CRISPR/Cas System

CRISPR/Cas targeting is based on genomic regions described as clustered regularly interspaced short palindromic repeats (CRISPR). These repeats were found in bacteria and archaea in 1990[77,78] and were later identified as part of an adaptive immune response in procaryotes[79,80]. The CRISPR locus contains short (semi) palindromic repeats that can have an extrachromosomal origin, e.g. phage DNA. Infected bacteria can gain immunity by storing parts of the infiltrators DNA in the CRISPR locus. Bolotin and colleagues found nuclease motifs in genes associated with the CRISPR locus and described the endonuclease today known as the CRISPR associated protein 9 (Cas9)[81].

It was later demonstrated that CRISPR RNAs (crRNA)[82] and trans-activating crRNA (tracrRNA)[83] are expressed from the CRISPR locus. Together they are targeting the Cas9 protein to exogenous DNA. The labs of Jennifer Doudna and Emmanuelle Charpentier published the underlying mechanism for Cas9 mediated cleavage in 2012. Moreover, they reduced the complexity of the system by creating a single guide RNA (sgRNA) that is sufficient to target Cas9 to a specific DNA sequence[76]. Lastly, they identified the two amino acid mutations D10A/H840A that create catalytically dead Cas9 (dCas9)[76].

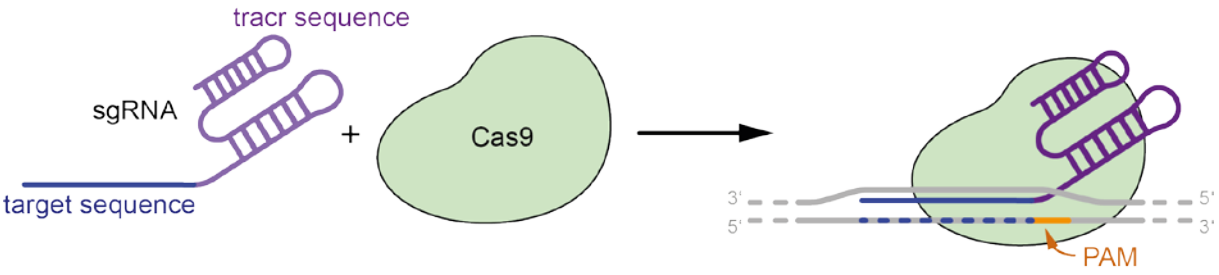


Figure 4-1: sgRNA dependent targeting of Cas9
CRISPR associated protein 9 (Cas9) forms a complex with a single guide RNA (sgRNA). The sgRNA consists of a target-specific sequence and a loop-forming transactivating CRISPR (tracr) sequence that facilitates binding with Cas9. The DNA target sequence requires a 3' protospacer adjacent motif (PAM) for Cas9 binding. PAM and tracr sequence vary among Cas9 proteins from different bacterial species.

For imaging, dCas9 is linked to a fluorophore and co-expressed with a sgRNA targeting the locus of interest. The sgRNA is composed of (i) a tracrRNA sequence forming a loop structure that interacts with the dCas9 protein and (ii) a guide sequence that determines the target of dCas9. The genomic target sequence is around 20 bp long and has to be flanked by a protospacer adjacent motif (PAM) which is necessary for full binding of Cas9 or dCas9[76] (Figure 4-1). Both, the tracrRNA sequence and the required PAM are specific for Cas9 proteins from different bacterial species. For the most commonly used Cas9 from *Streptococcus pyogenes (S.py.)* the motif is 5'-NGG-3' [83]. For *Staphylococcus aureus (S.au.)* the PAM is NNGRR[84], where "N" is any nucleotide and "R" is a purine base.

Besides its use in live cell imaging, the CRISPR/Cas9 system has transformed many procedures in modern research. The Cas9 nuclease is an established tool to generate transgenic cells and organisms[85,86] and multiple studies have used it to correct mutations in cells derived from patients with genetic diseases[87–90]. The inactive dCas9 is also employed for base editing without introducing DSB. Specific point mutations can be established in a genome by targeting the genetically engineered deaminases cytosine base editor or adenosine base editor (ABE) to a locus[91,92]. In addition, the discovery of other Cas proteins, like Cas13, has allowed specific editing and targeting of RNA molecules[93–95].

Although the CRISPR/Cas system simplified the targeting of fluorophores to specific DNA sequences, there are remaining difficulties associated with live cell imaging that require consideration.

4.1.3. Achieving high signal intensity and stability over background

Live cell imaging of a genomic locus with fluorescently labeled dCas9 or other sequence-specific proteins is associated with two main challenges that need to be resolved.

(i) Intensity and photostability of the signal: The fluorescence signal needs to be strong enough to be detected by a microscope. Therefore, imaging of a genomic locus usually requires multiple fluorophores to achieve sufficient brightness. Not only would a single fluorophore be hard to detect, it would also photo bleach due to its low photo stability. This becomes even more problematic when the goal is to image a biological process over an extended period.

(ii) Signal to background ratio: The fluorescence at the target locus needs to be stronger than the background fluorescence. Since there will be free untargeted dCas9-fluorophor complexes in the nucleus, a local enrichment needs to be achieved at the target site.

There have been different approaches to resolve these problems. Signal intensity and stability is improved by recruiting multiple fluorophores. The easiest way to achieve this is by imaging repetitive regions. As described above, these could be endogenous repeats or ectopic repeats integrated at the target locus. With this approach, one fluorophore is recruited per bound dCas9 protein. Other methods recruit multiple fluorophores per targeted dCas9 protein, thereby increasing signal intensity and stability. This can be accomplished by extending the sgRNA with additional stem loops that recruit RNA binding proteins fused to fluorophores (Figure 4-2A). Two examples are the MS2 and PP7 stem-loops that are bound by the MS2 or PP7 coat protein (MCP, PCP), respectively[96,97]. Similarly, the Pumilio/Fem3 mRNA-binding factor (PUF) can be used to recruit multiple fluorophores to specific sequence motifs in an extended sgRNA[98].

An alternative approach is to recruit multiple fluorophores via a polypeptide scaffold fused to the dCas9 protein (Figure 4-2B). Tanenbaum and colleagues developed such a polypeptide tag for signal amplification, called SunTag. It is based on the interaction between a 19 amino acid (aa) long peptide of the general control nonderepressible protein (GCN4) and an antibody single-chain variable fragment (scFv) that specifically recognizes this epitope (scFv-GCN4). The tag contains 24 repeats of the GCN4 peptide and recruits the respective number of fluorophores linked to the scFv-GCN4[99].

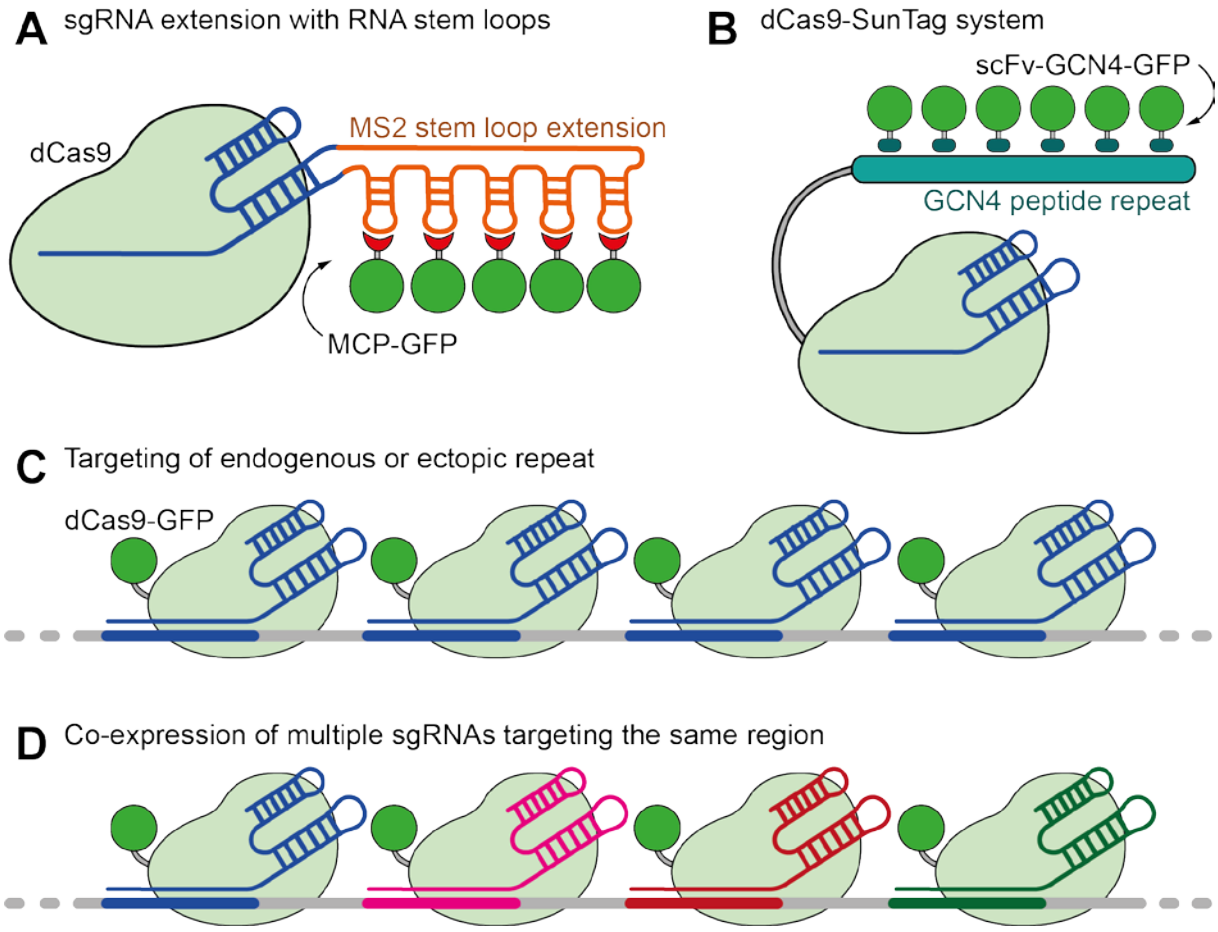


Figure 4-2: Signal amplification in dCas9 imaging
A) sgRNA can be extended by an additional loop forming sequence that recruits RNA binding proteins fused to a fluorophore. Examples for these RNA-protein interactions are the MS2-MCP or the PBS-PUF system. **B)** SunTag fused to dCas9. Up to 24 GCN4 peptide epitopes recruit multiple single chain variable fragment (scFV) antibodies fused to a fluorophore[99]. **C)** Signal amplification by imaging of a repetitive sequence that can be targeted multiple times using an individual sgRNA. The repeat can be endogenous, e.g. telomeres, or an ectopic sequence integrated by gene knock-in. **D)** Expression of multiple sgRNAs targeting the same locus for recruitment of multiple dCas9-fluorophores. This usually requires between 20 – 36 simultaneously expressed sgRNAs[100–102].

These strategies can greatly enhance the signal intensity and stability but they are not sufficient to increase the signal over background. Eventually, a bright and stable signal at the target locus is only detectable if it is enriched over the unbound fluorophore complexes in the nucleus. Hence, also imaging methods employing the CRISPR/Cas system often rely on integrating ectopic, repetitive elements into the target locus[103] (Figure 4-2C).

Other labs have achieved imaging of non-repetitive regions by expressing multiple sgRNAs targeting the same locus (Figure 4-2D). The lab of Joanna Wysocka employed 36 sgRNAs to image different regulatory elements like enhancers or promotors[100]. Others have reported imaging of a non-repetitive region of the *MUC4* gene using 20-26 sgRNAs[101,102].

The expression of a large number of sgRNAs often requires specific cloning techniques, which allow the assembly of multiple sgRNA expression cassettes. For this purpose, the Wysocka lab developed a chimeric array of gRNA oligonucleotides (CARGO) that delivers twelve sgRNAs on one plasmid. This cloning strategy involves two ligation steps and one restriction reaction[100]. An alternative single step cloning strategy for multiplexed sgRNA expression is string assembly gRNA cloning (STAgR). Here, up to eight gRNA cassettes are assembled in a single, restriction-free ligation reaction[104].

Although, the Wysocka lab has reported live cell imaging of non-repetitive regions, they employed a dCas9 directly fused to GFP. The dCas9-GFP approach cannot provide a signal stable enough for extended imaging periods. Instead, in their study the dCas9-GFP was employed to briefly measure an increase in mobility of a cis regulatory element upon transcriptional activation. To monitor a non-repetitive genomic region over an extended period, a system for signal amplification, like the SunTag or extended sgRNA loops, would be required. However, the sgRNA extensions can massively increase the length of the employed sgRNAs. Accordingly, providing 16 MS2 loops for recruitment of additional fluorophores increases the length of a 100 bp sgRNA to nearly 1 kb[105]. Since imaging of a non-repetitive region requires the expression of many sgRNAs, it seems favorable to employ the SunTag for signal amplification instead of sgRNA extensions. However, although the SunTag can amplify the signal of targeted dCas9 for one locus, simultaneous imaging of a second region would require another amplification tag to increase and stabilize the fluorescence signal. Hence, establishing an analogue polypeptide tag could greatly enhance the CRISPR/Cas toolset for imaging of multiple genomic regions.

4.1.4. Aim

Extended live cell imaging of repetitive genomic regions can be achieved by targeting dCas9-SunTag with one sgRNA. In contrast, imaging of non-repetitive loci requires the expression of multiple sgRNAs. The recruitment of multiple fluorophores with the SunTag system increases signal intensity and stability and allows continuous imaging with minor photobleaching.

The aim of this project was to image two independent non-repetitive genomic regions for extended periods, e.g. during cell division or a differentiation process. This would require a second orthogonal dCas9 system for targeting as well as a second signal amplification tag that stabilizes the signal during continuous imaging.

I planned to establish imaging with the dCas9-SunTag system in HeLa cells as well as mouse embryonic stem cells (mESCs) using dCas9 orthologues from *S.py.* and *S.au.* for separate targeting of two loci. To facilitate imaging with both systems in mESCs, I planned to generate mESC lines stably expressing the py.dCas9-SunTag. As imaging of a non-repetitive locus would require the simultaneous expression of 20-36 sgRNAs, I explored two plasmid systems that allow for multiplexed delivery and expression of sgRNAs. Finally, in collaboration with the lab of [redacted], a polypeptide Tag based on the interaction of a short Huntingtin (HTT) peptide and a specific variable domain light chain (V_L) antibody was designed. I tested the performance of this amplification tag during dCas9 imaging in human and mouse cells.

Ultimately, independent targeting of the SunTag and the alternative amplification tag by *S.py.* and *S.au.* dCas9 could facilitate simultaneous live cell imaging of two genomic loci.

4.2. Results

4.2.1. Live cell imaging of repetitive loci using dCas9-SunTag

I employed the dCas9-SunTag system to image genomic regions with high signal intensity and stability. To establish the method I performed live cell imaging of repetitive loci in HeLa cells. The three required components: the dCas9-SunTag, a sgRNA and the scFv-GCN4-fluorophore were expressed from separated plasmids, which I modified from the original vectors published by Tanenbaum and colleagues[99]. Cells were imaged 48 h after transfection using the Opera Phenix screening microscope at 37 °C and 5% CO_2 (Figure 4-3A). Delivering all components on separate plasmids allowed easy testing of different combinations of sgRNAs or fluorophores. Moreover, imaging in a 96-well format allowed simultaneous optimization of many parameters (i.e. cell density, plasmid amount, transfection time).

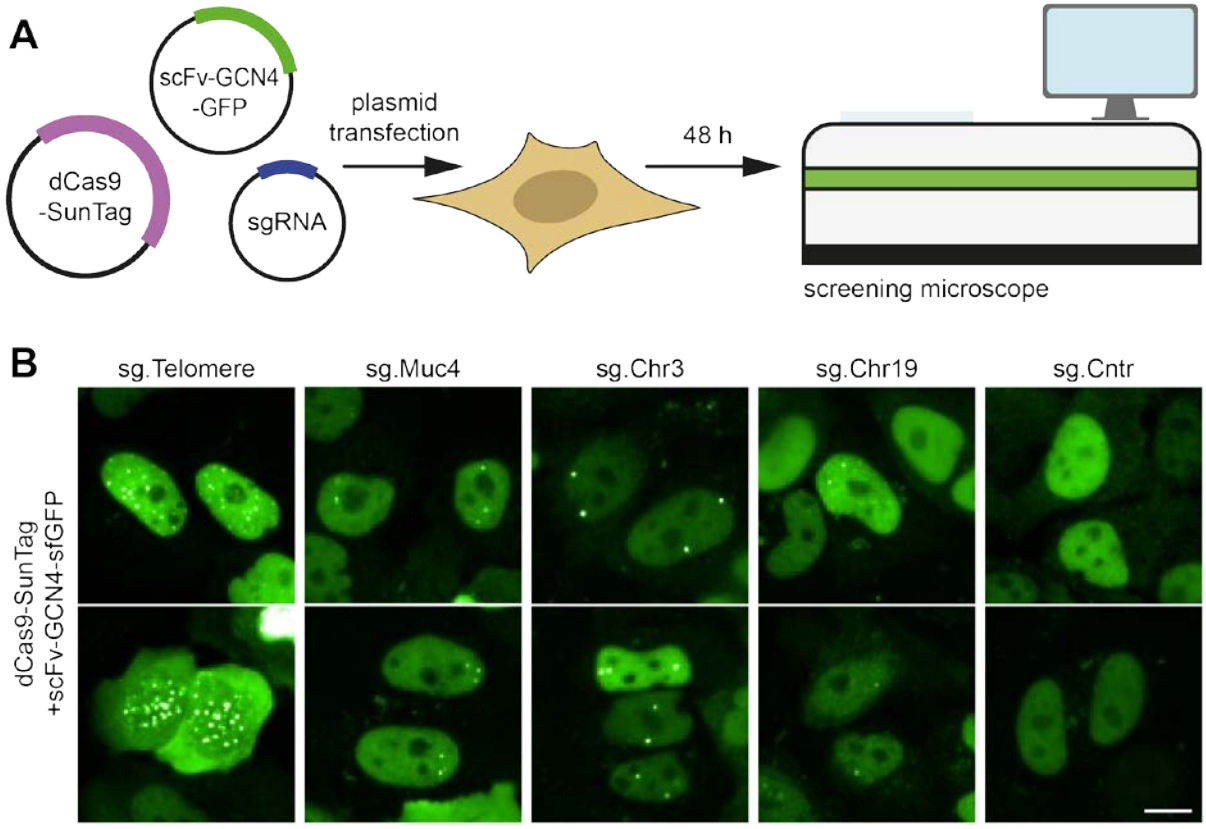


Figure 4-3: Live cell imaging of repetitive regions in human cells
A) Imaging approach using the SunTag system. HeLa cells were transfected with dCas9-SunTag, scFv-GCN4-GFP and sgRNA on separate plasmids in a 96-well plate. After 48 h, cells were imaged at 37 °C, 5% CO_2 with the Opera Phenix microscope. **B)** HeLa cells transfected with SunTag system and indicated sgRNA. Random fields of view were examined for nuclei containing spots with Hoechst33342 staining as reference (not shown). Images show representative foci in the respective sample. Scale bar: 10 µm

As expected, live cell imaging of telomeres was achieved in HeLa cells (Figure 4-3B) and other easily transfectable cells, like U2OS (not shown). Similarly, sgRNAs targeting the highly repetitive *MUC4* locus or subtelomeric repeats on chromosomes 3, 14 and 19 resulted in two to three detectable foci per nucleus. This number of spots is expected, as HeLa cells have heterogeneous, often triploid karyotypes[106] that result in more than two target sequences per nucleus. The detected spots were sgRNA dependent (not shown) and expression of a non-targeting control sgRNA did not produce any spots (Figure 4-3B).

To facilitate simultaneous live cell imaging of two genomic regions, I established an orthologous *S.au.* dCas9 construct for SunTag imaging. Therefore, I introduced point mutations preventing the nicking activity of active *S.au.* Cas9 and fused it to the published SunTag. Moreover, I generated sgRNA plasmids that contain the species-specific scaffold sequence and match the respective PAM after the target sequence. Overexpression of the generated au.dCas9-SunTag system with telomere and *MUC4* targeting sgRNA resulted in the expected spot patterns, respectively (Figure 4-4). The targeting with the *S.au.* system worked as efficient as the established py.dCas9-SunTag. Importantly, spot formation with either *S.au.* or *S.py.* dCas9-SunTag was only observed when the corresponding species-specific sgRNA was transfected (Figure 4-4). This indicates that there is no crosstalk between the different sgRNAs and Cas9 orthologues. Consequently, simultaneous live cell imaging of two loci could be achieved when combining each dCas9 orthologue with separate fluorophores.

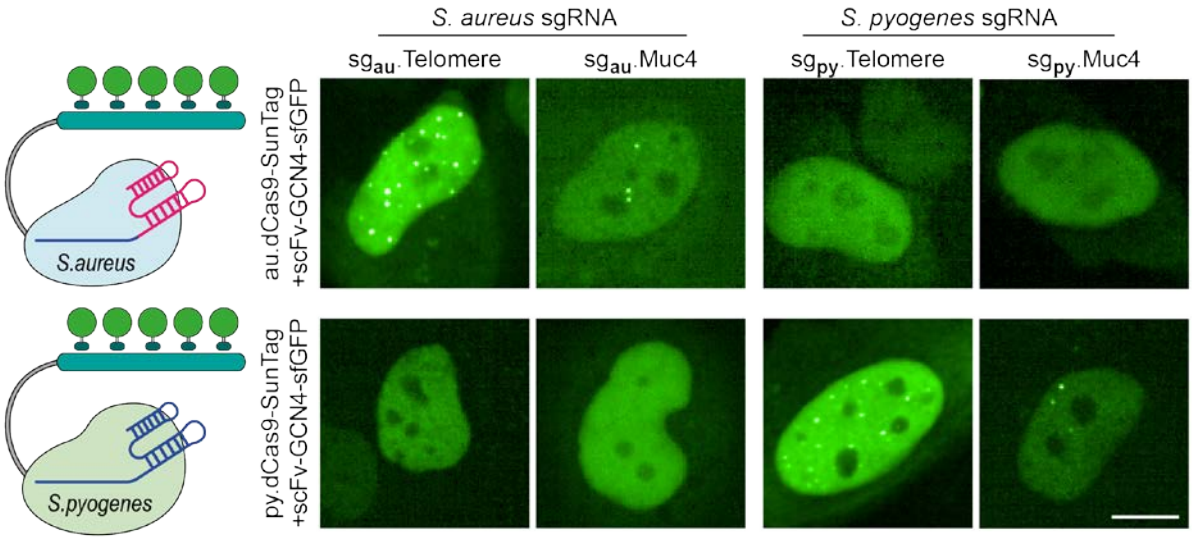


Figure 4-4: *S.au.* and *S.py. sgRNAs* show now cross-reactivity with the opposite dCas9-system
dCas9 fusion proteins from *S.au.* and *S.py.* were independently targeted via specific sgRNAs. HeLa cells were transfected with dCas9-SunTag, scFV-GCN4-GFP and the indicated sgRNA. Random fields of view were examined for nuclei containing spots, with Hoechst33342 staining as reference (not shown). Images show representative foci in the respective sample. Scale bar: 10 µm

4.2.2. Live cell imaging of repetitive loci in mESCs

Before working on an amplification tag analogue to the SunTag, I optimized live cell imaging in mESCs using the *S.py.* dCas9 system. Transfection of all imaging components into mESC was very inefficient, and hence, spots corresponding to genomic regions were rare in these cells. As the final aim was to use two analogous dCas9-Imaging systems simultaneously, I reduced the components required for transfection. Therefore, I generated mESCs expressing the established *S.py.* dCas9-SunTag from the *Rosa26* locus. The coding sequence was combined with a silencing resistant CAG promoter[107] and homology arms matching the *Rosa26* locus to allow targeted integration and stable expression. The knock-in components also encoded BFP and G418 resistance for fluorescence-activated cell sorting (FACS) and antibiotic selection, respectively. The knock-in workflow is shown in (Figure 4-5A). During FACS 600,000 cells were sorted positive for BFP. Subsequently, 300 single mESC colonies were isolated manually of which 98 were showing telomere spots in an imaging screen. I confirmed the correct integration of the dCas9-SunTag into the *Rosa26* locus in 25 cell lines by genotyping PCR (not shown). The following experiments were performed in multiple clones with similar results, however, for conciseness I will only present images of clone mESC-C8.

For imaging in mouse cells, I generated specific sgRNAs targeting major and minor satellite repeats and the *Akap6* gene. The satellite repeats should provide multiple thousand binding sites per chromosome and the repeat in the *Akap6* gene provides 87 binding sites matching guide sequence and PAM[96]. Since the telomere repeat sequence is conserved in humans and mouse, the telomere sgRNA was not changed.

Telomeres and minor satellite repeats showed defined and bright spots, whereas the major satellite repeats were visible as slightly larger patches with less contrast over the nuclear background. The sgRNA targeting the *Akap6* locus did not result in visible spots (Figure 4-5B). Notably, this sgRNA was previously used for imaging of *Akap6* in mouse fibroblast NIH3T3 cells[96].

The simplified transfection of the stable cell lines also rendered overnight imaging of stem cells more feasible. Transfected cells were imaged overnight in 20 min intervals, without loss of the fluorescence signal. In mESCs, telomeres were imaged through mitosis over a course of 6 h (Figure 4-5C).

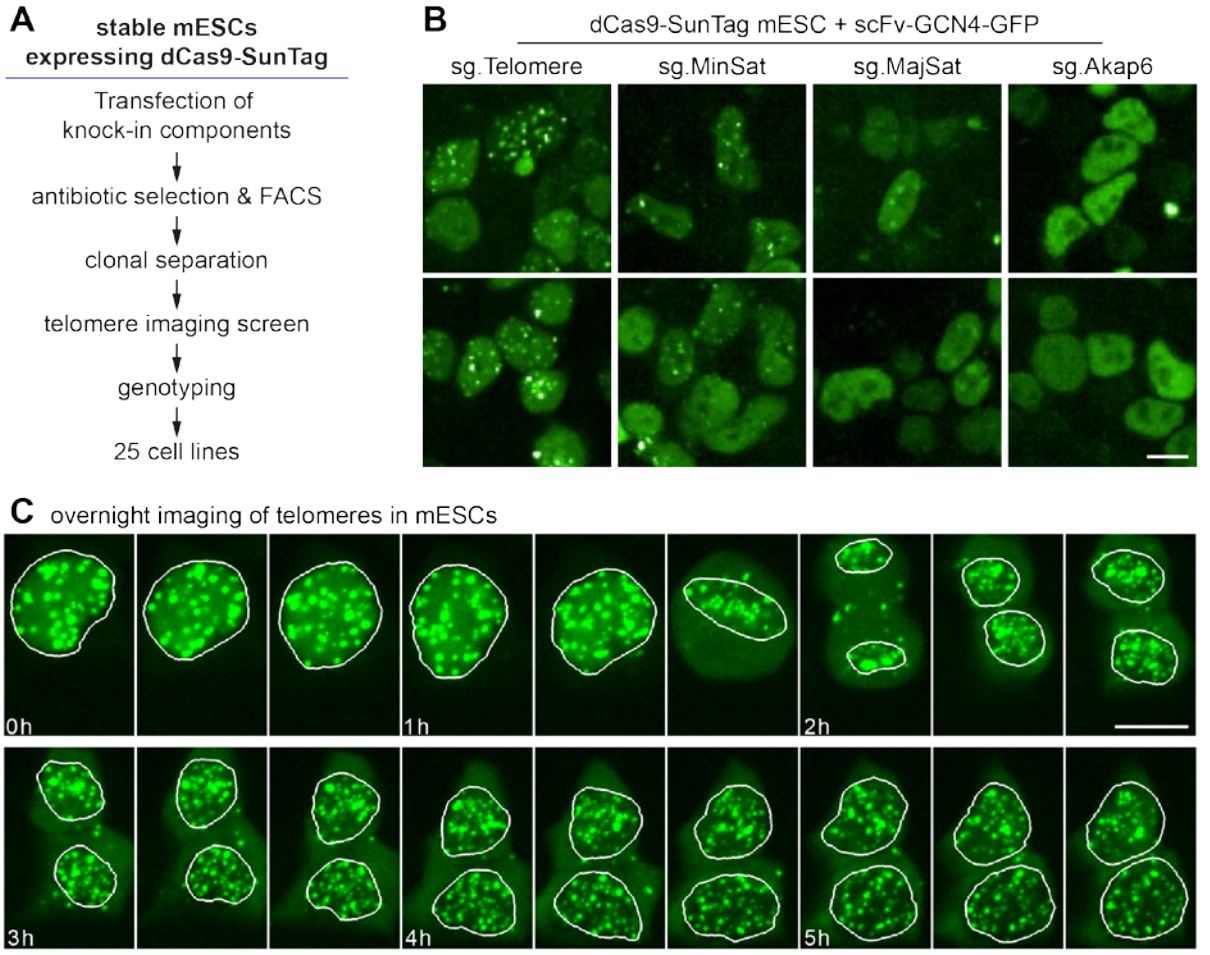


Figure 4-5: Stable dCas9-SunTag mESCs allow imaging of repetitive regions
A) Flowchart showing the knock-in procedure and selection processes to generate mESCs stably expressing the dCas9-SunTag from the *Rosa26* locus. **B)** Live cell imaging of mESC clone C8 after transfection of scFv-GCN4-sfGFP and indicated sgRNA. Telomere and minor satellite repeat sgRNAs resulted in strong foci. Major satellite repeat sgRNA yielded larger, fainter patches. No foci were detected with the *Akap6* sgRNA. Random fields of view were examined for nuclei containing spots, with Hoechst33342 staining as reference (not shown). Images show representative foci in the respective sample. Scale bar: 10 µm **C)** Overnight imaging of mESC-C8 after transfection of scFv-GCN4-sfGFP and telomere sgRNA. Images were taken in 20 min intervals. Nuclei were stained with SiRDNA for reference (not shown) and are outlined with a white line. Scale bar: 10 µm

4.2.3. Expression of multiple gRNAs interferes with spot formation

For imaging of non-repetitive loci multiple sgRNAs are required. The lab of Joanna Wysocka has imaged different promoter and enhancer regions in mESCs using dCas9-GFP and up to 36 gRNAs[100]. These sgRNAs were delivered on multiple CARGO plasmids, each expressing 12 sgRNAs. The Wysocka lab kindly provided the CARGO plasmids targeting *Fgf5* enhancer, *Fgf5* promoter and *Tbx3* promoter for own testing. In addition, I generated similar gRNA-arrays targeting the same three regions using STAgR cloning[104]. However, neither the CARGO plasmids from the Wysocka lab, nor the cloned STAgR constructs resulted in detectable spots at the *Fgf5* enhancer (Figure 4-6A) or the promoters (data not shown). After imaging, the cells

were lysed and expression of a random subset of *Fgf5* enhancer sgRNAs was monitored by reverse transcription quantitative PCR (RT-qPCR) (Figure 4-6B). sgRNAs from the STAgR plasmids were expressed in similar amounts as the telomere sgRNA from the control plasmid. Despite equimolar transfection of the STAgR and CARGO arrays, most of the CARGO sgRNAs were expressed 2-5x lower (Figure 4-6B).

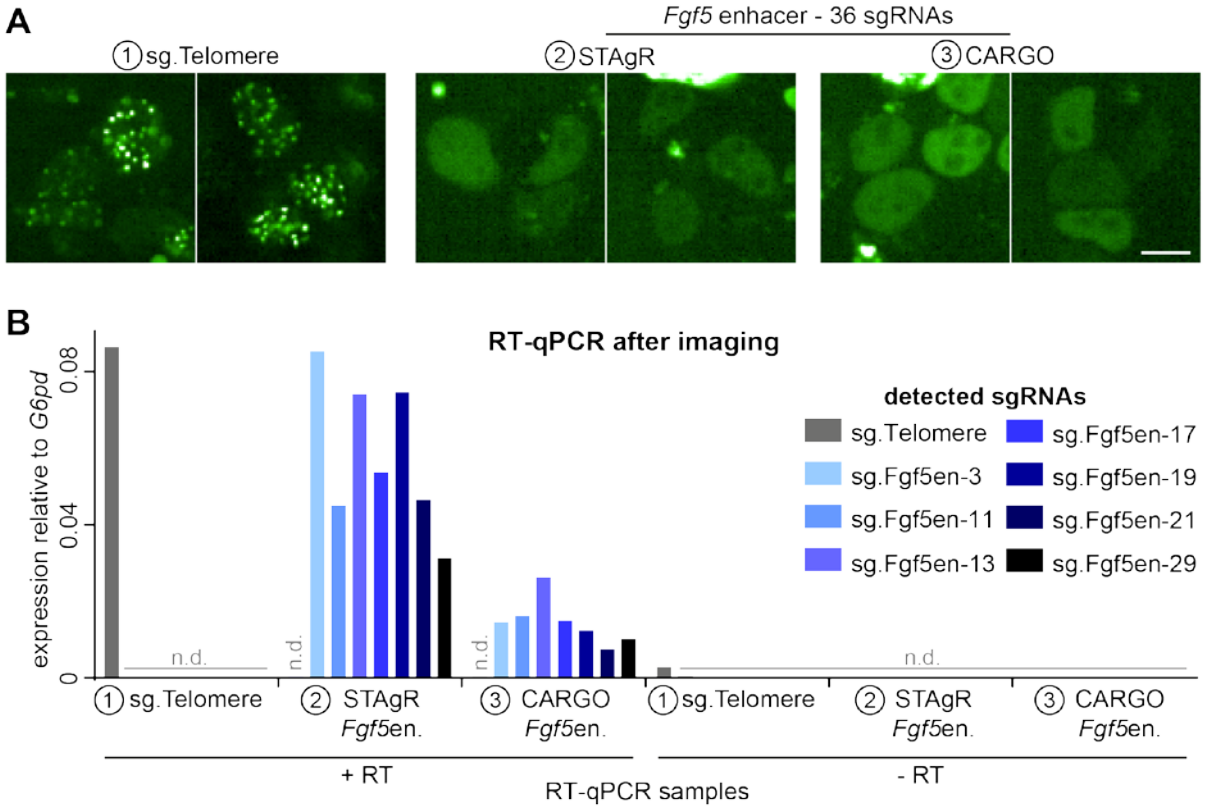


Figure 4-6: Imaging of non-repetitive regions fails despite delivery of 36 sgRNAs
A) mESCs (C8) stably expressing the dCas9-SunTag transfected with scFV-GCN4-sfGFP and different gRNA plasmids. Standard telomere sgRNA plasmid (1), five STAgR plasmids (2) or three CARO plasmids (3); (2) and (3) deliver 36 sgRNAs targeting the *Fgf5* enhancer[100]. Random fields of view were examined for nuclei containing spots. Images show representative foci in the respective sample. Scale bar: 10 µm **B)** RT-qPCR of RNA from imaged HeLa cells detecting the respective sgRNA, normalized to *G6pd*. RT: Reverse Transcriptase; n.d.: not detected

Next, I tested if the multiplexed sgRNA expression would allow imaging of a repetitive control region. Therefore, I generated a STAgR plasmid expressing the telomere sgRNA with 7 gRNAs targeting the *Fgf5* enhancer. Surprisingly, the telomere sgRNA expressed from the STAgR plasmid did not result in any detectable spots. Telomere spots were only visible when expressing the sgRNA from a standard sgRNA plasmid as established before. Notably, the expression of the telomere sgRNA, as detected by RT-qPCR, was similar in both samples (Figure 4-7A).

Lastly, the co-expression of sgRNAs from separate plasmids also affected the detection of telomere spots. When co-expressing another sgRNA for *S.py.* dCas9 no telomere spots were observed. In contrast, expression of a sgRNA specific for *S.au.* Cas9, i.e. a different scaffolding sequence, did not affect the telomere spots (Figure 4-7B). Overall, my results indicate that competition between sgRNAs interferes with imaging, when the sgRNAs contain tracrRNA sequences derived from the same bacterial species.

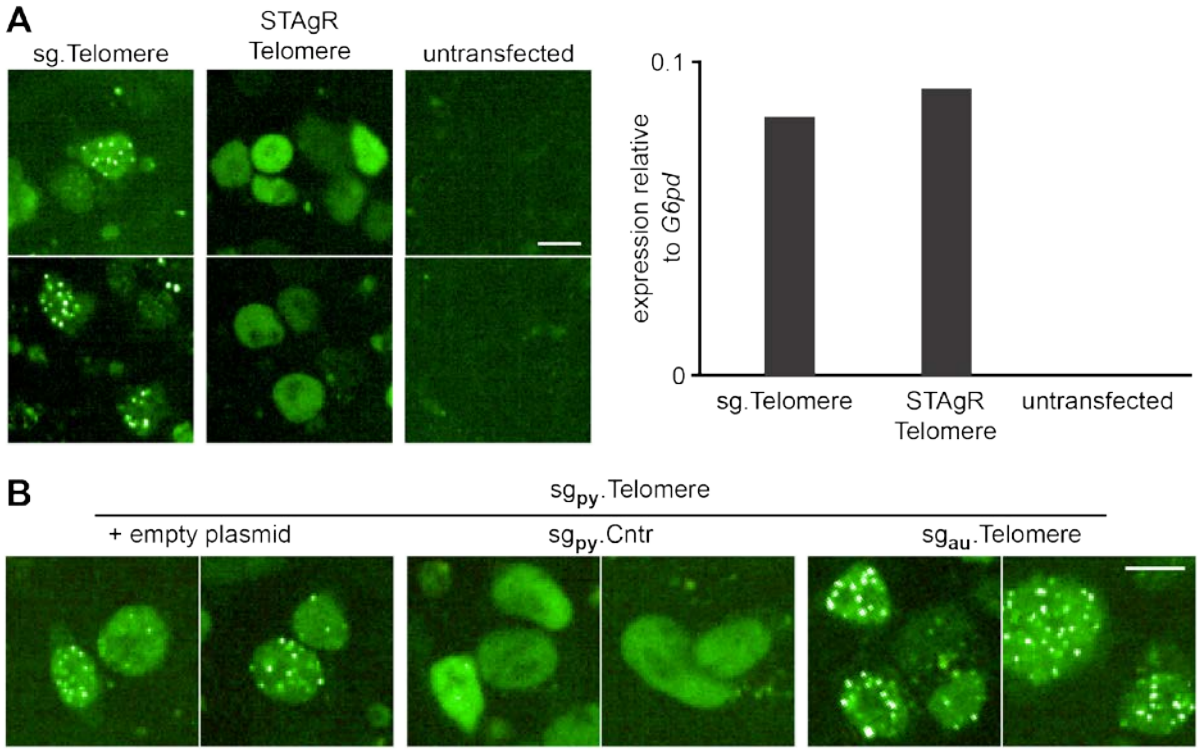


Figure 4-7: Expression of multiple sgRNAs interferes with spot formation
A) Live cell imaging and sgRNA expression in transfected and untransfected cells. mESCs (C8) stably expressing the dCas9-SunTag transfected with scFv-GCN4-sfGFP and telomere sgRNA from standard or STAgR plasmid. The expression of telomere sgRNA relative to *G6pd* was measured by RT-qPCR after imaging. Scale bar: 10 µm **B)** mESCs (C8) expressing the dCas9-SunTag, scFv-GCN4-sfGFP and a *S.py.* specific telomere sgRNA. Cells were co-transfected with empty (non-coding) plasmid or non-targeting sgRNAs specific for *S.py.* or *S.au.* Random fields of view were examined for nuclei containing spots. Images show representative foci in the respective sample. Scale bar: 10 µm

4.2.4. The novel HttTag is not suitable for live cell imaging

For simultaneous imaging of two genomic regions, I required a second tool for signal amplification. The ideal scenario comprises a polypeptide tag, analogue to the SunTag, based on the interaction of a short peptide with a minimal antibody fragment that does not require disulfide bonds for maturation. The lab of Arne Skerra described such a minimal peptide antibody interaction of a 14 amino acid peptide of HTT, bound by a V_L antibody (V_L-HTT)[108]. In

collaboration with [redacted] the "HttTag" for signal amplification was designed, which contains 20 HTT-peptide repeats recruiting fluorophores-V_L-HTT fusion proteins. [redacted] and [redacted] performed the cloning and initial test expressions of the HttTag. A comparison of both systems for targeting and signal amplification is shown in Figure 4-8A.

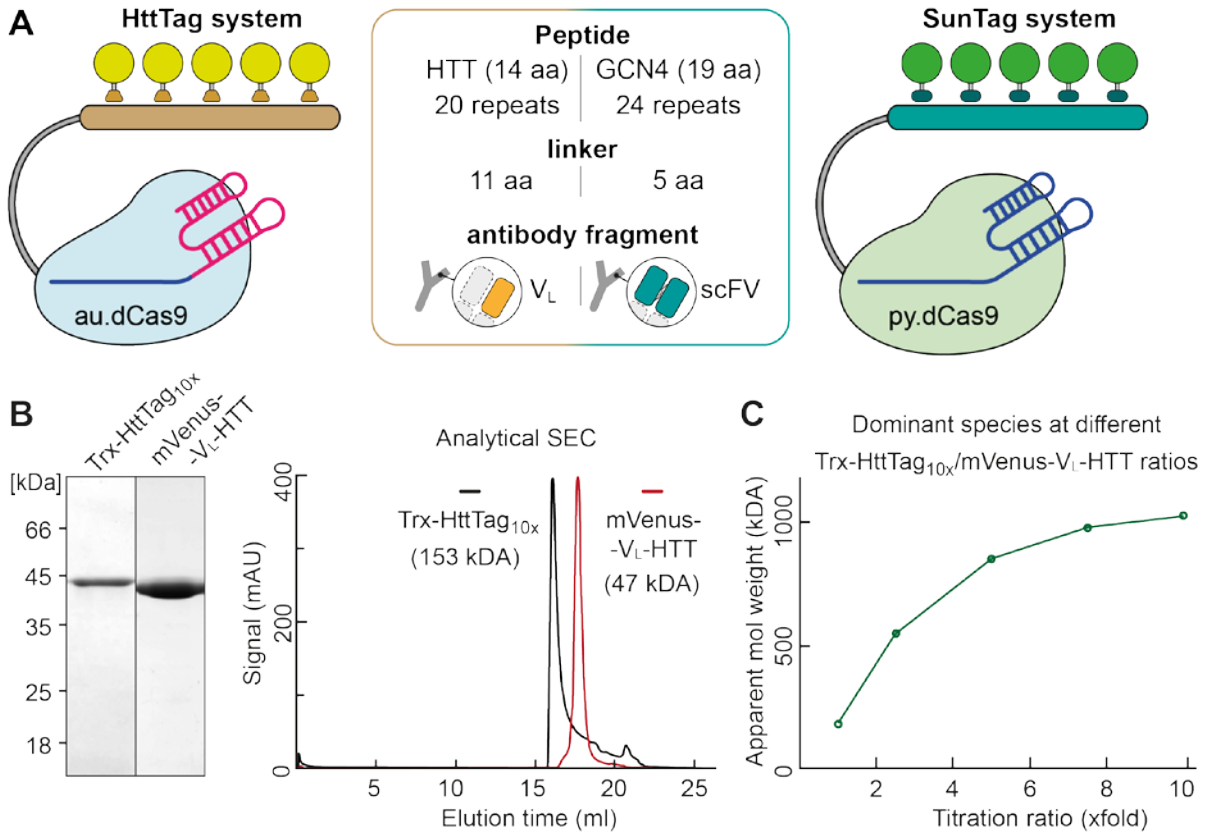


Figure 4-8: The new HttTag for signal amplification recruits fluorophores *in vitro*
A) Scheme showing the analogue signal amplification Tags for live cell imaging. As an example for targeting, the HttTag and SunTag are shown in combination with *S.au.* and *S.py.* dCas9, respectively. Features of both polypeptide tags and their corresponding antibody fragments are shown in the centered box. aa: amino acid **B)** SDS-Page and analytical size exclusion chromatography (SEC) of HttTag fusion proteins. The apparent molecular mass (in brackets) was estimated from the elution fraction. For analytical purpose, the dCas9-HttTag (20 HTT-peptides) was simplified to 10 HTT-peptides fused to thioredoxin (Trx-$HttTag_{10x}$). Proteins ran separately; mAU: milli absorbance units **C)** SEC analysis after titration of Trx-$HttTag_{10x}$ and mVenus-V_L-HTT to estimate complex formation. Note: the apparent molecular weight does not match the theoretic calculated weight as complex formation changes the running behavior in SEC **B)** & **C)** were performed by collaborators in the lab of [redacted].

In vitro assays performed by [redacted] revealed that a simplified $HttTag_{10x}$ (HttTag containing 10 peptide repeats) and the mVenus-V_L-HTT were not aggregating or forming oligomers (Figure 4-8B). Moreover, titration of the $HttTag_{10x}$ with increasing amounts of

mVenus-V_L-HTT revealed that the binding was nearly saturated at a molar ratio of 1:5. This means that adding more than five mVenus-V_L-HTT molecules per $HttTag_{10x}$ did not increase the size of the formed complex, indicating that only half of the ten available epitopes recruited fluorophores (Figure 4-8C). In contrast, the SunTag was previously shown to recruit scFv-GCN4-fluorophores to all available epitopes[99]. Nevertheless, I proceeded with testing the new HttTag in HeLa cells.

In overexpression experiments of the au.dCas9-HttTag system, no sgRNA-specific spots were observed. When expressing the HttTag in combination with Venus-V_L-HTT and telomere or *MUC4* sgRNAs, I observed multiple foci (Figure 4-9A, top row). Although the foci sometimes resembled telomere spots, I observed similar patterns upon sg.MUC4 transfection. Therefore, the spots likely represent aggregates. This aggregation was also observed with other fluorophores that potentially form dimers, like EGFP (not shown). In contrast, monomeric Venus (mVenus) did not show the previously observed aggregates. However, I also did not detect clear foci that could correspond to telomere or *MUC4* staining (Figure 4-9A, bottom row). In some rare cases, I observed nuclei with a faint staining of the telomeres when expressing the respective sgRNA and mVenus-V_L-HTT (Figure 4-9A, bottom left).

Genomic foci were also not detectable when combining the HttTag with the *S.py.* dCas9 targeting system, shortening the HttTag to 10 repeats, using alternative monomeric fluorophores or targeting other repeat regions (negative data not shown). This suggests that the HttTag system is not recruiting sufficient amounts of fluorophores when expressed in cells. To confirm this independently of the dCas9, I targeted the $HttTag_{10x}$ to mitochondria via fusing it to mito-mCherry the targeting domain of mitochondrial protein mitoNEET linked to mCherry[99,109]. For comparison, I expressed a similar $SunTag_{10x}$ fusion protein containing ten GCN4 peptide repeats. Expectedly, both amplification tags localized to the mitochondria, as shown by the mCherry signal. However, only the $SunTag_{10x}$ recruited the respective scFv-GCN4-sfGFP, while the HttTag failed to target mVenus-V_L-HTT to the mitochondria (Figure 4-9B). Taken together, the HttTag did not recruit the V_L-HTT *in vivo* and showed a tendency for aggregation that is not observed with the published SunTag system. Therefore, the HttTag cannot be used for imaging.

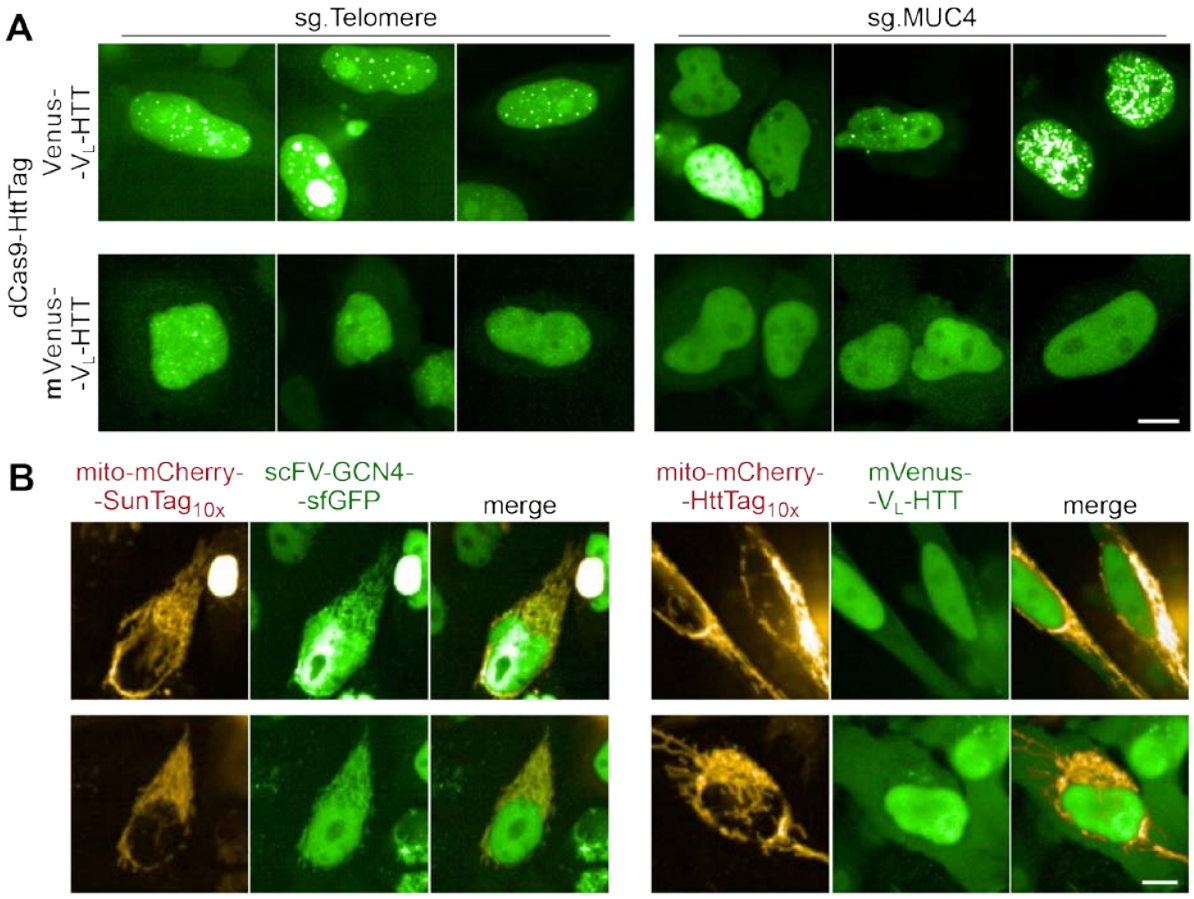


Figure 4-9: The HttTag aggregates and is not recruiting V$_L$-HTT-fusion proteins *in vivo*
A) HeLa cells transfected with au.dCas9-20xHttTag system with indicated fluorophore-V$_L$-HTT and *S.au.* sgRNA. Top: Aggregation was detected in samples with standard (i.e. dimerizing) fluorophore Venus. Bottom: monomeric mVenus-V$_L$-HTT did not show aggregates or specific spots. Random fields of view were examined for nuclei containing spots or aggregates with Hoechst33342 staining as reference (not shown). Images show representative nuclei in the respective sample. Scale bar: 10 µm **B)** HeLa cells transfected with SunTag$_{10x}$ or HttTag$_{10x}$ fused to mito-mCherry[99] (mitochondrial targeting domain of mitoNEET[109] fused to mCherry) co-expressed with the respective antibody-fluorophore fusion protein. mCherry channel shows localization of the polypeptide tags to the mitochondria. The green channel represents the localization of the corresponding antibody fragment. Merge shows colocalization of amplifaction tag and antibody fragment. Random fields of view were examined. Images show representative nuclei in the respective sample. Scale bar: 10 µm

4.3. Discussion

In this work, I aimed to establish live cell imaging of non-repetitive genomic regions in human and mouse cells. In addition, I planned to develop a system for signal amplification analogous to the previously published SunTag. This could facilitate stable dual color imaging in living cells and allow visualization of independent genomic regions. Utilizing such a system would enable investigation of dynamic long-range chromatin interactions on a single cell level, for example through mitosis or upon gene activation. Other methods like FISH or 3C cannot reveal these dynamics, as they rely on fixation of the chromatin state or bulk analysis of a cell population.

4.3.1. Generation of a stable cell line expressing dCas9-SunTag

I successfully generated stable mESC lines expressing the dCas9-SunTag from the *Rosa26*[110] locus to circumvent the transfection of the 11 kb plasmid encoding the fusion protein. The dCas9-SunTag was coupled with a CAG promoter that is resistant to silencing via methylation[107]. This ensured stable expression and should allow expression of the transgene even during differentiation of the cells into other cell types. Other potential safe harbor loci that are used for gene insertion in mouse include Tigre, Hrpt1 and Cd6[111].

Importantly, to integrate the dCas9 fusion protein I delivered active Cas9 protein as mRNA to avoid any integration of the active nuclease during antibiotic selection[112]. Similarly, the *Rosa26* sgRNA was transfected as *in vitro* transcribed RNA to prevent stable integration of the sgRNA that could interfere with the downstream imaging. The dCas9-SunTag sequence is 8.6 kb long and was expected to integrate only poorly into the mESC genome, as transgenes integrate less efficient with increasing size[113]. Therefore, I combined antibiotics selection, FACS, monoclonal selection, a telomere imaging screen, and genotyping to identify clones with correct integration of the dCas9-SunTag (Figure 4-5). I generated 25 mESC lines stably expressing the construct, representing 8% of the 300 monoclonal colonies isolated after FACS.

4.3.2. Orthogonal Cas9 systems allow imaging of repetitive loci

I performed imaging of repetitive regions in the generated dCas9-SunTag mESC lines as well as in human cancer cells. The high photostability of the amplification tag allowed imaging over extended periods even with weaker fluorescence signals (Figure 4-5).

In addition to the widely used *S.py.* Cas9 system, I also generated au.dCas9-SunTag constructs for imaging. As reported previously[114] there was no cross-reactivity of sgRNAs derived from *S.py.* or *S.au.* with Cas9 proteins from the opposite species (Figure 4-4). This suggests that the tracr sequences from *S.py.* and *S.au.* are not interchangeable, which consequently allows imaging with both dCas9 orthologues simultaneously. The longer PAM sequence required for *S.au.* binding, reduces the number of available targets making it less flexible than the widely used *S.py.* Cas9. Conversely, *S.au.* Cas9 is 23% smaller than the *S.py.*

orthologue resulting in slightly easier transfection. Recently, even smaller Cas variants have been engineered that can reduce the size of Cas-fusion proteins even further[115,116].

In HeLa cells, I successfully imaged telomeres, the *MUC4* locus and subtelomeric repeats on Chromosome 3, 14 and 19 (Figure 4-3). In human somatic cells telomere length is in the range of 5 – 15 kb[117]. However, in HeLa cells telomere length is usually in the lower range of 5 kb and can become shortened to below 1 kb[118,119]. Assuming a Cas9 footprint of 25 bp[120], up to 40 binding sites could be present per 1 kb telomere repeat. The repeats on Chr3, Chr14, Chr19 each contain around 100 repeats of a >30 bp sequence[121]. Finally, the human *MUC4* sgRNA has 32 potential target sites that perfectly match the genomic sequence. However, Cas9 binding is also possible with up to two mismatches between sgRNA and its DNA target[122]. In case of *MUC4*, allowing only one mismatch would more than double the potential binding sites to around 80 (data not shown).

In mESCs the highly repetitive telomeres and minor and major satellite repeats were imaged (Figure 4-5). Compared to human, mouse telomeres are up to 10x longer and were therefore robustly detected as bright spots[123]. The major satellite repeats span 6 Mb with a 234 bp repeat unit, and the minor satellite repeats span around 0.6-1.2 Mb with a 120 bp segment[96]. Even though the major satellite repeats should provide around 3x more binding sites, the minor satellite foci were detected more readily (Figure 4-5). The faint, less sharp signal observed for the major satellite repeats are likely related to the more spread-out distribution of the repeat unit as compared to the more condensed minor satellite. The repeat at the *Akap6* locus should provide 87 binding sites[96] but was not detectable in the stable or transiently transfected mESCs. This is surprising, as the selected sgRNA was previously used in NIH3T3 cells (mouse fibroblasts)[96] and provides a similar number of binding sites as the human *MUC4* sgRNA. The lack of detectable spots could be related to the increased background in the more condensed nucleus or less accessible chromatin in mESCs. In addition, mESCs grow in dome shaped colonies, which can produce more background fluorescence in the Z-dimension as compared to adherent cell lines growing in a flat monolayer.

4.3.3. Expression of multiple gRNAs is not resulting in detectable spots

The Wysocka lab previously reported live cell imaging of regulatory elements in mESCs by expressing 36 sgRNAs targeting the same region. I tried to confirm this in my stable mESCs by expressing the same sgRNAs from the original CARGO- or newly generated STAgR-plasmids (Figure 4-6). Surprisingly, there were no detectable spots in the stable mESCs or in cells separately transfected with all dCas9-SunTag components. The STAgR plasmids were fully sequenced and expression of the sgRNAs from CARGO and STAgR vectors was confirmed by RT-qPCR (Figure 4-6).

The mESCs used by the Wysocka lab express a doxycycline inducible dCas9-GFP fusion[100], in contrast to the constitutively expressed dCas9-SunTag in my experiments. It is likely that modulating the expression of dCas9-GFP helps to reduce background noise. Moreover, the inducible system allows fine-tuning of the ratio between sgRNAs and available dCas9-fusion proteins. Notably, I also tested imaging with dCas9-GFP, various fluorophores and nuclear localization signals combined with the SunTag system, or background reduction by fractional photo bleaching, but did not detect sgRNA dependent spots (data not shown).

It has been hypothesized that expression of multiple sgRNA can happen in a pulsating fashion leading to unsynchronized sgRNA pools[124]. Although this cannot be excluded, I observed that the RNAs were expressed uniformly in the transfected cell population (Figure 4-6). Indeed, single cells might only express isolated sgRNAs, however, my data suggest that the problem is not related to asynchronous sgRNA expression but rather to multiple sgRNAs competing for dCas9 binding. When expressing the telomere sgRNA from the context of a STAgR backbone, the telomere foci were gone. Since I detected normal levels of the telomere sgRNA by RT-qPCR, a pulsating sgRNA expression should allow detection of telomere spots at least in a subset of cells (Figure 4-7). However, I did not detect any nuclei containing telomere spots using the screening microscope, indicating that the imaging might be impaired due to internal competition of the sgRNAs. This was confirmed by co-expressing a standard telomere *S.py.* sgRNA plasmid with a non-targeting sgRNA from *S.au.* or *S.py.*. Whenever two sgRNAs for *S.py.* dCas9 were expressed, no spots were detectable, however, co-expression of *S.au.* specific gRNAs did not affect spot formation (Figure 4-7). This suggests that sgRNAs compete for Cas9 binding via the species-specific stem loop. It does not support a model where sgRNAs compete for expression by the transcription machinery.

Competition between different sgRNAs was also reported for Cas9-mediated genome editing[125]. It might be possible that the presence of different sgRNAs is affecting the dissociation rate of established sgRNA-Cas9 riboprotein complexes. Evolutionary, a bacterial Cas9 fighting of viral DNA might want to switch between different available sgRNAs, especially if the current sgRNA is not resulting in productive cutting of the targeted DNA. Interestingly, it was also reported that intracellular RNAs inhibit the assembly of sgRNA-Cas9 complexes[126]. Using the Cas9 system in cells with a transcriptionally hyperactive genome, like mESCs[127], might therefore be unfavorable.

The competition between sgRNAs might be resolved if more dCas9 protein was available for binding in the cell. However, the CMV and CAG promoters employed in this study should result in sufficient expression of the dCas9-SunTag. In the stable mESCs I did not confirm biallelic integration of the CAG-dCas9-SunTag cassette at the *Rosa26* locus. Hence, the dCas9-

SunTag expression might not be at the level that could be achieved with two integrated expression cassettes. However, I also did not observe spots upon transient transfection of the dCas9-SunTag system in combination with multiple sgRNAs (not shown). This indicates that the sgRNA competition is not a problem exclusive to the generated mESC lines. Nevertheless, it highlights the advantage of an inducible system, which would provide control over the dCas9-SunTag expression levels.

4.3.4. The new HttTag is not recruiting V_L-HTT fragments *in vivo*

Even though imaging of non-repeat regions was not successful, I aimed to establish an additional amplification tag for parallel dual color imaging of two repetitive genomic regions. In collaboration with the lab of [redacted], the HttTag was designed that recruits V_L-HTT-fusion proteins to a peptide array based on HTT. SEC analysis indicated that the amplification tag recruits V_L-HTT -fusion proteins *in vitro*. However, only half of the epitopes were bound by antibody fragments (Figure 4-8C). This could be explained by steric hindrance interfering with binding of all epitopes. Even though the amplification tag principally worked *in vitro*, it did not function when expressed in HeLa or mESCs (Figure 4-9). In combination with standard fluorophores, I regularly observed aggregation of the HttTag system, which was prevented when using monomeric fluorophores. However, there were no detectable foci when targeting the HttTag to genomic regions like telomeres or *MUC4* using dCas9. Finally, I tested if the HttTag allowed targeting of mitochondria by fusing it to mito-mCherry. Even though the HttTag localized to the mitochondria without visible aggregation, there was no recruitment of the V_L-HTT-fusion protein.

It is likely that the V_L-HTT is not folding properly when expressed in cells. Therefore, when fused to monomeric fluorophores no recruitment to the HTT-peptide was detected. In contrast, using dimer-forming fluorophores resulted in visible aggregation of the HttTag-system. The dimer formation might affect folding of the attached V_L-HTT, thereby promoting aggregation. In this case, the fluorophore-V_L-HTT might actually fold into a functional protein binding the poly-HttTag. However, dimerization of fluorophores complexed with different HttTags could result in the formation of large unwanted poly-HttTag complexes or aggregates. Similarly, this mechanism might affect the published SunTag, which was reported to form aggregates at high concentrations of the antibody-fluorophore[99].

The HttTag contains the amino acids 5-18 of the HTT N-terminus. The aggregation of HTT in Huntington disease is associated with expansion of a poly-glutamine downstream of the N-terminal amino acids 1-17[108,128]. Therefore, aggregation of the HttTag system cannot directly be attributed to the utilized peptide sequence. Nevertheless, the first 17 amino acids of HTT can promote aggregation of full length HTT *in vitro* and in cells[129,130]. The ubiquitously

expressed endogenous HTT[131,132] could interact with the “trigger-peptide” and form insoluble aggregates. This would also explain why the Tag shows no aggregation *in vitro*, as this process requires the endogenous full-length HTT. However, the overexpression of the HttTag fused to mito-mCherry did not show notable aggregation in the respective channel, which suggests that a functional V_L-HTT is involved in the aggregation process.

Recently, the lab of Marvin Tanenbaum published a new amplification tag orthogonal to the original SunTag. This tag, referred to as MoonTag, is based on the interaction of a gp41 peptide and a corresponding anti-gp41 nanobody. The tag only achieves 50% targeting efficiency as compared to the SunTag, but was nevertheless successfully employed to visualize translation of mRNA molecules in living cells[133]. The MoonTag could be used in combination with *S.au.* dCas9 and the established *S.py.* dCas9-SunTag system to image two repetitive genomic regions.

As an alternative for signal amplification extended sgRNA stem loops could be used to recruit more fluorescent proteins. Similarly, more stable organic dyes or nanoparticles could improve signal intensity or stability. However, none of these methods will overcome the general problem of background fluorescence when the aim is to image a non-repetitive locus. Next to the real targeted signal, there will also be untargeted dCas9-fluorophores floating in the nucleoplasm or screening the DNA for its target sequence.

Even though there are labs that have achieved imaging of non-repetitive regions by expressing an array of sgRNAs, these approaches currently do not seem to be robust. As demonstrated in this work, even the use of published sgRNAs in the same cell type is not reliably producing spots. Moreover, generating and validating the plasmids targeting a new locus with >30 sgRNAs is very laborious. An alternative approach could be the integration of an ectopic repeat that can be targeted by multiple dCas9 proteins with a single sgRNA. Even though an ectopic repeat is likely affecting the local chromatin dynamics, the same might be true for targeting of >30 dCas9-SunTags, each recruiting 24 additional antibody-fluorophores. Hence, integrating an ectopic stretch of DNA could be a valid solution of the background issue.

Very recently, Clow and colleagues performed live cell imaging of a non-repetitive locus with dCas9 employing only one sgRNA, extended with 15 PUF binding sites that recruit the respective number of PUF-fluorophores[134]. Surprisingly, there seems to be no other sgRNA in the nucleus that is forming a complex with the PUF-fluorophores and could result in noticeable background. The authors speculate that only the sgRNA in a complex with dCas9 at the target site is stable, and hence, only the bound dCas9-sgRNA complex recruits 15 fluorophores. Although the foci in this study represent a single dCas9 binding event recruiting 15 fluorophores, the presented spots are suspiciously large and variable in their size suggesting

that a much larger number of dCas9s or fluorophores is present in the detected spot. It would be interesting to understand why the same system was previously only applied to image highly repetitive regions[98] and it needs to be evaluated how robust the method is in other studies.

Overall, I was able to image multiple highly repetitive regions in human cells and mESCs. In addition, I generated various lines of mESCs stably expressing the py.dCas9-SunTag from the *Rosa26* locus. These cannot only be used for imaging but the integrated dCas9-SunTag system could also be used for targeted DNA methylation and demethylation, or targeted gene activation, as was demonstrated before[135–138]. In summary, imaging of non-repetitive loci was not achieved and the new HttTag was not functional *in vivo*. However, the established dCas9-SunTag cell lines provide a flexible model system that allows targeting of functional proteins to manipulate and study specific genomic regions.

Part-II

5. Mapping deoxyinosine in genomic DNA

5.1. Introduction

In humans, there are only a few reported epigenetic DNA modifications. Some of them well characterized, like 5mC and 5hmC, and others more controversial, like m6A or 8oxoG[139]. This is in stark contrast to RNA where more than 150 base modifications are reported. Many of them contain regulatory functions in transfer RNA (tRNA), ribosomal RNA (rRNA) and messenger RNA (mRNA)[140–142]. In analogy to epigenetic DNA modifications these are collectively described as the epitranscriptome[143].

One of the most abundant and best characterized RNA modifications is riboinosine (rI), the deamination product of adenosine[144–146] (Figure 5-1A). The DNA equivalent of rI, deoxyinosine (dI) was classically regarded as a DNA damage because it is mutagenic. Only recently, dI was associated with regulatory functions outside of DNA damage. I will introduce the origins and functional roles of rI and dI in the following chapters.

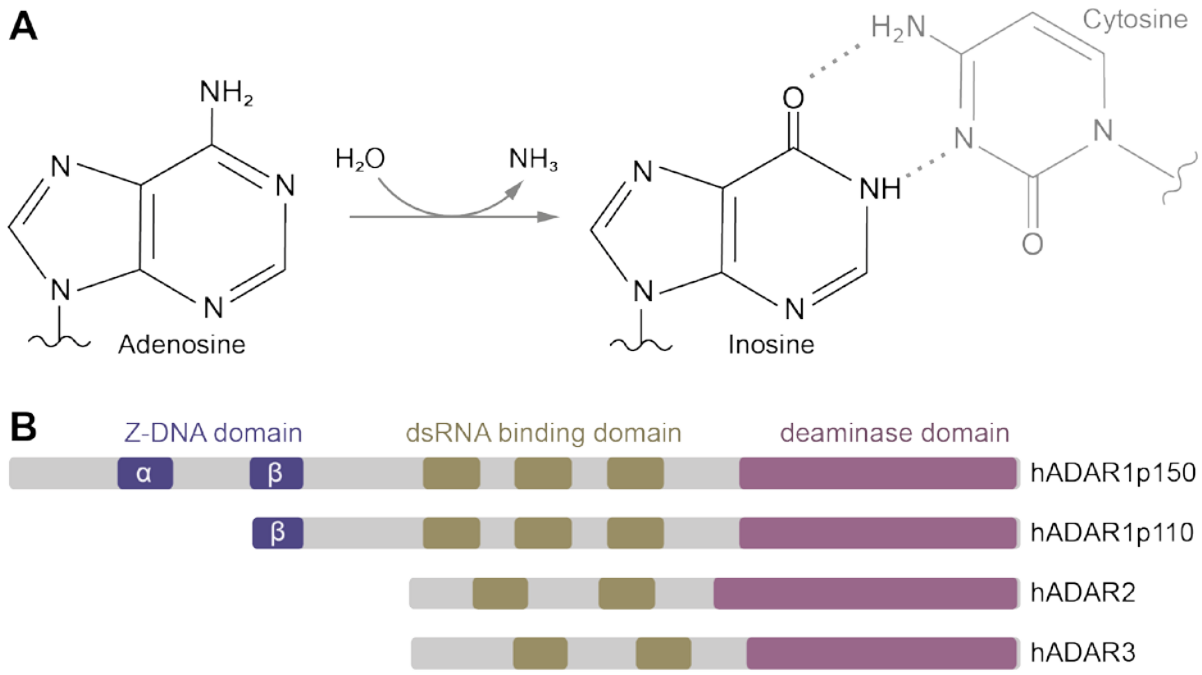


Figure 5-1: ADAR generates inosine in RNA by hydrolytic deamination
A) Hydrolytic deamination of (deoxy)adenosine produces (deoxy)inosine. Inosine preferentially base-pairs with cytosine. **B)** Human adenosine deaminases acting on RNA (ADAR). ADAR1 isoforms p150 and p110 contain two or one Z-DNA binding domains, respectively. All ADARs have a dsRNA binding domain as well as a C-terminal deaminase domain. In contrast to ADAR1 and 2, ADAR3 has no described enzymatic activity. Domain structure according to uniprot entries P55265, P78563, Q9NS39.

5.1.1. The functional role of inosine in RNA

In mammals, the hydrolytic deamination of adenosine is catalyzed by adenosine deaminases acting on RNA (ADAR)[147]. Multiple proteins of the ADAR family are described in humans: ADAR1[148], ADAR2[149] and ADAR3[150] (Figure 5-1B). In addition, there are specific adenosine deaminases acting on tRNA (ADAT)[151].

ADARs are conserved in vertebrates but are not present in yeast and plants[152]. In contrast, ADATs are also present in yeast[147]. All ADAR isoforms contain a dsRNA binding domain (dsRBD) that facilitates the interaction with dsRNA[153]. A deaminase domain, facilitating the deamination of adenosine to inosine, is located at the C-terminus of all ADAR isoforms. Even though the deaminase domain is present in ADAR3, there is no reported catalytic activity for this isoform[154].

For ADAR1 two isoforms are reported, full-length ADAR1p150 and a shorter ADRA1p110[155]. ADAR1p150 contains two N-terminal Z-DNA binding domains Zα and Zβ[156]. In the shorter ADARp110 only the Zβ domain is present (Figure 5-1B). Both ADAR1 isoforms shuttle between the nucleus and cytoplasm, whereas ADAR2 is mostly localizing to the nucleus[157,158]. ADAR1p110 is ubiquitously expressed, while ADAR1p150 is interferon inducible[159]. ADAR2 is expressed in all tissues, with high expression in brain and arteries[158,160]. ADAR3 is mostly expressed in the brain[161] and is thought to modulate RNA editing by inhibition of ADAR2[162].

The deamination of A to rI by ADARs and ADAT is changing the base-pairing properties at the edited position. In contrast to adenosine, rI preferentially base-pairs with cytosine instead of thymine[163,164]. This alteration has many known effects on RNA biology.

In tRNAs the presence of rI in the wobble position allows decoding of multiple mRNA codons[165,166]. Moreover, the introduction of rI in the coding region of the mRNA itself can alter the translated amino acid sequence as the translation machinery will perceive it as a guanosine. In humans, around 80 of these mRNA recoding events are known, which affect the function of the coded protein[158]. Most of the recoding events are associated with proteins related to functions of the nervous system[167,168] and are most frequently edited by ADAR2[163]. For example, in the mRNA of glutamate receptor 2 (GRIA2) a glutamine codon is modified into an arginine codon, resulting in changed Ca^{2+}-permeability of the associated AMPA receptor complex[169]. Similarly, in the α3 subunit of the $GABA_A$ receptor isoleucine to methionine recoding reduces receptor trafficking to the cell surface[170]. Lysine to arginine editing in the DNA repair enzyme NEIL1 is mediated by ADAR1 and changes its repair efficiency towards different substrates[171].

The majority of A to rI editing occurs in non-coding RNAs where it affects splicing[172], maturation of circular RNAs[173] and microRNA (miRNA) binding[174,175]. These mostly repetitive non-coding

regions are primarily targeted by ADAR1[163]. ADAR-mediated editing can also affect the maturation of primary miRNAs into pre miRNA. Here, A to I editing can block cleavage by the nuclear RNAse III DROSHA[176].

Recently, it was shown that deamination of dsRNA by ADAR1 is required to avoid activation of the cellular immune response[177]. A to rI editing allows the cell to recognize the dsRNA as own and to distinguish it from viral RNA. Unmodified dsRNA is sensed by melanoma differentiation-associated protein 5 (MDA5) that initiates an interferon dependent autoinflammatory response and results in cell death[178].

A to rI editing is also implicated in many diseases, highlighting the essential role of the ADAR enzymes. Hypo- and hyper-editing of RNA is reported for various human cancers and often these editing alterations are proposed to negatively affect patient survival[179–182]. Changes in A to I editing were also observed in Alzheimer's disease[183], autism[184], epilepsy[185] and other diseases related to disorders of the central nervous system[186–188].

ADAR2 knockout mice are prone to seizure and die young which was linked to hypo-editing of the Q/R site in the GRIA2 pre-mRNA[189]. In humans, the same target is showing reduced editing in patients with amyotrophic lateral sclerosis[190,191].

ADAR1 knockout in mice is embryonically lethal and associated with aberrant activation of interferon signaling. This severe outcome is related to ADAR1's role in labeling endogenous dsRNAs as "non-viral" to prevent activation of MDA5 mediated inflammation[177]. Similarly, in human ADAR1 mutations are associated with Aicardi Goutieres syndrome, an autoimmune disease associated with increased interferon signaling[192]. Most of the eleven reported mutations are located in the deaminase domain of ADAR1. However, three are located in the Zα Z-DNA binding domain of ADAR1p150[193]. This is an interesting observation, as one would rather expect a mutation in the domain required for binding to dsRNA. The structure of Z-DNA and its relation to ADAR1 will be discussed in chapter 5.1.4.

5.1.2. Occurrence and repair of deoxyinosine in genomic DNA

In contrast to rI, dI was classically considered a DNA damage introduced by spontaneous deamination[194] or passive integration during replication[195]. dI occurs at a rate of 1-10/10^6 bases in mammalian cells[196–198], equal to 3000 – 30,000 dI per human cell. Due to its preference to pair with cytosine it can result in an A>G transition mutation[199] (Figure 5-1A). Spontaneous DNA deamination is promoted by non-neutral pH and high nitric oxide levels. However, adenosine deamination is 40x slower than the rather common C or mC deamination. Spontaneous dI generation is reported to be 200-fold higher in ssDNA as compared to dsDNA since base-pairing protects from deamination[194].

Free dA nucleosides can be deaminated to dI by adenosine deaminase (ADA)[200]. Free dI is further processed into hypoxanthine by purine nucleoside phosphorylase (PNP)[201,202]. The hypoxanthine is than thought to be salvaged into deoxyinosine monophosphate (dIMP) and deoxyinosine triphosphate (dITP)[203] thereby increasing the pool of free dITP. dITP levels could also increase by spontaneous deamination of dATP. Subsequently, DNA polymerases integrate free dITP into the DNA during replication[195].

dI in genomic DNA (gDNA) can be repaired by (i) base excision repair (BER) or (ii) alternative excision repair (AER) (Figure 5-2). (i) BER is initiated by N-methylpurine DNA glycosylase (MPG), which releases hypoxanthine and leaves an abasic (AP) site. The AP site is than incised by AP endonuclease and further processed by short- or long-patch BER[204]. (ii) The AER pathway is initiated by Endonuclease V (EndoV). In contrast to MPG, EndoV does not release the base from the DNA but incises the second phosphodiester bond 3' of dI[205] (Figure 5-2). The removal of dI requires further processing by a 3' exonuclease activity. In *E. coli* this was shown to be facilitated by DNA polymerase I[206]. The downstream processing in mammalian cells is still unknown[207,208].

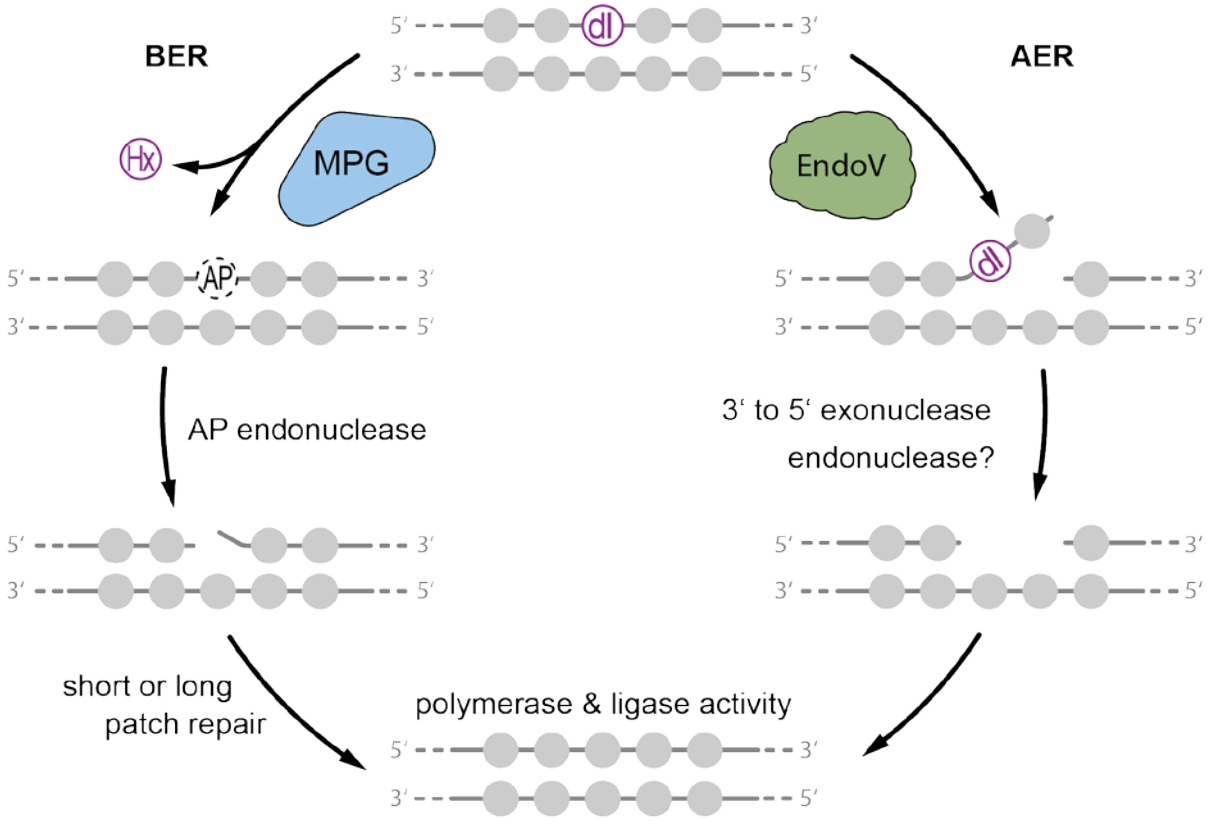


Figure 5-2: Repair of dI by MPG and Endonuclease V
Left: Methylpurine DNA glycosylase (MPG) initiates base excision repair (BER). MPG cleaves the N-glycosidic bond of dI to release hypoxanthine (Hx) from the DNA backbone. The abasic site (AP) is subsequently processed by AP endonuclease and repaired by short- or long patch BER. **Right:** Alternative excision repair (AER) is initiated by Endonuclease V (EndoV). EndoV nicks the second phosphodiester bond 3' of dI. The downstream processing steps in mammalian cells are not fully understood. Modified from Kuraoka[209]

Human and *E. coli* EndoV cleave double-stranded (ds) and single-stranded (ss) DNA containing dI. In addition, they show activity towards RNA[209,210]. While human EndoV has a strong preference for ssRNA, *E. coli* EndoV shows activity toward ss and dsRNA[211]. Nicking activity of EndoV requires Mg^{2+} and replacing Mg^{2+} with Ca^{2+} renders *E. coli* EndoV inactive, while enabling binding to its substrate[205,207]. This strategy was previously applied to enrich rI containing RNA species[212,213].

5.1.3. Deoxyinosine in DNA – more than a damage?

Until recently, no enzyme was described that could specifically deaminate dA to dI in DNA. However, in 2017 first *in vitro* evidence showed that dI might be enzymatically introduced into gDNA by ADARs. dA in DNA:RNA hybrids is deaminated by a hyperactive ADAR2 deaminase domain in the context of a dA:rC mismatch. Moreover, *in vitro* editing is observed with full length wild type ADAR2 when using long DNA:RNA hybrids[214]. In vivo DNA:RNA hybrids are mostly found in the context of R-loops, which are three stranded nucleic acid structures composed of the hybrid and a displaced ssDNA. I will provide a more detailed introduction on R-loops in the next chapter 5.1.4.

The potential activity of ADAR on DNA in DNA:RNA hybrids was further supported by multiple studies analyzing mutation profiles of genomic sequences that showed increased A>G and T>C transitions. This is the expected mutation profile that would occur after dI incorporation followed by replication. Steele and Lindley proposed a role for ADAR-mediated A-to-I DNA editing of the DNA:RNA hybrid at the transcription bubbles during somatic hypermutation of immunoglobulins (Ig)[215]. Tasakis and colleagues observed that multiple myeloma patients, with increased ADAR1 expression, acquired new DNA mutations at known ADAR1 editing sites[216]. Similarly, increased A>G transitions at ADAR editing sites were reported for human and drosophila by Popitsch *et al.*[217].

In 2021, the lab of Kuzuko Nishikura reported ADAR1p110 activity on telomeric R-loops formed between canonical and non-canonical telomere repeats. In cancer cells deamination of dA:rC mismatches in these R-loops facilitated their resolution by RNase H2[62].

Most recently, Tang *et al.* showed that ADAR1&2 process DNA:RNA hybrids during chromothrypsis, a severe mutation event with thousands of chromosome rearrangements. Their work suggests that dI and the subsequent action of DNA repair enzymes lead to fragmentation of micronuclear chromosomes[63].

These findings were further supported by mass spectrometry data from our own lab. [redacted] detected dI enrichment in DNA:RNA hybrids in various mammalian cell lines using liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) . After DNA:RNA immunoprecipitation (DRIP) with the hybrid-specific S9.6 antibody, the relative dI

signal was strongly enriched as compared to the input material (Figure 5-3). Notably, the signal was sensitive to RNase H, but not to Nuclease S1. This indicates that dI is present in the hybrid forming DNA strand and not just in the displaced ssDNA (Figure 5-3B). A. K.-Baumgärtner also showed that ADAR1 overexpression boosts the dI/dA DRIP enrichment while ADAR1 knockdown reduces it[64]. Since the DRIP LC-MS/MS measurements are not revealing the sequence context, it remains unresolved which hybrids show the enrichment.

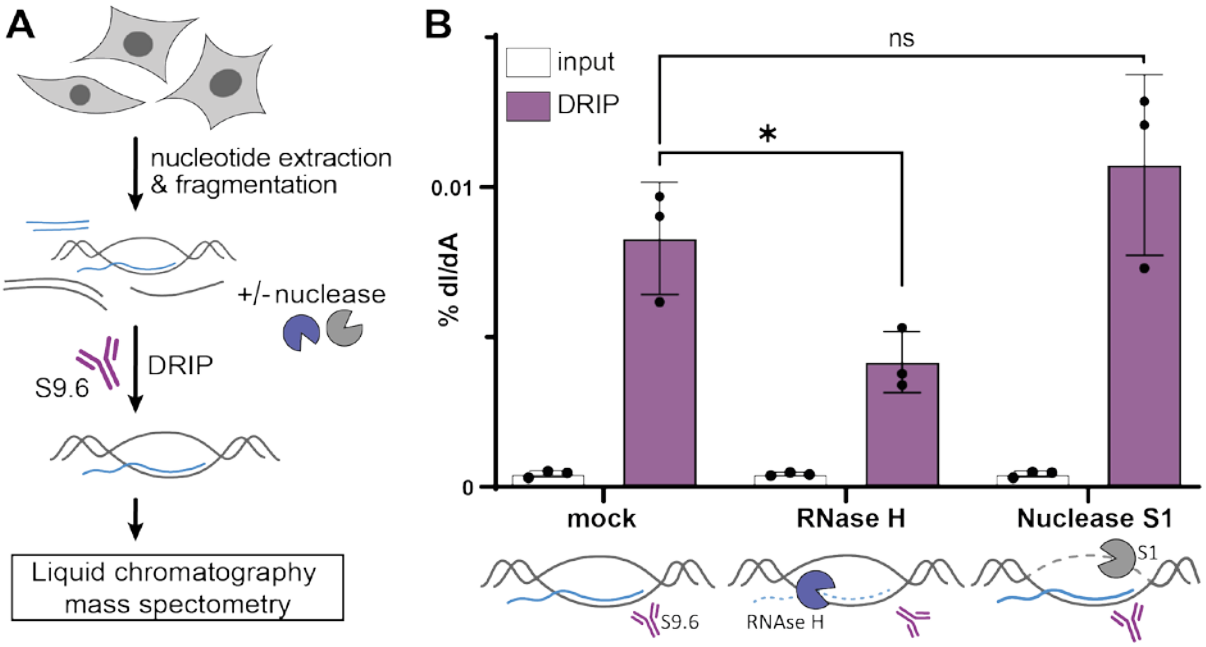


Figure 5-3: dI is enriched in RNA:DNA hybrids
A) DNA:RNA immune precipitation (DRIP) method. Nucleic acids are purified and treated with different nucleases. S9.6 antibody pulldown is performed and material is subjected to LC-MS/MS. **B)** Relative dI levels determined by stable isotope dilution LC-MS/MS in input and DRIP-enriched DNA of HEK293T cells. DRIP samples treated with no nuclease show enrichment of dI/dA compared to input material. The enrichment is sensitive to hybrid removal by RNase H, but not to removal of the DNA single strand by nuclease S1. Data are shown as mean ±SD, ns= not significant, * = $p<0.05$ according to paired t-test. Data from Dr. Anne K.-Baumgärtner[64].

5.1.4. R-loops and the conformation of nucleic acid helixes

R-loops are three stranded nucleic acid structures composed of a DNA:RNA hybrid and a displaced ssDNA. R-loops form in *cis* during transcription by RNA polymerases and their formation and stability is promoted through secondary structures established by the displaced non-coding DNA strand, e.g. G-quadruplexes[218]. R-loops can also form in *trans* when an RNA transcribed at another locus invades a dsDNA, displacing one of the DNA strands. This process can be facilitated by specific proteins like RAD51 recombinase[219] or CRISPR-Cas nucleases[76]. There are also non-R-loop DNA:RNA hybrids, which are found at telomeres or during replication where the RNA is required for telomere extension or lagging strand synthesis, respectively[218].

R-loops can affect gene expression by modulating promoter methylation[220,221] or recruiting proteins regulating the chromatin state[222,223]. They also affect transcription termination[224] through torsional stress[225], polymerase backtracking[226] and polymerase pausing[227,228]. Moreover, R-loops are involved in DSB repair[229], chromatin compaction at peri-/centromeric regions[230] and telomere biology[62,231,232].

Due to the various roles R-loops have in biological processes, they require tight regulation and the accumulation of R-loops can be a harmful threat to the genome[233–235]. R-loops are degraded by RNA-specific nucleases like RNase H[236], but also DNA endonucleases like XPF and XPG[237]. Several helicases can unwind RNA:DNA hybrids, including Senataxin[227,238], Aquarius[237] and DEAD-Box RNA helicases[239]. Finally, RNA base modifications like m6A and rI, can affect R-loop stability[218,240].

The recent report, that ADAR1 mediated DNA deamination of telomeric R-loops affects their resolution by RNase H2 highlights the potential role of dI in R-loop regulation[62]. However, it also raises the question if other DNA:RNA hybrids are targeted and which sequence feature facilitates the ADAR mediated deamination. Most likely the helix conformation of the DNA:RNA hybrid affects ADAR binding and its activity. In the following, I will introduce how the (i) Z- and (ii) A-helix conformation facilitate ADAR binding to its substrates.

(i) As stated before, ADAR1p150 contains a Zα binding domain that facilitates binding to Z-forming dsDNA. This Z-DNA is a left-handed helix that is formed preferentially by purine-pyrimidine dinucleotide repeats[241,242]. One helix turn is composed of six dinucleotide repeats, i.e. 12 bp. In contrast, the standard conformation of dsDNA is the B-helix, containing 10 bp in a right-handed helix turn. The Z-form helix is more slender than the B-form and shows a less smooth arrangement of the sugar-phosphate backbone (Figure 5-4A)[243].

The ADAR1 Zα domain stabilizes the Z conformation and promotes switching of B- into Z-helixes. Interestingly, switching an DNA:RNA hybrid into Z-formation is kinetically more favorable than switching dsDNA or dsRNA[244]. Z-DNA is stabilized by negative supercoiling through transcription[245–247], however, the biological significance of Z-DNA is not fully understood. Previous studies have shown that Z-DNA prevents nucleosome formation[248] and can regulate gene expression[249–252]. Moreover, increased levels of Z-DNA were observed in lupus erythematosus[253] and Alzheimer's patients[254].

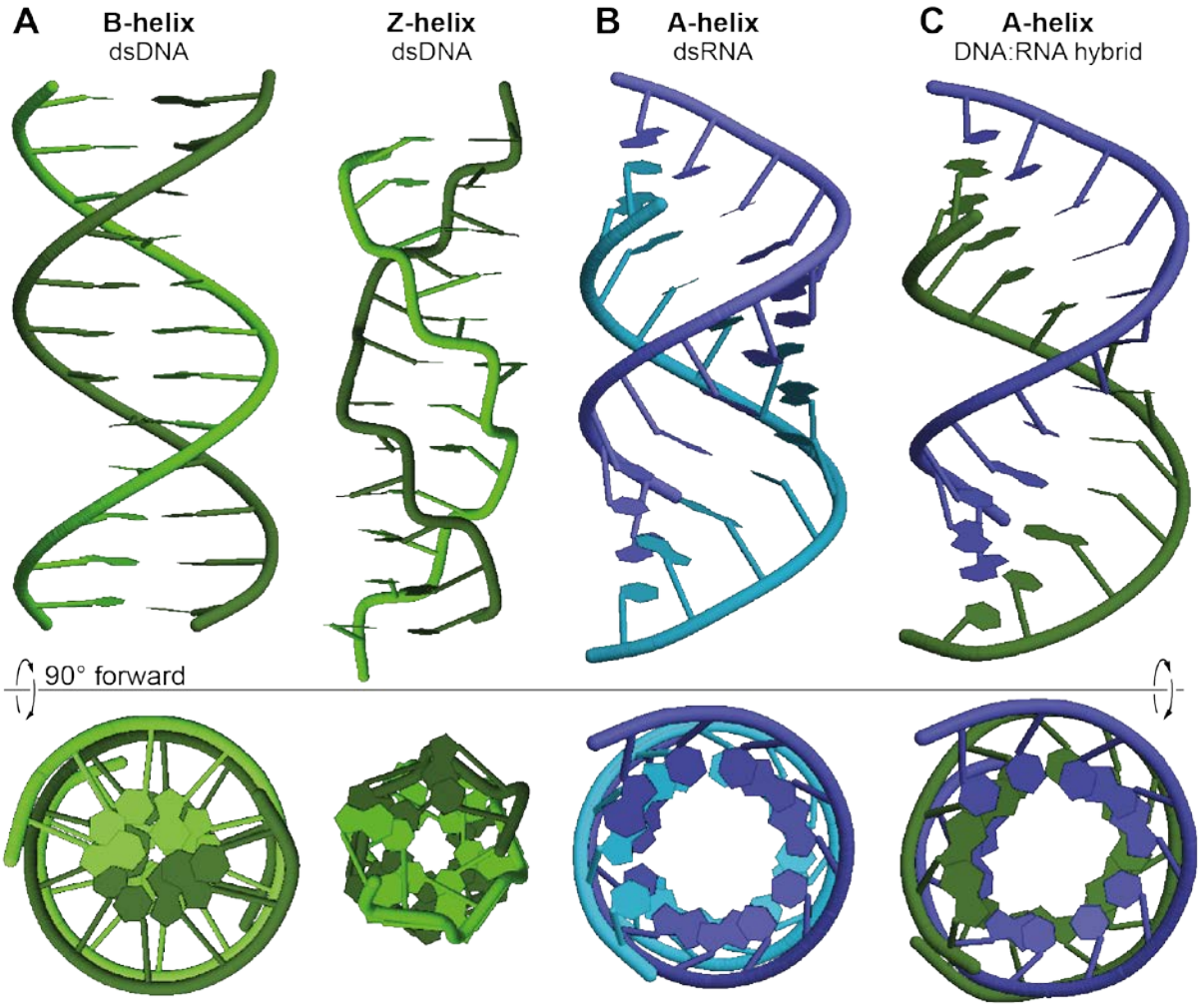


Figure 5-4: Conformation of nucleic acid helixes
Potential helix conformations of nucleic acid double strands. Top row represents vertical view of helix and bottom row represents top view after tilting helix by 90°. **A)** Left: Canonical right-handed dsDNA B-helix. Sequence: $(ATCG)_3$. Right: Left-handed Z-helix formed in sequences containing purine dinucleotide repeats. Sequence $(GC)_6$. **B)** Canonical dsRNA A-helix. Sequence: $(AUCG)_3$ **C)** DNA:RNA hybrids most frequently form A-helixes similar to dsRNA. DNA-Sequence: GAATCAGGTGTC; (NDB ID:4WKJ[255]). DNA: green; RNA: blue. Based on Heinemann and Roske[243]. Models were generated with Web 3DNA 2.0 and color adjusted with Adobe Photoshop.

(ii) dsRNA, in contrast to dsDNA, usually establishes A-form helixes. Here, the right-handed helix turn contains 12 bp (Figure 5-4B). ADARs can bind these A-helixes via its dsRNA binding domain. However, this domain is also forming sequence dependent hydrogen bonds with the substrate, indicating that the interaction is not only facilitated by the conformation of the dsRNA[256]. Since the Zα domain of ADAR1 is required for efficient processing of dsRNA, some sequence portions of dsRNA targets likely form Z-helixes[257].

DNA:RNA hybrids, like dsRNA, mostly form A-helixes containing 12 bp per helix turn[255,258–260] (Figure 5-4C). This structural similarity between dsRNA and DNA:RNA hybrids could enable binding and editing by ADARs.

5.1.5. Aim

Recent publications as well as data from this lab show that genomic dI is established by ADAR proteins acting on the DNA portion of DNA:RNA hybrids. This suggests that dI could have yet unknown roles in regulation of specific R-loops. As there is no data on the global distribution of dI, I planned to develop a dI sequencing method that allows mapping of genomic dI in different cell types.

I explored an antibody- and an EndoV-based method to enrich dI-containing gDNA for downstream analysis. Efficient dI enrichment with these methods could be monitored by LC-MS/MS or qPCR detection of dI containing reference sequences.

For detection of dI by NGS, a sequencing method would be required that allows reliable amplification of dI containing DNA. The identification of A>G transition sites, associated with the altered base-pairing of dI compared to dA, could enable mapping of dI with base resolution. With this new method, I planned to map genomic dI in different cell types. Subsequently, potential dI sites could be monitored after manipulation of ADAR or dI-repair enzymes, MPG and EndoV.

A dI sequencing method might reveal specific R-loops containing dI, dI sequence context and dI sites outside of R-loops. Finally, dI mapping could be an important tool to explore proteins involved in dI-formation and -removal and help to identify functional roles of dI in gDNA.

5.2. Results

5.2.1. dI-antibody weakly enriches genomic dI

Genomic dI levels are very low and specific enrichment steps are required for downstream analysis like NGS. Therefore, I validated a commercial inosine antibody that could be used to enrich genomic dI or to quantify dI complementary to the LC-MS/MS detection. In dot blot experiments the antibody was specifically detecting dI containing PCR amplicons, without cross reactivity towards other base modifications (Figure 5-5A). Next, I tested the sensitivity of the antibody by gradually diluting dI-containing and unmodified PCR amplicons. To estimate the dI content of each dilution step relative dI levels in the modified and unmodified amplicons were quantified by LC-MS/MS. The limit of quantification by dI dot blot was reached upon 125x dilution of the dI-amplicon, which corresponds to 0.02% dI/dA according to the LC-MS/MS quantification. This is more than 100x higher than the expected genomic dI level in the range of 0.0005% dI/dA (Figure 5-5B). Therefore, the dI dot blot is not suitable to quantify genomic dI levels. Nevertheless, I established a DNA immunoprecipitation (DIP) protocol to evaluate if the antibody is sufficient to enrich genomic dI. The DIP was performed on denatured ssDNA with dI- and control spike-ins added to the purified gDNA. After the DIP the enrichment efficiency was estimated by detecting the spike-in sequences via qPCR. The input recovery of the dI spike-in was around 30% compared to 0.1% of the control sequence (Figure 5-5C). However, when performing the DIP only with gDNA without spike-in, the dI enrichment in the DIP sample was only 5x over the input material when measured by LC-MS/MS (Figure 5-5D). Notably, the dI spike-in contained three dI sites per 120 bp, and hence, has much higher relative dI levels compared to gDNA. Overall, the dI antibody was not sensitive enough to quantify genomic dI and only allowed for weak enrichment of dI by DIP. Therefore, I explored other methods to enrich genomic dI.

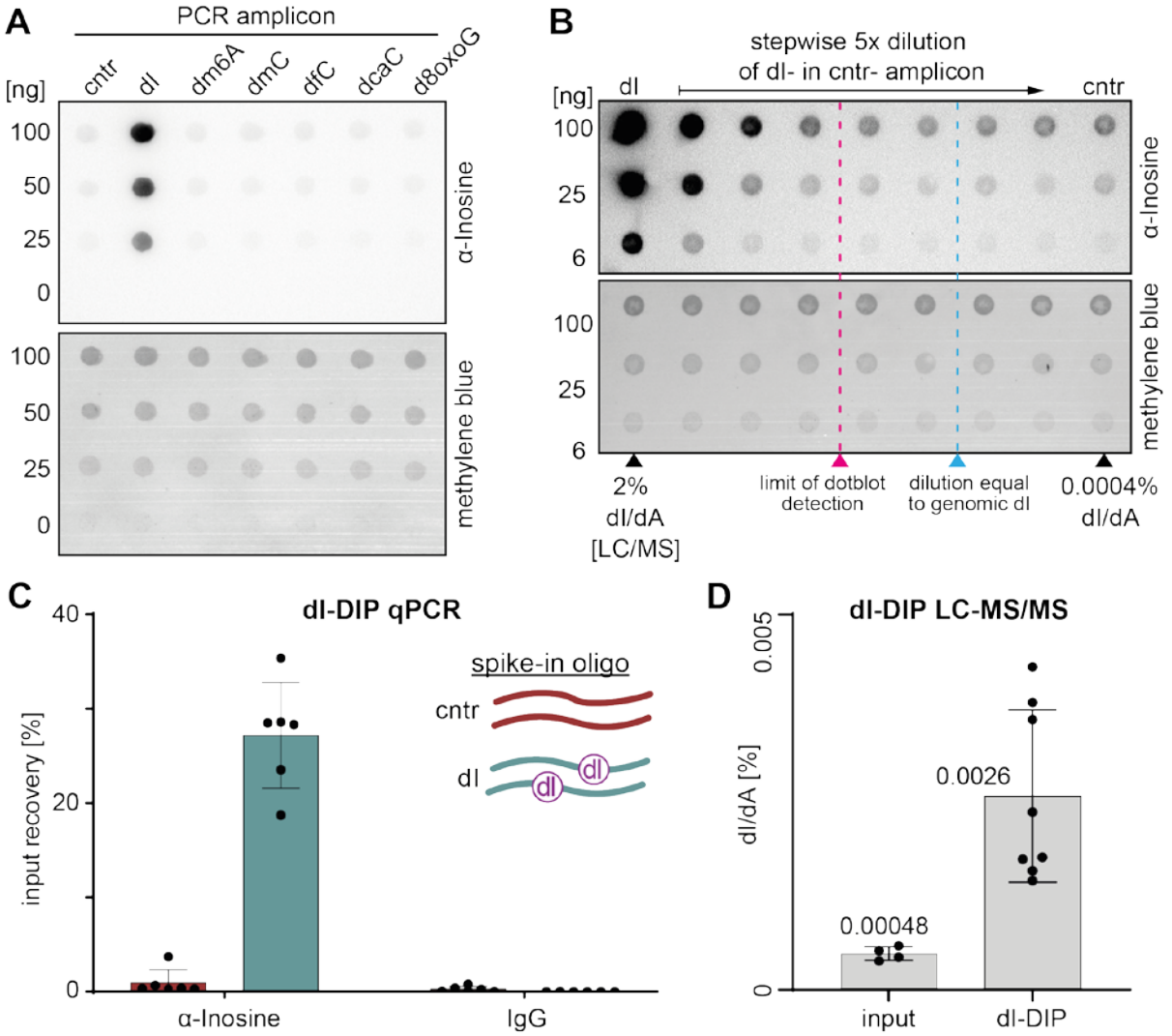


Figure 5-5: Inosine antibody weakly enriches genomic dI
A) dI dot blot of PCR amplicons containing indicated modified nucleotides. Methylene blue staining shows equal loading. **B)** dI dot blot of dI amplicon titrated in unmodified control amplicon. 5x dilution steps. dI content of amplicons was measured by LC-MS/MS to estimate dI of each dilution step. Red line indicates limit of antibody detection; Blue line indicates dilution step that should correspond to genomic dI – between 1:3000 and 1:15000. Methylene blue staining shows equal loading. **C)** dI-DIP qPCR on ss gDNA with unmodified- and dI- spike-in sequences. Spike-in sequences are detected by qPCR. Data are shown as mean ±SD. **D)** Relative dI levels determined by stable isotope dilution LC-MS/MS in input (n=4) and DIP (n=8) HEK293T samples. dI-DIP was performed without spike-ins to avoid detection of ectopic dI. Data are shown as mean ±SD. LC-MS/MS performed by [redacted].

5.2.2. dI EndonucleaseV–enrichment (dIVe) enriches genomic dI

As an alternative to the antibody-based dI enrichment, I adapted an EndoV-based pulldown protocol that was previously used to enrich rI containing mRNA. Even though the absolute levels of dI in a cell are around 5-15x lower than rI[64], I reasoned that a similar approach could be used on gDNA. By replacing Mg^{2+} with Ca^{2+} in the EndoV reaction, the repair enzyme is rendered inactive and only binds inosine without nicking the target sequence. (Figure 5-6A).

To examine if the nicking activity was abolished after Mg^{2+} replacement, I performed an EndoV activity assay with or without Mg^{2+}. The nicking activity on dI- and control-PCR amplicons was analyzed by running and quantifying a denaturing DNA polyacrylamide gel. As expected, EndoV nicked dI containing DNA only when supplemented with Mg^{2+} but not when supplemented with Ca^{2+} (Figure 5-6B).

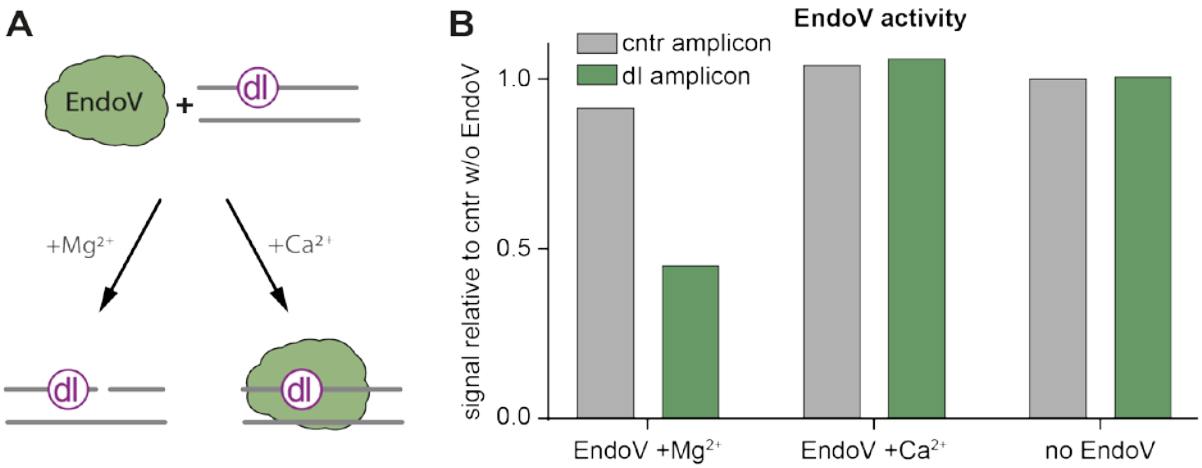


Figure 5-6: EndonucleaseV nicking requires Mg2+
A) Endonuclease V nicks 3' of dI in presence of Mg2+. Substitution of Mg2+ with Ca2+ impedes nicking activity but allows binding to the target sequence. B) EndoV activity assay. dI containing PCR amplicon was incubated with EndoV in indicated reaction buffers. DNA was purified and analyzed on denaturing polyacrylamide gel to detect nicked material. Signal was quantified using ImageJ and normalized to non-digested control sample.

Next, I established a pulldown protocol to enrich genomic dI: dI-EndonucleaseV enrichment (dIVe) (Figure 5-7A). A detailed protocol of dIVe is provided in the methods section. In short, gDNA is purified and RNA is removed by RNase A and RNase I treatment. DNA is fragmented to an average size of 200-300 bp and the sample is incubated with MBP-EndoV. The DNA-EndoV complexes are subsequently purified using magnetic anti-MBP beads. All steps are performed in the presence of the deaminase inhibitor pentostatin to block aberrant deamination by contaminating deaminases. As a reference, I included modified and unmodified spike-in sequences, which would allow direct qPCR readout of the enriched sequences. Compared to the published rI enrichment protocol, I reduced the amount of EndoV enzyme 50x and included additional steps for blocking of the magnetic beads in order to reduce unspecific binding.

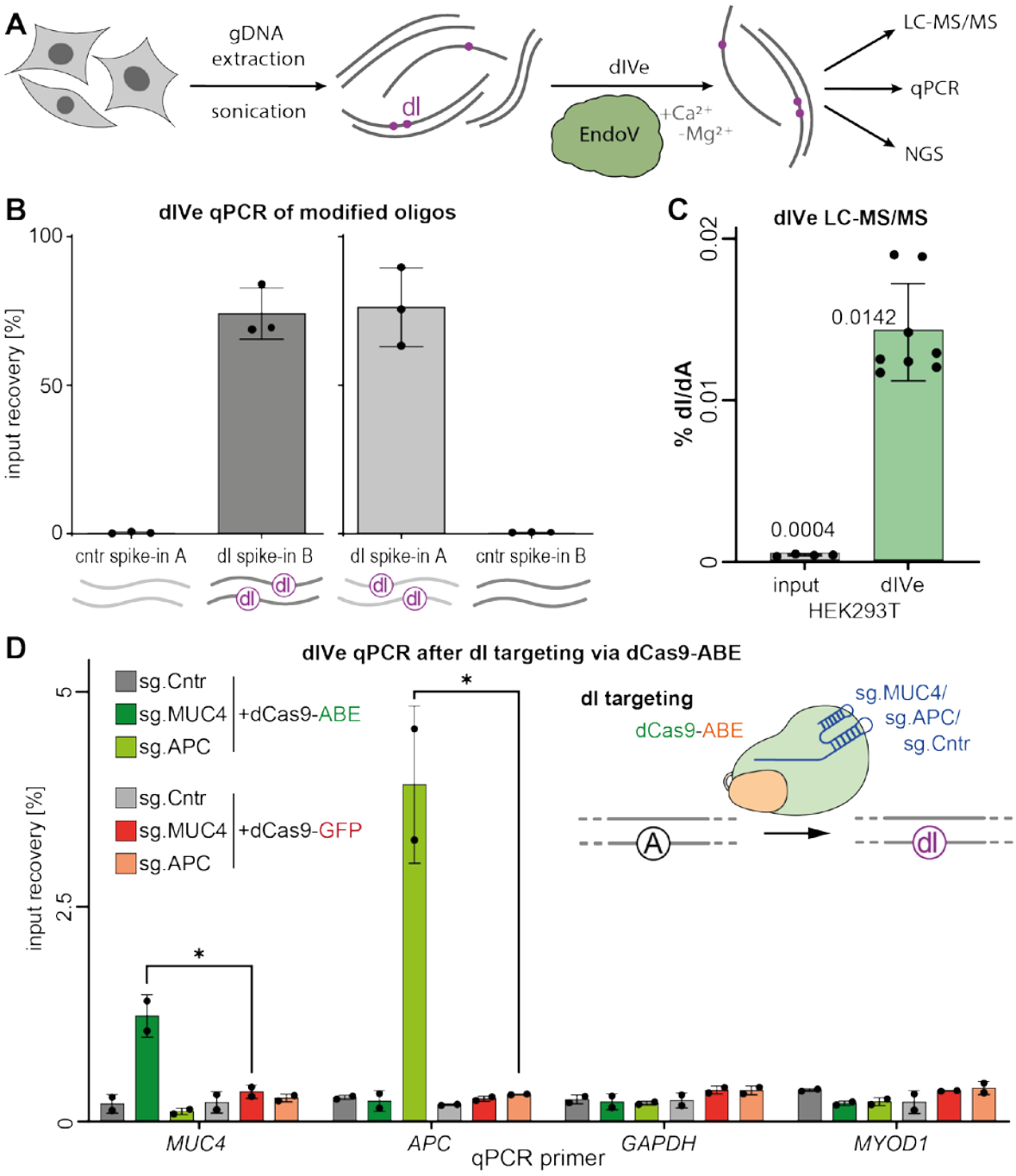


Figure 5-7: dIVe enriches genomic dI:
A) dIVe scheme. gDNA is extracted and fragmented and MBP-EndoV binding is performed in the absence of Mg^{2+}. EndoV is captured by anti-MBP magnetic beads. The purified material can then be investigated by qPCR, LC-MS/MS or NGS. **B)** Meta-analysis of 6 experiments showing input recovery of spike-in sequences in dIVe qPCR. Graphs represent 2x 3 independent experiments each containing multiple samples with unmodified spike-in A and modified spike-in B or vice versa. Each dot represents the average spike-in enrichment per experiment (n≥3). **C)** Relative dI levels determined by stable isotope dilution LC-MS/MS in input (n=4) and dIVe (n=8) HEK293T samples. dIVe pulldown was performed without spike-in sequences to avoid detection of ectopic dI. Data are shown as mean ±SD. LC-MS/MS performed by [redacted]. **D)** dIVe qPCR after dI-targeting in HEK293T. Cells were transfected with dCas9-ABE or -GFP and the indicated sgRNA. dIVe was performed as described before and enrichment of genomic regions was detected by indicated qPCR primers. Data are shown as mean ±SD. * = $p<0.05$, according to t-test.

In dIVe qPCR the input recovery of an unmodified spike-in-A was around 0.05% compared to 75% input recovery of the dI containing spike-in-B. This enrichment was only dependent on dI, as exchanging the dI content between the spike-ins completely reversed the input recoveries (Figure 5-7B). When measuring the enrichment of genomic dI by LC-MS/MS, I observed up to 35x more dI/dA in dIVe compared to the input sample, making dIVe much more efficient than the dI-DIP (Figure 5-7C, Figure 5-5D). As a positive control for genomic dI, independent of the spike-ins, I employed a targeted deamination system, using dCas9 fused to an adenosine base editor (ABE)[261]. Notably, the deamination is specifically occurring in the displaced ssDNA of the dCas9-induced R-loop at the sgRNA target site[91]. Due to the precision of the dCas9-ABE, this approach is frequently used to introduce A>G transitions via dI incorporation.

I used the system to introduce dI at specific loci and tested if the increased dI levels were detectable by dIVe qPCR. Targeting ABE to the *MUC4* or *APC* locus resulted in a specific increase of the dIVe qPCR signal, respectively. In contrast, control regions, like *GAPDH* or *MYOD1*, were not affected and targeting of dCas9-GFP did not increase the dIVe-qPCR signal at any of the tested loci (Figure 5-7D). Taken together, dIVe allows enrichment of dI containing gDNA and further analysis by qPCR and LC-MS/MS.

5.2.3. dIVe signal is sensitive to WGA and MPG treatment

I further validated the dIVe method by investigating if the enrichment could be reversed by *in vitro* removal of genomic dI. First, whole genome amplification (WGA) is a method that amplifies the entire genome from small amounts of gDNA. The amplified product is a direct copy of the genome that does not contain site-specific DNA modifications, as the polymerase reaction only contains dATP, dCTP, dGTP and dTTP. After introducing dI at the *MUC4* and *APC* locus WGA was performed. The specific dIVe enrichment at both deaminated loci was completely lost upon WGA (Figure 5-8A). Notably, the input recovery of the control loci *GAPDH* and *MYOD1* was increased in the WGA samples, indicating that the relative dI level might be elevated through WGA. dI quantification by LC-MS/MS showed that the relative levels of dI/dA were almost 10x increased in the WGA compared to the gDNA sample (Figure 5-8B). This could be related to the alkaline denaturing performed during the WGA and the random integration of dITPs that might contaminate the WGA reaction.

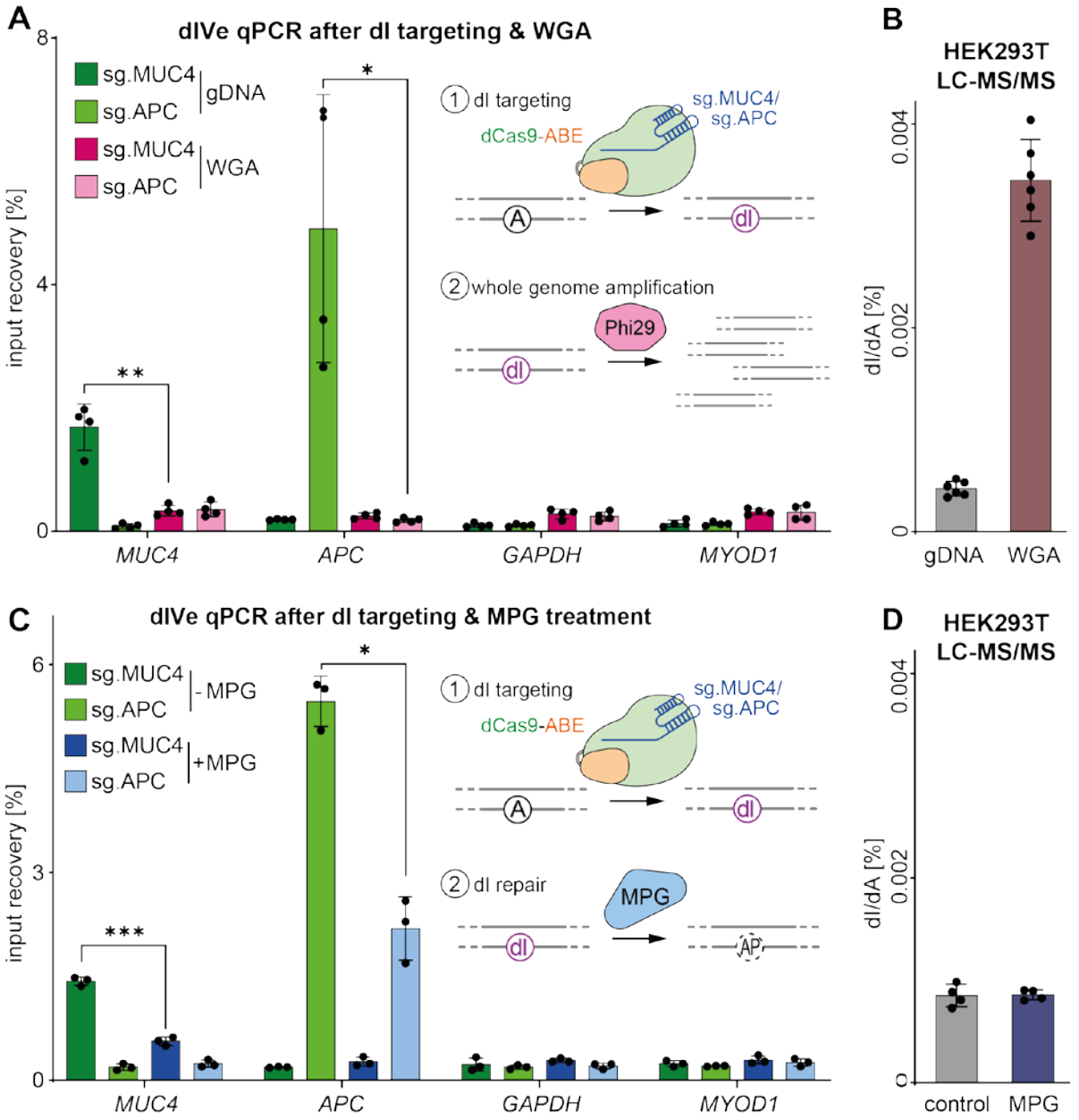


Figure 5-8: dIVe signal is reduced upon removal of dI

A) dIVe qPCR after dI-targeting and whole genome amplification (WGA) in HEK293T. Cells were transfected with dCas9-ABE and the indicated sgRNA. WGA was performed on 20 ng purified gDNA. gDNA and WGA were used for dIVe and enrichment of genomic regions was detected by indicated qPCR primers. **B)** Relative dI levels in HEK293T gDNA and WGA determined by stable isotope dilution LC-MS/MS. Independent experiment (no transfection). LC-MS/MS measurement was performed by [redacted]. Data are shown as mean ±SD. **C)** dIVe qPCR after dI-targeting and *in vitro* MPG repair in HEK293T cells. Cells were transfected with dCas9-ABE and the indicated sgRNA. Purified gDNA was treated with MPG or as a control and then subjected to dIVe qPCR. All data are shown as mean ±SD **D)** Relative dI levels in HEK293T gDNA after MPG treatment, determined by stable isotope dilution LC-MS/MS. gDNA was incubated with *E. coli* MPG for 1 h at 37 °C without enzyme (control) or with MPG. LC-MS/MS measurement was performed by [redacted]. Data are presented as mean ±SD. **A/C)** * = p<0.05, ** = p<0.01, *** = p<0.001, according to paired t-test.

Second, *in vitro* repair of genomic dI was performed using recombinant *E. coli* MPG. The repair reaction reduced the dIVe qPCR signal after dCas9-ABE targeting by around 60% (Figure 5-8C). Increasing the amount of MPG or the reaction time did not further reduce the dIVe signal. Instead, it resulted in overall higher input recovery in all tested regions, including the controls *GAPDH* and *MYOD1* (not shown). This could be related to unspecific deamination of the gDNA due to the extended incubation time. The addition of more recombinant MPG might also increase the contamination with deaminases and hence exhaust the deaminase inhibitor pentostatin. When quantifying the global dI levels in control and MPG treated samples by LC-MS/MS no significant change was observed (Figure 5-8D). This indicates that the MPG treatment is sufficient to reduce local dI enrichment, as seen in the qPCR, but has no detectable effect on the global dI pool.

Taken together, WGA and MPG treatment expectedly reduced the dI-specific signal, and hence, confirmed the specificity of the dIVe. In principle, both treatments can serve for validation of yet to discover dI sites in a genome (see below). However, the global increase of dI/dA in whole genome amplified DNA might limit the usefulness of WGA for certain applications; e.g. as a control for dIVe Sequencing.

5.2.4. Establish dIVe-sequencing protocol

As dIVe allowed specific enrichment of genomic dI, I planned to sequence the enriched material in order to identify genomic dI sites. This would require the generation of dIVe-sequencing (dIVe-seq) libraries through adapter ligation and subsequent amplification of the DNA fragments enriched by dIVe. However, while establishing the dI dot blot and dIVe method I observed that certain polymerases failed to incorporate dITPs during PCR or did not amplify dI containing material. This was further evaluated to ensure correct library preparation for dIVe-seq. I designed a 100 bp spike-in sequence as a unmodified, hemi- or fully modified double strand to test different DNA polymerases. The high fidelity polymerase Q5 (NEB) with 3´ to 5´ exonuclease activity failed to amplify the fully modified template and only amplified 50% of the hemi-modified target. In contrast, a standard *Taq* polymerase or the Phusion-U (PhuU) polymerase, which is engineered to accept dU (i.e. deaminated cytosine), did amplify all templates equally (Figure 5-9A).

Likewise, I tested if dI in the context of gDNA would block the amplification by Q5 or if the effect was specific for the highly modified dI spike-ins. Therefore, I targeted dCas9-ABE to the established *MUC4* and *APC* regions and measured the recovery of the loci in dIVe qPCR, using Q5 or PhuU for the amplification. Notably, only the *APC*-specific qPCR primers span the ABE target site, while the *MUC4* primers are both 5’ of the expected deamination site (Figure 5-9B, cartoons below graph). The dIVe Q5-qPCR signal at the deaminated *APC* locus was

decreased to 50% of the dIVe PhuU-qPCR. As the ABE is specifically deaminating one of the DNA strands between the selected *APC* primers the Q5 can only amplify one of the strands, resulting in the 50% reduction. In contrast, detection of the *MUC4* locus was not affected, as the primers are not spanning the site of deamination (Figure 5-9B).

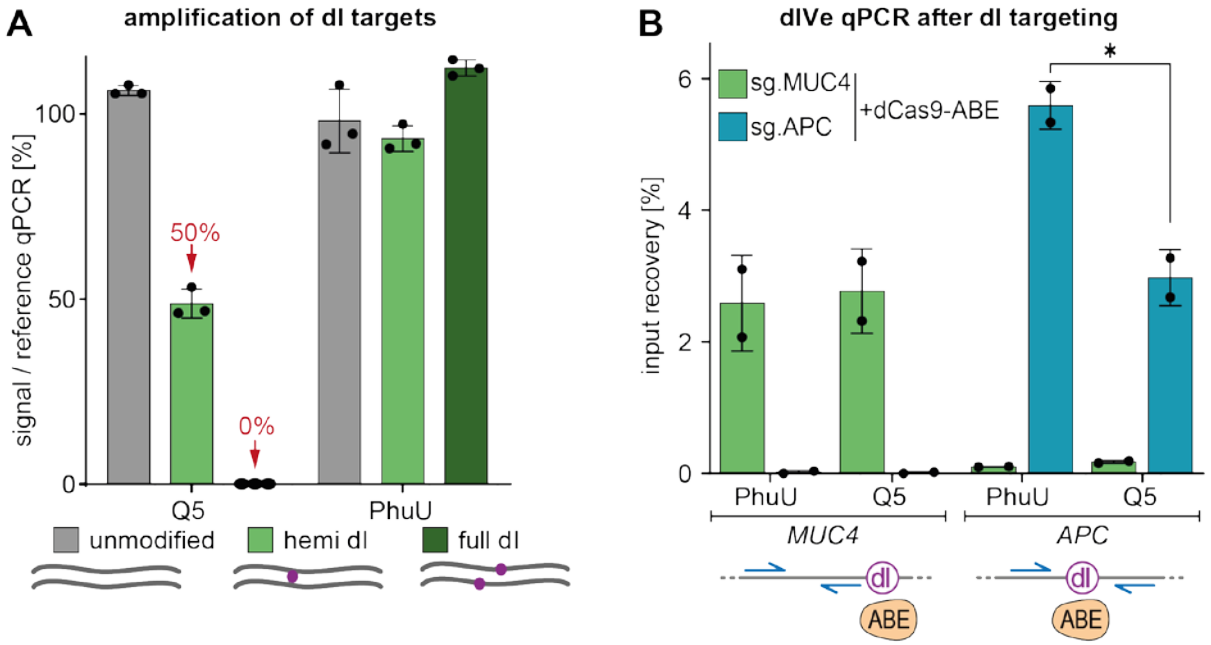


Figure 5-9: dI containing DNA is amplified by Phusion-U
A) qPCR detection of modified and control sequence using Q5 or Phusion U (PhuU) polymerase and EvaGreen dye. As a reference, the same qPCR was performed with a commercial SYBRgreen qPCR mastermix from Roche, containing a standard *Taq* polymerase. Data are shown as mean ±SD. **B)** EvaGreen qPCR with PhuU and Q5 after dI-targeting to *MUC4* and *APC* locus and dIVe in HEK293T. qPCR was performed with Q5 and PhuU. Cartoon is indicating deamination site relative to primer position (blue arrow) in target locus. Data are shown as mean ±SD. * = p<0.05, according to t-test.

Next, the library preparation for dIVe-seq was optimized with the Genomics Core Facility (CF) at IMB using 100 bp dI and control spike-in sequences as starting material. Two commercial high fidelity polymerases, Q5 and Phusion, were tested against the PhuU and a standard *Taq* polymerase. All polymerases successfully produced libraries from the unmodified control sequence. Expectedly, Q5 and Phusion failed to amplify the dI sequence, whereas the use of PhuU and *Taq* polymerase resulted in proper libraries for sequencing (Figure 5-10A, Phusion and *Taq* polymerase not shown).

As described before, dI favors base-pairing with cytosine (Figure 5-1A). Hence, during sequencing dI should be read at least partially like G. Indeed, when sequencing the initial libraries generated using PhuU and *Taq* polymerase, dI was mostly read as G. The sequence logo generated from all mapped reads showed that dI was detected as G in ≥95% of the reads for the PhuU (Figure 5-10B). The library generated with the *Taq* polymerase showed similar transitions levels of ~90% (not shown).

Taken together, PhuU and *Taq* polymerase facilitated the generation of sequencing libraries from dI containing material. Moreover, dI was specifically detected as guanosine during sequencing of dI spike-ins. For the following dIVe-seq experiments, the library preparation was performed with the PhuU polymerase.

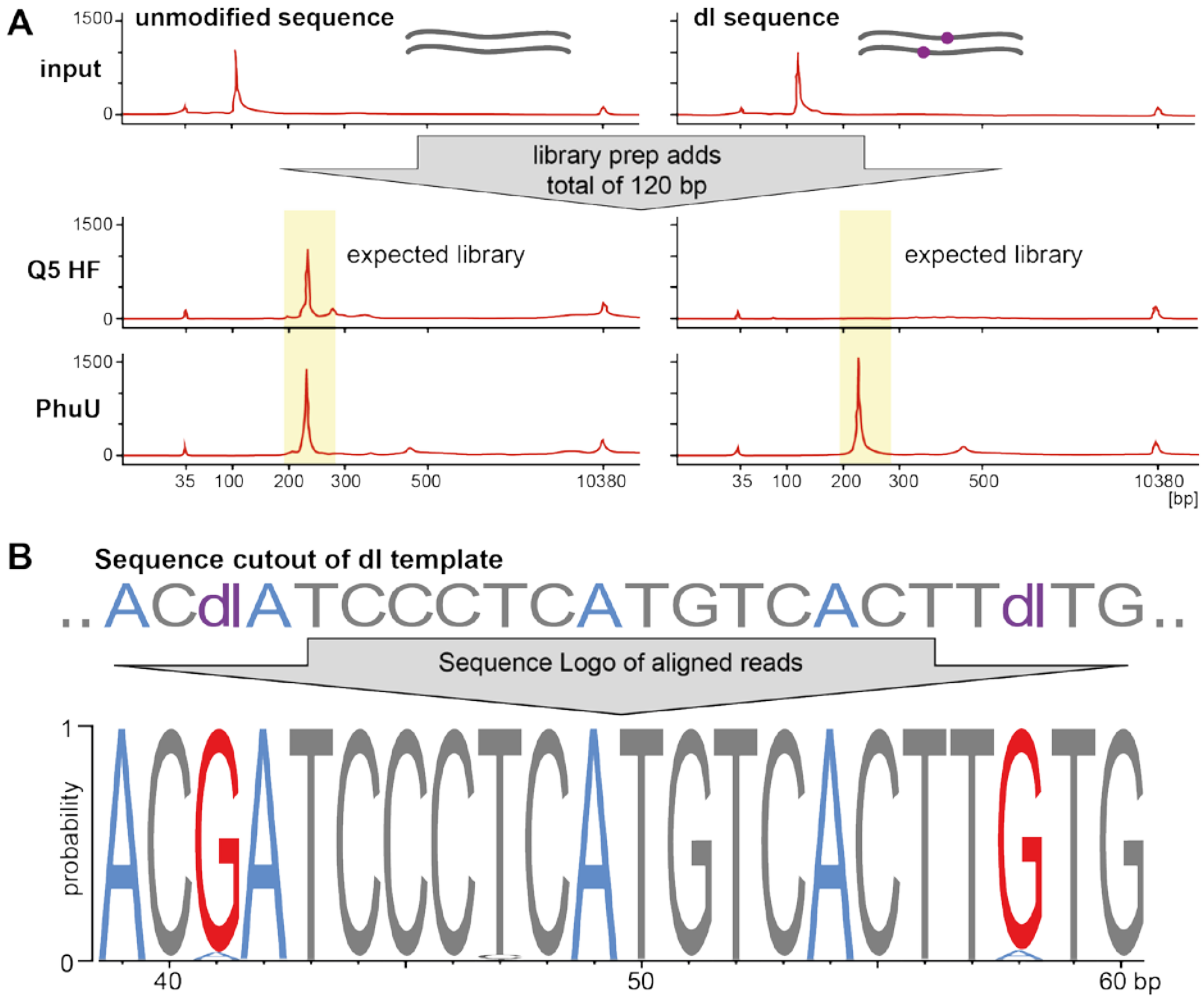


Figure 5-10: dI is perceived as G during dIVe-sequencing
A) Bioanalyzer profile after library preparation according to NEBNext Ultra II protocol using the supplied Q5 or the dI-permissive PhuU polymerase. Unmodified or dI containing sequence was used as a template. Genomic CF at IMB performed library preparation and Bioanalyzer analysis **B)** Sequence logo generated from reads mapping to a dI containing spike-in sequence. Original dI position (purple) is read as guanosine (red). Some reads are showing adenosine (small A below the large G). NGS data was generated with Illumina MiSeq platform after library preparation with PhuU polymerase. Library preparation was performed by IMB genomics CF. [redacted] performed initial analysis of the MiSeq data.

5.2.5. dIVe signal is enriched in $(TG)_n$ simple repeats in HEK and MEF cells

Next, I performed dIVe sequencing on gDNA of HEK293T and mouse embryonic fibroblast (MEF) cells. In addition to dIVe samples the input material and a mock sample were sequenced. For the mock controls, no EndoV was added to the sample, so the magnetic beads used to capture EndoV should only capture unspecifically bound sequences. The mock pulldown did not yield detectable amounts of DNA, hence, the respective samples were pooled and concentrated before library preparation. Consequently, the mock sample is not directly comparable to the input or dIVe samples and was not used for peak calling during the NGS analysis. However, the mock sample can still provide a general indication of unspecific enrichment related to the beads when observed in the genome browser. I mapped the dIVe-seq reads to hg38 or mm10 reference genomes and called peaks in dIVe samples over the input. The dIVe-seq was performed in biological quadruplicates and peaks that were common in all four replicates were used for downstream analysis. In HEK293T 371 common peaks were detected, whereas in MEF cells 931 peaks were common between replicates. Peaks in both cell lines were clustering in subtelomeric regions (Figure 5-11).

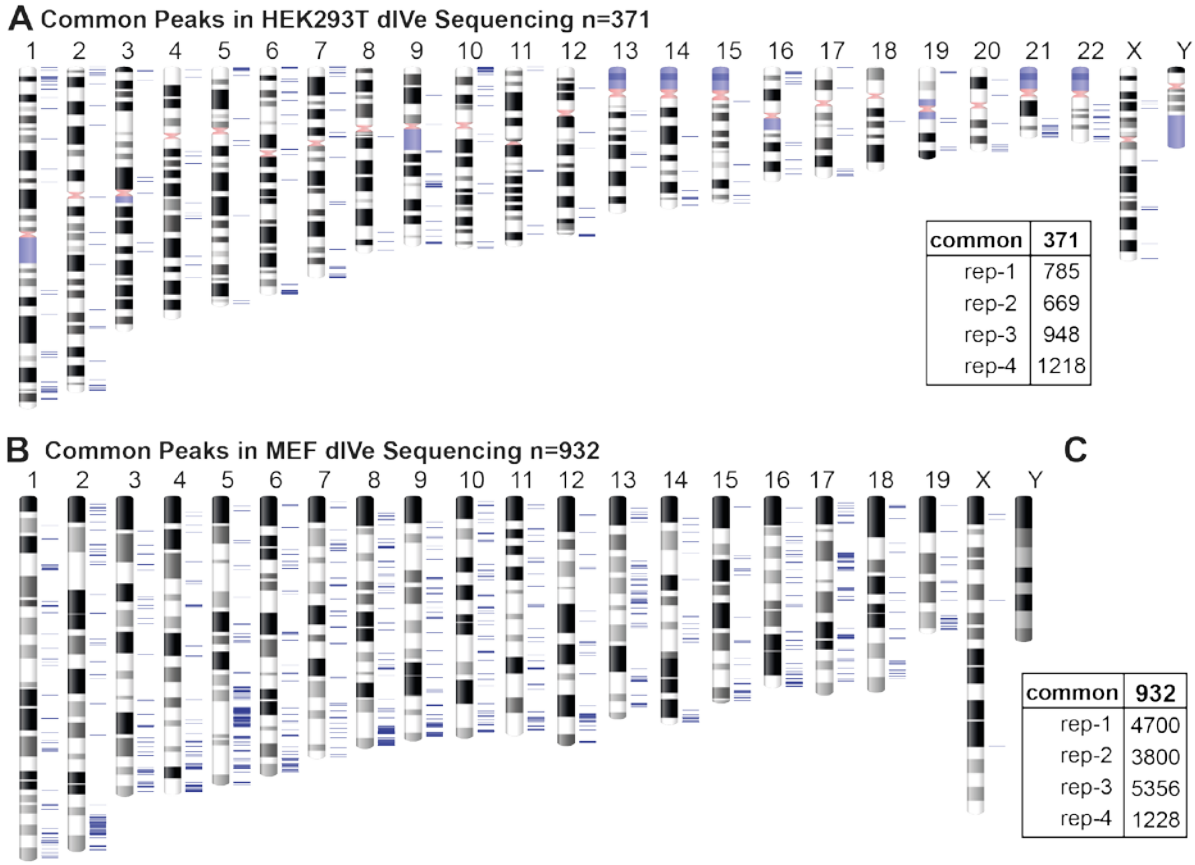


Figure 5-11: HEK293T and MEF dIVe-sequencing peaks cluster in subtelomeric regions
Ideogram representation of common dIVe peaks in **A)** HEK293T and **B)** MEF cells. Human centromeres are shown in red. Mouse telocentric chromosomes are aligned with the centromere at the top. Ideograms were generated with Genome Decoration Page[262] using hg38 or mm10 as reference assemblies. Peak numbers in individual replicates are shown in box

Next, I analyzed the overlap of the peaks with annotated gene features to identify potential functional implications of the dIVe peaks. In HEK293T the peak annotation revealed only minor differences compared to the expected random distribution. Notably, promoter and transcription start sites (TSS) as well as Transcription termination sites (TTS) were underrepresented (Figure 5-12A). Since dI is mutagenic, it seems logical that it is less abundant in these functionally important regions.

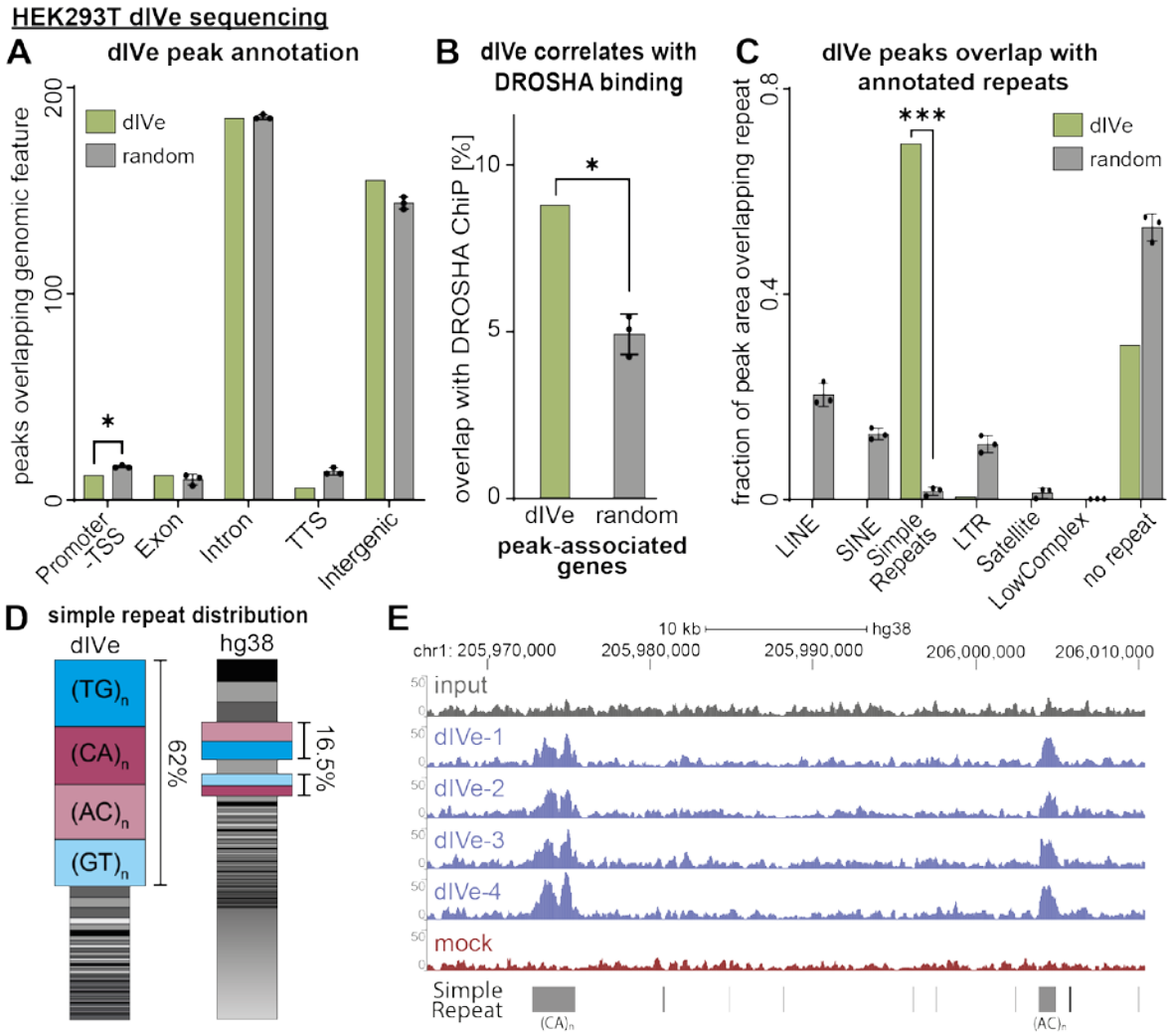


Figure 5-12: dIVe Sequencing peaks in HEK293T gDNA overlap with $(TG)_n/(CA)_n$ simple repeats
Analysis of dIVe-seq performed in HEK293T cells. **A)** dIVe-seq peak annotation with gene features (UCSC, hg38, GENCODE V39). TSS/TTS: transcription start/termination site. Common dIVe peaks (green), randomized peak set (grey). Peak randomization was performed in triplicates. **B)** Gene-set analysis of genes overlapping dIVe peaks. Genes in common dIVe peaks were extracted with AnnoMiner[263] and analyzed with EnrichR[264] to identify overlap with published gene-sets. As control, the analysis was performed with 3 randomized peak sets. A significant enrichment with a DROSHA ChiP dataset derived from HeLa cells[265] was observed **C)** dIVe-seq peak annotation with known repeats (UCSC, hg38, RepeatMasker). LINE/SINE: long/short interspersed nuclear element; LTR: long terminal repeat. Common dIVe peaks (green), randomized peak set (grey). Peak randomization was performed in triplicates. **A/B/C)** * = p<0.05, ** = p<0.01, *** = p<0.001, according to t-test. **D)** Distribution of simple repeats in dIVe peaks (left) and human genome (right). $(TG)_n$: blue; reverse complement $(CA)_n$: red; other repeat motifs shown in grey. **E)** UCSC genome browser of dIVe-seq tracks. Peaks overlap with annotated simple repeats.

As a large fraction of peaks overlapped with introns, I analyzed if the dIVe peaks enriched a specific group of genes. First, I identified the genes that overlap with dIVe peaks using AnnoMiner[263], a webtool for annotation of sequencing data. Subsequently, I analyzed the obtained list of genes using EnrichR, a tool that correlates a set of genes with publicly available gene-set libraries and identifies overlaps[266–268]. Interestingly, genes overlapping with the dIVe peaks showed an overlap with DROSHA associated genes. These DROSHA associated genes were previously identified by DROSHA chromatin immunoprecipitation (ChiP) sequencing in Hela cells[265]. As introduced before, DROSHA is involved in miRNA maturation and ADAR mediated A to I editing of RNA targets can block DROSHA processing[176]. Even though only 9% of the dIVe associated genes overlap with the DROSHA associated genes, this is significantly higher than the expected overlap with a randomized dataset (Figure 5-12B).

Interestingly, looking at the peak distribution over repetitive elements, 65% of all peaks overlapped with simple repeats. This is in stark contrast to the expected distribution of randomized peaks across repetitive elements. Notably, there was nearly no overlap of dIVe peaks with other repeat types (Figure 5-12C). Next, I analyzed which types of simple repeats were overlapping with the peaks and observed a strong enrichment of $(CA)_n$/$(AC)_n$/$(TG)_n$/$(GT)_n$ repeats. These dinucleotide repeats all represent variations of the same sequence, shifted by one position or reverse complemented (Figure 5-12D). The global analysis was further confirmed by zooming into single peaks using the UCSC genome browser. Peaks were specifically appearing over $(CA)_n$ simple repeats with only background signal in input or mock samples (Figure 5-12E).

Similar results were observed in the dIVe-seq in MEF cells (Figure 5-13). The common peaks occurred less frequent across important functional gene features, like promoter-TSS and exons. As seen in HEK293T dIVe samples, there was a strong enrichment of simple repeat sequences. Overall, more than 50% of the common peak area overlapped with annotated simple repeats. In more than 80% these simple repeats were $(TG)_n$ or the reverse $(CA)_n$ dinucleotide repeats (Figure 5-13C). In addition to simple repeats, MEF dIVe peaks also enriched satellite repeats. Finally, gene set enrichment analysis of the MEF dIVe peaks using AnnoMiner and EnrichR did not reveal notable overlaps with public gene-set libraries (not shown).

In summary, dIVe-seq analysis performed in HEK293T and MEF cells delivered similar results. dIVe peaks were enriched across $(TG)_n$/$(CA)_n$ simple repeat sequences, suggesting that dI is enriched in these dI nucleotides in human and mouse.

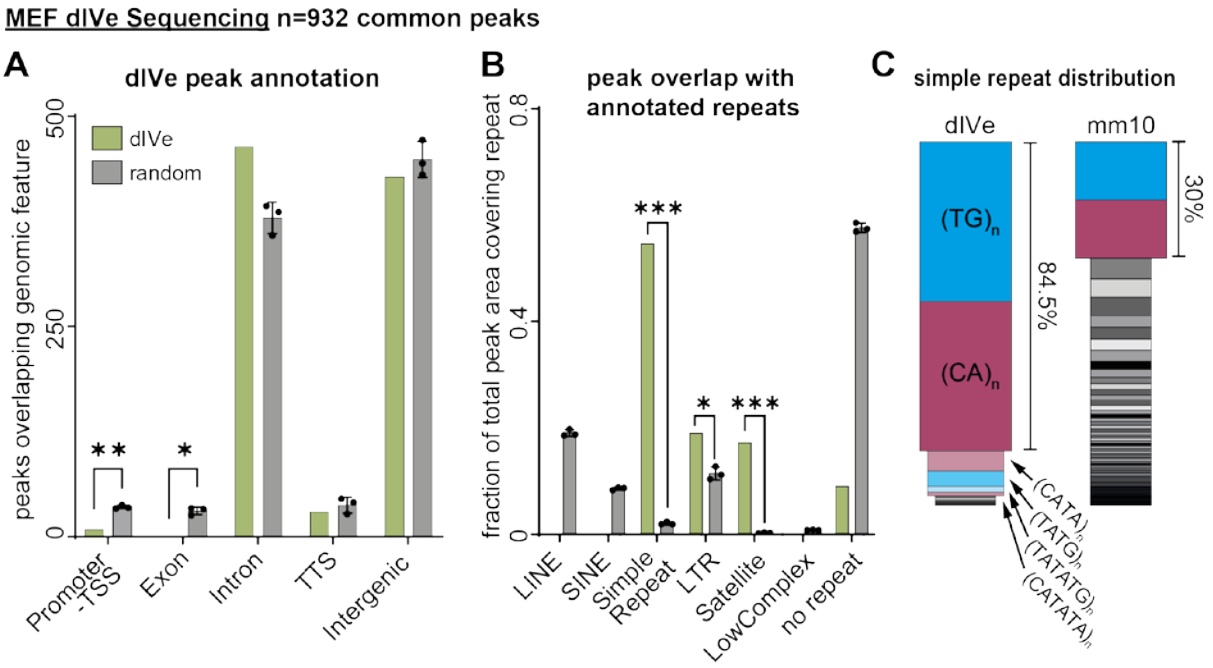


Figure 5-13: dIVe sequencing peaks in MEF cells overlap with $(TG)_n$/$(CA)_n$ simple repeats
dIVe sequencing analysis of MEF cells. **A)** dIVe-seq peak annotation with gene features (UCSC, mm10, GENCODE VM23). TSS/TTS: transcription start/termination site. Common dIVe peaks (green), randomized peak set (grey). Peak randomization was performed in triplicates. **B)** dIVe-seq peak annotation with known repeats (UCSC, mm10, RepeatMasker). LINE/SINE: long/short interspersed nuclear element; LTR: long terminal repeat. Common dIVe peaks (green), randomized peak set (grey). Peak randomization was performed in triplicates. **A/B)** * = p<0.05, ** = p<0.01, *** = p<0.001, according to t-test. **C)** Right) Distribution of simple repeats in dIVe peaks (left) and mouse genome (right). $(TG)_n$: blue; reverse complement $(CA)_n$: red; other repeat motifs shown in grey.

5.2.6. Manipulations of potential dI regulators are not affecting dIVe signal

For further experiments, I focused on the HEK293T cells, as they would allow easier manipulation by knockdown or overexpression of potential dI effectors. Based on the peaks from the dIVe-seq, I designed primers that would allow investigation of these potential dI-sites by qPCR. Since most of the peaks were located in repetitive regions, designing specific primers was not possible for the majority of the peaks. However, I managed to design 17 primer pairs with specific hybridization probes, targeting human dIVe peak sites (hdIp-01 to 17). Using these primers, I tested if the peak regions would be sensitive to manipulation of potential dI effectors or dI-removal.

First, I generated and overexpressed hyperactive ADAR1 isoforms p110 and p150. Enzymatic hyper-activity of the generated ADAR mutants was confirmed by LC-MS/MS analysis of rI levels in global RNA (not shown). However, there was no increase in the input recovery of the dIVe peak sites in dIVe qPCR with separate or combined overexpression. Against

expectations, combined overexpression of p150(+) and p110(+) seemed to reduce the signal, however, this tendency was also seen at the control locus *GAPDH* (Figure 5-14A, left). Therefore, I conclude that none of the tested regions was sensitive to enzymatic activity of ADAR1.

Second, I tested if the combined knockdown of the dI repair enzymes MPG and EndoV (siRepair) followed by hyperactive ADAR1p150 overexpression would result in increased dIVe qPCR signals. The knockdown efficiency was confirmed by RT-qPCR and overexpression was validated by GFP expression (not shown). Although I observed fluctuations in the input recovery of the dIVe qPCR targets, there was no significant increase in either of the peak regions (Figure 5-14A, right). Other gain and loss of function experiments were performed, including overexpression of inactive ADAR1p110/p150 and single knockdowns of ADAR1, EndoV, MPG or MSH2. However, none of these manipulations reproducibly affected the tested primer regions in dIVe qPCR (data not shown).

As the input recovery with the standard dIVe protocol was very low and often only 3-5x higher than the control loci *GAPDH* and *MYOD1*, I tested the dIVe protocol on ss gDNA. In this protocol, the gDNA was briefly heat denatured before the EndoV incubation. Surprisingly, I observed a strong increase in the input recovery of the tested dIVe peaks when performing ss-dIVe qPCR. Using denatured gDNA for dIVe resulted in 5x higher input recovery of the selected dI targets without affecting the control regions. However, also with the ss-dIVe protocol, the tested dI manipulations did not change the input recovery (Figure 5-14B).

In addition to the gain and loss of function experiments, I tested the previously established *in vitro* removal of dI by WGA or MPG treatment to reduce the signal at potential dI-target sites. The WGA did not reduce the signal, but rather increased the input recovery of all tested regions including negative controls GAPDH and MYOD1 (Figure 5-14C, magenta). This was observed earlier when establishing the WGA treatment after dCas9-ABE targeting (Figure 5-8). As described before, LC-MS/MS analysis before and after WGA revealed that the dI levels in the WGA samples were around 8x higher than in gDNA (Figure 5-8B). Hence, WGA might not be suitable to reduce the input recovery at the tested dIVe peak sites.

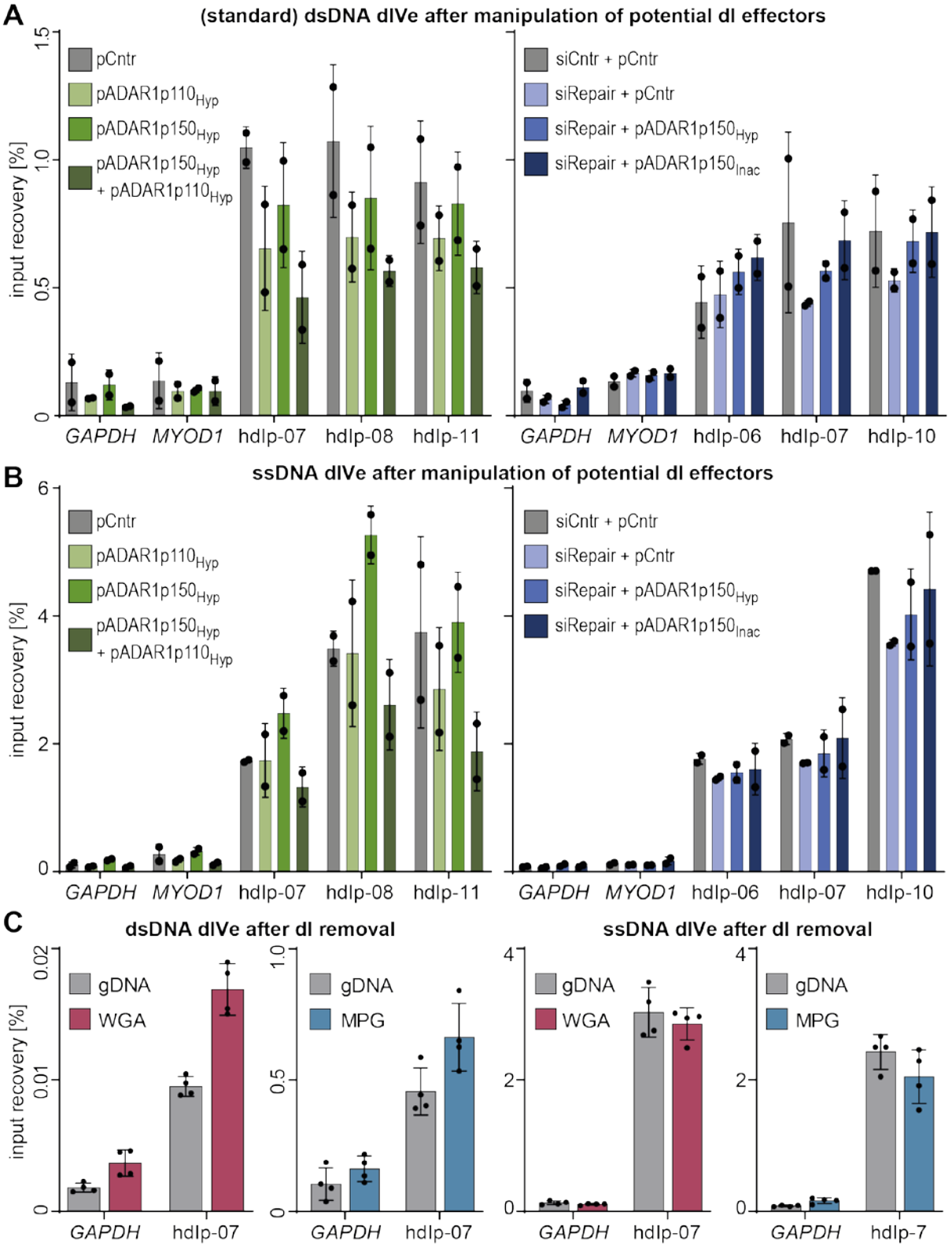


Figure 5-14: dIVe qPCR signal of selected targets is not affected by dI-manipulation
dIVe qPCR in HEK293T cells **A)** qPCR of standard (ds) dIVe or **B)** ss dIVe samples. Primers were designed based on common peaks derived from initial sequencing. Relative enrichment of specific genomic targets is given as percent input recovery. left: ADAR1 overexpression; right: combined EndoV and MPG knockdown (siRepair) followed by ADAR1 overexpression. hdIp: human deoxyinosine peak; hyp: hyperactive. **C)** ds-dIVe or ss-dIVe qPCR after dI removal by MPG treatment or whole genome amplification (WGA). All data are shown as mean ±SD.

Finally, I performed MPG treatment of gDNA to see if it would reduce the dIVe qPCR input recovery. The treatment was performed according to the optimized protocol that was previously reducing dIVe input recovery of deaminated *APC* and *MUC4* to around 60%. Surprisingly, MPG treatment did not reduce the input recovery of the dIVe peak-sites (Figure 5-14C, blue). This was independent of ds or ss-dIVe. In the ds-dIVe I also observed a slight increase of signal at the control loci.

Taken together, I did not detect changes in the dIVe signal upon manipulating ADAR1, EndoV, or MPG or upon *in vitro* removal of dI. The selected sites might either not contain dI or the qPCR might fail to detect them due to their repetitiveness or nicks and breaks associated with the dI sequences. However, the fact that neither WGA nor MPG reduced the dIVe signal significantly, rather suggests that the selected peaks were false positives. Therefore, I tried to identify real dI-peaks by analyzing the A>G mutation signature that should be associated with genomic dI.

5.2.7. dI Peaks are enriched in A>G transitions

As shown before, dI is read as G during NGS. I hypothesized that an increase in A>G transitions should be detectable in the dIVe-seq peaks. The following analysis was performed using the HEK293T dIVe-seq data. To monitor transition mutations, I screened the aligned reads for variants at a specific base position using VarScan, a tool for variant detection in NGS datasets[269]. The output file produces a list of all detected mutations, and specifies the mutation frequency as well as the read depth (i.e. base coverage) at each detected mutation site (Figure 5-15A). For the analysis, I employed a minimum coverage of 4 reads and quantified transitions in the dIVe and input samples, both in the peak area as well as over the entire genome. Subsequently, the mutation count in the dIVe sample was normalized to the respective amount in the input to monitor if a certain mutation was increased. To ensure that observed changes were specific to the peak, the dIVe/input ratio in the peak area was normalized further to the same ratio calculated from the entire genome. When combining all mutations independent of their frequency, there was no clear enrichment (not shown). For a more detailed analysis, I grouped the mutations according to the detected mutation frequencies. This resulted in four groups (I-IV) with different mutation frequencies: I (0-24.9%), II (25-49.9%), III (50-74.9%) and IV (75-99.9%). In group-I and -IV no increase of A>G transitions was observed. However, I detected a mild increase of A>G transitions in group-II and a stronger increase in group-III (Figure 5-15B). Notably, I also observed an enrichment of G>A transitions in these groups. The G>A transition could be linked to deamination of (methyl-) cytosine to T or dU, respectively. It is worth mentioning, that group-I contained the majority of all detected mutations, hence, effects that were observed in groups-II and-III were masked in the initial mutation analysis (not shown).

Next, I tested if a similar enrichment of A>G transitions could be detected in publicly available data. I employed public single nucleotide polymorphism (SNP) annotation files of the human reference genome (UCSC, hg38, ds153). When I analyzed the number of annotated SNPs in the dIVe peak areas, I detected a significant enrichment of A>G SNPs over the expected distribution in randomized intervals (Figure 5-15C).

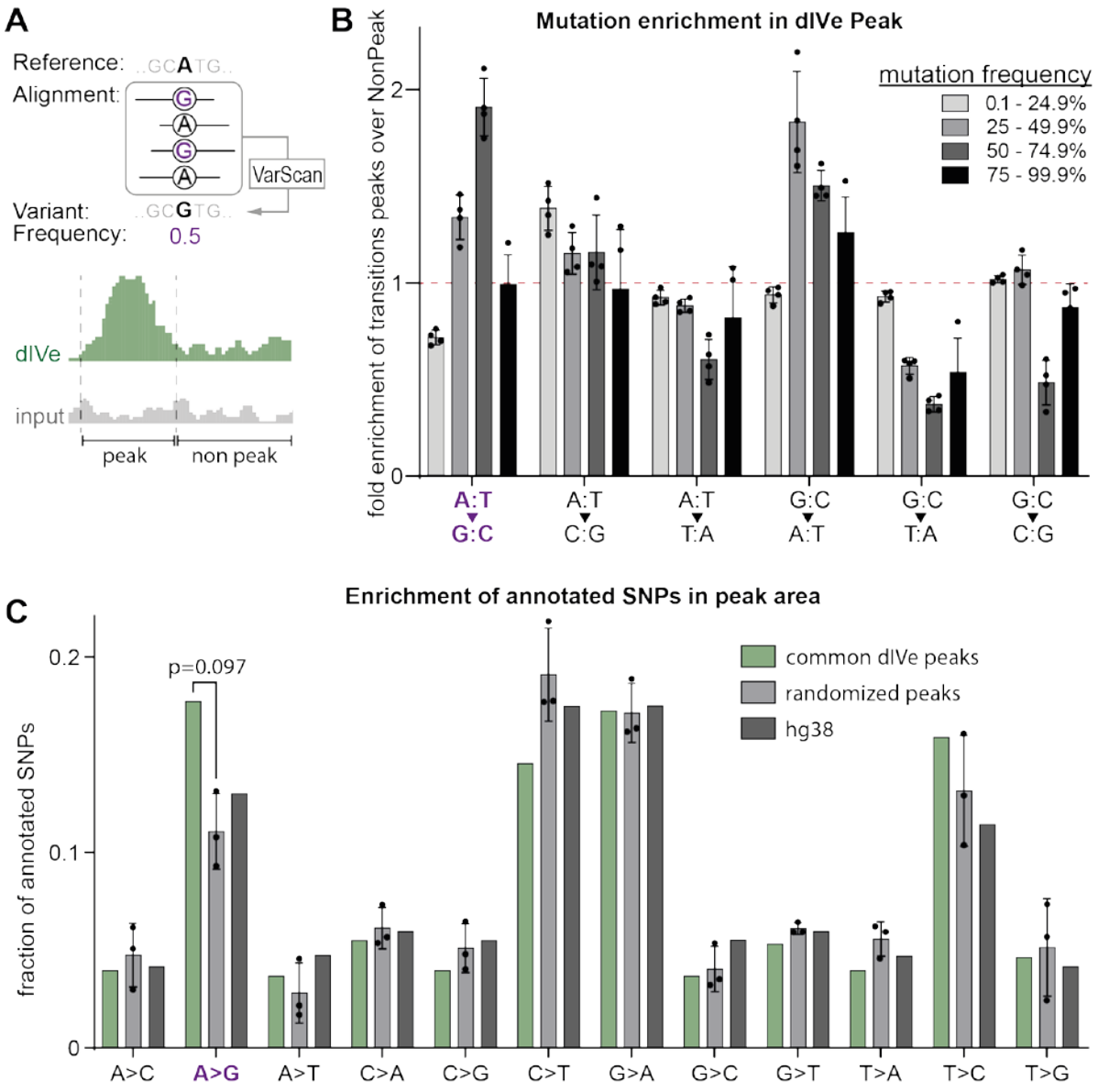


Figure 5-15: dIVe peaks are enriched in A>G transitions
Analysis of HEK293T dIVe-seq data **A)** Analysis scheme showing detection of point mutations. Aligned reads were used to generate pileup files and detect nucleotide variations using VarScan. Variants were detected in peak areas of dIVe and input samples as well as non-peak areas for comparison. Variants were grouped in quartiles according to their mutation frequency. A mutation frequency of 75% means that three quarters of all reads covering a specific base show the same point mutation **B)** Fold enrichment of indicated mutation in peak over non-peak area grouped by mutation frequency. dIVe sample was normalized to input sample (red line). Each dot represents a replicate. **C)** Relative occurrence of annotated single-nucleotide polymorphisms (SNP) in common dIVe peaks (green), randomized dIVe peaks (light grey) and human genome (dark grey). Peak randomization was performed in triplicates. SNP data from UCSC, hg38, db SNP153.

Overall, I observed an increase of A>G transitions in the reads covering the dIVe-seq peaks. This is confirmed by public datasets that contain significantly more A>G SNPs in these areas. Together this suggests, that at least a subset of the dIVe-seq peaks contain dI, which is detected as guanosine during sequencing.

5.2.8. dI related mutations occur in a (GTAT)/(ATAC) motif

Next, I investigated if the medium-frequency A>G transitions (group-I & -II) were enriched specific sequences. I first filtered A:T > G:C transitions that were detected in 3 out of 4 dIVe replicates with a mutation frequency between 25-75%. This resulted in 128 transition sites that were associated with 61 unique dIVe peaks (Figure 5-16A). I isolated a 10 bp window surrounding these 128 A>G transitions and performed a motif search. The common motif found at these transition sites was (GTAT) or the reverse (ATAC) (Figure 5-16B). Although the sequence context reflected the previously observed $(TG)_n$ and $(CA)_n$ repeats, the A>G transition seemed to occur in the $(TG)_n$ strand and not in the adenosine rich $(CA)_n$ strand. The identified motif suggested that adenosine deamination takes place in a GTATG motif (or the reverse CATAC) (Figure 5-16B, Supp. Figure 5-1). Notably, this motif was also detected when analyzing the distribution of simple repeats in the MEF dIVe-seq data (Figure 5-13C). Since larger repeat stretches are classified together, single interruptions by (GTAT) or (ATAC) are often not represented in the public repeat masker files. Hence, the initial simple repeat analysis mainly identified $(TG)_n/(CA)_n$ dinucleotide in the dIVe peaks.

Since dI was previously reported in telomeric R-loops, I expected the enrichment of telomere sequences in the dIVe-seq data. However, due to the repetitiveness telomere sequences are not mapped during NGS data analysis. [redacted] established an unbiased motif search in all sequenced reads. In his analysis, all reads were considered without initial mapping to the reference genome. Hence, reads are included that otherwise have been excluded due to too many dI-derived mutations or because they are derived from unmappable genomic regions.

By analyzing nucleotide repeat patterns in the dIVe reads compared to the input, [redacted] detected several repeat patterns. This included variants of the $(TG)_n/(CA)_n$ repeats and the $(TGTA)_n/(ATAC)_n$ motif observed in my previous analysis. Importantly, the telomere repeat $(TTAGGG)_n$ was identified by this approach, in line with previous reports of dI in telomeres[62] (Figure 5-16C). Taken together these results not only confirm the presence of dI in telomeres, but in addition, suggest that dI is present in a (TATG) motif contained in $(TG)_n$ simple repeats.

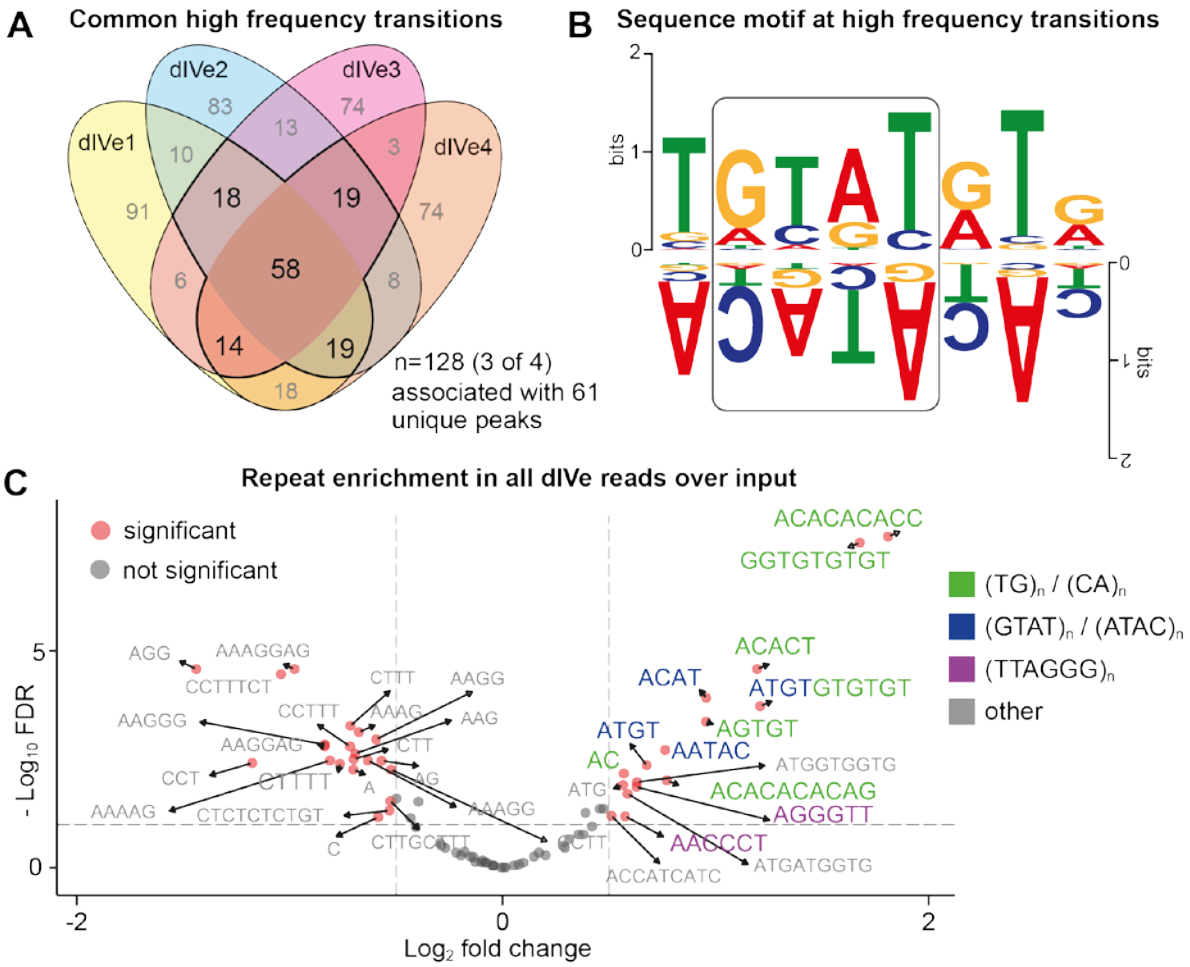


Figure 5-16: High transition mutations are associated with $(GTAT)_n$ motif
Analysis of HEK293T dIVe-seq data **A)** Venn diagram showing common medium frequency (25-74.9%) A:T > G:T transitions. **B)** Motif analysis of all transitions ±5 bp. Motif search was performed with MEME Chip. Upper and lower panel show forward and reverse motif, respectively. **C)** De-novo repeat enrichment analysis developed by [redacted]. Enrichment of specific nucleotide sequences is calculated in dIVe over input samples. Notable enriched repeat types are grouped by colour green: $(TG)_n$ repeat; blue: $(GTAT)_n$; purple: telomere repeat; FDR: false discovery rate.

5.2.9. MPG treatment reduces dIVe signals at a subset of peaks

As shown above, employing qPCR analysis to validate dIVe-seq peaks in HEK293T cells did not provide any promising hits (Figure 5-14). To exclude problems related to the qPCR method or the selected primers, I performed a preliminary dIVe-seq after MPG treatment in HEK293T cells. dIVe samples were prepared in triplicates and corresponding conditions were pooled to reduce sequencing costs. IMB genomics CF performed library preparation and sequencing and [redacted] performed the initial data processing and read mapping.

Next, I inspected the regions associated with the 128 A:T>G:C transition sites, identified from the previous dIVe-seq analysis (Figure 5-16A). The dIVe signal was slightly reduced after MPG treatment in several peaks (Figure 5-17A).

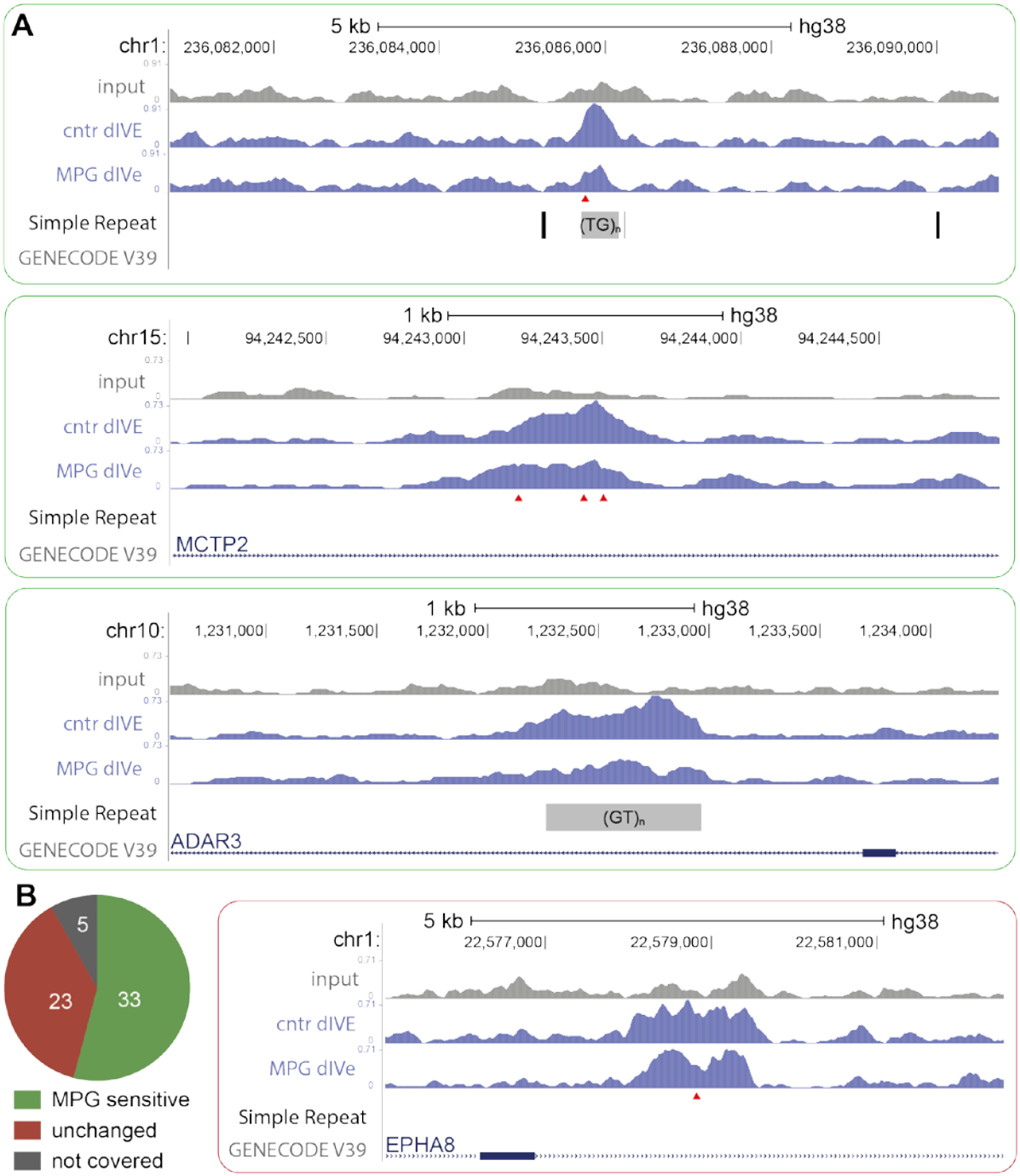


Figure 5-17: MPG treatment reduces dIVe signal at selected A:T>G:C transition sites
dIVe-seq of HEK293T gDNA after MPG treatment **A)** Genome browser tracks of dIVe sequencing after control and MPG treatment. First two boxes show example peaks associated with medium frequency mutations (red arrowhead) and dIVe signal reduced upon MPG treatment. Third box shows peak at intronic region of ADAR3 that is not associated with a detected mutation but is sensitive to MPG treatment. **B)** Left: Summary of all peaks associated with medium frequency (25-75%) transitions, identified in (Figure 5-16A). Right: Peak not reacting to MPG treatment. More examples are shown in Supp. Figure 5-2/Supp. Figure 5-3. Red arrowhead indicates site of identified transition mutation.

In line with the previous qPCR results after dI targeting and MPG treatment (Figure 5-8B), only a partial reduction of the dIVe signal was detectable. Overall, half of the selected peaks showed a signal reduction whereas the remaining peaks were not affected by MPG or had very low read coverage in treated and untreated samples (Figure 5-17B, Supp. Figure 5-2/Supp. Figure 5-3). Interestingly, a peak in the intronic region of *ADAR3* was also sensitive to MPG treatment, even though I did not detect high frequency mutations at these sites. This suggests that dI is also present in peaks that do not contain A>G transitions with increased mutation frequency. In such peaks, dI might be more distributed and less biased to a specific base position.

Taken together the data from this dIVe-seq suggest that some of the identified high frequency mutations are MPG-sensitive and therefore contain genomic dI. However, for conclusive results further dIVe-seqs after MPG treatment should be performed.

5.3. Discussion

In this work, I established dIVe, a method to enrich and map genomic dI. I validated the method using spike-in oligonucleotides and genomic positive regions. dIVe-seq was employed to map potential dI sites in HEK293T and MEF cells. For validation of potential dI sites, I established WGA and MPG treatment for *in vitro* removal of dI. Combined, in the future these approaches might allow to unambiguously and base-specifically identify and investigate dI in gDNA.

5.3.1. Investigating genomic dI

In RNA inosine is a well characterized modification and various sequencing strategies have been employed for base resolution mapping of rI in RNA including antibody-based pulldowns, chemical modifications as well as EndoV based enrichment. The relative abundance of dI/dA is around 3x lower in DNA as compared to rI/rA in total RNA[64]. In addition, rI is very abundant in specific RNA species, e.g. tRNAs[270,271] and can therefore be detected reliably.

In DNA the relative dI abundance is lower and due to its mutagenic potential dI is likely rapidly repaired and therefore mostly transient. However, recent studies[62,63] and unpublished data from our lab[64] suggest that dI is enriched in specific R-loops, where it is placed through ADAR activity and could regulate R-loop stability. To identify genomic regions containing increased levels of dI, I tested two methods for dI enrichment: a commercial antibody and inactive EndoV. Although the antibody enriched modified spike-in sequences and was specific for dI in dot blot experiments, I found only 5x enrichment of genomic dI by LC-MS/MS. In contrast, using inactive EndoV enzyme I enriched genomic dI 35x (Figure 5-5D, Figure 5-7C). Noteworthy, the dI-antibody was recently used to measure dI during chromotrypsis[63], although my results

suggest that the antibody is not sensitive enough to quantify endogenous dI levels (Figure 5-5B). As it also detects rI, traces of rI rich RNA species might additionally distort antibody-based experiments, e.g. in immunofluorescence.

To obtain a positive control of genomic dI for the dIVe procedure, I targeted the optimized adenosine deaminase ABE to different loci using the dCas9 system. The dIVe enrichment was dependent on a specific sgRNA as well as the deaminase and it was reversible by WGA or *in vitro* repair with MPG (Figure 5-8A,C). In principle, WGA can be applied to confirm the presence of DNA modifications[272,273] as it produces a copy of the genome without the site-specific modifications. However, the WGA procedure increased global levels of dI 5-10x as revealed by LC-MS/MS measurements (Figure 5-8B). This is likely due to random spontaneous deamination of the amplified material or due to random integration of dITPs that are present in the WGA mastermix. dITPs could be produced either by deaminase contamination or due to spontaneous deamination promoted by the harsh alkaline denaturing involved in WGA. Since I detected an overall increase of the dIVe signal after WGA, even at control regions like *GAPDH* and *MYOD1*, the integration of dITP is likely random (Figure 5-8A). Moreover, in NGS data I observed that WGA did not uniformly amplify the input material, especially in the simple repeat regions detected by dIVe-seq (not shown). This is in line with studies comparing different WGA methods, which all show sequence bias during amplification[274–276]. Repetitive sequences like telomeres or centromeres are often not properly amplified, due to the limited compatibility with randomized primers, which results from the uniformity of these regions[277]. Simple repeat regions are likely depleted in a similar way during amplification. Therefore, WGA is not suitable to verify the dIVe signal or other modifications that are enriched in highly repetitive sequences.

I did not achieve complete removal of the dIVe signal using *in vitro* MPG treatment (Figure 5-8C). This repair reaction was performed for 1 h, while longer MPG treatment led to increased input recoveries in dIVe qPCR at all tested loci (data not shown). Typically, recombinant enzymes purified from *E. coli* are contaminated with deaminases[64,278]. Therefore, using more enzyme or longer incubation times can artificially increase the dI signal. Even though I always included the deaminase inhibitor pentostatin in the experiments, this might not be sufficient to block the activity of ectopic deaminases completely. In a pilot dIVe-seq after MPG treatment, the signal was slightly reduced in half of the regions of interest (Figure 5-17). However, the signal reduction was less efficient than the decrease observed after dI targeting by dCas9-ABE and subsequent MPG treatment (Figure 5-8B). dCas9-ABE introduces dI efficiently at specific target regions, resulting in short sequences with dI spots in a large population of cells[261]. In contrast, endogenous dI could be more spread out and will be less abundant.

Consequently, MPG treatment might affect endogenous dI levels differently compared to induced dI hotpots, leading to less strong effects in dIVe-seq.

In collaboration with the IMB genomics CF, I established a dI-compatible library preparation for NGS. The protocol is based on the NEBNext ultra II Kit in combination with a PhuU polymerase, which is optimized to amplify DNA containing deaminated cytosine, dU. I reasoned that this polymerase would also accept deaminated adenosine, which was confirmed on spike-in sequences as well as gDNA and while establishing the library preparation (Figure 5-9, Figure 5-10). High fidelity polymerases contain a 3´ to 5´ exonuclease activity and recognize dI in the template as a damage, leading to degradation of the newly synthesized DNA strand. Polymerases without 3´ to 5´ exonuclease activity, like *Taq* polymerase, or modified proofreading, like PhuU, amplified the dI template. Initial NGS of libraries generated from dI containing spike-in confirmed that dI was efficiently read as guanosine when using PhuU for library preparation (Figure 5-10B). This supports previous studies that also reported efficient pairing of dI with cytosine[279]. My results highlight the importance of custom library preparation when investigating dI. Standard polymerases containing proofreading and 3´to 5´ exonuclease activity will not properly amplify dI containing material and hence skew the library towards non-dI sequences. Therefore, utilizing public datasets for correlation analysis with the dIVe-seq data, cannot provide conclusive results without detailed knowledge of the library generation. Correlating the dIVe peaks with DRIP-seq data could reveal dI rich sites that are prone to form R-loops. Ideally, such DRIP-seq datasets should be generated with the dI compatible library preparation, to avoid the loss of dI containing sequences.

5.3.2. dIVe sequencing in human and mouse cells

dIVe-seq revealed around 350 common peaks in human and 900 peaks in mouse cells, often with subtelomeric localization (Figure 5-11). dIVe likely also enriches telomere and centromere sequences, however, these are not mappable by standard NGS pipelines and require specific long read sequencing approaches[280]. Nevertheless, an unbiased de-novo motif assembly of all reads revealed an enrichment of the (TTAGGG) telomere repeat, confirming the previous report from the Nishikura lab[62] (Figure 5-16C).

The identified dIVe peaks were less frequent in promoter-TSS as well as TTS and were mostly located in introns or intergenic (Figure 5-12A, Figure 5-13A). Since dI is mutagenic, it would probably be unfavorable to have high levels of dI in coding regions of the genome. However, in intergenic regions dI could potentially affect transcription of associated genes. This idea is supported by the overlap between genes enriched in HEK293T dIVe-seq with genes that are associated with the miRNA processing factor DROSHA (Figure 5-12B). DROSHA is not only involved in RNA processing but also modulates gene expression through binding of promoter

proximal elements[281,282]. In RNA processing it was shown that ADAR binding can block further processing by DROSHA[283]. A similar mechanism would be conceivable on the DNA level, where ADAR and DROSHA could compete for target binding.

Further analysis of the peak regions revealed that most sites were overlapping with annotated simple repeats (Figure 5-12C, Figure 5-13B). Like telomeres, simple repeats are strongly associated with R-loops[284], suggesting that the dI found in DRIP-LC-MS/MS experiments[64] was derived from R-loops located at these repeat regions. Notably, there was little overlap with other repetitive elements like LINEs or LTRs. Likewise, the dIVe peaks did not enrich SINE Alu family repeats, although Alu elements are known ADAR targets in RNA[285–287].

The simple repeats covered by dIVe peaks had a strong bias for variants of $(TG)_n/(CA)_n$, simple repeats (Figure 5-12D, Figure 5-13C). These dinucleotides are known to adopt left-handed Z-form helixes, in contrast to the standard right-handed B-form helix[257,288,289] (Figure 5-4A). Strikingly, one of the most popular binders of Z-DNA is ADAR1[156,290,291]. Hence, in the $(TG)_n/(CA)_n$ simple repeats, ADAR1 could bind to preexisting Z-helixes or stabilize their formation. Notably, these helixes could be formed by dsDNA or DNA:RNA hybrids. As discussed in the Introduction, DNA:RNA hybrids are energetically more prone to adopt a Z-helix[244] and are much more likely to be deaminated than dsDNA[214]. Moreover, the negative supercoiling that is associated with DNA unwinding during transcription, is promoting the formation of Z-form helixes. Hence, transcribed simple repeats might flip into a Z-conformation while R-loops are formed, which are subsequently deaminated by ADARs.

So far, only *in silico* data support the presence of R-loops at the identified peaks[284]. As discussed before, public R-loop mapping data is usually not generated with polymerases that allow dI amplification. To confirm the presence of R-loops at specific peak sites it would be beneficial to generate DRIP-sequencing data with the established dI-library protocol or verify interesting dIVe peaks in DRIP qPCR if suitable primers can be designed.

5.3.3. Shortcomings of the dIVe method

Due to the repetitiveness and similarity of the dIVe peak sequences, designing qPCR primers was challenging. I generated 17 primer pairs with a corresponding hybridization probe, however, none of the tested regions was sensitive to ADAR1 overexpression or EndoV and MPG manipulation (Figure 5-14). Similarly, whole genome amplification and *in vitro* repair by MPG did not reduce the signal. Therefore, the 17 selected sites most likely represent false positive regions that do not contain dI. However, it is conceivable that there are fundamental problems with the qPCR detection of dI in the $(TG)_n/(CA)_n$ repeats. (i) In case the dI is usually present in the context of a strand break/ nick, due to ongoing repair, the amplification could fail. The dIVe-seq might be less affected due to the initial adaptor ligation during library

preparation. This ligation step could facilitate amplification of fragments that are not detected when only using two fixed qPCR primers. (ii) The enriched DNA could also contain too many modified/ deaminated bases that interfere with primer binding. (iii) The actual signal could also be masked by unspecific amplification if the selected primers falsely amplify similar simple repeat sequences. Finally, subsequent mutation analysis showed that the selected primer targets were not enriched in A>G transitions (data not shown), further suggesting that they do not contain dI.

It is possible that several of the dIVe peaks are based on unspecific binding of EndoV to sequences that do not contain dI. These could be minor substrates like AP sites, mismatches or hairpins[292]. However, it is unlikely that EndoV is simply binding $(TG)_n/(CA)_n$ rich sequences as this should result in much more peaks overlapping the numerous simple repeats in the human or mouse genome. The enriched sites could also represent unspecific binding of EndoV to ssDNA which was previously reported for *E. coli* EndoV[293]. However, another study suggests very low affinity to unmodified ssDNA in the presence of Ca^{2+} [294]. Still, the dIVe-specificity might be improved by using less enzyme or harsher washing conditions.

During LC-MS/MS analysis of input and dIVe samples, I also observed remaining RNA contamination (not shown). In the presence of Ca^{2+} EndoV preferentially binds ssRNA over dsRNA[212], which could sequester EndoV from binding to the DNA substrate and therefore impair the dI enrichment. Theoretically, EndoV could also bind to rI in RNA that is hybridized to a DNA strand, thereby enriching DNA sequences that do not contain dI. However, in this case WGA amplification should have removed any signal in the qPCR as there is no RNA in these samples. Moreover, when doing an initial denaturing step before the dIVe, DNA:RNA hybrids should be dissolved leading to a weaker signal. Instead, when performing ss-dIVe the input recovery in qPCR increased (Figure 5-14B,C) and in ss-dIVe sequencing the absolute number of peaks was increased by 30x compared to ds-dIVe (Supp. Figure 5-4B). The surprisingly high number of peaks in the ss-dIVe-seq would rather argue for unspecific binding of EndoV to ssDNA and against problems associated with RNA contamination.

Nevertheless, gDNA samples could be treated with RNase cocktails after purification to reduce the RNA contamination. During the established dIVe protocol, RNase A and RNase I were used for gDNA purification. Other ribonucleases like RNase III and H could be incorporated to specifically remove dsRNA and RNA in DNA:RNA hybrids, respectively. It might also be helpful to repeat the RNAse treatment after the initial gDNA purification step. However, as discussed before recombinant proteins typically are contaminated with deaminases that could introduce ectopic dI, especially during extended incubation times. Additional RNase treatments will require careful re-optimization and monitoring of dI levels by LC-MS/MS.

It will be important to confirm potential dI sites through gain or loss of function experiments (e.g. ADAR overexpression or knockdown of ADARs, EndoV or MPG) as well as *in vitro* dI repair by MPG treatment.

5.3.4. Detection of dI associated transition mutations

As described above, dI is read as guanosine in dIVe NGS pipeline (Figure 5-10B). Therefore, I analyzed the presence of transition mutations in the peak area. Due to the very high number of unspecific mutations, I grouped the mutations according to their frequency at the respective base position (Figure 5-15A,B). Most mutations were present in group-I, with a mutation frequency <25%, and were likely related to random polymerase errors and spontaneous DNA damages (absolute values not shown). I did not observe an increase of A>G transitions in this fraction, but rather a minor decrease compared to the non-peak area. One would expect that the A>G transitions in the peak are less random, if dI is introduced at specific sites, thus the decrease of A>G in group-I seems logical. Similarly, A>G transitions in group-IV (75-99.9%) were not changed. These mutations likely reflect manifested mutations that differ from the reference genome and are present in all cells of the HEK293T population. However, in group-II and –III (25-74.9%) I observed more A:T > G:C transitions in the peak area, supporting the presence of dI (Figure 5-15B). Surprisingly, I also detected an increase of the opposite G:C > A:T transition. This transition could be related to dU, i.e. cytosine deamination. dU is read as T, effectively producing a C>T transition during sequencing. If the dI peaks represent general deamination hot spots there might be simultaneous adenosine and cytosine deamination. Importantly, EndoV is not enriching dU containing DNA sequences (unpublished data from [redacted]). Therefore, dU could be associated with the dIVe peaks, but should not be the cause of the dIVe enrichment.

Interestingly, I also detected that independent of the sequencing data, the dIVe peak regions are enriched in annotated A>G SNPs (Figure 5-15C). There are more A>G SNPs in the peak area than in random peaks or the entire human genome. One could speculate that the A>G transitions other studies have reported actually relate to dI enrichment in these regions. In general, finding transition mutations related to dI requires an annotated A at the site of interest. Assuming the base position is annotated with a G in the reference genome, dI would not produce a detectable mutation. This could happen if dI is incorporated erroneously opposite to cytosine during replication.

5.3.5. Potential roles of dI in R-loop forming microsatellites

When I extracted and analyzed the sequence context of the medium frequency of A>G transitions (25-74.9%; group-II & -III), I found that they often overlap with TGTA or ATAC motifs (Figure 5-16B). This motif was also observed when analyzing all sequenced reads

independent of their mappability (Figure 5-16C). In a DNA:RNA hybrid this repeat could easily form a rC:dA mismatch; either by slippage of a *cis* R-loop or formation of an R-loop in *trans*. A:C mismatches are a preferred substrate of ADAR[62,214] and could be resolved by deamination, creating a dI:C match (Figure 5-18A).

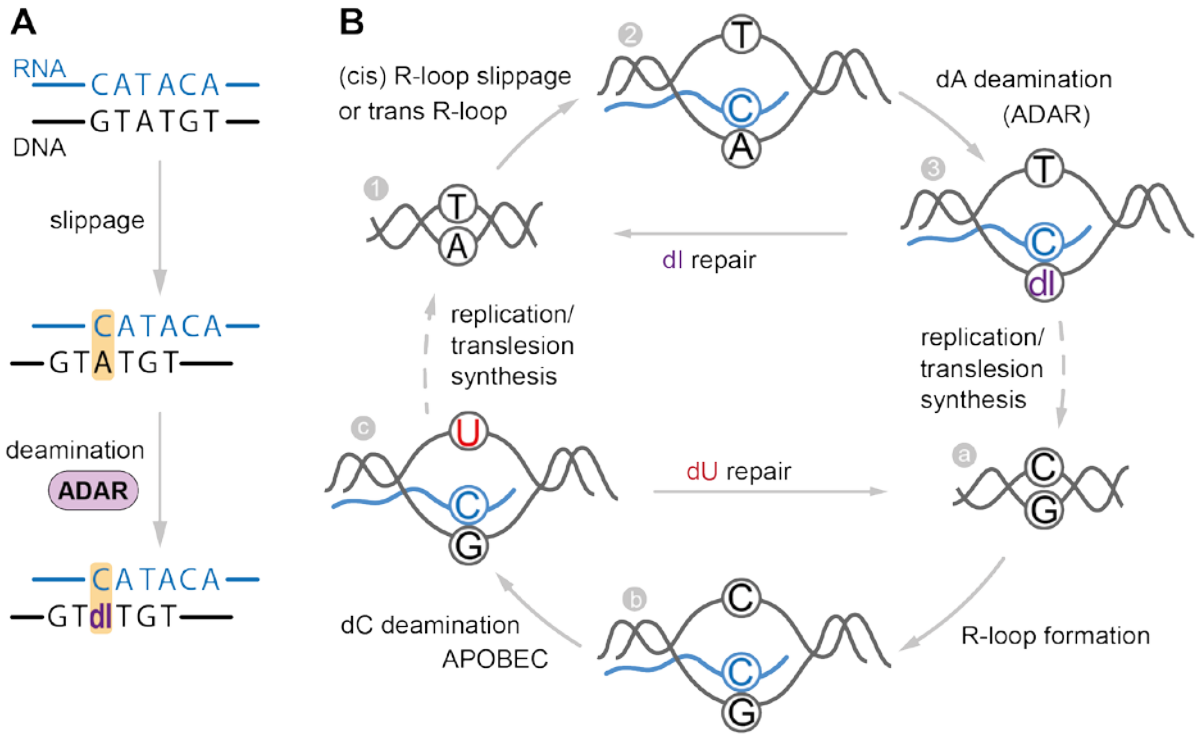


Figure 5-18: ADAR mediated deamination of GTATGT motifs in simple repeats
A) R-loop slippage at a GTATGT motif due to sequence similarity of the surrounding $(TG)_n$ simple repeat. The arising dA:rC mismatch (orange) is a preferred substrate of ADAR. Enzymatic deamination by ADAR is producing a dI:rC match. **B)** 1-3: Introduction and repair of dI in a mismatched R-loop. Slippage of a cis R-loop or formation of an imperfectly matched trans R-loop could produce a dA:rC mismatch. ADAR mediated deamination resolves the mismatch by creating a dI:rC match. dI could than recruit the repair machinery to restore dA. Through translesion synthesis an A:T>G:C transition mutation could manifest at the dI site. a-c: R-loop formation makes cytosine in displaced single strand vulnerable to spontaneous or APOBEC mediated deamination. Analogous to dI, the generated dU can be repaired by BER or result in a G:C>A:T transition mutation. Similarly, 5mC deamination could directly produce a G:T mismatch that establishes a transition mutation if not repaired before replication (not shown).

The ATAC motif in RNA was shown to interact with hnRNP L[295] which is also a strong binder of $(CA)_n$ containing RNA[296]. hnRNP L is a known interactor of ADAR1 and could hence mediate its recruitment to R-loops, rich in $(rCrA)_n$ and $(dTdG)_n$[297]. After deamination of a mismatched dA:rC into dI:rC, repair enzymes MPG and EndoV could initiate repair. In some cases the deamination might also establish a permanent A>G transition in the genomic sequence, i.e. by replication through translesion synthesis[298] (Figure 5-18B)

Similar to adenosine deamination, R-loops could promote cytosine deamination. However, this is more likely to happen in the displaced ssDNA of the R-loop. ssDNA itself is more prone to

deamination[299,300] but is also targeted by cytosine deaminases AID/APOBEC[301], which could produce dU in the ssDNA (Figure 5-18B). dU could be repaired by BER or mismatch repair pathways but might also result in transition mutations through translesion synthesis and replication[302]. In the same way deamination of 5mC in the ssDNA would result in an mC>T transition. CpA pairs are prone for non-canonical, i.e non-CpG, cytosine methylation. This CpA methylation is most prevalent in undifferentiated cells[303–305] but was also reported for human[306] and mouse brain tissue[307]. Deamination of these mC sites would produce a mC>T transition. The resulting T:G mismatch could be removed by mismatch repair[308], however, replication without repair would result in a C:G>T:A transition.

One could speculate that the dI-rich microsatellites are prone to deamination events and cycle between A:T <> T:A base pairs (Figure 5-18B). This would likely be a rare event, which does not happen during every cell cycle. However, through multiple cell cycles dI and dU could counterbalance their respective mutations.

Additionally, the deamination in the $(TG)_n/(CA)_n$ repeats could help to maintain repeat length by recruiting repair enzymes. This might counterbalance repeat expansions, which are usually associated with replication of dinucleotide repeats[309–313]. At the same time repair enzymes are associated with repeat instability[314–316] which is typically increasing with the length of a repeat[317]. The recruitment of dI repair enzymes like MPG and EndoV might promote repeat instability and facilitate contraction of expanded repeats[318]. This would be comparable to previously reported cytosine deamination at R-loops in $(CAG)_n$ repeats, which was implicated in repeat contraction in yeast[56]. Similarly, dI repair might result in repeat contraction.

Notably, classical neuropathological microsatellite expansions, e.g. in Huntington's disease, are usually related to specific trinucleotide repeat expansions[319]. These microsatellites are forming secondary structures[320] that promote repeat expansion dependent on DNA replication, recombination and repair[319,321,322]. Hence, this specific subset of microsatellites behaves differently from the microsatellites observed in the dIVe peaks.

Interruptions of perfect repeats were shown to counteract repeat expansion and stabilize repeats[323,324]. The introduction of an A>G transition could hence be beneficial in a repeat region. By managing the repeat length, dI could also affect expression and splicing of associated genes. The enhancer-like properties of $(TG)_n/(CA)_n$ stretches in human cells have been described as early as 1984[325]. Additionally, an intronic $(TG)_n$ repeat in the Human type I collagen alpha2 was shown to regulate it's transcription in combination with a 5' flanking $(CA)_n/(GC)_n$ repeat[326]. Furthermore, splicing of the cystic fibrosis transmembrane regulator is modulated by a $(TG)_n\,(T)_m$ repeat in intron 8[327]. Likewise, a $(GT)_n$ repeat in Na^+/Ca^{2+}

exchanger 1 (*NCX1*) gene facilitates splicing[328] and in the endothelial nitric oxide synthase gene (eNOS) a $(CA)_n$ repeat functions as intronic splicing enhancer[329].

Lastly, dI could affect the resolution of R-loops as has been reported for telomeric R-loops[62]. Similarly, dI could affect sensitivity to nucleases or helicases, e.g. by changing the stability of the hybrid. As dI is perceived as a G the deamination could potentially affect DNA binding proteins that are sensitive to specific sequence motifs. Additionally, the formation of Z-helixes in the $(TG)_n$ repeats could stall transcription by RNA polymerases[330–332]. This could stabilize R-loops, which would promote deamination of adenosine and cytosine in the hybrid forming DNA and the displaced ssDNA, respectively. Through repeat expansion a sequence could become even more prone to form Z-helixes, consequently increasing R-loop formation and deamination through ADAR. Consequently, dI might affect R-loop resolution as well as repeat stability through initiation of DNA repair (Figure 5-19).

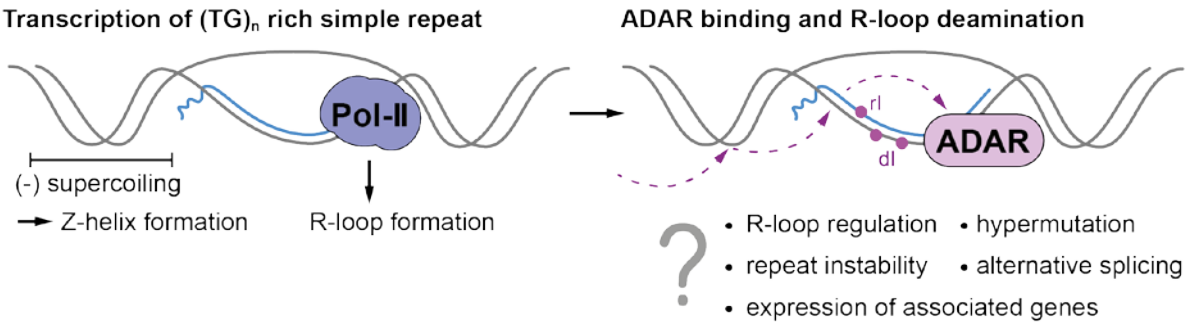


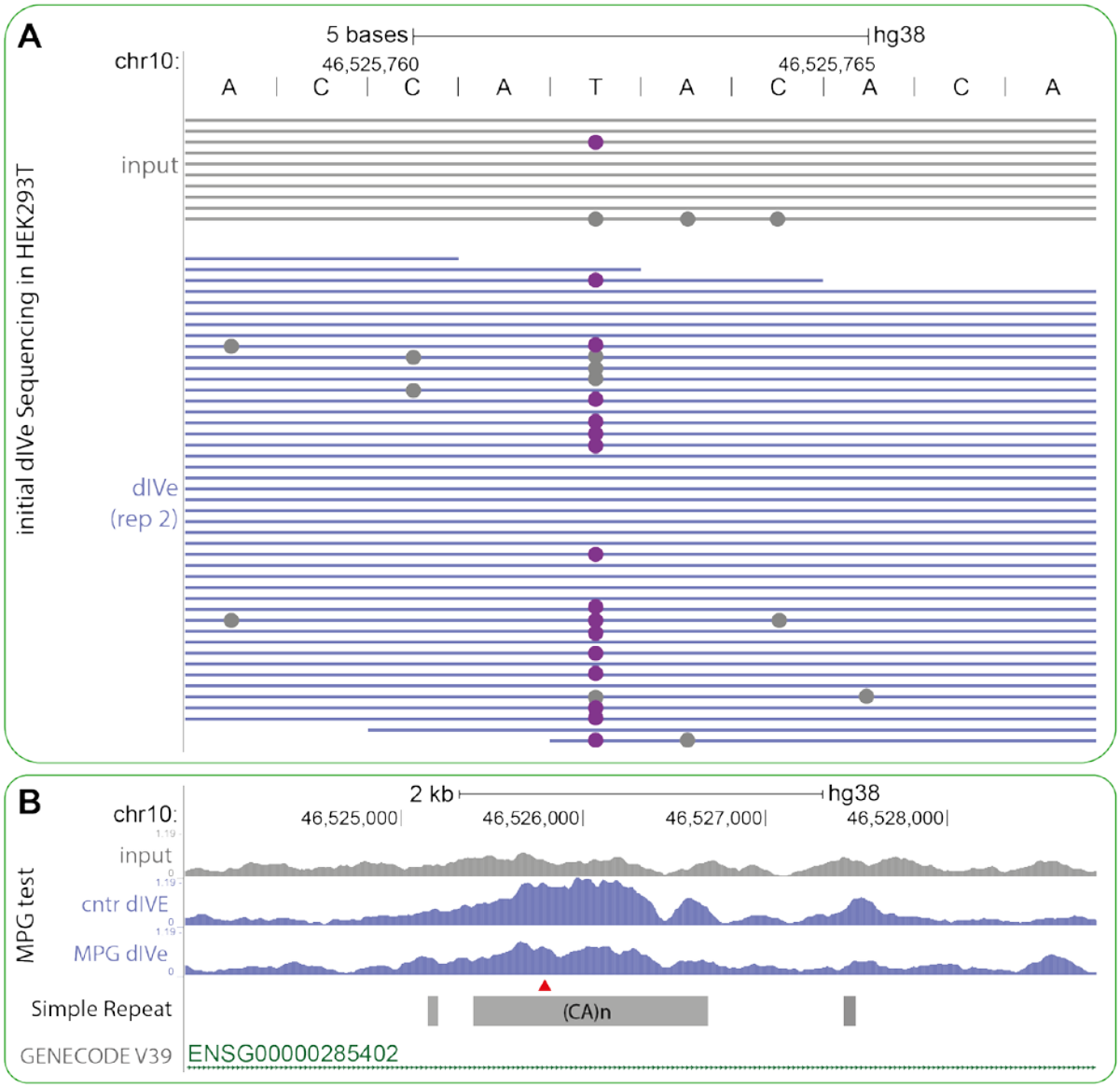
Figure 5-19: Model of ADAR mediated hybrid deamination in $(TG)_n$ simple repeats
Transcription increases negative (-) supercoiling upstream of Polymerase II (Pol-II). In a $(TG)_n$ rich sequence this can result in Z-helix formation of the dsDNA or the DNA:RNA hybrid formed during transcription. Z-helix formation could promote binding of ADAR1, which would further stabilize the Z-helix. ADAR could than deaminate adenosine in the RNA and the DNA portion of the hybrid. The presence of dI could directly affect the stability of the R-loop or result in increased mutation rates. Recruitment of the dI repair machinery could increase the repeat instability and result in repeat contraction (or expansion). The length of the $(TG)_n$ simple repeat could ultimately affect splicing events or the expression of associated genes.

Further experiments will be necessary to see if ADAR is required for dI incorporation at the described simple repeats. Currently, I cannot exclude that the dI enrichment is due to increased incorporation of dITPs at these sites. This could happen via salvage of free inosine and hypoxanthine in the cell[203,333]. If the dI sites are amplified by error prone polymerases the incorporation of damaged nucleotides could be elevated[334,335]. Unfortunately, this mechanism could also indirectly increase genomic dI after ADAR overexpression. ADAR overexpression would lead to increased rI and hypoxanthine pools that could result in increased

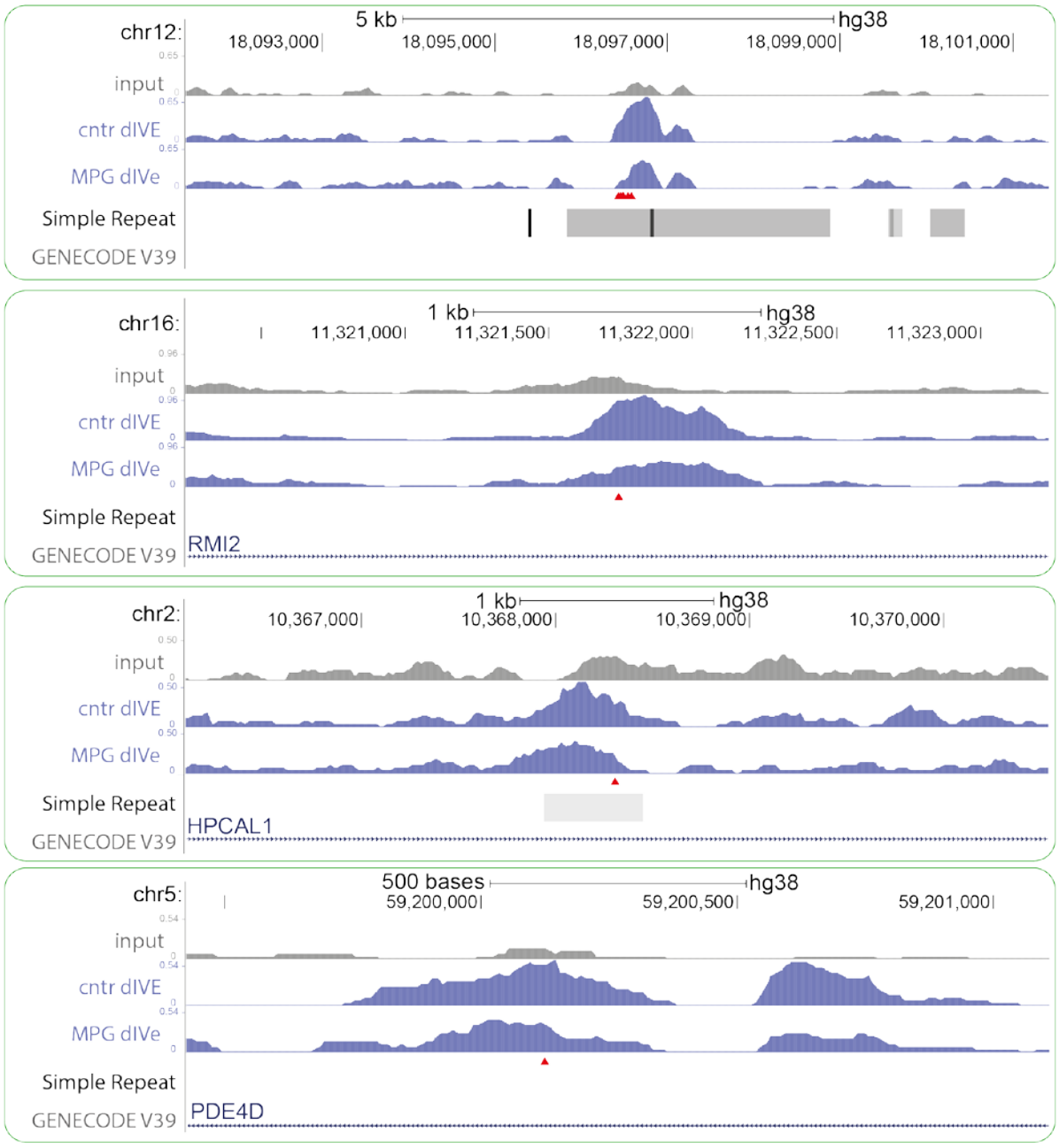
misincorporation of dI. Hence, it will be important to confirm the direct enzymatic activity of ADAR on identified targets by employing *in vitro* experiments decoupled from the salvage pathway.

Even though there is evidence that simple repeats form R-loops in human and mouse[284], there is currently no data directly confirming the presence of R-loops at the identified microsatellites regions. Ideally, R-loop mapping techniques, like DRIP-sequencing, will be combined with the established dI-library preparation to ensure amplification of dI-containing regions. It would also be interesting to obtain strand specific DRIP-sequencing data to identify which DNA strand is forming the hybrid structure. Further, optimization of the method and the dIVe-seq analysis pipeline will help identifying correct peaks. The established dI-targeting by dCas9-ABE as well as dI removal by MPG will allow to verify potential dI-sites.

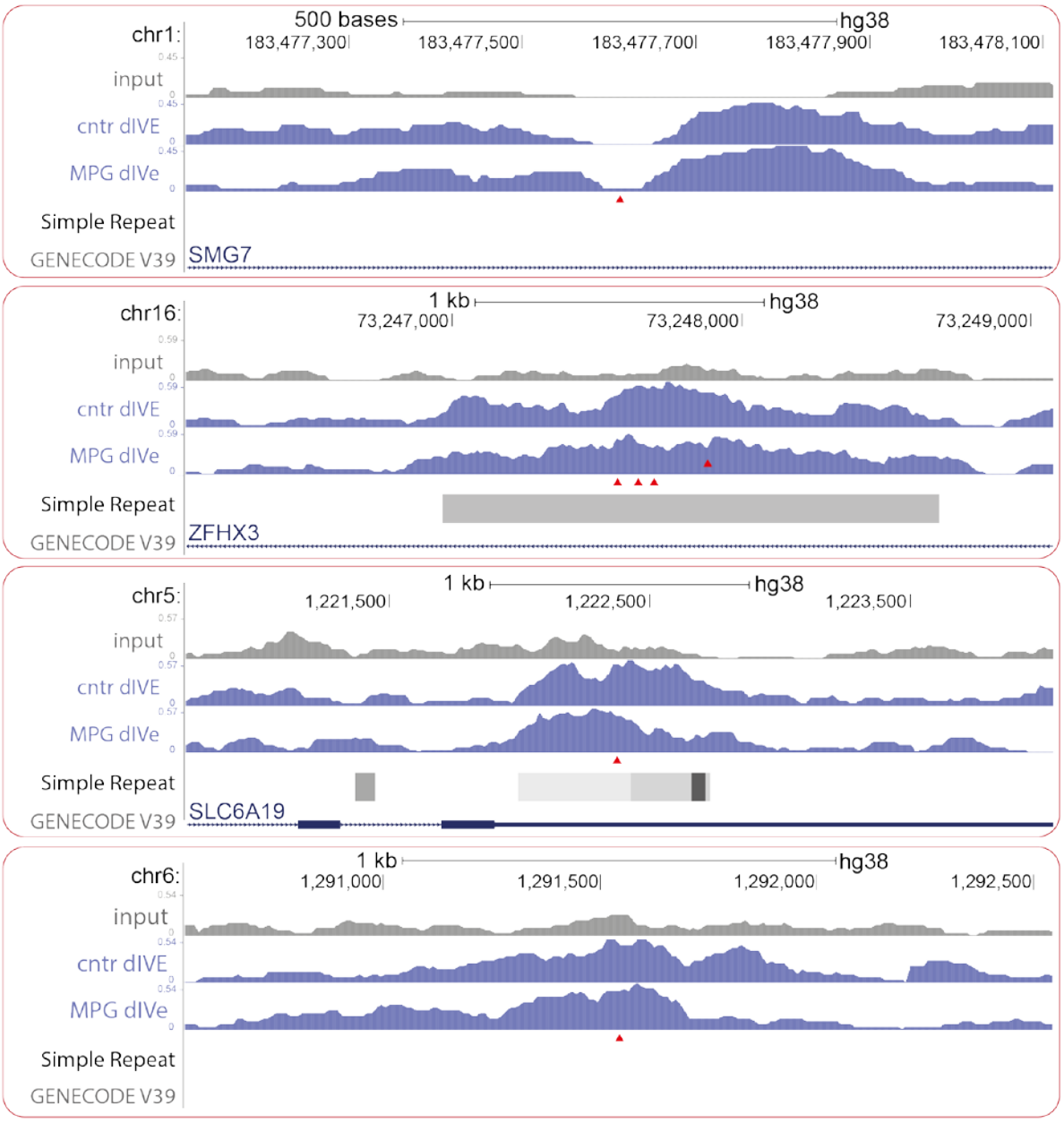
5.4. Supplementary Figures



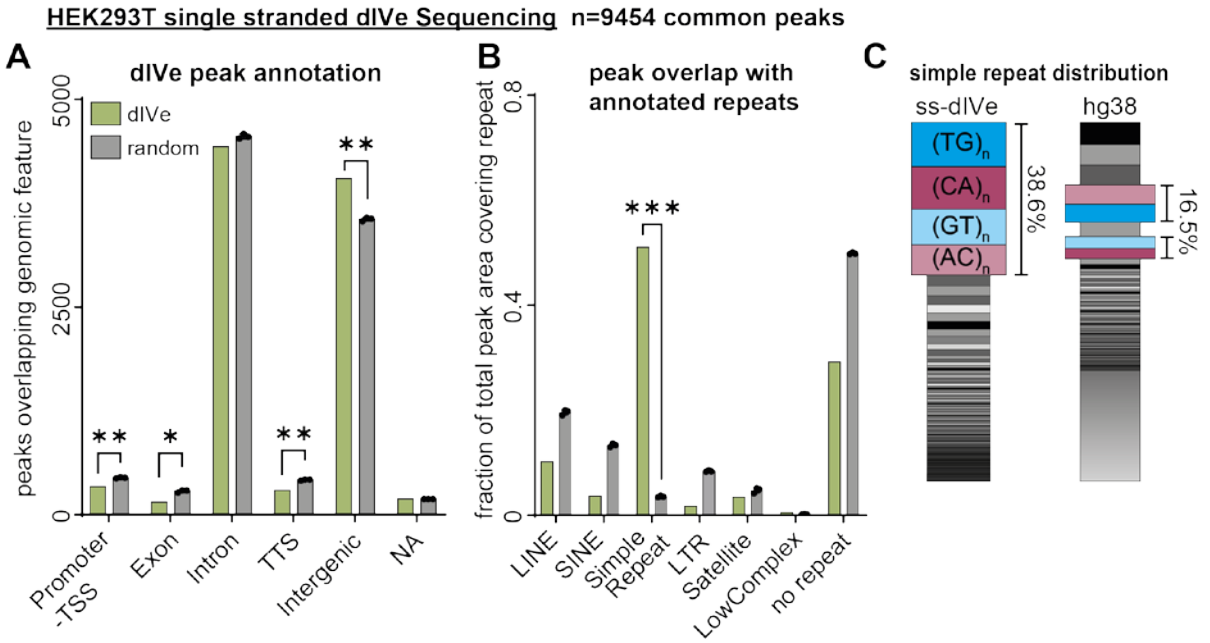
Supp. Figure 5-1: A:T > G:C transitions are associated with (ATAC)/(GTAT) motif
A) dIVe sequencing data from HEK293T cells. Genome browser track showing mapped reads of input and replicate 2 in peak region containing increased T>C transition (reverse complement of A>G). Purple circles represent T>C transitions, grey circles any other mutation in read. **B)** Zoomed out version of same region in independent dIVe sequencing after MPG treatment. Red arrowhead indicates position of identified medium frequency transition.



Supp. Figure 5-2: dIVe signal at selected A>G transitions are affected by MPG treatment
dIVe-seq of HEK293T gDNA after MPG treatment. Genome browser tracks of dIVe-seq after control and MPG treatment. Selected regions show areas that show reduced dIVe signal in the MPG sample. Red arrowhead indicates position of identified medium frequency transition.



Supp. Figure 5-3: dIVe signal after MPG treatment is not reduced at all A>G transitions
dIVe-seq of HEK293T gDNA after MPG treatment. Genome browser tracks of dIVe-seq after control and MPG treatment. Selected regions show areas that are not affected by MPG treatment. Red arrowhead indicates position of identified medium frequency transition.



Supp. Figure 5-4: ss dIVe-seq in HEK293T results in 30x more peaks than ds dIVe-seq
ss dIVe sequencing analysis of HEK293T cells. **A)** dIVe-seq peak annotation with gene features (UCSC, hg38, GENCODE V39). TSS/TTS: transcription start/termination site; NA: no annotated repeat. Common dIVe peaks (green), randomized peak set (grey). Peak randomization was performed in triplicates. **B)** dIVe-seq peak annotation with known repeats (UCSC, hg38, RepeatMasker). LINE/SINE: long/short interspersed nuclear element; LTR: long terminal repeat. Common dIVe peaks (green), randomized peak set (grey). Peak randomization was performed in triplicates. **A/B/C)** * = p<0.05, ** = p<0.01, *** = p<0.001, according to t-test. **C)** Distribution of simple repeats in dIVe peaks (left) and human genome (right). $(TG)_n$: blue; reverse complement $(CA)_n$: red; other repeat motifs shown in grey.

6. Material and Methods

6.1. Material

6.1.1. Equipment

agarose gel chambers (Bio-Rad); bacterial incubators (Thermo Scientific); bacterial shaker (Infors); Bioanalyzer (Agilent); Bioruptor pico/plus (Diagenode); cell counter (Bio-Rad); cell culture incubators (Thermo Scientific); centrifuges (Thermo Scientific); refrigerated centrifuges (Thermo Scientific); ChemiDoc XRS+ System (Bio-Rad); Cryo-Safe Cooler (Belart); Dotblot apparatus (Bio-Rad); DynaMag-2 Magnet rack (Invitrogen); Electronic Multichannel pipette (Sartorius); Electronic Pipette (Sartorius); Fireboy burner (Integra); freezer -20°C (Liebherr), -80°C (Liebherr), -50°C (Sanyo); fridges (Liebherr); ice maker (Wessamat); laminar flow hoods (Dometic); LightCycler 480 (Roche); magnetic stirrer (Heidolph); microscope (Leica); microcentrifuges (Heraeus); microwave oven (Sharp); multichannel pipettes (Sartorius); multidispenser pipette (Eppendorf); Nanodrop 2000 spectrophotometer (Thermo Scientific); Opera Phenix Microscope (Perkin Elmer); orbital shaker (Neolab); PAGE midigel chambers (Bio-Rad); PCR thermocyclers (Biometra); pH meter (Mettler Toledo); pipet boy (Integra); pipettes (Eppendorf); power supplies (Bio-Rad); precision balance (Neolab); Qubit 2.0 (Thermo Fisher); SpeedVac concentrator (Eppendorf); test tube rotator (Neolab); ThermoMixer compact (Eppendorf); Trans-Blot Turbo (Bio-Rad); ultrapure water purification system (Millipore); UV crossliner (Stratagene); vortexer (Scientific industries); waterbath (Neolab)

6.1.2. Lab supplies

Bio-Dot SF Filter Paper (Bio-Rad); cell culture dishes/ flasks/ multiwell plates (TPP); cell counting slides (Bio-Rad); Clear Bottom CellCarrier-96 Black (Perkin Elmer); Combitips (Eppendorf); cryo tubes (Greiner Bio-One); DNA LoBind 1.5 ml Tubes (Eppendorf); extra thick blot filter paper (Bio-Rad); Multiwell Plate 384, LC480 (Roche); micro tube for bioruptor (Roche); Nitrocellulose membrane (VWR); PCR test tube strips (Biozym); PCR plates, 96-well (Biozym); Pipette filter tips (Starlab); PVDF transfer membrane (Neolab); Serological pipettes (Sarstedt); silicone sealing mats (nerbe); test tubes 0.5/1.5/2 ml (Eppendorf); test tubes 15/50 ml (Sarstedt); Trans-Blot Turbo (Bio-Rad); Trans-Blot Turbo Midi PVDF (Bio-Rad); tubes with cell-strainer cap (Falcon); wide bore tips (VWR)

6.1.3. Chemicals

agarose (Biozym); acetate (Sigma-Aldrich); acrylonitrile (Sigma-Aldrich); ampicillin (Sigma-Aldrich); BSA bovine serum albumin (Sigma-Aldrich); DTT, dithiothreitol (Sigma-Aldrich); EDTA (Sigma-Aldrich); ethanol (Sigma-Aldrich); ethidium bromide (Roth); glycerol (Sigma-

Aldrich); Glycogen (Fermentas); HCl, hydrochloric acid (Sigma-Aldrich); isopropanol (Sigma-Aldrich); kanamycin disulfate salt (Sigma-Aldrich); $MgCl_2$, magnesium chloride (Sigma-Aldrich); methanol (Sigma-Aldrich); nuclease-free water (Qiagen); potassium hydroxide (Sigma-Aldrich); Pentostatin (Sigma-Aldrich); skim milk powder (Sigma-Aldrich); NaCl, sodium chloride (Sigma-Aldrich); SDS, sodium dodecyl sulfate (Sigma-Aldrich); NaOH, sodium hydroxide (Sigma-Aldrich); $Na_2HPO_4 \cdot 2H_2O$, sodium phosphate dibasic dihydrate (Sigma-Aldrich); $NaH_2PO_4 \cdot H_2O$, sodium phosphate monobasic monohydrate (Sigma-Aldrich); TEA, triethylamine (Sigma-Aldrich); Tris base (Sigma-Aldrich); Tris HCl (Sigma-Aldrich); Triton X-100 (Sigma-Aldrich); Tween-20 (Sigma-Aldrich)

6.1.4. Enzymes

BbsI-HF (NEB); dCas9 mRNA (A29378; ThermoFisher); Endonuclease V (NEB); Gibson Cloning Master Mix (Protein Production CF, IMB Mainz); hAAG/MPG (NEB); OneTaq Polymerase (NEB); PhusionU Hot Start (Thermo Scientific); Phusion HF DNA Polymerase (NEB); Proteinase K (Qiagen); Q5 Polymerase (NEB); restriction enzymes for subcloning (NEB); Reverse Transcriptase (Protein Production CF, IMB Mainz); RNase 1 (Thermo Scientific); RNAse A (Qiagen); Taq Polymerase (homemade); T4 DNA ligase (NEB)

6.1.5. Reagents

Anti-MBP magnetic beads (NEB); buffer EB (Qiagen); DNA gel loading dye 6x (Thermo Scientific); DNA Ladder 100bp/ 1kb (Thermo Scientific); dITP (Thermo Scientific); dNTP Set (Thermo Scientific); dOxoGTP (TriLink Biotechnologies); dmCTP (TriLink Biotechnologies); dhmCTP (TriLink Biotechnologies); dfCTP (TriLink Biotechnologies); dcaCTP (TriLink Biotechnologies); EvaGreen 20x (Biotium); gelred nucleic acid stain (Sigma); GeneAmp 10X PCR Buffer II & MgCl2 (Thermo Scientific); Protein-G dynabeads (Invitrogen); random primers (Invitrogen); Ribolock RNAse inhibitor (Thermo Scientific); SuperSignal West Femto/Pico (Thermo Scientific); yeast tRNA (Invitrogen)

6.1.6. Kits

Accel-NGS 1S Plus DNA Library Kit (Bioscience); ChIP DNA clean and concentrator (Zymo Research); DNeasy Blood & Tissue kit (Qiagen); LightCycler 480 Probes Master (Roche); LightCycler 480 SYBR Green Master (Roche); MEGAscript T7 Transcription Kit (Thermo Scientific); miRNeasy MiniKit (Qiagen); NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB); PCR Mycoplasma Test Kit I/C (Promokine); Qiaprep Miniprep/ Midiprep kit (Qiagen); Qiaquick Gel extraction kit (Qiagen); QIAquick PCR purification kit (Qiagen); QIAquick Nucleotide Removal Kit (Qiagen); QIAshredder (Qiagen); Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific); Qubit ssDNA Assay Kit (Thermo Fisher Scientific); REPLI-g Mini Kit, whole genome amplification (Qiagen); RNeasy Mini Kit (Qiagen); RNAse-Free DNase Set (Qiagen)

6.1.7. Reagents and Kits for cell culture

CRISPR mRNA (GeneArt Fisher Scientific; A29378); 0.1% gelatine in ultrapure water (Millipore); L-Glutamin 100x (Lonza); MEM Non-essential amino acids 100x (Gibco); Penicillin/Streptomycin 100x (Lonza); sodium pyruvate 100x (Gibco); DMEM, high glucose (Life technologies); DMEM, high glucose, w/o phenol red (Gibco); dimethylsulfoxide (Sigma-Aldrich); ES-grade FBS (PAN Biotech); fetal bovine serum (Lonza); Geneticin G418 (Fisher Scientific); LIF (Protein Production CF, IMB Mainz); ß-mercaptoethanol (Sigma-Aldrich); OptiMEM (Gibco); DPBS (w/o Ca_{2+}/Mg_{2+})(Fisher Scientific); QuikChange II Mutagenesis Kit (Agilent); Trypsin 0.05% (Fisher Scientific); Trypsin 0.25% (Fisher Scientific); Lipofectamine MessangerMax (Invitrogen); Lipofectamine RNAiMAX (Invitrogen); Lipofectamine 2000 (Invitrogen); DAPI (Sigma-Aldrich); SiR-DNA (Spirochrome); X-tremeGENE 9 transfection reagent (Roche)

6.1.8. Antibodies

Anti-Inosine, polyclonal, rabbit (PM098, MBL International);

Anti-Rabbit IgG, polyclonal, goat (111-035-144, dyanova)

IgG, polyclonal, rabbit (ab171870, abcam)

6.1.9. siRNAs

siADAR1 SMARTpool, human (M-008630-01-0005, Dharmacon)

siADAR2 SMARTpool, human (M-009263-01-0005, Dharmacon)

siCntr (D-001210-01-20, Dharmacon)

siEndoV SMARTpool, human (M-018799-01-0005, Dharmacon)

siMPG SMARTpool, human (M-005146-01-0005, Dharmacon)

6.1.10. Software

LightCycler480 (Roche); ImageLab (Bio Rad); ImageJ; Adobe Illustrator & Photoshop; InkScape; SnapGene Viewer; Microsoft Office; Mendeley; usegalaxy.eu; GraphPad Prism

6.1.11. Buffers and Solutions

Solution	components
bead block buffer	10% 10x dIVe-Stock, 0.1% BSA, 50 ng/µl yeast tRNA in H2O
bead wash buffer	10% 10x dIVe-Stock, 0.1% BSA, in H2O
DIP block buffer	10% 10x DIP stock; 100 nM Pento, 1 mM DTT, 0.5% Triton X-100, 0.1% BSA, 50 µg/µl yeast tRNA in H2O
DIP buffer	10% 10x IP stock; 100 nM Pento, 1 mM DTT, 0.5% Triton X-100, 0.01% BSA in H2O
dIVe-Stock (10x)	190 mM Tris HCL pH 7.5; 1.4 M NaCl; 10 mM CaCl
dIVe-Buffer (2x)	20% 10x EndoV-Stock; 14 % Glycerol; 0.02% BSA; 200 nM Pento; 300 nM Dtt; in H2O (always prepare freshly!)
dIVe-Buffer (1x)	10% 10x EndoV-Stock; 7 % Glycerol; 0.01% BSA; 100 nM Pento; 150 nM Dtt; in H2O (always prepare freshly!)
DIP stock (10x)	100 mM NaPO Buffer, pH 7.4, 1.4 M NaCl, H2O
DMEM+++	DMEM, 1% gluthamine, 10% FBS, 1% penicillin/streptomycin
Imaging medium	DMEM w/o phenol red, 1% gluthamine, 10% FBS, 1% penicillin/streptomycin
Luria broth (LB)	For 1 l: 10 g bactotryptone, 5 g yeast extract, 10 g NaCl pH 7.0, in 100 ml water, autoclaved
mESC medium	DMEM, 15% PANsera, 1% penicillin/streptomycin, 1% non-essential amino acids, 1% glutamine, 1% sodium pyruvate, 0.7‰ β-mercaptoethanol, 0.1% homemade LIF
NaPO Buffer (0.2 M, pH to 7.4)	19 ml 0.2 M NaH_2PO_4, 81 ml 0.2 M Na_2HPO_4, in 100 ml water
ProteinaseK-buffer	50 mM Tris-HCl pH 8, 10 mM EDTA, 0.5 % SDS, in H2O
PBS-T	1x PBS, 0.1 % Tween-20
PD digest buffer	1.5 µg/µl Proteinase K; 125 nM Pentostain in ProtK Buffer
SSC (20x)	3 M sodium chloride, 0.3 M sodium citrate
TE buffer	10 mM Tris-Hl pH 8.0, 1 mM EDTA, autoclaved

6.1.12. Oligonucleotides

6.1.12.1. RT-qPCR primers and probes

Target	Forward primer	Reverse primer	Probe
ADAR1	TTCGAGAATCCCAAACAAGG	CTGGATTCCACAGGGATTGT	Roche UPL 39
ADAR2	GTGTAAGCACGCGTTGTACTG	CGTAGTAAGTGGGAGGGAACC	Roche UPL 42
EndoV	TCTTGTGGATGGAAACGGGG	AAGTTTCTTGGCCACCCCAA	Roche UPL 87
MPG	GGGTCCGAGTCCCACGAA	TCACGTCTGACGATGGACGG	Roche UPL 83
GAPDH	GCATCCTGGGCTACACTGAG	AGGTGGAGGAGTGGGTGTC	Roche UPL 82
TBP	GAACATCATGGATCAGAACAACA	ATAGGGATTCCGGGAGTCAT	Roche UPL 87
Tbp	GGGGAGCTGTGATGTGAAGT	CCAGGAAATAATTCTGGCTCA	Roche UPL 97
G6pd	GAAAGCAGAGTGAGCCCTTC	CATAGGAATTACGGGCAAAGA	Roche UPL 78

6.1.12.2. dIVe qPCR primers and probes

Target	Forward primer	Reverse primer	Probe
spike-in-S	GTGTAGGATGCGTAAGTATTTAG	ACCATGCGAATCTTTGCATT	-
spike-in-A	TGACGGAGGCTATTCGTTTGT	CTCGACTCCACCTCCGTTTT	Roche UPL 29
spike-in-B	GCAGCTTCGCTTATCTAGGTTG	AATTAGGCGCAGTCGTTTGA	Roche UPL 32
spike-in-C	GCAACGTCGAGACGTGTAAT	TGCGTTTAAGATATAATCCTGAGTG	/56-FAM/AGTATGGCC/ZEN/TGGCGGTACGT/3IABkFQ/
GAPDH	GAAGGGCTTCGTATGACTGG	CTTAAGGCATGGCTGCAACT	Roche UPL 1
MYOD1	GGATATCAGGGACGCGTTT	GATCCTGGCCCAAACCTC	Roche UPL 24
MUC1	TGGCATCAGTCTTGGTGCT	ACCTCTGCCAGGGCTACC	Roche UPL 83
MUC4	GTCTCCATCACCCAGCTTTC	GTCTGGAAGGATCCATGGTG	Roche UPL 106
APC	CGGCATCTTGTGCTAATCCT	AACCGCACAGCCTGCCTA	Roche UPL 57
hdIp-01	AGGAACGAACACACCCACAC	GCGTGCATCTCTATCGGTTT	/56-FAM/CAGACACAC/ZEN/CCATATGAACCCG/3IABkFQ/
hdIp-02	TCAGTTGCGTGTGTTGGTGT	CGACACACGTGCCTACACAT	/56-FAM/TCGGTTTGT/ZEN/GTGTGCATGTGTG/3IABkFQ/
hdIp-03	GTAGGGGGTGTGTTTCGTGT	ACGCATCCCATACAGCACA	/56-FAM/TGTTTGTGT/ZEN/TGTGGATATGTGTTGTG/3IABkFQ/
hdIp-04	GTTGGTGTGAATGCAAATGG	CCACTCACACAACCCACACT	/56-FAM/TTGCTGTGA/ZEN/ATGGTGGGGGA/3IABkFQ/
hdIp-05	ACACACACCACCCATACTGA	TGGAGGGTCTGAGAGTATGTGTAT	/56-FAM/CACATACAC/ZEN/ACAATACACTCACATGCA/3IABkFQ/
hdIp-06	TGTGGAGCTTAACCCTGTGA	CATGCACATATGCACCTACACA	/56-FAM/TGTGCACAT/ZEN/GCATGTGTATGTATGTG/3IABkFQ/

hdIp-07	CCATGCACACACTCACACAA	GGCATGTGCATTCCTGTGTA	/56-FAM/TGCATGCAT/ZEN/ATGTATGACATGTGAGC/3IABkFQ/
hdIp-08	TCACATGCACACATCACATATAG	GGGTGGTGCATGGTGTGTAT	/56-FAM/CACACATAC/ZEN/ACCACAAACACAGCACA/3IABkFQ/
hdIp-09	GGGTGTATTTGCGTGTGGAT	GCACACATACACAGTTGCACA	/56-FAM/TGCACACAC/ZEN/GTGCATACATGC/3IABkFQ/
hdIp-10	TACTACCCCCACACCCTTCA	TGCTGTGTGTGGTATGTGTGA	/56-FAM/CACACAACA/ZEN/GAAACGTACACACACA/3IABkFQ/
hdIp-11	TGGCATGCAGTACGTGTTAC	CACATTGCCTACACCACACA	/56-FAM/TGAGGTGTG/ZEN/TGTGGTACGTGTGTT/3IABkFQ/
hdIp-12	GTGTGGGGTAGTGTGTGTGAT	TCACATGCTACACTACATACATGCT	/56-FAM/TTGTGTGCA/ZEN/TGCAGTGGTGTG/3IABkFQ/
hdIp-13	AGTGATGTGCGTGCGTAGG	CATTCAGTGCACACCACACA	/56-FAM/TTGCATGTG/ZEN/TGTGTTGAGTGTTGTG/3IABkFQ/
hdIp-14	TGCACATGCCACACAGAGTA	CGTGTGTCCATGTGTTTGGT	/56-FAM/TGCAGGTAC/ZEN/ATATACCACATGTCACA/3IABkFQ/
hdIp-15	GCAGTGTGGTGTCTGAGGAG	ACATGCAGAAGCCACACACA	/56-FAM/ACCACACAC/ZEN/CATGCACGTGC/3IABkFQ/
hdIp-16	TCACACCCTCACACATTCATT	GCGTGTGAATGTGCATGACT	/56-FAM/TGCTCACAC/ZEN/ACATTTGCACACTCA/3IABkFQ/
hdIp-17	ACATCCATGCCACGCAAAT	TGTGGCATATGTGTGATGTGT	/56-FAM/CACACACCA/ZEN/CACACGCCAC/3IABkFQ/

6.1.12.3. Spike-in oligos

Oligonucleotide		Sequence
spike-in S (100 bp)	S-Fwd-dI	gtgtaggatgcgtaagtattt/ideoxyI/gcccactaaagttagtagatgttatgttatgccc/ideoxyI/gccacgtgctaaacggccccaatgca/ideoxyI/agattcgcatggtggg
	S-Rev-dI	cccaccatgcgaatctttgcattggggccgtttagc/ideoxyI/cgtggctgggcataacataac/ideoxyI/tctactaactttagtgggct/ideoxyI/aatacttacgcatcctacac
	S-Fwd-Cntr	gtgtaggatgcgtaagtatttagcccactaaagttagtagatgttatgttatgcccagccacgtgctaaacggcccaatgcaaagattcgcatggtggg
	S-Rev-Cntr	cccaccatgcgaatctttgcattggggccgtttagcacgtggctgggcataacataacatctactaactttagtggctaaatacttacgcatcctacac

spike-in A (120 bp)	A-Fwd-Cnt	gtaacacacacttctttttgacggaggctattcgtttgtaagcaaaactaacgaggcagaag ggtcacgggtctacatctgaacctatagatcctagaaaaacggaggtggagtcgagaa
	A-Rev-Cnt	ttctcgactccacctccgtttttctaggatctataggttcagatgtagacccgtgacccttctgcctcgttagttttgct tacaaacgaatagcctccgtcaaaaagaagtgtgtgttac
	A-Fwd-dI	gtaacacacacttctttttgacggaggctattcgtttgtaagca/ideoxyI/aactaacgaggcagaagggtca cgggtctac/ideoxyI/tctgaacctatagatcctagaaaaacggaggtggagtcgagaa
	A-Rev-dI	ttctcgactccacctccgtttttct/ideoxyI/ggatctataggttcagatgtagacccgtg/ideoxyI/cccttctg cctcgttagttttgcttacaaacgaatagcctccgtcaaaaag/ideoxyI/agtgtgtgttac
spike-in B (120 bp)	B-Fwd-Cnt	gtggcggtgtaagcagcttcgcttatctaggttgaaccacaatccctcatgtcacttatgttctgctcccggaccttt gcgtctagttggcacagtcaaacgactgcgcctaattactgct
	B-Rev-Cnt	agcagtaattaggcgcagtcgtttgactgtgccaactagacgcaaggtccgggagcagaacataagtgaca tgagggattgtggttcaacctagataagcgaagctgcttacaccgccac
	B-Fwd-dI	gtggcggtgtaagcagcttcgcttatctaggttgaaccac/ideoxyI/atccctcatgtcactt/ideoxyI/tgttc tgctcccggaccttgcgtct/ideoxyI/gttggcacagtcaaacgactgcgcctaattactgct
	B-Rev-dI	agcagtaattaggcgcagtcgtttgactgtgccaactagacgca/ideoxyI/ggtccgggagcagaacata agtgacatg/ideoxyI/gggattgtggttcaacctagataagcgaagctgcttacaccgccac
spike-in C (87 bp)	C-Fwd-dI	ctgcaacgtcgagacgtgtaattctttctgaat/ideoxyI/ctacatagt/ideoxyI/tggcctggcggtacgtac actcaggattatatcttaaacgcag
	C-Rev-Cntr	Ctgcgtttaagatataatcctgagtgtacgtaccgccaggccatactatgtagtattcagaaagaattacacgt ctcgacgttgcag

6.1.12.4. sgRNA cloning

Target (Cas9 species)	Target sequence	Forward oligo.	Reverse oligo.
Rosa26 (S.py.)	ACTCCAGTCTTTCTAGAA GA	CACCGACTCCAGTCTTT CTAGAAGA	AAACTCTTCTAGAAAGAC TGGAGTC
APC (S.py.)	CACGCATAGTAAAGAGT CGG	CACCGCACGCATAGTAA AGAGTCGG	AAACCCGACTCTTTACTAT GCGTGC
MUC1 (S.py.)	GCTCCACCGCCCCCCCA GCCCA	CACCGGCTCCACCGCC CCCCCAGCCCA	AAACTGGGCTGGGGGGG CGGTGGAGCC
MUC4 repeat (S.py.) for imaging	GTGGCGTGACCTGTGGA TGCTG	CACCGGTGGCGTGACC TGTGGATGCTG	AAACCAGCATCCACAGGT CACGCCACC
MUC4 non-repeat (S.py.) for dCas9-ABE	AACAGAGGGCCAGAGA GCAGCC	CACCGAACAGAGGGCC AGAGAGCAGCC	AAACGGCTGCTCTCTGGC CCTCTGTTC
Telomere (S.py.)	GTTAGGGTTAGGGTTAG GGTTA	CACCGGTTAGGGTTAG GGTTAGGGTTA	AAACTAACCCTAACCCTA ACCCTAACC
Telomere (S.au.)	GGTTAGGGTTAGGGTTA GGGT	CACCGGGTTAGGGTTA GGGTTAGGGT	AAACACCCTAACCCTAAC CCTAACCC
MUC4 (S.au.)	GAAGTGTCGGTGACAGG AAG	CACCGGAAGTGTCGGT GACAGGAAG	AAACCTTCCTGTCACCGA CACTTCC

mMajSat (S.py.)	CAAGAAAACTGAAAATCA	CACCGCAAGAAAACTGAAAATCA	AAACTGATTTTCAGTTTTCTTGC
mMinSat (S.py.)	ACACTGAAAAACACATTCGT	CACCGACACTGAAAAACACATTCGT	AAACACGAATGTGTTTTTCAGTGTC
Akap6 (S.py.)	CACAGTGCTCAGGGGACCTGG	CACCGCACAGTGCTCAGGGGACCTGG	AAACCCAGGTCCCCTGAGCACTGTGC
hChr3 (S.py.)	TCCTCTGTATGATATCACAG	CACCGTCCTCTGTATGATATCACAG	AAACCTGTGATATCATACAGAGGAC
hChr14 (S.py.)	GCTGTCCCCCGCCCCCAGCT	CACCGGCTGTCCCCCGCCCCCAGCT	AAACAGCTGGGGGCGGGGGACAGCC
hChr19 (S.py.)	GAGGAGGGAAGC	CACCGGAGGAGGGAAGC	AAACGCTTCCCTCCTCC

6.1.12.5. *Fgf5* enhancer sgRNA sequences

The following sequences were published by the lab of Joanna Wysocka[100]. They were ordered as primers for sequencing and as forward primers for RT-qPCR detection of the *Fgf5* sgRNAs.

Target ID	**Sequence**
*Fgf5*_enhancer_gRNA1	GGGGACTCTTATTGAAGAAT
*Fgf5*_enhancer_gRNA2	GCTCCATCCACTGTTTGCCA
*Fgf5*_enhancer_gRNA3	TAGCATGGCTTTCTTCTGAG
*Fgf5*_enhancer_gRNA4	AGTGTCAGGGGGTCTCAAGC
*Fgf5*_enhancer_gRNA5	GTATTTCCTTTCAGGGACAG
*Fgf5*_enhancer_gRNA6	TACTTTTTTTAAAAATTATT
*Fgf5*_enhancer_gRNA7	TGATCCCGAGACTTAACAAC
*Fgf5*_enhancer_gRNA8	ACTCATAACATTGTATTTTG
*Fgf5*_enhancer_gRNA9	GAACATTAAAAGATAATGTG
*Fgf5*_enhancer_gRNA10	AGCCTTCAAGGAATGTATGA
*Fgf5*_enhancer_gRNA11	AAAAAGTGATTGGGTGTTTG
*Fgf5*_enhancer_gRNA12	GAAAACTTGCTGAATGACAGTAAAGA
*Fgf5*_enhancer_gRNA13	GACATGGGAGATAGGGCTACAACTGGG
*Fgf5*_enhancer_gRNA14	GCTCCATCAGATCTCCCTCTTCCCTA
*Fgf5*_enhancer_gRNA15	GCATACAAGGGTGGTCCAAGCCCCC
*Fgf5*_enhancer_gRNA16	GCAGTCTCAACTAACCTTATAGAGTCC
*Fgf5*_enhancer_gRNA17	GGGGACCCCTGTGGTTGAATTGGAAGA
*Fgf5*_enhancer_gRNA18	GCCCAATGGTCTGATGCAAGTATCTGC
*Fgf5*_enhancer_gRNA19	GAAGGAAAGGAGATGAGTGGCCCAACT
*Fgf5*_enhancer_gRNA20	GCTTAAAAGCAGTGCTGAGACCAAAA
*Fgf5*_enhancer_gRNA21	GAAGAGTCCAGCTATAGTAGATAGAC
*Fgf5*_enhancer_gRNA22	GAGTCCTGAGAGTCTCTCTCCTCCT
*Fgf5*_enhancer_gRNA23	GGTAGCCTTGGAGGTGGGACATGGGG
*Fgf5*_enhancer_gRNA24	GTGTTCTCTTATCCTTACATTTTGAA
*Fgf5*_enhancer_gRNA25	GATGAAAGACTTGCCTAATGTCAGA

*Fgf5*_enhancer_gRNA26	GAGACCAGTGGATATGCAACTTATCCA
*Fgf5*_enhancer_gRNA27	GACATATGCTAGCTGCTCAAAAATTTT
*Fgf5*_enhancer_gRNA28	GAGTGGGTGTCTTGCTCATACGGAAG
*Fgf5*_enhancer_gRNA29	GGTACCTAAACCAGCTCATTAATGCC
*Fgf5*_enhancer_gRNA30	GATTTGCCTAGTGACCTATTTTTATGA
*Fgf5*_enhancer_gRNA31	GTGAACCGGTCGGTAGGTATCAGTGG
*Fgf5*_enhancer_gRNA32	GACTGAGTTTTTCTGTCCCCCTACAG
*Fgf5*_enhancer_gRNA33	GACAGAGAATTTAATGAGAAACACAT
*Fgf5*_enhancer_gRNA34	GGAAATTTATCAGATTTTCATACAGGG
*Fgf5*_enhancer_gRNA35	GTGGATGCACAGCCATAATTGCTCTC
*Fgf5*_enhancer_gRNA36	GTAAATATATAGCGTGTGTCAGCGTG

The following reverse primer was employed for RT-qPCR detection of sgRNA expression

reverse_qPCR_primer	GCACCGACTCGGTGCCACTT

6.1.12.6. Primers for STAgR cloning

Fragment		Forward primer	Reverse primer
Fgf5 enhancer 1	1A	ACTCATAACATTGTATTTTGGTTTTAGAGCTAGAAATAGCAAGTT	ATTCTTCAATAAGAGTCCCCCGGTGTTTCGTCCTTTCC
	1B	GGGGACTCTTATTGAAGAATGTTTTAGAGCTAGAAATAGCAAGTT	TGGCAAACAGTGGATGGAGCCGGTGTTTCGTCCTTTCC
	1C	GCTCCATCCACTGTTTGCCAGTTTTAGAGCTAGAAATAGCAAGTT	CTCAGAAGAAAGCCATGCTACGGTGTTTCGTCCTTTCC
	1D	TAGCATGGCTTTCTTCTGAGGTTTTAGAGCTAGAAATAGCAAGTT	GCTTGAGACCCCCTGACACTCGGTGTTTCGTCCTTTCC
	1E	AGTGTCAGGGGGTCTCAAGCGTTTTAGAGCTAGAAATAGCAAGTT	CTGTCCCTGAAAGGAAATACCGGTGTTTCGTCCTTTCC
	1F	GTATTTCCTTTCAGGGACAGGTTTTAGAGCTAGAAATAGCAAGTT	AATAATTTTTAAAAAAAGTACGGTGTTTCGTCCTTTCC
	1G	TACTTTTTTTAAAAATTATTGTTTTAGAGCTAGAAATAGCAAGTT	GTTGTTAAGTCTCGGGATCACGGTGTTTCGTCCTTTCC
	1H	TGATCCCGAGACTTAACAACGTTTTAGAGCTAGAAATAGCAAGTT	CAAAATACAATGTTATGAGTCGGTGTTTCGTCCTTTCC
Fgf5 enhancer 2	2A	GCAGTCTCAACTAACCTTATAGAGTCCGTTTTAGAGCTAGAAATAGCAAGTT	CACATTATCTTTTAATGTTCCGGTGTTTCGTCCTTTCC
	2B	GAACATTAAAAGATAATGTGGTTTTAGAGCTAGAAATAGCAAGTT	TCATACATTCCTTGAAGGCTCGGTGTTTCGTCCTTTCC
	2C	AGCCTTCAAGGAATGTATGAGTTTTAGAGCTAGAAATAGCAAGTT	CAAACACCCAATCACTTTTTCGGTGTTTCGTCCTTTCC
	2D	AAAAAGTGATTGGGTGTTTGGTTTTAGAGCTAGAAATAGCAAGTT	TCTTTACTGTCATTCAGCAAGTTTTCCGGTGTTTCGTCCTTTCC
	2E	GAAAACTTGCTGAATGACAGTAAAGAGTTTTAGAGCTAGAAATAGCAAGTT	CCCAGTTGTAGCCCTATCTCCCATGTCCGGTGTTTCGTCCTTTCC
	2F	GACATGGGAGATAGGGCTACAACTGGGGTTTTAGAGCTAGAAATAGCAAGTT	TAGGGAAGAGGGAGATCTGATGGAGCCGGTGTTTCGTCCTTTCC
	2G	GCTCCATCAGATCTCCCTCTTCCCTAGTTTTAGAGCTAGAAATAGCAAGTT	GGGGGCTTGGACCACCCTTGTATGCCGGTGTTTCGTCCTTTCC
	2H	GCATACAAGGGTGGTCCAAGCCCCCGTTTTAGAGCTAGAAATAGCAAGTT	GGACTCTATAAGGTTAGTTGAGACTGCCGGTGTTTCGTCCTTTCC

Fgf5 enhancer 3	3A	GTGTTCTCTTATCCTTACATTTTGAAGTTTTAGAG CTAGAAATAGCAAGTT	TCTTCCAATTCAACCACAGGGGTCCCCCGGTGT TTCGTCCTTTCC
	3B	GGGGACCCCTGTGGTTGAATTGGAAGAGTTTTA GAGCTAGAAATAGCAAGTT	GCAGATACTTGCATCAGACCATTGGGCCGGTGT TTCGTCCTTTCC
	3C	GCCCAATGGTCTGATGCAAGTATCTGCGTTTTA GAGCTAGAAATAGCAAGTT	AGTTGGGCCACTCATCTCCTTTCCTTCCGGTGTT TCGTCCTTTCC
	3D	GAAGGAAAGGAGATGAGTGGCCCAACTGTTTTA GAGCTAGAAATAGCAAGTT	TTTTGGTCTCAGCACTGCTTTTAAGCCGGTGTTT CGTCCTTTCC
	3E	GCTTAAAAGCAGTGCTGAGACCAAAAGTTTTAGA GCTAGAAATAGCAAGTT	GTCTATCTACTATAGCTGGACTCTTCCGGTGTTT CGTCCTTTCC
	3F	GAAGAGTCCAGCTATAGTAGATAGACGTTTTAGA GCTAGAAATAGCAAGTT	AGGAGGAGAGAGACTCTCAGGACTCCGGTGTTT CGTCCTTTCC
	3G	GAGTCCTGAGAGTCTCTCTCCTCCTGTTTTAGAG CTAGAAATAGCAAGTT	CCCCATGTCCCACCTCCAAGGCTACCCGGTGTT TCGTCCTTTCC
	3H	GGTAGCCTTGGAGGTGGGACATGGGGGTTTTA GAGCTAGAAATAGCAAGTT	TTCAAAATGTAAGGATAAGAGAACACCGGTGTTT CGTCCTTTCC
Fgf5 enhancer 4	4A	GACTGAGTTTTTCTGTCCCCCTACAGGTTTTAGA GCTAGAAATAGCAAGTT	TCTGACATTAGGCAAGTCTTTCATCCGGTGTTTC GTCCTTTCC
	4B	GATGAAAGACTTGCCTAATGTCAGAGTTTTAGAG CTAGAAATAGCAAGTT	TGGATAAGTTGCATATCCACTGGTCTCCGGTGTT TCGTCCTTTCC
	4C	GAGACCAGTGGATATGCAACTTATCCAGTTTTAG AGCTAGAAATAGCAAGTT	AAAATTTTTGAGCAGCTAGCATATGTCCGGTGTT TCGTCCTTTCC
	4D	GACATATGCTAGCTGCTCAAAAATTTTGTTTTAG AGCTAGAAATAGCAAGTT	CTTCCGTATGAGCAAGACACCCACTCCGGTGTT TCGTCCTTTCC
	4E	GAGTGGGTGTCTTGCTCATACGGAAGGTTTTAG AGCTAGAAATAGCAAGTT	GGCATTAATGAGCTGGTTTAGGTACCCGGTGTTT CGTCCTTTCC
	4F	GGTACCTAAACCAGCTCATTAATGCCGTTTTAGA GCTAGAAATAGCAAGTT	TCATAAAAATAGGTCACTAGGCAAATCCGGTGTT TCGTCCTTTCC
	4G	GATTTGCCTAGTGACCTATTTTTATGAGTTTTAG AGCTAGAAATAGCAAGTT	CCACTGATACCTACCGACCGGTTCACCGGTGTT TCGTCCTTTCC
	4H	GTGAACCGGTCGGTAGGTATCAGTGGGTTTTAG AGCTAGAAATAGCAAGTT	CTGTAGGGGGACAGAAAAACTCAGTCCGGTGTT TCGTCCTTTCC
Fgf5 enhancer 5	5A	GTAAATATATAGCGTGTGTCAGCGTGGTTTTAGA GCTAGAAATAGCAAGTT	ATGTGTTTCTCATTAAATTCTCTGTCCGGTGTTTC GTCCTTTCC
	5B	GACAGAGAATTTAATGAGAAACACATGTTTTAGA GCTAGAAATAGCAAGTT	CCCTGTATGAAAATCTGATAAATTTCCCGGTGTT TCGTCCTTTCC
	5C	GGAAATTTATCAGATTTTCATACAGGGGTTTTAG AGCTAGAAATAGCAAGTT	GAGAGCAATTATGGCTGTGCATCCACCGGTGTT TCGTCCTTTCC
	5D	GTGGATGCACAGCCATAATTGCTCTCGTTTTAGA GCTAGAAATAGCAAGTT	CACGCTGACACACGCTATATATTTACCGGTGTTT CGTCCTTTCC

6.1.13. Oligonucleotides for subcloning

Target plasmid	Forward primer	Reverse primer
PT084	GAAGCTTGGATCCAGGTGGA	AAGCGGCCGCTTAACTAGTTTATTCG GTTACCGTGAAGGTTTTGG
PT118/119/120	GAAGCTTGGATCCAGGTGGAGGTGGAAGCGGTG GATCTCAGCCTGTGCTGACACAGAGC	GCCGCTTAACTAGTTCACACTTTCCG CTTTTTCTTAGGAGCGGCG

6.1.14. Sanger sequencing primers

Primer ID	Sequence
F1ori-F	GTGGACTCTTGTTCCAAACTGG
LKO.1 5'	GACTATCATATGCTTACCGT
pCAG-F	GCAACGTGCTGGTTATTGTG
SFFV-F	AAAGAGCTCACAACCCCTCA
SP6	ATTTAGGTGACACTATAG
SV40pro-F	TATTTATGCAGAGGCCGAGG

SV40pA-R	GAAATTTGTGATGCTATTGC
T3	ATTAACCCTCACTAAAGGGA
T7	TAATACGACTCACTATAGGG
CMV-F	CGCAAATGGGCGGTAGGCGTG
oPT017, sgRNA sequencing	GGCCTATTTCCCATGATTCCT

6.1.15. QuikChange Primers

Target mutation	Forward primer	Reverse primer
Cas9_H557A	CACGCTTCTGGGGATAATAGCATCGACCT CGTAGTTGAAT	ATTCAACTACGAGGTCGATGCTATTATCCC CAGAAGCGTG
Cas9_D10A	TACATTCTGGGGCTGGCCATCGGGATTAC AAGC	GCTTGTAATCCCGATGGCCAGCCCCAGAA TGTA
dVenus	GGGGTCTTTGCTCAGCGCGGACTGGTAGC TCAGG	CCTGAGCTACCAGTCCGCGCTGAGCAAAG ACCCC
ADAR1_E1008Q (hyperactivation)	CACCAAGGTGGAGAACGGACAGGGCACAA TCCCT	AGGGATTGTGCCCTGTCCGTTCTCCACCTT GGTG
ADAR1_E912A (inactivation)	CAATGACTGCCATGCAGCAATAATCTCCCG GAGAG	CTCTCCGGGAGATTATTGCTGCATGGCAG TCATTG

6.1.16. Plasmids

Plasmids and their IDs used in IMD database, lab book and Niehrs lab database are given below. Plasmids that start with “PT” were only used for cloning and were not used for experiments shown in this thesis. Plasmid maps are deposited in Niehrs Lab database.

Name	IMB ID	Lab book ID	Niehrs ID	origin	Ref.
au.dCas9-HttTag (20 HTT repeats)	3794	PT133	259	gibson cloning (GA22), IMBID:2901+2977+1431	
au.dCas9-SunTag (24 GCN4 repeats)	2977	PT117	201	gibson cloning (GA21), IMBID:3790+1882	
CARGO_Fgf5enh.-1 (sgRNA 1-12)	3802	PT191	267	Wysocka Lab	[100]
CARGO_Fgf5enh.-2 (sgRNA 13-24)	3803	PT192	268	Wysocka Lab	[100]
CARGO_Fgf5enh.-3 (sgRNA 25-36)	3804	PT193	269	Wysocka Lab	[100]
dCas9-ABE	3758	Ph040	235	addgene:124447	[261]
dCas9-SunTag (*S.py.*)	1882	PT040	179	subcloning: IMBID:1321+897 (AgeI+EcoRI)	
GFP-V_L-HTT	1875	PT031	172	gibson cloning (GA3), IMBID:377+1429	
mito-mCherry-HttTag$_{10x}$	3751	PT278	228	gibson cloning (GA42) IMBID:3794+1882+1431 & subcloning with IMBID:3746 (RsrII+XhoI)	
mito-mCherry-SunTag$_{10x}$	3746	PT249	223	addgene, 60914	[99]
mVenus-V_L-HTT	2979	PT119	203	subcloning: IMBID:2927 (BamHI+SpeI)	
p.STAgR-backbone	2992	PT150	209	gibson cloning (GA25), IMBID:1861	
p.STAgR-insert	2991	PT139	208	gibson cloning (GA24), IMBID:1861	
pADAR1p110	3523	Ph032	277	addgene, 117928	[336]
pADAR1p110$_{Hyp}$	3754	Ph036	231	QuikChange of pADAR1p110	
pADAR1p150	3522	Ph031	276	addgene, 117927	[336]
pADAR1p150$_{Hyp}$	3752	Ph034	229	QuikChange of pADAR1p150	

pADAR1p150$_{Inac}$	3753	Ph035	230	QuikChange of pADAR1p150	
pBLKS- (empty vector)	3787	PT002	250	Stratagene	[337]
pNeo	3811	PT010	281	Dr. Andrea Schäfer	
pRosa-CAG-dCas9-SunTag	3800	PT174	265	see method section	
PT006_pHRdSV40-dCas9-24xGCN4-NLS	1321	PT006	251	addgene:60910	[99]
PT011_pPuro	897	PT011	12	addgene:11349	[338]
PT012_pX330-sgCntr-py.dCas9	903	PT012	6	addgene:42230	[339]
PT081_au.Cas9-EGFP	1920	PT081	186	addgene:64709 (Kiran Musunuru; unpublished)	
PT082_pSaGuide	1923	PT082	187	addgene:64710 (Kiran Musunuru; unpublished)	
PT084_pαGCN4-sfGFP	1928	PT084	189	subcloning: PT022 (BamHI+NotI)	
PT091_au.dCas9-HttTag	2901	PT091	195	gibson cloning (GA16); IMBID:3790+1882+1431	
PT092_pmVenus	2902	PT092	196	addgene:27794	[340]
PT094_pVenus	3789	PT094	254	QuikChange, PT092	
PT101_mVenus-V$_L$-HTT	2927	PT101	198	gibson cloning (GA20), PT084+PT092	
PT102_pCloning-Venus-V$_L$-HTT	2928	PT102	199	gibson cloning (GA20), PT084+PT094	
PT103_au.dCas9-EGFP	3790	PT103	255	QuikChange, PT081	
py.dCas9-EGFP	3363	PT200	280	addgene:51023	[101]
scFv-GCN4-GFP	1320	PT022	252	addgene:60906	[99]
sg.MajSat	3739	PT176	216	sgRNA cloning, PT020	[96]
sg.MinSat	3740	PT177	217	sgRNA cloning, PT020	[96]
sg.Rosa26	3738	PT160	215	sgRNA cloning, PT020	
sgAkap6	3741	PT178	218	sgRNA cloning, PT020	[96]
sg$_{au}$.Cntr	1934	PT087	192	subcloning: IMBID:1861+ addgene:64710 (NdeI+HindIII)	
sg$_{au}$.MUC4	2981	PT129	205	sgRNA cloning, PT087	
sg$_{au}$.Telomere	3791	PT107	256	sgRNA cloning, PT087	
sg$_{py}$.APC	3788	PT076	253	sgRNA cloning, PT020	[341]
sg$_{py}$.Chr14	3749	PT273	226	sgRNA cloning, PT020	[121]

sg_{py}.Chr19	3750	PT274	227	sgRNA cloning, PT020	[121]
sg_{py}.Chr3	3748	PT272	225	sgRNA cloning, PT020	[121]
sg_{py}.Cntr	1861	PT020	166	subcloning: IMBID:903+ 3787 (XmaI+KpnI)	
sg_{py}.MUC4	1881	PT037	178	sgRNA cloning, PT020	[101]
sg_{py}.MUC4 (non rep.)	3792	PT124	257	sgRNA cloning, PT020	
sg_{py}.Telomere	1879	PT035	176	sgRNA cloning, PT020	[101]
STAgR_Fgf5enh.-1 (sgRNA 1-8)	3795	PT151	260	STAgR	[100,104]
STAgR_Fgf5enh.-2 (sgRNA 9-16)	3796	PT152	261	STAgR	[100,104]
STAgR_Fgf5enh.-3 (sgRNA 17-24)	3797	PT153	262	STAgR	[100,104]
STAgR_Fgf5enh.-4 (sgRNA 25-32)	3798	PT168	263	STAgR	[100,104]
STAgR_Fgf5enh.-5 (sgRNA 33-36)	3799	PT169	264	STAgR	[100,104]
STAgR-Telomere	3801	PT184	266	subcloning: IMBID:3797+ 1879 (AgeI+KasI)	
Venus-V_L-HTT	2980	PT120	204	subcloning: IMBID:2928 (BamHI+SpeI)	

6.2. Methods

6.2.1. General molecular biology

General molecular biology methods including preparation of chemically competent XL1-blue *E. coli* bacteria, plasmid amplification in *E. coli*, spectrophotometric quantification of DNA and RNA, restriction digests, DNA ligations, PCR, and agarose gel electrophoresis were carried out as previously described[342].

6.2.1.1. Plasmid maintenance and purification.

For long-term storage, bacterial culture was mixed with 33% glycerol and stored at -80 °C. To recover the plasmids, the frozen glycerol stock was scraped with a pipette tip, which was then transferred to 5 ml LB medium + antibiotic for overnight incubation at 37 °C.

For plasmid preparation, QIAGEN Mini and Midi Kits were used according to manufacturer's protocol. DNA amount and purity were estimated on a Nanodrop 2000 spectrophotometer. Plasmid sequences were confirmed by Sanger sequencing at GATC Biotech AG (until 08/2018) and StarSeq GmbH.

6.2.1.2. RNA isolation, cDNA generation

RNA was isolated using the QIAGEN RNeasy MiniKit according to manufacturer's protocols. DNA was digested on column using DNase I according to the provided optional protocol including the following modification: Column was soaked with 80 µl DNase I solution; briefly centrifuged, soaked with another 80 µl DNase I solution and incubated for 25 min at RT. Further steps were performed according to manufacturer's instructions. RNA was eluted in 30 µl RNase free water. The concentration was measured on a NanoDrop 2000 spectrophotometer.

Complementary DNA (cDNA) was generated using reverse transcriptase provided by IMB protein production CF according to supplemented protocols. 1 µg RNA was mixed with 2 µl 100 µM dN6 primer and 1 µl 10 mM dNTPs in a final volume of 12 µl. The mix was denatured 5 min at 65 °C and cooled on ice for 2 min. 4 µl 5x FS buffer, 2 µl 0.1 M DTT, 1 µl Ribolock and 0.2 µl reverse transcriptase were added and samples were incubated for 10 min at 25 °C, 90 min 42 °C and 5 min 72 °C. The resulting cDNA was diluted 1:6 in TE buffer for qPCR.

6.2.1.3. Real-time quantitative PCR

Real-time quantitative PCR (RT-qPCR) was conducted in a 384-well format using the Roche LightCycler480 System. It is based on detection of a specific PCR product using short hydrolysis probes or the DNA dye SYBR Green. Gene-specific primers and corresponding probes were determined using the Roche Universal ProbeLibrary Assay Design Center or primer3 (open source) & probeBAGEL (custom software provided by [redacted]). 5 µl cDNA or diluted pulldown sample was mixed with 5.5 µl 2x ProbesMaster, 0.12 µl probe, and 0.12 µl

primer mixture containing forward and reverse primer at 50 mM each. For assays that did not allow binding of a specific probe; Roche SYBR Green qPCR was performed: 5 µl cDNA was mixed with 5.5 µl 2x SYBR Green I Master, and 0.12 µl primer mixture containing forward and reverse primer at 50 mM each. The following settings were used for RT-qPCR:

Roch Probes Master qPCR Assay

	PCR step	temperature	time	Ramp rate
1	Initial Denaturation	95 °C	10 min	4.8 °C/s
2	Denaturation	95 °C	10 s	4.8 °C/s
3	Annealing	60 °C	15 s	2.5 °C/s
4	Elongation (signal acquisition per cycle)	72 °C	10 s	4.8 °C/s
	Go to step 2 and repeat 49 cycles			
5	Cooling	40 °C	30 s	2.5 °C/s

SYBR Green qPCR Assay

	PCR step	temperature	time	Ramp rate
1	Initial Denaturation	95 °C	10 min	4.8 °C/s
2	Denaturation	95 °C	10 s	4.8 °C/s
3	Annealing	58 °C	15 s	2.5 °C/s
4	Elongation (signal acquisition per cycle)	72 °C	10 s	4.8 °C/s
	Go to step 2 and repeat 49 cycles			
5	Denaturing	95 °C	10 s	4.8 °C/s
6	Melting Start	58 °C	1 min	2.5 °C/s
7	Melting Target (signal acquisition per 5 °C)	97 °C		0.11 °C/s
8	Cooling	40 °C	30 s	2.5 °C/s

Cp values and expression levels were calculated using LightCycler 480 software release 1.5.1.62. All samples were measured in technical replicates and normalized to *TBP*, *GAPDH,* or *G6pd* expression.

6.2.2. Non-standard cloning techniques

6.2.2.1. sgRNA plasmids for Cas9 targeting

sgRNA plasmid cloning was based on material provided by addgene.org[343]. *S.py.* and *S.au.* Cas9 sgRNAs were designed using web tools provided by the Broad Institute[344]. The sequences were ordered as oligonucleotides including the appropriate overhangs for cloning (forward: 5'-CACCGFFFFFFFFFFFFFFFF-3'; reverse: 5'-AAACRRRRRRRRRRRRC-3'). The oligonucleotides were annealed in a PCR block by heating to 95°C for 5 min and then switching off the block, to cool down to room temperature. Plasmids sg_{py}.Cntr (PT020) for *S.py.* and sg_{au}.Cntr (PT087) for *S.au.* were digested using BbsI. Annealed oligonucleotides were ligated into digested plasmids using T4 ligase in presence of the restriction enzyme BbsI. 2 µl

of ligation was used for transformation of 50 µl X10 blue cells. Positive clones were identified by control digest using BbSI and sequencing with primer oPT017.

6.2.2.2. Site directed mutagenesis

Point mutations were introduced using Agilent QuikChange II XL Kit according to supplier's protocol. In brief, mutagenesis primers were designed using Agilent webtool[345]. Whole plasmid was amplified introducing the desired mutation and the original plasmid was subsequently removed by DpnI digest. 2 µl QuikChange reaction was used to transform 45 µl XL10-Gold ultracompetent cells, provided with the Kit.

6.2.2.3. Gibson Cloning

Gibson cloning allows restriction free assembly of multiple overlapping fragments in a one-step reaction[346]. Overlapping fragments were designed using the NEBuilder Assembly Tool[347] and generated by restriction digests or PCR. All fragments were gel purified using QIAGEN gel purification Kit (supplier's protocol) before assembly. 0.05 pmol of backbone was combined with 2-6 fold amount of insert fragments in 10 µl H2O. 10 µl 2x GibsonMix was added and incubated for 1-2 h at 50 °C. 3 µl of assembly reaction was used to transform 50 µl XL1 blue bacteria.

6.2.2.4. STAgR Cloning

STAgR cloning is a specialized version of Gibson cloning that allows assembly of up to eight gRNA cassettes in a single plasmid[104,346]. To generate a plasmid, unique gRNA sequences fused to adapter sequences are used to amplify a constant gRNA cassette containing hU6 promoter, gRNA scaffold and a PolII terminator sequence. The eight DNA fragments (A-H) contain the variable gRNA sequences (1-8) in the overlapping regions. The backbone fragment-A contains gRNA-8 at 5' and gRNA-1 at 3'. The insert fragment-B contains sgRNA-1 at 5' and gRNA-2 at 3', fragment-C contains sgRNA-2 at 5' and gRNA-3 at 3', etc. Fragment-H has gRNA-7 at 5' and gRNA-8 at 3', generating a closed plasmid if all fragments align correctly. Primers were designed as suggested by the authors[104]. Plasmid p.STAgR-backbone (PT150) was used as template for fragments A and p.STAgR-insert (PT139) as template for fragments B-H. Due to the large number of fragments and their sequence similarity, the reaction is not very efficient. Colony PCR was performed, to identify plasmids containing the correct inserts. For sanger sequencing primers binding the variable gRNA regions (i.e. the guide sequence) were employed to confirm proper assembly of all gRNA cassettes.

6.2.3. Cell culture

6.2.3.1. HEK293T and HeLa cell culture

Cells were cultured in DMEM+++ medium at 37°C, 5% CO_2 and 21% O_2. Cells were passaged every other day, once reaching approximately 80% confluency, by washing cells with PBS once, loosening cells with 0.25% trypsin for 3 minute at 37°C, quenching trypsin with DMEM+++ and plating 1/8th on a new cell culture dish.

6.2.3.2. Primary MEF cell maintenance

MEF cells were maintained in DMEM +10% FBS, 1% penicillin/streptomycin and 1% glutamine at 37°C, 5% CO_2, 5% O_2 and 85% relative humidity. Cells were passaged every 2-3 days, once reaching approximately 80% confluency, by washing cells with PBS once, loosening cells with 0.25% trypsin for 3 minute at 37°C, quenching trypsin with DMEM+++ and plating 1/8th on a new, cell culture dish.

6.2.3.3. mESC cell culture

mESCs (E14tg2a) were maintained on pregelatinized culture dishes in mESC medium at 37 °C 5% CO_2 and 21% O_2. Culture medium was changed every day and cells were passaged every other day. Culture dishes were gelatinized by incubating them for 15 min with 5 ml 0.4% gelatin in ultrapure water.

For passaging, cells were washed with PBS and incubated 3 min with 2 ml of 0.05% trypsin/EDTA at 37 °C. Trypsin was quenched by adding mESC medium and colonies were dissociated into single cells by gentle titruation. The cells were pelleted 3 min at 300 xg, resuspended in fresh culture medium and split 1:8 to a pregelatinized culture dish.

6.2.3.4. Cell freezing

To freeze cells from a culture dish they were dissociated as described before. After quenching and removing trypsin HEK293T, HeLa and MEF cells were resuspended in FBS + 10% DMSO; mESC were resuspended in 6 ml PANsera +10% DMSO. The suspension was mixed and distributed to four 1.5-ml-freezing-vials. Vials were slowly frozen in an isopropanol-based freezing container or styrofoam box at -80 °C. To thaw cells they were resuspended in 10 ml prewarmed medium. DMSO containing medium was removed by centrifugation at 300 xg for 3 min and all cells were plated on a culture dish

6.2.3.5. Plasmid Transfection

HEK293T were transfected with Xtreme Gene9 (Roche) directly after cell passaging. For harvesting after 48 h 1.8 x10^6 cells were plated in culture medium w/o P/S on a 10-cm-dish. For 72 h transfections, 3.3 x10^5 cells were plated. Between 6-15 µg DNA was prepared in 500 µl OptiMEM and 3 µl XtremeGene9 per µg DNA (e.g.18 µl for 6 µg) was prepared in 500 µl

OptiMEM. Both solutions incubated for 5 min and mixed together by carefully pipetting up and down. After 20 min incubation at room temperature 1000 µl transfection solution was added dropwise to the cells and distributed by gentle swirling.

mESCs were transfected with Lipofectamine 2000 (Thermo-Fischer). For a 6 cm-dish 0.5-0.75x10^6 cells were plated in culture medium w/o P/S and let settle for 4 h. 6 µg DNA was prepared in 250 µl OptiMEM and 18 µl Lipofectamine 2000 (3 µl/µg DNA) was prepared in 250 µl OptiMEM. Both solutions were incubated for 5 min and mixed together by carefully pipetting up and down. After 20 min incubation at room temperature 500 µl transfection solution was added dropwise to the cells and distributed by gentle swirling. Cells were harvested after 48 h.

6.2.3.6. siRNA Transfection

siRNAs were ordered as Smart Pools from Dharmacon and resuspended in 250 µl to a stock concentration of 20 µM. HEK293T cells were transfected with Lipofectamine RNAiMAX. For harvesting after 72 h 3.3 x10^5 cells were plated in culture medium w/o P/S on a 10-cm-dish. 6 µl 20 µM siRNA was prepared in 500 µl OptiMEM and 36 µl RNAiMax (6 µl/µg DNA) was prepared in 500 µl OptiMEM. Both solutions were incubated for 5 min and mixed together by carefully pipetting up and down. After 20 min incubation 1000 µl transfection solution was added dropwise to the cells and distributed by gentle swirling.

6.2.4. Generation of stable cell lines expressing dCas9-SunTag

6.2.4.1. Template for homology directed repair and sgRNA

For stable expression of dCas9-SunTag in mESCs a template for homology directed repair into the *Rosa26* locus was generated. In short, *Rosa26* homology arms (5' and 3') and CAG-promoter (spilt in 5' & 3' for efficient amplification) were extracted by PCR and assembled in pBLKS- (PT002) backbone (BamHI + XhoI digested) by gibson cloning. Primers used to generate fragments for gibson cloning are given below.

Amplicon	Primer	Sequence	Template
Homology-*Rosa26*-5'	Fwd	cggccgctctagaactagtgGCGGCAGGCCCTCCGAGC	mESC gDNA
	Rew	ttatgtaacgCTAGAAAGACTGGAGTTGCAGATCACGAGGGAAG	
Homology-*Rosa26*-3'	Fwd	ggaattcacccccaagaagaagcgcaaggtgggacgcgtCTAGAAGATGGGCGGGAG	mESC gDNA
	Rev	gctgggtaccgggccccccctCACATTTAGACCAGCAATAAC	
CAG-5'	Fwd	gtctttctagCGTTACATAACTTACGGTAAATGGCCCGCCT	addgene 107270
	Rev	cagcgactccCCGCCCGCCGCGCGCTTC	
CAG-3'	Fwd	cggcgggcggGGAGTCGCTGCGACGCTG	addgene 107270
	Rev	tgcgcttcttcttgggGGTGAATTCCTGCAGCCCGGG	

The assembled plasmid served as backbone for the dCas9-SunTag insert extracted from dCas9-SunTag (PT040) by subcloning using restriction enzymes MluI-HF and SbfI-HF. The resulting plasmid pRosa-CAG-dCas9-SunTag (PT174) contained the *Rosa26* homology regions flanking the CAG-dCas9-SunTag sequence.

Rosa26 gRNA plasmid sg.Rosa26 was generated as described above. This plasmid was used as PCR template to add T7 promoter required for *in vitro* transcription using the following primers:

Forward: GAAATTAATACGACTCACTATAGACTCCAGTCTTTCTAGAAGA
Reverse: AAAAAAGCACCGACTCGGTGCC

PCR product was gel purified (QIAGEN) and used as template for T7 MegaScript Kit according to suppliers protocol. Reaction was cleaned up by filling to 100 µl and proceeding with miRNA Mini Kit (QIAGEN), according to provided protocol. Final elution was done in RNAse free water and concentration was determined using Nanodrop 2000.

6.2.4.2. Transfection of knock-in components

0.5 x10^6 cells were seeded in pregelatinized 6 cm dish and transfected after 4 h with Lipofectamine 2000: 0.6 µg pNeo and 5.4 µg pRosa-CAG-dCas9-SunTag were transfected as described above. After 24 h medium was changed and GeneArt Cas9 mRNA (Fisher Scientific) and Rosa26 sgRNA were transfected using Lipofectamine MessangerMax: 5 µg Cas9 mRNA and 1 µg IVT sgRNA were transfected with 18 µl MessangerMax reagent according to suppliers protocol. As a control, the same transfections were performed w/o the pNeo plasmid. After another 24 h medium was changed to mESC culture medium + 200 µg/ml G418. Cells were maintained in culture for 10 days under G418 selection before FACS sorting. Control cells w/o pNeo plasmid died during this period.

6.2.4.3. FACS

Cells were sorted 10 days after transfection and G418 selection. Cell sorting was conducted by the Cytometry CF of the IMB. mES cells were detached and dissociated as described before and resuspended in 2 ml culture medium. Sorting was conducted in a Becton Dickinson FACSAria III SORP equipped with a 100 µm nozzle. All measured events were gated by forward scatter-area (FSC-A) and side scatter area (SSC-A) to exclude cell debris. Cell doublets were detected by measuring SSC-A towards SSC width. BFP was detected after excitation with a violet laser (VL 405 nm), passing a 450/50 bandpass filter. Cells were gated by area of fluorescence signal relative to FSC-A. Cells were sorted into a 96-well plate containing preconditioned mESC culture medium. Around 6000 cells were sorted with 100 cells being combined per well.

6.2.4.4. Monoclonal selection

After FACS cells were expanded and grown to large separate colonies in 15 cm dishes. The appropriate size was reached 17 days after transfection (7 days after sorting). The monoclonal colonies were picked under a microscope using a 20 µl Eppendorf pipette. Each colony was carefully detached from the dish by bouncing into it with the pipette and gently sucking it into the tip. The colony was then transferred to a gelatinized 96-well-plate containing 100 µl culture medium (+ 200 µg/ml G418). 288 colonies were picked and stored in the cell-culture incubator overnight. After 18 h the colonies were washed with 100 µl PBS and separated into single cells with 50 µl trypsin at 37 °C. Trypsin was quenched after 3 min by adding 200 µl culture medium. Suspension was mixed and 200 µl was transferred to a new gelatinized 96-well-plate.

To identify successful integration of dCas9-SunTag, cells were transfected with sg.Telomere and scFv-GCN4-sfGFP in 96-well format. Cells were imaged with Opera Phenix as described in next chapter. Cells showing telomere spots in all transfected cells were considered positive, expanded and stored for further experiments.

6.2.5. Live cell imaging using dCas9-SunTag

Cells were seeded in 96-well plates (CellCarrier-96 Black, Perkin Elmer). HeLa cells were plated in 100 µl at $1.5x10^5$ cells/ml (15,000 cells per well) 24h before transfection. 300 ng DNA was prepared for transfection using XtremeGene9 (supplier's protocol). For mESC imaging, 20,000 cells were plated per well, 5 hours before transfection on gelatinized plate and transfection was performed using Lipofectamin 2000 (supplier's protocol). Each reaction was prepared with 300 ng DNA. 10% of the pipetted transfection Mix (XtremeGene9 or Lipofectamine) was added per well, resulting in 30 ng plasmid DNA per well. The rations were: 5 ng dCas9-SunTag, 5 ng gRNA, 5 ng antibody-fluorophore, 15 ng empty vector. According to the experiment design, the empty vector was substituted by another Cas9-SunTag system or additional sgRNAs. Imaging was performed 48 h after transfection using Opera Phenix screening microscope (Perkin Elmer). 45 min before imaging temperature and CO_2 control were adjusted to 37 °C and 5% CO_2 for live cell imaging. In between, cells were stained with Hoechst 33342 (1:10000) in imaging medium for 15 min and subsequently washed 3x 5 min with imaging medium.

Set-up and imaging was performed according to Opera Phenix user manual provided by IMB Microscopy CF. Imaging was performed in confocal mode. The following laser lines and emission filters were used for fluorophore detection:

mCherry: excitation 561 nm; emission filter 570-630 nm
EGFP/sfGFP/(m)Venus: excitation 488 nm; emission filter 520-550 nm
Hoechst3342: excitation 405 nm; emission filter 435 – 550 nm
SiR DNA: excitation 640 nm; emission filter 650 – 760 nm

Even though the Opera Phenix allows imaging of fluorophores in neighboring emission bands, the separate channels were scanned sequentially to exclude crosstalk. For imaging 40x or 63x water objectives were utilized. Laser intensity for excitation was chosen between 50-80%, exposure time around 200 ms per channel. Hoechst was imaged using 60% intensity for 100 ms. Laser intensity and exposure time was adjusted for each experiment and objective type. At least eight random fields of view were chosen per well and imaged in stacks of 3-6 planes, with a distance of 1 µm between planes.

Time courses were imaged in 15-20 min intervals over a course of 6-8 h. Images were uploaded to local server and analyzed using Columbus data storage and analysis system from Perkin Elmer. Interesting fields of view were chosen using Columbus and exported for further processing in ImageJ.

6.2.6. *In vitro* investigation of the HttTag

In vitro experiments were performed by [redacted] in the lab of [redacted]. In short, fusion proteins were overexpressed and purified from *E. coli.* His-Trx-HttTag$_{10x}$-HA and mVenus-V$_L$-HTT-His fusion proteins were purified with immobilized metal affinity chromatography (IMAC) and ion exchange chromatography (IEX). SDS PAGE was performed as described before[348]. Analytical size exclusion chromatography (SEC) was performed on a superpose 6 HR 10/30 column using the ÄKTA Purifier 10 liquid chromatography system (GE Healthcare) with PBS as a running buffer. Purified proteins or titration reactions (Trx-HttTag$_{10x}$ + mVenus-V$_L$-HTT) were applied to the column and respective elution volumes were measured. The apparent molecular masses were estimated by interpolation from a calibration curve obtained with reference proteins.

6.2.7. Detection and mapping of genomic dI

6.2.7.1. gDNA isolation for dIVe and dI quantification by LC-MS/MS

DNA was isolated using the DNeasy Blood & Tissue kit from QIAGEN according to the manufacturer's instructions with the following modifications: all buffers and solutions were supplemented with 100 nM pentostatin to avoid unspecific deamination by free adenosine deaminases. Up to 5x10^6 cells were resuspended in 200 µl PBS and incubated with 200 µl lysis buffer AL, 20 µl Proteinase K; 4 µl RNAseA for 45 min at 56 °C, shaking at 900 rpm.

Samples were then cooled to RT for 15 min and incubated with 1 µl RNAse 1 for 25 min. After adding 200 µl of 100% ethanol, lysates were transferred to the spin column and further steps were performed according to manufacturer's instructions. The gDNA was eluted with 2x 100 µl dI-buffer and concentration was determined using Nanodrop 2000.

6.2.7.2. Quantitative measurement of deoxyinosine

dI and rI detection by LC-MS/MS was established and performed by [redacted]. In brief, DNA or RNA was degraded to nucleosides with 0.003 U nuclease P1, 0.02 U snake venom phosphodiesterase and 0.2 U alkaline phosphatase. Up to 4 µg, or 1 µg of degraded DNA or RNA was injected into LC system, respectively. Nucleosides were separated on an Agilent 1290 Infinity Binary LC system (Agilent Technologies) with a 15 cm ReproSil 100 C18 column (Jasco) and detected by a triple-quadrupole mass spectrometer (Agilent 6490, Agilent Technologies). Running solutions were 5 mM ammonium acetate, pH 6.9 (A) and acetonitrile (B). Separations were performed with the following gradient: 1 min 0% B, 19 min linear increase to 10% B, 1 min 10% B, 5 min 50% B. The flow rate was 20 min 0.5 ml/min, 1 min gradual increase from 0.5 ml/min to 1 ml/min, 9 min 1 ml/min and 3 min 0.5 ml/min[64].

For LC-MS/MS based dI quantification, the samples were first mixed with the isotopic standard mix containing, 15 N 5 -dA and 15 N 4 -dI. Quantification of highly abundant dA was performed using 100× diluted samples. Data was analyzed with Agilent MassHunter Quantitative Analysis software v.B.09.00 (Agilent technologies) using isotopic standards to confirm the peak identity. Areas of the integrated peaks were exported into Microsoft Excel with which the areas were normalized to the area of the corresponding isotopic standard. Absolute amounts of the nucleosides were calculated using linear interpolation from a standard curve. Linear interpolation was performed using the two closely matching data points from the standard curve. Isotopic standards were spiked into the mixture of isotopic standards to normalize for ionization variability. The standard curve for every nucleoside was prepared to cover the amount of the corresponding nucleoside in the DNA sample analyzed. All described steps were performed by [redacted].

6.2.7.3. Preparation of dI spike-in oligonucleotides

DNA spike-ins were dissolved to 10 µM in 400 µl dI-buffer. 20 µl forward and reverse oligonucleotide were mixed with 5 µl 20x SSC buffer and 55 µl H_2O. Oligonucleotides were annealed in a PCR block by heating to 95 °C for 5 min and then switching off the block to cool down to room temperature. dI and control spike-ins were diluted 1:1; aliquoted to 6 µl and stored at -20 °C. Aliquots were only used once to avoid unspecific deamination during freeze thaw cycles.

6.2.7.4. Generation of modified PCR amplicons

For each 50 µl PCR reaction, 1 µl 20 ng/µl PT022 template, 5 µl 10x GeneAmp Buffer, 3 µl 25 mM $MgCl_2$, 1 µl 50 µM primer mixture (forward: GAAGCTTGGATCCAGGTGGA; reverse: GGGTTTTACCGTTCAGG), 38 µl H2O, 0.5 µl Taq Polymerase (IMB CF), and 2 µl 5mM dNTP solution were mixed. The 5 mM dNTP solution was pipetted separately for each sample, with equal amounts of dATP/CTP/GTP/TTP. To incorporate modified nucleotides, 20% of a standard dNTP were substituted with a modified version, as shown below:

Modified base	Substituted dNTP
dI	dGTP
m6dA	dATP
mC/hmC/fC/caC	dCTP
8oxoG	dTTP

PCR amplification was performed with the following setting

	PCR step	Temperature	Time
1	Initial Denaturation	95 °C	2 min
2	Denaturation	95 °C	30 s
3	Annealing	61 °C	30 s
4	Elongation	72 °C	70 s
	Go to step 2 and repeat 29 cycles		
5	Final Elongation	72 °C	5 min
6	STOP		

PCR amplicons were purified by QIAGEN PCR purification Kit and concentrations were determined by Nanodrop 2000.

6.2.7.5. Deoxyinosine dot blot

Serial dilutions between 1 and 0.25 ng/µl in 150 µl were prepared in 6x SSC buffer. For blotting, a nitrocellulose membrane and whatman papers were soaked in 20x SSC buffer for 10 min and assembled in a Bio-Dot Apparatus with the membrane on top of 2 whatman papers. The wells were washed 2x with 200 µl 20xSSC buffer by applying vacuum to all wells. To ensure consistent loading all wells were filled with equal volumes for washing and sample loading. Empty wells were filled with 20x SSC for compensation. Samples were denature 3 min at 95°C and put on ice for 10 min. 100 µl of the samples were loaded to the wells by applying vacuum. The membrane was dried for 10 min at room temperature and subsequently UV-crosslinked at 3x1400 kJ. After crosslinking, the membrane was blocked for 1 h with 5% BSA in PBS-T and then incubated on a sample rocker at 4 °C overnight with 1:5000 dI antibody in 1% BSA

PBS-T. After washing 3x 5 min at room temperature with PBS-T, the secondary antibody (goat-anti-rabbit-HRP conjugated antibody from Dianova) was diluted 1:20000 in 1% BSA PBS-T and incubated for 1 h on a sample rocker. The membrane was washed 3x 10 min with PBS-T and signals were developed with SuperSignal West Pico or Femto Chemiluminescent Substrate and analyzed using a ChemiDoc with Image Lab software. Membrane was washed 3x 5 min with PBS-T and stained with methylene blue to confirm equal loading.

6.2.7.6. dI DIP

After gDNA purification material was sonicated at 50 ng/µl in a total volume of 220 µl using the Bioruptor pico for 9 cycles 30/30 s on/off, using 1.5 ml bioruptor tubes. Sufficient sonication was confirmed by checking 200 ng of sample on 1% agarose gel.

1:90000 dilution was generated from 2 µM annealed spike-in oligos. 200 µl fragmented DNA was mixed with 1 µl spike-in dilution and 20 µl IP buffer (w/o triton). Samples were denatured for 5 min at 95 °C and directly put on ice for 10 min. All following steps were conducted on ice and with ice-cold buffers and solutions. 78 µl IP buffer (w/o triton), 1 µl 100 µM pentostatin, 1 µl 1M DTT, 25 µl 20% Triton X-100, 10 µl 1% BSA and 665 µl H2O were to added each sample. 25 µl of each sample were taken as input samples and stored at 4°C. Anti-inosine antibody was diluted 1:1 in DIP buffer and 2 µl were transferred to each tube. The same volume of IgG antibody was added to the IgG control samples. Samples were rotated at 10 rpm overnight at 4 °C. In addition, 50 µl ProteinG Dynabeads were prepared per sample. Beads were washed 2x 5 min in DIP buffer. Subsequently, beads were incubated in DIP block buffer at 4 °C overnight.

After 16 h DIP buffer was prepared freshly and beads were washed 2x for 5 min using DIP buffer. Beads were resuspended in initial volume using 1x DIP buffer and 50 µl beads were transferred to each pulldown sample. Samples were rotated for 2 h at 4 °C, 10 rpm. Beads were recovered using magnetic rack and were washed 4x 10 min with 1000 µl 1x DIP buffer at 4°C. Subsequently, 200 µl fresh PD digest buffer was added to DIP and input samples. All samples were incubated for 60 min in ThermoMixer at 56 °C, 900 rpm. Samples were briefly inverted and spun down at 500 xg for 5 s to collect all liquid at the bottom at the tube. Dynabeads were collected with magnetic rack and supernatant was transferred to fresh 1.5 ml DNA-low-bind tube. Input samples were directly used for further processing. Eluted DNA fragments were purified by Zymo Chip grade PCR purification, supplementing all buffers with pentostatin. In brief, 1 ml binding buffer was added to digested samples. 2x 650 µl were loaded to each column and washed 2x with 250 µl washing buffer. To remove remaining buffer, columns were centrifuged at max speed. Samples were eluted in 9 µl dI-buffer. Material was than stored at -20 °C or directly used for qubit, qPCR or NGS library prep.

6.2.7.7. dIVe: dI EndonucleaseV enrichment

After gDNA purification material was sonicated at 50 ng/µl in a total volume of 220 µl using the Bioruptor pico for 9 cycles 30/30 s on/off, using 1.5 ml bioruptor tubes. Sufficient sonication was confirmed by checking 200 ng of sample on 1% agarose gel.

2x dIVe buffer was prepared freshly directly before use. 1:90000 dilution was generated from 2 µM annealed spike-in oligos. 10 µg fragmented DNA was mixed with 300 µl H2O + 500 µl 2x dIVe buffer. 1 µl spike-in dilution was added per sample. EndoV (NEB) was diluted 1:10 in 1x dIVe buffer; and 2 µl of dilution was added to each pulldown sample. Samples were mixed carefully and 25 µl of each sample was taken as input sample. Samples were rotated at 10 rpm overnight at 4 °C. In addition, 50 µl anti-MBP bead slurry was prepared per sample. Beads were washed 2x in dIVe wash buffer. Subsequently beads were incubated in dIVe block buffer at 4 °C overnight.

After 16 h samples were transferred to rotating wheel, at room temperature. Accordingly, input samples were taken from fridge and stored at room temperature. 1x dIVe buffer was prepared freshly and beads were washed 2x for 5 min using 1x dIVe buffer. Beads were resuspended in initial volume using 1x dIVe buffer and 50 µl beads were transferred to each pulldown sample. Samples were rotated for another 2 h at RT, 10 rpm. Beads were recovered using magnetic rack and were washed 4x 10 min with 1000 µl 1x dIVe buffer. Subsequently, 200 µl fresh PD digest buffer was added to dIVe- and input samples. All samples were incubated for 60 min in ThermoMixer at 56 °C, 900 rpm. Samples were briefly inverted and spun down at 500 xg for 5 s to collect all liquid at the bottom at the tube. Dynabeads were collected with magnetic rack and supernatant was transferred to fresh 1.5 ml DNA-low-bind tubes. Input samples were directly used for further processing. Eluted DNA fragments were purified by Zymo Chip grade PCR purification, supplementing all buffers with pentostatin. In brief, 1 ml binding buffer was added to digested samples. 2x 650 µl were loaded to each column and washed 2x with 250 µl washing buffer. To remove remaining buffer, columns were centrifuged at max speed. Samples were eluted in 9 µl dI-buffer. Material was than stored at -20 °C or directly used for qubit, qPCR or NGS library prep.

6.2.7.8. dI sensitive RT-qPCR

RT-qPCR was conducted in a 384-well format using the Roche LightCycler. Gene-specific primers were determined using primer3 webtool. Each sample was amplified with Q5 and PhuU master mix. Q5 is sensitive to dI and therefore results in relatively lower signal compared to PhuU signal. 5 µl diluted dIVe sample were mixed with 2.2 µl 5x polymerase buffer, 0.22 primer mixture containing forward and reverse primer at 50 mM each; 0.22 µl 10 mM dNTPs, 0.1 µl Polymerase, 0.55 µl EvaGreen, 2.7 µl H_2O.

qPCR assay was performed with the same setting as the standard SYBR assay (6.2.1.3), however, the annealing temperature was increased to 60 °C.

6.2.7.9. *In vitro* removal of dI by WGA or MPG treatment

WGA was performed with REPLI-g Kit from Qiagen, according to suppliers' protocol with the following modifications: all buffers and solutions were supplemented with 100 nM pentostatin to avoid unspecific deamination by free adenosine deaminases. In short 20 ng purified gDNA were chemically denatured with denaturing buffer (D1) for 3 min at room temperature. Denaturing was stopped with neutralization buffer (N1) and WGA was started by adding the provided polymerase and WGA buffer. Samples were incubated for 16 h at 30 °C. Polymerase was heat-inactivated for 5 min at 65 °C and WGA samples were purified by ethanol precipitation (according to supplementary protocol from Qiagen). Pellets were resolved in 100 µl dI-buffer and the concentration was determined by NanoDrop 2000. WGA samples were stored at -20 °C for subsequent sonication and dIVe.

MPG treatment was performed on sonicated gDNA using commercial enzyme from NEB (hAAG). 100 µl of fragmented gDNA (50ng/µl) was mixed with 25 µl ThermoPol reaction buffer, 0.25 µl pentostatin, 120 µl H2O and 2 µl MPG enzyme. Samples were incubated for 1 h at 37 °C. As a control, the same reaction was performed w/o MPG enzyme. The samples were purified with QIAquick Nucleotide Removal Kit (Qiagen) according to suppliers' protocol with the following modifications: all buffers and solutions were supplemented with 100 nM pentostatin. The final elution was done in 50 µl dI-buffer and the concentration was determined by NanoDrop 2000. Samples were stored at -20 °C for subsequent dIVe. Since the material was sonicated before the MPG treatment, no additional sonication was performed before dIVe.

6.2.7.10. Establishing dI library preparation

Modified and unmodified DNA oligonucleotides were annealed and subjected to library preparation with NEBNext II Ultra Kit. Steps were performed according to manufacturer's protocol, but with different polymerases and the respective buffer: Q5-HF (NEB; provided in the KIT), OneTaq (NEB), Phusion (NEB), and Phusion Uracil (ThermoFisher). Library quality was checked by bioanalyzer and qubit. Since Q5 and Phusion polymerases failed to produce libraries, only OneTaq and PhuU libraries were sequenced. Sequencing was performed using MiSeq (Illumina) sequencing platform with a Nano v2, 300 cycle Kit. [redacted] performed the initial data analysis. Transition mutations were analyzed by mapping reads to the oligo sequence and representing the sequence conservation as sequence logo.

For ssDNA library preparation Accel NGS 1S Kit was tested with PhuU polymerase (Thermo Fisher). Successful library preparation was confirmed by bionalyzer and qubit assay before initial sequencing (data not shown).

6.2.7.11. dIVe sequencing

Library preparation and sequencing was performed by IMB Genomics CF. ds-dIVe samples were analyzed by qubit dsDNA HS assay and bioanalyzer before library preparation. For library preparation 2 ng starting material was employed using the NEBNext Ultra II Kit for Illumina. Library preparation was performed according to manufacturer's protocol substituting the Q5-HF polymerase and buffer with the PhuU polymerase and buffer from Thermo Fisher. After purification, the library quality was verified by bioanalyzer and qubit measurement. Sequencing was performed using Illumina NextSeq500 platform with high output 150 cycle kit, with either single end, 150 cycles, or paired end, 2x75 cycles.

ss-dIVe samples were analyzed by qubit ssDNA assay and bioanalyzer before library preparation. For library prep 3 ng starting material was utilized using the Accel-NGS® 1S Plus DNA Library Kit (Bioscience) substituting the included polymerases and corresponding buffers with the PhuU polymerase and buffer from Thermo Fisher. After purification, the library quality was verified by bioanalyzer and qubit measurement. Sequencing was performed using Illumina NextSeq500 platform with high output 150 cycle kit, in paired end mode (2x75 cycles).

6.2.8. NGS data processing

Bioinformatics analysis was performed by myself using the open source galaxy platform (usegaalxy.eu). Additionally, data was cross-validated by [redacted]. The data presented in this thesis represents my own analysis, if not stated otherwise. Repeat analysis in unmapped reads was performed by [redacted].

6.2.8.1. Read mapping and annotation

Low quality reads and adapters were removed using Trim Galore![349]. Quality was checked using FastQC[350]. Reads were mapped with Bowtie2[351,352] using hg38 or mm10 as reference assemblies. Peak calling of dIVe samples over input was performed with a minimal mapping quality of 3, using MACS2[353,354].

Peak intersections and annotations were generated using bedtools multiple intersect[355], Upset diagram[356], Homer - AnnotatePeaks[357]. Bedtools coverage[355], was used to determine the overlap of dIVe peaks with annotated repeats from repeat masker file (UCSC).

6.2.8.2. Generation of shuffled control data

Control peaks were generated with bedtools shuffleBed[355]. Reverence peaks were shuffled over hg38 or mm10 ChromosomeInfo genome file from UCSC, with exclusion of "hg38/mm10: gaps". Features were kept on same chromosome while shuffling. Shuffling and corresponding analysis was done in triplicates.

6.2.8.3. Transition analysis and SNP counting

After read trimming and mapping, variants were detected using samtools mpileup[358] and VarScan mplieup[269]. Variants were determined for the peak area requiring a minimal coverage of four reads. As a non-peak control, a random 1% fraction of all detected variants was selected. Mutation frequencies were grouped in four groups 0.1-24.9; 25-49.9%; 50-74.9%; 75-99.9%. MEME ChiP[359] was used for de-novo motif detection in selected Peaks. SNPs in peak and control regions were counted using Bedtools intersect intervals[355] with hg38 dbSnp153 single nucleotide variants (UCSC).

6.2.8.4. Tandem repeat analysis of unmapped reads

The tandem repeat analysis was established and performed by [redacted]. In brief, dIVe-seq reads were trimmed and then for each read, a de novo search for repeats was performed using Phobos[360]. The detected repeats were output in alphabetic normal form, and then filtered by repeat region length (>=0.60bp) and repeat perfection (>=90%). The longest remaining repeat per read was selected and all discovered repeats were grouped and their numbers counted. EdgeR[361–363] was used to normalize samples and to analyze the fold change of a specific repeat in pulldown over controls.

7. References

1. Bickmore, W. A. The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.* **14**, 67–84 (2013).
2. Anania, C. & Lupiáñez, D. G. Order and disorder: abnormal 3D chromatin organization in human disease. *Brief. Funct. Genomics* **19**, 128–138 (2020).
3. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature* **292**, (2001).
4. Parada, L. A., McQueen, P. G., Munson, P. J. & Misteli, T. Conservation of Relative Chromosome Positioning in Normal and Cancer Cells. *Curr. Biol.* **12**, 1692–1697 (2002).
5. Ronneberger, O. *et al.* Spatial quantitative analysis of fluorescently labeled nuclear structures: problems, methods, pitfalls. *Chromosome Res.* **16**, 523–562 (2008).
6. Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10**, 211–219 (2001).
7. Cremer, M. *et al.* Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Res.* **9**, 541–567 (2001).
8. Küpper, K. *et al.* Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma* **116**, 285–306 (2007).
9. Boyle, S., Rodesch, M. J., Halvensleben, H. A., Jeddeloh, J. A. & Bickmore, W. A. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. *Chromosom. Res.* **19**, 901–909 (2011).
10. Van Koningsbruggen, S. *et al.* High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol. Biol. Cell* **21**, 3735–3748 (2010).
11. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).
12. Spector, D. L. & Lamond, A. I. Nuclear speckles. *Cold Spring Harb. Perspect. Biol.* **3**, 1–12 (2011).
13. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
14. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
15. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
16. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
17. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

18. Parelho, V. *et al.* Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell* **132**, 422–433 (2008).

19. Wang, H., Han, M. & Qi, L. S. Engineering 3D genome organization. *Nature Reviews Genetics* **22**, 343–360 (2021).

20. Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nat. 2008 4517180* **451**, 796–801 (2008).

21. Davidson, I. F. & Peters, J.-M. Genome folding through loop extrusion by SMC complexes. *Nat. Rev. Mol. Cell Biol.* **22**, 445 (2021).

22. Amano, T. *et al.* Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev. Cell* **16**, 47–57 (2009).

23. Williamson, I. *et al.* Developmentally regulated Shh expression is robust to TAD perturbations. *Dev.* **146**, (2019).

24. Paliou, C. *et al.* Preformed chromatin topology assists transcriptional robustness of Shh during limb development. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12390–12399 (2019).

25. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305–320 (2017).

26. Busslinger, G. A. *et al.* Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* **544**, 503–507 (2017).

27. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 996–1001 (2014).

28. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).

29. Szabo, Q. *et al.* Regulation of single-cell genome organization into TADs and chromatin nanodomains. *Nat. Genet.* **52**, 1151–1157 (2020).

30. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: Advances through genomics. *Nature Reviews Genetics* **10**, 161–172 (2009).

31. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).

32. Zhang, T., Cooper, S. & Brockdorff, N. The interplay of histone modifications - writers that read. *EMBO Rep.* **16**, 1467–1481 (2015).

33. Paull, T. T. *et al.* A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage. *Curr. Biol.* **10**, 886–895 (2000).

34. Arnaudo, A. M. & Garcia, B. A. Proteomic characterization of novel histone post-translational modifications. *Epigenetics and Chromatin* **6**, 1–7 (2013).

35. Tan, M. *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146**, 1016–1028 (2011).

36. Gruenbaum, Y., Cedar, H. & Razin, A. Substrate and sequence specificity of a eukaryotic DNA methylase. *Nature* **295**, 620–622 (1982).

37. Bird, A. P. CpG-Rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).

38. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).

39. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72–77 (2017).

40. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).

41. Wu, H. & Zhang, Y. Reversing DNA methylation: Mechanisms, genomics, and biological functions. *Cell* **156**, 45–68 (2014).

42. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–5 (2009).

43. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–3 (2011).

44. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472–9 (2013).

45. Song, C. X. & He, C. Potential functional roles of DNA demethylation intermediates. *Trends Biochem. Sci.* **38**, 480–484 (2013).

46. Song, C. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678–691 (2013).

47. Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S. & Heintz, N. MeCP2 binds to 5hmc enriched within active genes and accessible chromatin in the nervous system. *Cell* **151**, 1417 (2012).

48. Lu, X. *et al.* Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Research* **25**, 386–389 (2015).

49. Boulias, K. & Greer, E. L. Means, mechanisms and consequences of adenine methylation in DNA. *Nature Reviews Genetics* **23**, 411–428 (2022).

50. Musheev, M. U., Baumgärtner, A., Krebs, L. & Niehrs, C. The origin of genomic N 6-methyl-deoxyadenosine in mammalian cells. *Nat. Chem. Biol.* **16**, 630–634 (2020).

51. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.* **58**, 235–263 (2017).

52. Chon, J., Field, M. S. & Stover, P. J. Deoxyuracil in DNA and disease: Genomic signal or managed situation? *DNA Repair* **77**, 36–44 (2019).

53. Gorini, F., Scala, G., Cooke, M. S., Majello, B. & Amente, S. Towards a comprehensive view of 8-oxo-7,8-dihydro-2'-deoxyguanosine: Highlighting the intertwined roles of DNA damage and epigenetics in genomic instability. *DNA Repair* **97**, 103027 (2021).

54. Shu, X. *et al.* Genome-wide mapping reveals that deoxyuridine is enriched in the human centromeric DNA article. *Nat. Chem. Biol.* **14**, 680–687 (2018).

55. Bryan, D. S., Ransom, M., Adane, B., York, K. & Hesselberth, J. R. High resolution mapping of modified DNA nucleobases using excision repair enzymes. *Genome Res.* **24**, 1534–1542 (2014).

56. Su, X. A. & Freudenreich, C. H. Cytosine deamination and base excision repair cause R-loop–induced CAG repeat fragility and instability in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E8392–E8401 (2017).

57. Yu, S.-L., Lee, S.-K., Johnson, R. E., Prakash, L. & Prakash, S. The stalling of transcription at abasic sites is highly mutagenic. *Mol. Cell. Biol.* **23**, 382–388 (2003).

58. Kathe, S. D., Shen, G. P. & Wallace, S. S. Single-stranded breaks in DNA but not oxidative DNA base damages block transcriptional elongation by RNA polymerase II in HeLa cell nuclear extracts. *J. Biol. Chem.* **279**, 18511–18520 (2004).

59. Lühnsdorf, B., Epe, B. & Khobta, A. Excision of uracil from transcribed DNA negatively affects gene expression. *J. Biol. Chem.* **289**, 22008–22018 (2014).

60. Ding, Y., Fleming, A. M. & Burrows, C. J. Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J. Am. Chem. Soc.* **139**, 2569–2572 (2017).

61. Poetsch, A. R., Boulton, S. J. & Luscombe, N. M. Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *Genome Biol.* **19**, 1–23 (2018).

62. Shiromoto, Y., Sakurai, M., Minakuchi, M., Ariyoshi, K. & Nishikura, K. ADAR1 RNA editing enzyme regulates R-loop formation and genome stability at telomeres in cancer cells. *Nat. Commun.* **12**, 1–18 (2021).

63. Tang, S., Stokasimov, E., Cui, Y. & Pellman, D. Breakage of cytoplasmic chromosomes by pathological DNA base excision repair. *Nature* (2022). doi:10.1038/s41586-022-04767-1

64. Kijonka-Baumgärtner, A. Investigation / Characterization of base modifications in mammalian DNA. (University Mainz, 2021).

65. Van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nature Biotechnology* **28**, 1089–1095 (2010).

66. Molenaar, C. *et al.* Visualizing telomere dynamics in living mammalian cells using PNA probes. *EMBO J.* **22**, 6631–6641 (2003).

67. Fang, Y. & Spector, D. L. Centromere positioning and dynamics in living Arabidopsis plants. *Mol. Biol. Cell* **16**, 5710–5718 (2005).

68. Bronshtein, I. *et al.* Loss of lamin A function increases chromatin dynamics in the nuclear interior. *Nat. Commun.* **6**, 1–9 (2015).

69. Straight, A. F., Belmont, A. S., Robinett, C. C. & Murray, A. W. GFP tagging of budding yeast chromosomes reveals that protein–protein interactions can mediate sister chromatid cohesion. *Curr. Biol.* **6**, 1599–1608 (1996).

70. Michaelis, C., Ciosk, R. & Nasmyth, K. Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. *Cell* **91**, 35–45 (1997).

71. Lindhout, B. I. *et al.* Live cell imaging of repetitive DNA sequences via GFP-tagged polydactyl zinc finger proteins. *Nucleic Acids Res.* **35**, e107–e107 (2007).

72. Miyanari, Y., Ziegler-Birling, C. & Torres-Padilla, M. E. Live visualization of chromatin dynamics

with fluorescent TALEs. *Nat. Struct. Mol. Biol.* **20**, 1321–1324 (2013).

73. Thanisch, K. *et al.* Targeting and tracing of specific DNA sequences with dTALEs in living cells. *Nucleic Acids Res.* **42**, e38–e38 (2014).

74. Beerli, R. R. & Barbas, C. F. Engineering polydactyl zinc-finger transcription factors. *Nature Biotechnology* **20**, 135–141 (2002).

75. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509–1512 (2009).

76. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–822 (2012).

77. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J. Bacteriol.* **169**, 5429–33 (1987).

78. Mojica, F. J., Juez, G. & Rodríguez-Valera, F. Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Mol. Microbiol.* **9**, 613–21 (1993).

79. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–82 (2005).

80. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–63 (2005).

81. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–61 (2005).

82. Brouns, S. J. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–4 (2008).

83. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–7 (2011).

84. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015).

85. Wang, F. & Qi, L. S. Applications of CRISPR Genome Engineering in Cell Biology. *Trends Cell Biol.* **26**, 875–888 (2016).

86. Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164**, 29–44 (2016).

87. Park, C. Y. *et al.* Functional Correction of Large Factor VIII Gene Chromosomal Inversions in Hemophilia A Patient-Derived iPSCs Using CRISPR-Cas9. *Cell Stem Cell* **17**, 213–220 (2015).

88. Schwank, G. *et al.* Functional Repair of CFTR by CRISPR/Cas9 in Intestinal Stem Cell Organoids of Cystic Fibrosis Patients. *Cell Stem Cell* **13**, 653–658 (2013).

89. Dabrowska, M. & Olejniczak, M. Gene therapy for huntington's disease using targeted endonucleases. *Methods Mol. Biol.* **2056**, 269–284 (2020).

90. Zhao, Z. *et al.* CRISPR knock out of programmed cell death protein 1 enhances anti-tumor activity of cytotoxic T lymphocytes. *Oncotarget* **9**, 5208–5215 (2017).

91. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nat. 2015 5337603* **533**, 420–424 (2016).

92. Gaudelli, N. M. *et al.* Programmable base editing of T to G C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).

93. Liu, L. *et al.* The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. *Cell* **170**, 714-726.e10 (2017).

94. Abudayyeh, O. O. *et al.* RNA targeting with CRISPR–Cas13. *Nature* **550**, 280–284 (2017).

95. Gootenberg, J. S. *et al.* Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a and Csm6. *Science* **360**, 439–444 (2018).

96. Fu, Y. *et al.* CRISPR-dCas9 and sgRNA scaffolds enable dual-colour live imaging of satellite sequences and repeat-enriched individual loci. *Nat. Commun.* **7**, 11707 (2016).

97. Ma, H. *et al.* CRISPR-Sirius: RNA scaffolds for signal amplification in genome imaging. *Nat. Methods* **15**, 928–931 (2018).

98. Cheng, A. W. *et al.* Casilio: A versatile CRISPR-Cas9-Pumilio hybrid for gene regulation and genomic labeling. *Cell Research* **26**, 254–257 (2016).

99. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* **159**, 635–646 (2014).

100. Gu, B. *et al.* Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science* **359**, 1050–1055 (2018).

101. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).

102. Shao, S. *et al.* Multiplexed sgRNA Expression Allows Versatile Single Nonrepetitive DNA Labeling and Endogenous Gene Regulation. *ACS Synth. Biol.* **7**, 176–186 (2018).

103. Ochiai, H., Sugawara, T. & Yamamoto, T. Simultaneous live imaging of the transcription and nuclear position of specific genes. *Nucleic Acids Res.* **43**, e127–e127 (2015).

104. Breunig, C. T. *et al.* One step generation of customizable gRNA vectors for multiplex CRISPR approaches through string assembly gRNA cloning (STAgR). *PLoS One* **13**, e0196015 (2018).

105. Qin, P. *et al.* Live cell imaging of low-and non-repetitive chromosome loci using CRISPR-Cas9. *Nat. Commun.* **8**, (2017).

106. Heneen, W. K. & Heneen, W. K. HeLa cells and their possible contamination of other cell lines: Karyotype studies. *Hereditas* **82**, 217–247 (1976).

107. Hitoshi, N., Ken-ichi, Y. & Jun-ichi, M. Efficient selection for high-expression transfectants with a novel eukaryotic vector. *Gene* **108**, 193–199 (1991).

108. Schiefner, A. *et al.* A disulfide-free single-domain VL intrabody with blocking activity towards huntingtin reveals a novel mode of epitope recognition. *J. Mol. Biol.* **414**, 337–355 (2011).

109. Colca, J. R. *et al.* Identification of a novel mitochondrial protein ('mitoNEET') cross-linked specifically by a thiazolidinedione photoprobe. *Am. J. Physiol. - Endocrinol. Metab.* **286**, 252–260 (2004).

110. Zambrowicz, B. P. *et al.* Disruption of overlapping transcripts in the ROSA βgeo 26 gene trap strain leads to widespread expression of β-galactosidase in mouse embryos and hematopoietic cells. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 3789–3794 (1997).

111. Yoshiki, A., Ballard, G. & Perez, A. V. Genetic quality: a complex issue for experimental study reproducibility. *Transgenic Res.* 1–18 (2022).

112. Wurm, F. M. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotechnol* **22**, 1393–1398 (2004).

113. Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–548 (2015).

114. Chen, B. *et al.* Expanding the CRISPR imaging toolset with Staphylococcus aureus Cas9 for simultaneous imaging of multiple genomic loci. *Nucleic Acids Res.* **44**, 75 (2016).

115. Ma, D., Peng, S., Huang, W., Cai, Z. & Xie, Z. Rational Design of Mini-Cas9 for Transcriptional Activation. *ACS Synth. Biol.* **7**, 978–985 (2018).

116. Xu, X. *et al.* Engineered miniature CRISPR-Cas system for mammalian genome regulation and editing. *Mol. Cell* **81**, 4333-4345.e4 (2021).

117. Pickett, H. A., Henson, J. D., Au, A. Y. M., Neumann, A. A. & Reddel, R. R. Normal mammalian cells negatively regulate telomere length by telomere trimming. *Hum. Mol. Genet.* **20**, 4684–4692 (2011).

118. Kahl, V. F. S. *et al.* Telomere Length Measurement by Molecular Combing. *Front. Cell Dev. Biol.* **8**, 493 (2020).

119. Baur, J. A., Zou, Y., Shay, J. W. & Wright, W. E. Telomere position effect in human cells. *Science* **292**, 2075–2077 (2001).

120. Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* **156**, 935–949 (2014).

121. Guo, D. G. *et al.* CRISPR-based genomic loci labeling revealed ordered spatial organization of chromatin in living diploid human cells. *Biochim. Biophys. Acta - Mol. Cell Res.* **1866**, 118518 (2019).

122. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–32 (2013).

123. Calado, R. T. & Dumitriu, B. Telomere Dynamics in Mice and Humans. *Semin. Hematol.* **50**, 165–174 (2013).

124. Wu, X., Mao, S., Ying, Y., Krueger, C. J. & Chen, A. K. Progress and Challenges for Live-cell Imaging of Genomic Loci Using CRISPR-based Platforms. *Genomics, Proteomics and Bioinformatics* **17**, 119–128 (2019).

125. Miao, Y. *et al.* Overcoming diverse homologous recombinations and single chimeric guide RNA competitive inhibition enhances Cas9-based cyclical multiple genes coediting in filamentous fungi. *Environ. Microbiol.* **23**, 2937–2954 (2021).

126. Mekler, V., Minakhin, L., Semenova, E., Kuznedelov, K. & Severinov, K. Kinetics of the CRISPR-Cas9 effector complex assembly and the role of 3'-terminal segment of guide RNA. *Nucleic Acids Res.* **44**, 2837–2845 (2016).

127. Efroni, S. *et al.* Global Transcription in Pluripotent Embryonic Stem Cells. *Cell Stem Cell* **2**, 437–447 (2008).

128. Cornett, J. *et al.* Polyglutamine expansion of huntingtin impairs its nuclear export. *Nat. Genet.* **37**, 198–204 (2005).

129. Thakur, A. K. *et al.* Polyglutamine disruption of the huntingtin exon 1 N terminus triggers a complex aggregation mechanism. *Nat. Struct. Mol. Biol.* **16**, 380–389 (2009).

130. Tam, S. *et al.* The chaperonin TRiC blocks a huntingtin sequence element that promotes the conformational switch to aggregation. *Nat. Struct. Mol. Biol.* (2009). doi:10.1038/nsmb.1700

131. Sharp, A. H. *et al.* Widespread expression of Huntington's disease gene (IT15) protein product. *Neuron* **14**, 1065–1074 (1995).

132. Trottier, Y. *et al.* Cellular localization of the Huntington's disease protein and discrimination of the normal and mutated form. *Nat. Genet.* **10**, 104–110 (1995).

133. Boersma, S. *et al.* Multi-Color Single-Molecule Imaging Uncovers Extensive Heterogeneity in mRNA Decoding. *Cell* **178**, 458-472.e19 (2019).

134. Clow, P. A. *et al.* CRISPR-mediated multiplexed live cell imaging of nonrepetitive genomic loci with one guide RNA per locus. *Nat. Commun.* **13**, 1–10 (2022).

135. Zhang, X. *et al.* Gene activation in human cells using CRISPR/Cpf1-p300 and CRISPR/Cpf1-SunTag systems. *Protein and Cell* **9**, 380–383 (2018).

136. Huang, Y. H. *et al.* DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. *Genome Biol.* **18**, 1–11 (2017).

137. Hanzawa, N. *et al.* Targeted DNA demethylation of the Fgf21 promoter by CRISPR/dCas9-mediated epigenome editing. *Sci. Rep.* **10**, 1–14 (2020).

138. Morita, S., Horii, T., Kimura, M. & Hatada, I. Synergistic Upregulation of Target Genes by TET1 and VP64 in the dCas9–SunTag Platform. *Int. J. Mol. Sci. 2020, Vol. 21, Page 1574* **21**, 1574 (2020).

139. Sood, A. J., Viner, C. & Hoffman, M. M. DNamod: The DNA modification database. *J. Cheminform.* **11**, (2019).

140. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187–1200 (2017).

141. Boccaletto, P. *et al.* MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res.* **50**, D231–D235 (2022).

142. Huang, H., Weng, H., Deng, X. & Chen, J. RNA Modifications in Cancer: Functions, Mechanisms, and Therapeutic Implications. *Annual Review of Cancer Biology* **4**, 221–240 (2020).

143. Saletore, Y. *et al.* The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.* **13**, 175 (2012).

144. Slotkin, W. & Nishikura, K. Adenosine-to-inosine RNA editing and human disease. *Genome Med.* **5**, 1–13 (2013).

145. Picardi, E. *et al.* Profiling RNA editing in human tissues: Towards the inosinome Atlas. *Sci. Rep.* **5**, 1–17 (2015).

146. Alseth, I., Dalhus, B. & Bjørås, M. Inosine in DNA and RNA. *Curr. Opin. Genet. Dev.* **26**, 116–123 (2014).

147. Gerber, A. P. & Keller, W. RNA editing by base deamination: more enzymes, more targets, new mysteries. *Trends Biochem. Sci.* **26**, 376–384 (2001).

148. Kim, U., Wang, Y., Sanford, T., Zeng, Y. & Nishikura, K. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 11457–11461 (1994).

149. Melcher, T. *et al.* A mammalian RNA editing enzyme. *Nature* **379**, 460–464 (1996).

150. Chen, C. X. *et al.* A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA* **6**, 755 (2000).

151. Maas, S., Gerber, A. P. & Rich, A. *Identification and characterization of a human tRNA-specific adenosine deaminase related to the ADAR family of pre-mRNA editing enzymes*. *Biochemistry* **96**, (1999).

152. Jin, Y., Zhang, W. & Li, Q. Origins and evolution of ADAR-mediated RNA editing. *IUBMB Life* **61**, 572–578 (2009).

153. Ryter, J. M. & Schultz, S. C. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.* **17**, 7505 (1998).

154. Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annual Review of Biochemistry* **79**, 321–349 (2010).

155. Patterson, J. B. & Samuel, C. E. Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Mol. Cell. Biol.* **15**, 5376–5388 (1995).

156. Li, H. *et al.* Human genomic Z-DNA segments probed by the Zα domain of ADAR1. *Nucleic Acids Res.* **37**, 2737–2746 (2009).

157. Eisenberg, E. & Levanon, E. Y. A-to-I RNA editing — immune protector and transcriptome diversifier. *Nat. Rev. Genet. 2018 198* **19**, 473–490 (2018).

158. Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nature Reviews*

Molecular Cell Biology **17**, 83–96 (2016).

159. George, C. X. & Samuel, C. E. Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 4621–4626 (1999).

160. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

161. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580–585 (2013).

162. Oakes, E., Anderson, A., Cohen-Gadol, A. & Hundley, H. A. Adenosine Deaminase That Acts on RNA 3 (ADAR3) Binding to Glutamate Receptor Subunit B Pre-mRNA Inhibits RNA Editing in Glioblastoma. *J. Biol. Chem.* **292**, 4326–4335 (2017).

163. Tan, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550**, 249–254 (2017).

164. Martin, F. H., Castro, M. M., Aboul-Ela, F. & Tinoco, I. Base pairing involving deoxyinosine: implications for probe design. *Nucleic Acids Res.* **13**, 8927 (1985).

165. Su, A. A. H. & Randau, L. A-to-I and C-to-U editing within transfer RNAs. *Biochemistry (Moscow)* **76**, 932–937 (2011).

166. Gerber, A. P. & Keller, W. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science* **286**, 1146–1149 (1999).

167. Rosenthal, J. J. C. & Seeburg, P. H. A-to-I RNA editing: effects on proteins key to neural excitability. *Neuron* **74**, 432–439 (2012).

168. Hoopengardner, B., Bhalla, T., Staber, C. & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301**, 832–836 (2003).

169. Sommer, B., Köhler, M., Sprengel, R. & Seeburg, P. H. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11–19 (1991).

170. Danie, C., Wahlstedt, H., Ohlson, J., Björk, P. & Öhman, M. Adenosine-to-Inosine RNA Editing Affects Trafficking of the γ-Aminobutyric Acid Type A (GABAA) Receptor. *J. Biol. Chem.* **286**, 2031–2040 (2011).

171. Yeo, J., Goodman, R. A., Schirle, N. T., David, S. S. & Beal, P. A. RNA editing changes the lesion specificity for the DNA repair enzyme NEIL1. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 20715–20719 (2010).

172. Lev-Maor, G. *et al.* RNA-editing-mediated exon evolution. *Genome Biol.* **8**, 1–12 (2007).

173. Ivanov, A. *et al.* Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep.* **10**, 170–177 (2015).

174. Pinto, Y., Buchumenski, I., Levanon, E. Y. & Eisenberg, E. Human cancer tissues exhibit reduced A-to-I editing of miRNAs coupled with elevated editing of their targets. *Nucleic Acids Res.* **46**, 71–82 (2018).

175. Wang, Y. *et al.* Systematic characterization of A-to-I RNA editing hotspots in microRNAs across human cancers. *Genome Res.* **27**, 1112–1125 (2017).

176. Tomaselli, S. *et al.* ADAR enzyme and miRNA story: A nucleotide that can make the difference. *International Journal of Molecular Sciences* **14**, 22796–22816 (2013).

177. Liddicoat, B. J. *et al.* RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science* **349**, 1115–1120 (2015).

178. Dias Junior, A. G., Sampaio, N. G. & Rehwinkel, J. A Balancing Act: MDA5 in Antiviral Immunity and Autoinflammation. *Trends in Microbiology* **27**, 75–85 (2019).

179. Paz, N. *et al.* Altered adenosine-to-inosine RNA editing in human cancer. *Genome Res.* **17**, 1586–1595 (2007).

180. Ishiuchi, S. *et al.* Ca2+-Permeable AMPA Receptors Regulate Growth of Human Glioblastoma via Akt Activation. *J. Neurosci.* **27**, 7987–8001 (2007).

181. Han, L. *et al.* The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer Cell* **28**, 515–528 (2015).

182. Cesarini, V. *et al.* ADAR2/miR-589-3p axis controls glioblastoma cell migration/invasion. *Nucleic Acids Res.* **46**, 2045–2059 (2018).

183. Khermesh, K. *et al.* Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. *RNA* **22**, 290–302 (2016).

184. Eran, A. *et al.* Comparative RNA editing in autistic and neurotypical cerebella. *Mol. Psychiatry* **18**, 1041–1048 (2013).

185. Srivastava, P. K. *et al.* Genome-wide analysis of differential RNA editing in epilepsy. *Genome Res.* **27**, 440–450 (2017).

186. Silberberg, G., Lundin, D., Navon, R. & Öhman, M. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. *Hum. Mol. Genet.* **21**, 311–321 (2012).

187. Gal-Mark, N. *et al.* Abnormalities in A-to-I RNA editing patterns in CNS injuries correlate with dynamic changes in cell type composition. *Sci. Rep.* **7**, (2017).

188. Gurevich, I. *et al.* Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron* **34**, 349–356 (2002).

189. Higuchi, M. *et al.* Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406**, 78–81 (2000).

190. Kawahara, Y. *et al.* RNA editing and death of motor neurons. *Nat. 2004 4276977* **427**, 801–801 (2004).

191. Yamashita, T. & Kwak, S. Cell death cascade and molecular therapy in ADAR2-deficient motor neurons of ALS. *Neurosci. Res.* **144**, 4–13 (2019).

192. Rice, G. I. *et al.* Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature. *Nat. Genet.* **44**, 1243–1248 (2012).

193. Song, B., Shiromoto, Y., Minakuchi, M. & Nishikura, K. The role of RNA editing enzyme ADAR1

in human disease. *Wiley Interdiscip. Rev. RNA* **13**, e1665 (2022).

194. Karran, P. & Lindahl, T. Hypoxanthine in Deoxyribonucleic Acid: Generation by Heat-Induced Hydrolysis of Adenine Residues and Release in Free Form by a Deoxyribonucleic Acid Glycosylase from Calf Thymus. *Biochemistry* **19**, 6005–6011 (1980).

195. Myrnes, B., Guddal, P. H. & Krokan, H. Metabolism of dITP in hela cell extracts, incorporation into DNA by isolated nuclei and release of hypoxanthine from DNA by a hypoxanthine-DNA glycosylase activity. *Nucleic Acids Res.* **10**, 3693–3701 (1982).

196. Taghizadeh, K. *et al.* Quantification of DNA damage products resulting from deamination, oxidation and reaction with products of lipid peroxidation by liquid chromatography isotope dilution tandem mass spectrometry. *Nat. Protoc.* **3**, 1287–1298 (2008).

197. Pang, B. *et al.* Lipid peroxidation dominates the chemistry of DNA adduct formation in a mouse model of inflammation. *Carcinogenesis* **28**, 1807–1813 (2007).

198. Abolhassani, N. *et al.* NUDT16 and ITPA play a dual protective role in maintaining chromosome stability and cell growth by eliminating dIDP/IDP and dITP/ITP from nucleotide pools in mammals. *Nucleic Acids Res.* **38**, 2891–2903 (2010).

199. Yasui, M. *et al.* Miscoding Properties of 2′-Deoxyinosine, a Nitric Oxide-Derived DNA Adduct, during Translesion Synthesis Catalyzed by Human DNA Polymerases. *J. Mol. Biol.* **377**, 1015–1023 (2008).

200. Grunebaum, E., Cohen, A. & Roifman, C. M. Recent advances in understanding and managing adenosine deaminase and purine nucleoside phosphorylase deficiencies. *Curr. Opin. Allergy Clin. Immunol.* **13**, 630–638 (2013).

201. Wiginton, D. A., Coleman, M. S. & Hutton, J. J. Characterization of purine nucleoside phosphorylase from human granulocytes and its metabolism of deoxyribonucleosides. *J. Biol. Chem.* **255**, 6663–6669 (1980).

202. Ealick, S. E. *et al.* Three-dimensional structure of human erythrocytic purine nucleoside phosphorylase at 3.2 A resolution. *J. Biol. Chem.* **265**, 1812–1820 (1990).

203. Sakumi, K. *et al.* ITPA protein, an enzyme that eliminates deaminated purine nucleoside triphosphates in cells. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis* **703**, 43–50 (2010).

204. Krokan, H. E. & Bjørås, M. Base Excision Repair. *Cold Spring Harb. Perspect. Biol.* **5**, 1–22 (2013).

205. Yao, M., Hatahet, Z., Melamede, R. J. & Kow, Y. W. Purification and characterization of a novel deoxyinosine-specific enzyme, deoxyinosine 3' endonuclease, from Escherichia coli. *J. Biol. Chem.* **269**, 16260–16268 (1994).

206. Lee, C. C. *et al.* Endonuclease V-mediated deoxyinosine excision repair in vitro. *DNA Repair (Amst).* **9**, 1073–1079 (2010).

207. Mi, R., Alford-Zappala, M., Kow, Y. W., Cunningham, R. P. & Cao, W. Human endonuclease V as a repair enzyme for DNA deamination. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **735**, 12–

18 (2012).

208. Yasui, A. Alternative excision repair pathways. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

209. Kuraoka, I. Diversity of endonuclease V: From DNA repair to RNA editing. *Biomolecules* **5**, 2194–2206 (2015).

210. Morita, Y. *et al.* Human endonuclease V is a ribonuclease specific for inosine-containing RNA. *Nat. Commun. 2013 41* **4**, 1–10 (2013).

211. Vik, E. S. *et al.* Endonuclease V cleaves at inosines in RNA. *Nat. Commun. 2013 41* **4**, 1–7 (2013).

212. Knutson, S. D., Arthur, R. A., Johnston, H. R. & Heemstra, J. M. Selective Enrichment of A-to-I Edited Transcripts from Cellular RNA Using Endonuclease v. *ACS Appl. Mater. Interfaces* (2020).

213. Knutson, S. D. & Heemstra, J. M. EndoVIPER-seq for Improved Detection of A-to-I Editing Sites in Cellular RNA. *Curr. Protoc. Chem. Biol.* **12**, e82 (2020).

214. Zheng, Y., Lorenzo, C. & Beal, P. A. DNA editing in DNA/RNA hybrids by adenosine deaminases that act on RNA. *Nucleic Acids Res.* **45**, 3369–3377 (2017).

215. Steele, E. J. & Lindley, R. A. ADAR deaminase A-to-I editing of DNA and RNA moieties of RNA:DNA hybrids has implications for the mechanism of Ig somatic hypermutation. *DNA Repair (Amst).* **55**, 1–6 (2017).

216. Rafail Nikolaos Tasakis, A. *et al.* ADAR1 can drive Multiple Myeloma progression by acting both as an RNA editor of specific transcripts and as a DNA mutator of their cognate genes. *bioRxiv* 2020.02.11.943845 (2020).

217. Popitsch, N. *et al.* A-to-I RNA Editing Uncovers Hidden Signals of Adaptive Genome Evolution in Animals. *Genome Biol. Evol.* **12**, 345–357 (2020).

218. Niehrs, C. & Luke, B. Regulatory R-loops as facilitators of gene expression and genome stability. *Nature Reviews Molecular Cell Biology* **21**, 167–178 (2020).

219. Wahba, L., Gore, S. K. & Koshland, D. The homologous recombination machinery modulates the formation of RNA-DNA hybrids and associated chromosome instability. *Elife* **2013**, (2013).

220. Grunseich, C. *et al.* Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. *Mol. Cell* **69**, 426-437.e7 (2018).

221. Arab, K. *et al.* GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nature Genetics* **51**, 217–223 (2019).

222. Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nat. 2014 5167531* **516**, 436–439 (2014).

223. Chen, P. B., Chen, H. V., Acharya, D., Rando, O. J. & Fazzio, T. G. R loops regulate promoter-proximal chromatin architecture and cellular differentiation. *Nat. Struct. Mol. Biol.* **22**, 999–1007 (2015).

224. Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II

juggernaut. *Science* **352**, (2016).

225. Belotserkovskii, B. P. *et al.* Mechanisms and implications of transcription blockage by guanine-rich DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12816–12821 (2010).

226. Kireeva, M. L., Komissarova, N. & Kashlev, M. Overextended RNA:DNA hybrid as a negative regulator of RNA polymerase II processivity. *J. Mol. Biol.* **299**, 325–335 (2000).

227. Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. *Mol. Cell* **42**, 794 (2011).

228. Sanz, L. A. *et al.* Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63**, 167–178 (2016).

229. Ohle, C. *et al.* Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair. *Cell* **167**, 1001-1013.e7 (2016).

230. Castellano-Pozo, M. *et al.* R Loops Are Linked to Histone H3 S10 Phosphorylation and Chromatin Condensation. *Mol. Cell* **52**, 583–590 (2013).

231. Deng, Z., Norseen, J., Wiedmer, A., Riethman, H. & Lieberman, P. M. TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. *Mol. Cell* **35**, 403–413 (2009).

232. Balk, B. *et al.* Telomeric RNA-DNA hybrids affect telomere-length dynamics and senescence. *Nat. Struct. Mol. Biol.* **20**, 1199–1206 (2013).

233. Stork, C. T. *et al.* Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *Elife* **5**, (2016).

234. Zhang, X. *et al.* Attenuation of RNA polymerase II pausing mitigates BRCA1-associated R-loop accumulation and tumorigenesis. *Nat. Commun.* **8**, (2017).

235. Tan, S. L. W. *et al.* A Class of Environmental and Endogenous Toxins Induces BRCA2 Haploinsufficiency and Genome Instability. *Cell* **169**, 1105-1118.e15 (2017).

236. Wahba, L., Amon, J. D., Koshland, D. & Vuica-Ross, M. RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. *Mol. Cell* **44**, 978–988 (2011).

237. Sollier, J. *et al.* Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Mol. Cell* **56**, 777–785 (2014).

238. Cohen, S. *et al.* Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. *Nat. Commun.* **9**, (2018).

239. Cargill, M., Venkataraman, R. & Lee, S. Dead-box rna helicases and genome stability. *Genes* **12**, 1471 (2021).

240. Jimeno, S. *et al.* ADAR-mediated RNA editing of DNA:RNA hybrids is required for DNA double strand break repair. *Nat. Commun.* **12**, 1–15 (2021).

241. Wang, A. H. J. *et al.* Molecular structure of a left-Handed double helical DNA fragment at atomic

resolution. *Nature* **282**, 680–686 (1979).

242. Ho, P. S., Ellison, M. J., Quigley, G. J. & Rich, A. A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.* **5**, 2737–2744 (1986).

243. Heinemann, U. & Roske, Y. Symmetry in nucleic-acid double helices. *Symmetry* **12**, 737 (2020).

244. Bae, S. *et al.* Energetics of Z-DNA Binding Protein-Mediated Helicity Reversals in DNA, RNA, and DNA−RNA Duplexes. *J. Phys. Chem. B* **117**, (2013).

245. Peck, L. J., Nordheim, A., Rich, A. & Wang, J. C. Flipping of cloned d(pCpG)n.d(pCpG)n DNA sequences from right- to left-handed helical structure by salt, Co(III), or negative supercoiling. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 4560–4564 (1982).

246. Liu, L. F. & Wang, J. C. Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 7024 (1987).

247. Wittig, B., Dorbic, T. & Rich, A. The level of Z-DNA in metabolically active, permeabilized mammalian cell nuclei is regulated by torsional strain. *J. Cell Biol.* **108**, 755–764 (1989).

248. Garner, M. M. & Felsenfeld, G. Effect of Z-DNA on nucleosome placement. *J. Mol. Biol.* **196**, 581–590 (1987).

249. Rothenburg, S., Koch-Nolte, F. & Haag, F. DNA methylation and Z-DNA formation as mediators of quantitative differences in the expression of alleles. *Immunol. Rev.* **184**, 286–298 (2001).

250. Liu, R. *et al.* Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. *Cell* **106**, 309–318 (2001).

251. Zhang, J. *et al.* BRG1 interacts with Nrf2 to selectively mediate HO-1 induction in response to oxidative stress. *Mol. Cell. Biol.* **26**, 7942–7952 (2006).

252. Kotze, M. J. *et al.* Analysis of the NRAMP1 gene implicated in iron transport: association with multiple sclerosis and age effects. *Blood Cells. Mol. Dis.* **27**, 44–53 (2001).

253. Lafer, E. M. *et al.* Z-DNA-specific antibodies in human systemic lupus erythematosus. *J. Clin. Invest.* **71**, 314–321 (1983).

254. Suram, A., Rao, J. K. S., Latha, K. S. & Viswamitra, M. A. First evidence to show the topological change of DNA from B-DNA to Z-DNA conformation in the hippocampus of Alzheimer's brain. *NeuroMolecular Med.* **2**, 289–297 (2002).

255. Davisy, R. R., Shaban, N. M., Perrino, F. W. & Hollis, T. Crystal structure of RNA-DNA duplex provides insight into conformational changes induced by RNase H binding. *Cell Cycle* **14**, 668–673 (2015).

256. Barraud, P. & Allain, F. H. T. ADAR Proteins: Double-stranded RNA and Z-DNA Binding Domains. *Curr. Top. Microbiol. Immunol.* **353**, 35 (2012).

257. Herbert, A. Z-DNA and Z-RNA in human disease. *Communications Biology* **2**, (2019).

258. Nowotny, M. *et al.* Structure of human RNase H1 complexed with an RNA/DNA hybrid: insight into HIV reverse transcription. *Mol. Cell* **28**, 264–276 (2007).

259. Rychlik, M. P. *et al.* Crystal structures of RNase H2 in complex with nucleic acid reveal the mechanism of RNA-DNA junction recognition and cleavage. *Mol. Cell* **40**, 658–670 (2010).

260. Nowotny, M. *et al.* Specific recognition of RNA/DNA hybrid and enhancement of human RNase H1 activity by HBD. *EMBO J.* **27**, 1172–1181 (2008).

261. Smith, C. J. *et al.* Enabling large-scale genome editing at repetitive elements by reducing DNA nicking. *Nucleic Acids Res.* **48**, 5183–5195 (2020).

262. Genome Decoration Page. Available at: https://www.ncbi.nlm.nih.gov/genome/tools/gdp/. (Accessed: 26th July 2022)

263. AnnoMiner. Available at: http://chimborazo.ibdm.univ-mrs.fr/AnnoMiner/annominer.html. (Accessed: 25th July 2022)

264. Enrichr. Available at: https://maayanlab.cloud/Enrichr/. (Accessed: 25th July 2022)

265. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).

266. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, (2013).

267. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).

268. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

269. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

270. Paul, M. S. & Bass, B. L. Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *EMBO J.* **17**, 1120–1127 (1998).

271. Torres, A. G. *et al.* Human tRNAs with inosine 34 are essential to efficiently translate eukarya-specific low-complexity proteins. *Nucleic Acids Res.* **49**, 7011–7034 (2021).

272. Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B. & Nestor, C. E. *No evidence for DNA N6-methyladenine in mammals*. *Science Advances* **6**, (2020).

273. Clark, T. A. *et al.* Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.* **40**, e29 (2012).

274. Börgstrom, E., Paterlini, M., Mold, J. E., Frisen, J. & Lundeberg, J. Comparison of whole genome amplification techniques for human single cell exome sequencing. *PLoS One* **12**, (2017).

275. Ahsanuddin, S. *et al.* Assessment of REPLI-g multiple displacement whole genome amplification (WGA) techniques for metagenomic applications. *J. Biomol. Tech.* **28**, 46–55 (2017).

276. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).

277. REPLI-G Kits. Available at: https://www.qiagen.com/de-us/products/discovery-and-translational-research/pcr-qpcr-dpcr/preamplification/repli-g-kits/. (Accessed: 13th July 2022)

278. Dong, M., Wang, C., Deen, W. M. & Dedon, P. C. Absence of 2′-Deoxyoxanosine and Presence

of Abasic Sites in DNA Exposed to Nitric Oxide at Controlled Physiological Concentrations. (2003). doi:10.1021/tx034046s

279. Watkins, N. E. & SantaLucia, J. Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes. *Nucleic Acids Res.* **33**, 6258–6267 (2005).

280. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

281. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).

282. Gromak, N. *et al.* Drosha regulates gene expression independently of RNA cleavage function. *Cell Rep.* **5**, 1499–1510 (2013).

283. Heale, B. S. E. *et al.* Editing independent effects of ADARs on the miRNA/siRNA pathways. *EMBO J.* **28**, 3145–3156 (2009).

284. Zeng, C., Onoguchi, M. & Hamada, M. Association analysis of repetitive elements and R-loop formation across species. *Mob. DNA* **12**, 3 (2021).

285. Gabay, O. *et al.* Landscape of adenosine-to-inosine RNA recoding across human tissues. *Nat. Commun. 2022 131* **13**, 1–17 (2022).

286. Shiromoto, Y., Sakurai, M., Qu, H., Kossenkov, A. V. & Nishikura, K. Processing of Alu small RNAs by DICER/ADAR1 complexes and their RNAi targets. *RNA* **26**, 1801–1814 (2020).

287. Nichols, P. J. *et al.* Recognition of non-CpG repeats in Alu and ribosomal RNAs by the Z-RNA binding domain of ADAR1 induces A-Z junctions. *Nat. Commun.* **12**, 1–15 (2021).

288. Taboury, J. A. & Taillandier, E. Right-handed and left-handed helices of poly(dA-dC).(dG-dT). *Nucleic Acids Res.* **13**, 4469–4483 (1985).

289. Beknazarov, N., Jin, S. & Poptsova, M. Deep learning approach for predicting functional Z-DNA regions using omics data. *Sci. Rep.* **10**, (2020).

290. Rich, A. & Zhang, S. Z-DNA: The long road to biological function. *Nature Reviews Genetics* **4**, 566–572 (2003).

291. Herbert, A. *et al.* The Zα domain from human ADAR1 binds to the Z-DNA conformer of many different sequences. *Nucleic Acids Res.* **26**, 3486–3493 (1998).

292. Yao, M. & Kow, Y. W. Cleavage of insertion/deletion mismatches, flap and pseudo-Y DNA structures by deoxyinosine 3'-endonuclease from Escherichia coli. *J. Biol. Chem.* **271**, 30672–30676 (1996).

293. Zhang, Z., Jia, Q., Zhou, C. & Xie, W. Crystal structure of E. coli endonuclease V, an essential enzyme for deamination repair. *Sci. Rep.* **5**, 12754 (2015).

294. Hitchcock, T. M., Gao, H. & Cao, W. Cleavage of deoxyoxanosine-containing oligodeoxyribonucleotides by bacterial endonuclease V. *Nucleic Acids Res.* **32**, 4071–4080 (2004).

295. Hui, J. *et al.* Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J.* **24**, 1988–1998 (2005).

296. Hui, J., Reither, G. & Bindereif, A. Novel functional role of CA repeats and hnRNP L in RNA stability. *RNA* **9**, 931–936 (2003).

297. Orecchini, E. *et al.* ADAR1 restricts LINE-1 retrotransposition. *Nucleic Acids Res.* **45**, 155–168 (2017).

298. Jung, H., Hawkins, M. A. & Lee, S. Structural insights into the bypass of the major deaminated purines by translesion synthesis DNA polymerase. *Biochem. J.* **477**, 4797–4810 (2020).

299. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).

300. Wang, S. & Hu, A. Comparative study of spontaneous deamination of adenine and cytosine in unbuffered aqueous solution at room temperature. *Chem. Phys. Lett.* **653**, 207–211 (2016).

301. Siriwardena, S. U., Chen, K. & Bhagwat, A. S. Functions and Malfunctions of Mammalian DNA-Cytosine Deaminases. *Chemical Reviews* **116**, 12688–12710 (2016).

302. Békési, A., Holub, E., Pálinkás, H. L. & Vértessy, B. G. Detection of genomic uracil patterns. *International Journal of Molecular Sciences* **22**, 3902 (2021).

303. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5237–5242 (2000).

304. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).

305. Ziller, M. J. *et al.* Genomic distribution and Inter-Sample variation of Non-CpG methylation across human cell types. *PLoS Genet.* **7**, e1002389 (2011).

306. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23**, 555–567 (2013).

307. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**, 816–831 (2012).

308. Kunkel, T. A. & Erie, D. A. Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annu. Rev. Genet.* **49**, 291–313 (2015).

309. Schlötterer, C. & Tautz, D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20**, 211–215 (1992).

310. Bhargava, A. & Fuentes, F. F. Mutational dynamics of microsatellites. *Molecular Biotechnology* **44**, 250–266 (2010).

311. Murat, P., Guilbaud, G. & Sale, J. E. DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. *Genome Biol.* **21**, (2020).

312. Mansour, A. A., Tornier, C., Lehmann, E., Darmon, M. & Fleck, O. Control of GT Repeat Stability in Schizosaccharomyces pombe by Mismatch Repair Factors. *Genetics* **158**, 77–85 (2001).

313. Weber, J. L. & Wong, C. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**, 1123–1128 (1993).

314. Strand, M., Prolla, T. A., Liskay, R. M. & Petes, T. D. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**, 274–276 (1993).

315. Sia, E. A., Kokoska, R. J., Dominska, M., Greenwell, P. & Petes, T. D. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol. Cell. Biol.* **17**, 2851–2858 (1997).

316. Bichara, M., Pinet, I., Schumacher, S. & Fuchs, R. P. P. Mechanisms of Dinucleotide Repeat Instability in Escherichia coli. *Genetics* **154**, 533–542 (2000).

317. Wierdl, M., Dominska, M. & Petes, T. D. Microsatellite instability in yeast: Dependence on the length of the microsatellite. *Genetics* **146**, 769–779 (1997).

318. Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 10774–10778 (1998).

319. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).

320. Poggi, L. & Richard, G.-F. Alternative DNA Structures In Vivo : Molecular Evidence and Remaining Questions. *Microbiol. Mol. Biol. Rev.* **85**, (2021).

321. Pearson, C. E., Edamura, K. N. & Cleary, J. D. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet. 2005 610* **6**, 729–742 (2005).

322. Wells, R. D., Dere, R., Hebert, M. L., Napierala, M. & Son, L. S. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res.* **33**, 3785–3798 (2005).

323. Boyer, J. C., Hawk, J. D., Stefanovic, L. & Farber, R. A. Sequence-dependent effect of interruptions on microsatellite mutation rate in mismatch repair-deficient human cells. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **640**, 89–96 (2008).

324. Petes, T. D., Greenwell, P. W. & Dominska, M. Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. *Genetics* **146**, 491–498 (1997).

325. Hamada, H., Seidman, M., Howard, B. H. & Gorman, C. M. Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. *Mol. Cell. Biol.* **4**, 2622–2630 (1984).

326. Akai, J., Kimura, A. & Hata, R. I. Transcriptional regulation of the human type I collagen α2 (COL1A2) gene by the combination of two dinucleotide repeats. *Gene* **239**, 65–73 (1999).

327. Pagani, F. *et al.* Splicing factors induce cystic fibrosis transmembrane regulator exon 9 skipping through a nonevolutionary conserved intronic element. *J. Biol. Chem.* **275**, 21041–21047 (2000).

328. Gabellini, N. A polymorphic GT repeat from the human cardiac Na+Ca2+ exchanger intron 2 activates splicing. *Eur. J. Biochem.* **268**, 1076–1083 (2001).

329. Hui, J., Stangl, K., Lane, W. S. & Bindereiff, A. HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat. Struct. Biol.* **10**, 33–37 (2003).

330. Ditlevson, J. V. *et al.* Inhibitory effect of a short Z-DNA forming sequence on transcription elongation by T7 RNA polymerase. *Nucleic Acids Res.* **36**, 3163–3170 (2008).

331. Peck, L. J. & Wang, J. C. Transcriptional block caused by a negative supercoiling induced structural change in an alternating CG sequence. *Cell* **40**, 129–137 (1985).

332. Wölfl, S., Martinez, C., Rich, A. & Majzoub, J. A. Transcription of the human corticotropin-releasing hormone gene in NPLC cells is correlated with Z-DNA formation. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 3664–3668 (1996).

333. Pang, B. *et al.* Defects in purine nucleotide metabolism lead to substantial incorporation of xanthine and hypoxanthine into DNA and RNA. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2319–2324 (2012).

334. Hile, S. E., Wang, X., Lee, M. Y. W. T. & Eckert, K. A. Beyond translesion synthesis: Polymerase κ fidelity as a potential determinant of microsatellite stability. *Nucleic Acids Res.* **40**, 1636–1647 (2012).

335. Gadgil, R., Barthelemy, J., Lewis, T. & Leffak, M. Replication stalling and DNA microsatellite instability. *Biophys. Chem.* **225**, 38–48 (2017).

336. Galipon, J., Ishii, R., Suzuki, Y., Tomita, M. & Ui-Tei, K. Differential binding of three major human ADAR isoforms to coding and long non-coding transcripts. *Genes (Basel).* **8**, (2017).

337. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. Lambda ZAP: a bacteriophage lambda expression vector with in vivo excision properties. *Nucleic Acids Res.* **16**, 7583–7600 (1988).

338. Tucker, K. L. *et al.* Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. *Genes Dev.* **10**, 1008–1020 (1996).

339. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).

340. Koushik, S. V., Chen, H., Thaler, C., Puhl, H. L. & Vogel, S. S. Cerulean, venus, and venusY67C FRET reference standards. *Biophys. J.* **91**, L99–L101 (2006).

341. Song, Y. & Zhang, C. Hydralazine inhibits human cervical cancer cell growth in vitro in association with APC demethylation and re-expression. *Cancer Chemother. Pharmacol.* **63**, 605–613 (2009).

342. Sambrook, J. F. & Russell, D. W. *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, 2001).

343. Addgene: Zhang Lab CRISPR Page. Available at: http://www.addgene.org/crispr/zhang/. (Accessed: 21st July 2022)

344. CRISPick. Available at: https://portals.broadinstitute.org/gppx/crispick/public. (Accessed: 21st July 2022)

345. QuikChange Primer Design. Available at: https://www.agilent.com/store/primerDesignProgram.jsp. (Accessed: 21st July 2022)

346. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

347. NEBuilder. Available at: https://nebuilder.neb.com/#!/. (Accessed: 21st July 2022)

348. Fling, S. P. & Gregerson, D. S. Peptide and protein molecular weight determination by electrophoresis using a high-molarity tris buffer system without urea. *Anal. Biochem.* **155**, 83–88

(1986).

349. GitHub - FelixKrueger/TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. Available at: https://github.com/FelixKrueger/TrimGalore. (Accessed: 28th June 2022)

350. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 28th June 2022)

351. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, 1–10 (2009).

352. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods 2012 94* **9**, 357–359 (2012).

353. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, 1–9 (2008).

354. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).

355. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

356. Khan, A. & Mathelier, A. Intervene: A tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics* **18**, 1–8 (2017).

357. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).

358. GitHub - samtools/samtools: Tools (written in C using htslib) for manipulating next-generation sequencing data. Available at: https://github.com/samtools/samtools. (Accessed: 28th June 2022)

359. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).

360. Mayer, C. Phobos 3.3.12, 2006-2010,. (2006). doi:https://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.htm

361. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* **5**, 1438 (2016).

362. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

363. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

8. List of Abbreviations

(d)A	(deoxy)adenosine
(d)C	(deoxy)cytosine
(d)G	(deoxy)guanosine
(d)T	(deoxy)thymidine
3C	chromosome conformation capture
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
8oxoG	8-oxo-guanosine
aa	amino acid
ABE	adenosine base editor
ADAR	adenosine deaminases acting on RNA
ADAR1p110	adenosine deaminase acting on RNA 1, 110 kDa isoform
ADAR1p150	adenosine deaminase acting on RNA 1, 150 kDa isoform
ADAT	adenosine deaminases acting on tRNA
AER	alternative excision repair
Akap6	A-kinase anchoring protein 6
AP	abasic
APC	adenomatous polyposis coli
BER	base excision repair
BFP	blue fluorescent protein
BSA	bovine serum albumin
CARGO	chimeric array of gRNA oligonucleotides
Cas9	CRISPR-associated protein 9
Cd6	T-Cell Differentiation Antigen CD6
cDNA	complementary DNA
CF	core facility
CF	core facility
ChiP	chromatin immunoprecipitation
Cntr	control
CRISPR	clustered, regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
CTCF	CCCTC-binding factor
dCas9	catalytically dead Cas9
dI	deoxyinosine
dIMP	deoxyinosine monophosphate
DIP	DNA immunoprecipitation

dITP	deoxyinosine triphosphate
dIVe	dI-EndonucleaseV enrichment
dIVe-seq	dIVe-sequencing
DMEM	Dulbecco's Modified Eagle Medium
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNMT	DNA methyltransferases
DRIP	DNA:RNA immunoprecipitation
ds	double-stranded
DSB	double-strand breaks
dsRBD	dsRNA binding domain
DTT	dithiothreitol
dU	deoxyuracil
EDTA	ethylenediamine-tetraacetic acid
EndoV	endonuclease V
FACS	fluorescence-activated cell sorting
FCS-A	forward scatter - area
FDR	false discovery rate
Fgf5	fibroblast growth factor 5
FISH	fluorescence in situ hybridization
fwd	forward
G418	genticin
G6pd	glucose-6-phosphate dehydrogenase
GA	gibson cloning reaction
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GCN4	general control nonderepressible protein
gDNA	genomic DNA
GFP	green fluorescent protein
GRIA2	glutamate receptor 2
H2A/2B/3/4	Histone 2A/2B/3/4
H3K4/9/27/36	Histone 3 modified at lysine 4/9/27/36
hdIP	human dIVe peak
HEK293T	human embryonic kidney 293F/T cells
HeLa	human cervical cancer cell line (Henrietta Lacks)
hg38	human genome build 38
Hrpt1	hypoxanthine phosphoribosyl-transferase 1
HTT	huntingtin
Hx	hypoxanthine
IEX	ion exchange chromatography
Ig	immunoglobulin
IMAC	immobilized metal affinity chromatography

IMB	Insitute of Molecular Biology
LAD	laminar associated domains
LB	luria broth
LC-MS/MS	liquid chromatography coupled with tandem mass spectrometry
LIF	leukemia inhibitory factor
LINE	long interspersed nuclear elements
LTR	long terminal repeat
m6dA	N6-methyl-deoxyadenosine
MajSat	major satellite repeat
mAU	milli absorbance unit
MCP	MS2 coat protein
MDA5	melanoma differentiation-associated protein 5
me3	trimethylation
MeCP2	methyl-CpG binding protein 2
MEF	mouse embryonic fibroblast
MEM	minimum assential medium
mESC	mouse embryonic stem cell
MinSat	minor satellite repeat
miRNA	microRNA
mm10	mus musculus genome build 10
MPG	N-methylpurine DNA glycosylase
mRNA	messenger RNA
MUC4	mucin 4, cell surface associated
mVenus	monomeric Venus
MYOD1	myogenic differentiation 1
n.d.	not detected
NA	no annotated repeat
NGS	next-generation sequencing
PAGE	polyacrylamide gel electrophoresis
PAM	protospacer adjacent motif
PBS	phosphate buffered saline
PCP	PP7 coat protein
PCR	polymerase chain reaction
PhuU	Phusion U DNA polymerase
PUF	Pumilio/Fem3 mRNA-binding factor
qPCR	quantitative PCR
rev	reverse

rI	ribosinosine
RNA	ribonucleic acid
rpm	revolutions per minute
rRNA	ribosomal RNA
RT	reverse transcriptase
RT-qPCR	reverse transcription quantitative PCR
S.au.	*staphylococcus aureus*
S.py.	*streptococcus pyogenes*
scFv	single-chain variable fragment
SDS	sodium dodecyl sulfate
SEC	size-exclusion chromatography
sgRNA	single guide RNA
Shh	sonic hedgehog gene
SINE	short interspersed nuclear elements
siRNA	small interfering Ribonucleic Acid
SNP	single nucleotide polymorphism
ss	single-stranded
SSC-A	side scatter - area
STAgR	string assembly gRNA cloning
TAD	topologically associating domain
TALE	transcription activator–like effectors
Tbx3	T-Box transcription factor 3
TDG	thymine DNA glycosylase
Tet	ten-eleven-translocation
Tigre	Igs7
TP	triphosphate
tracrRNA	trans-activating CRISPR RNA
tRNA	transfer RNA
Trx	thioredoxin
TSS	transcription start sites
TTS	transcription termination sites
UPL	universal probe library (Roche)
V_L	variable domain light chain
w/o	without
WGA	whole genome amplification
ZFP	zinc finger proteins
γH2A.X	phosphorylated H2A variant X

9. Acknowledgements

10. Lebenslauf

[redacted]