Role of G-quadruplexes in targeting of somatic hypermutation

Master's Thesis

University of Turku

MSc Degree Programme in Biomedical Sciences

Drug Discovery and Development

May 2021

Eveliina Honkonen

Supervisor: Jukka Alinikula

Institute of Biomedicine

The originality of this thesis has been verified in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service

Abstract

UNIVERSITY OF TURKU
Institute of Biomedicine, Faculty of Medicine

HONKONEN, EVELIINA: Role of G-quadruplexes in targeting of somatic hypermutation
Master's Thesis, 55 p, 3 appendices
MSc Degree Programme in Biomedical Sciences/Drug Discovery and Development
May 2021

Formation of high-affinity antibodies towards antigens is ensured with a process called somatic hypermutation, which happens during affinity maturation of activated B-cells. During this process point mutations are introduced to variable regions of rearranged immunoglobulin genes. Activation-induced cytidine deaminase (AID) is the key enzyme in somatic hypermutation and it deaminates deoxycytidines in single-stranded DNA into deoxyuridine causing mutations. AID preferably binds structured instead of linear DNA and G-quadruplexes (G4s) can potentially be such structures. G-quadruplexes are formed when four guanines interact via Hoogsteen hydrogen bonds and rearrange into co-planar G-quartets which when stacked form the quadruplex structure. The purpose of this study was to find out if these G4s are connected to targeting of somatic hypermutation.

The research was done using GFP-based reporters in chicken DT40 B-cell lines. The cells were studied by flow cytometry in GFP reporter assay and by sequencing. Results from the study indicate that G4s could have an effect in targeting of somatic hypermutation. Removing the already existing G4s significantly reduced somatic hypermutation and changed the locations of mutations. In contrary, adding a strong G4 increased somatic hypermutations but the difference was not statistically significant. However, also in this case, sequencing showed a shift in the locations of mutations towards the G4 structure and an increase in the mutations around the area where the G4 was added.

The results suggest that when the requirements of locus-specific targeting of somatic hypermutation are met, G4s play a role in determining the exact locations of mutations within the gene.

Keywords: Somatic hypermutation, G-quadruplex, GFP loss

Table of contents

1. Introduction

1.1 B lymphocyte development and formation of antibody repertoire

B cells are essential for the normal functioning of the human immune system and the protection of the body against thousands of pathogens. They get their name as they are derived from the bone marrow or the bursa of Fabricius in birds in comparison to T cells that are derived from the thymus. In short B cells are a group of cells that express immunoglobulin receptors on their surface and use them to recognize antigens. When activated they differentiate into plasma cells and produce antibodies needed in the immune response. (Lebien and Tedder, 2008)

The development of B cells takes place at several distinctive stages based on their antigen receptor status. B cell development starts already in the foetal liver where the hematopoietic stem cells are formed and then continues in the primary lymphoid tissue, the bone marrow, where these hematopoietic precursor cells differentiate into immature B cells. The early stages of the development comprise of the rearrangement of the immunoglobulin (Ig) genes called VDJ recombination. During VDJ recombination, the variable (V), diversity (D) and joining (J) segments of the immunoglobulin heavy chain (IgH) are rearranged and combined with the rearranged V and J segments of the immunoglobulin light chain (IgL). All in all, this can produce antibodies recognising up to 10^{13} different antigens. The first part of VDJ recombination, the joining of the heavy chain V, D and J segments, happens in the stage of development where the cells are called pro-B cells (Pieper et al., 2013). When these cells start to express a so-called pre-B cell receptor (BCR) where the heavy chain is already rearranged, but the light chain is still a surrogate light chain, the cells will become pre-B cells (Winkler and Mårtensson, 2018). In the pre-B cells, the immunoglobulin light chain segments will be joined and combined with the heavy chain. At this point the heavy chains are with type μ constant regions and cells expressing IgM molecules will be formed. These are called immature B cells (Pieper et al., 2013). Immature B cells will also start to express IgD on their surface and can then exit the bone marrow (Lebien and Tedder, 2008). These transitional B cells will then move into the spleen where they will finalise the early development by differentiating into mature naive follicular or marginal zone (MZ) B cells. Figure 1 summarises the main developmental steps.

The early stages of B cell development are strongly regulated to avoid autoimmunity (Melchers, 2015; Pieper et al., 2013). There are four main checkpoints out of which three happen in the developmental stages in the bone marrow and the fourth one in the spleen. The first two checkpoints probe the functionality of the pre BCR (Melchers, 2015).

The first checkpoint probes the heavy chain for successful rearrangement. The heavy chain is expressed on the cell surface together with a surrogate light chain to make sure that it can bind a light chain. This leads to the formation of the pre BCR. Before the development can go on, the newly formed pre BCR is checked for autoreactivity at the second checkpoint. Only the unreactive B cells will go on forming the pre-B cells. The third checkpoint probes for the functionality of the rearranged light chain. When a fitting rearranged light chain is combined with the heavy chain, a finalized BCR is formed. This checkpoint is also for the autoreactivity and B cells with highly autoreactive BCRs will undergo apoptosis while the B cells with no autoreactivity can continue development. As mentioned, the fourth checkpoint happens in the spleen at the transition to follicular and MZ B cells. Here the possible autoreactivity is screened for the last time before the maturation of the B cells. (Melchers, 2015)

The marginal zone B cells in the spleen rapidly develop into plasma cells and start secreting IgM antibodies upon B cell activation by an antigen, thus forming the first-line defence against pathogens. In addition to this, the MZ B cells can undergo some class switch recombination (CSR) and produce type IgG and IgA antibodies as well. (Cerutti et al., 2013)

Follicular B cells on the other hand normally circulate in the body and when activated by an antigen, they will start to proliferate and form germinal centres (GC) in the spleen and lymph nodes. In the GCs, the B cells undergo CSR and somatic hypermutation (SHM) in a process called affinity maturation. As a result of this the GC B cells will form memory B cells and long-lived plasma cells that have the ability to produce antibodies with high affinity to specific antigens. (Sagaert et al., 2007)

The fifth checkpoint of B cell development takes place in the GC. The checkpoint makes sure that no autoreactive B cells are accidentally created in the process of SHM. Also, cells in which mutations have led to lower affinity antibodies will be discarded. (Melchers, 2015) These stages of B cell development are also summarized in Figure 1.

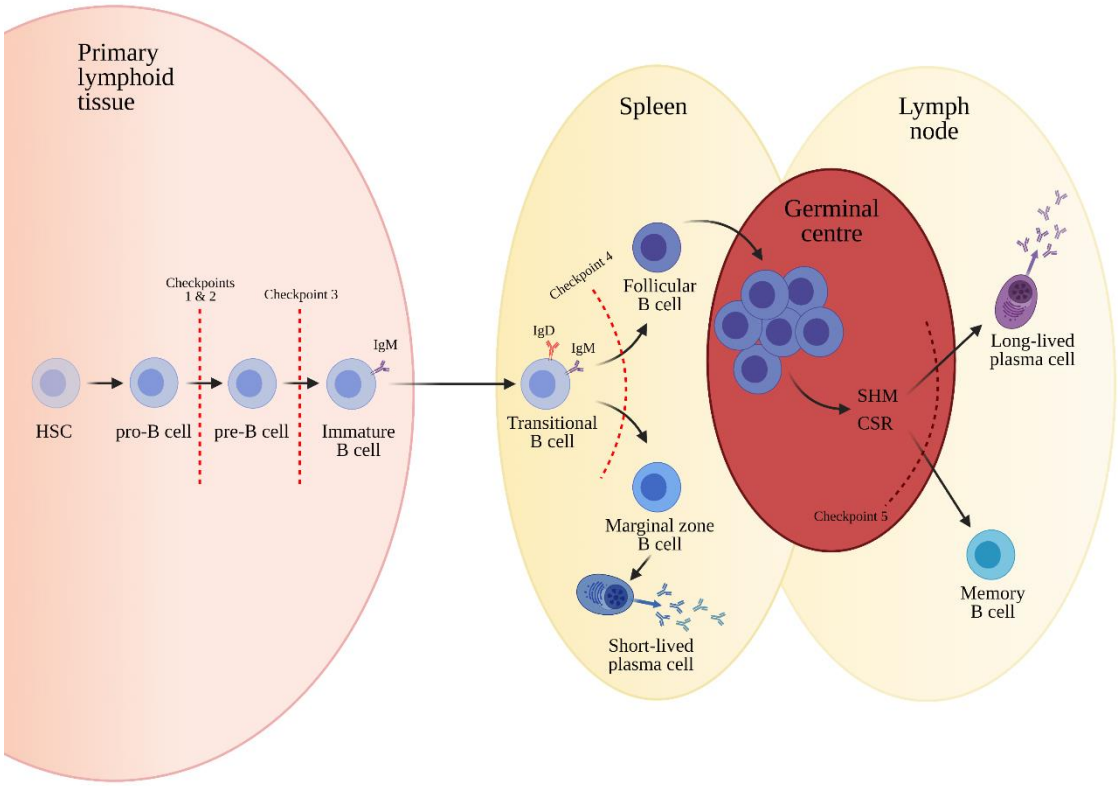


Figure 1. Summary of the main points of B cell development. HSC = hematopoietic stem cell, SHM = somatic hypermutation, CSR= class switch recombination

1.2 Somatic hypermutation

Somatic hypermutation happens during affinity maturation of antigen activated B cells in germinal centres where point mutations are introduced into the V-regions of immunoglobulins. This is an important phase of the B cell development and makes it possible for the memory B cells and long-lived plasma cells to recognise specific antigens and form antibodies with high affinity to them. An important enzyme in the hypermutation process is activation induced cytidine deaminase (AID) that causes the mutations by deaminating deoxycytidine into deoxyuridine. (Di Noia and Neuberger, 2007)

Most commonly the mutations observed in SHM are substitutions, but some deletions and insertions are observed as well (Briney et al., 2012). The mutations have been observed to start roughly 150 bp downstream from the IgV promoter and most of the mutations are found within the three complementary determining regions (CDR) out of which the CDR3 has the most diversity (Yeap et al., 2015; Neuberger and Milstein, 1995).

In general, somatic hypermutation process is thought to happen in two phases. In the first phase AID is recruited to the gene and DNA is deaminated. In the second phase the uracils added are removed and replaced by DNA repair mechanisms described later. (Kohler et al., 2012)

It is not understood how somatic hypermutation is targeted to IgV regions (Odegard and Schatz, 2006). What is curious is that these hypermutations caused by AID can also be found in some other non-Ig genes in AID expressing B cells, but the number of mutations in Ig genes during the process of SHM can be up to 1000 times higher than in the non-Ig genes (Kohler et al., 2012; Buerstedde et al., 2014). This leads to a thought that there must be some mechanism that targets SHM especially to immunoglobulin genes, but it is still not fully known what leads to this targeting of SHM. However, there are theories, and one idea is that there are specific cis-acting elements and trans-acting factors that recruit AID and all the machinery needed to SHM to Ig loci (Odegard and Schatz, 2006).

It has been proven that Ig enhancers play a role in determining the locus specificity of SHM and such sequences that are part of Ig enhancers or act like Ig enhancers in activating somatic hypermutation in neighbouring genes have been named diversification activators (DIVAC). Even though these DIVACs have been proven to activate SHM the exact mechanism by how they do it, is unknown. They do not increase the transcription of the mutated genes and it has been speculated that perhaps DIVACs promote the formation of a protein complex that guides AID to the transcription initiation complex of the mutated gene. Other plausible theories are that DIVAC bound factors recruit AID or that DIVACs induce changes in the elongation complex making DNA more accessible for AID. (Blagodatski et al., 2009; Buerstedde et al., 2014)

Immunoglobulin enhancers have also helped in determining that the mechanism of activation of SHM must be evolutionally conserved because human Ig lambda and IgH enhancers can stimulate SHM also in chicken B cells even though normally chicken cells use gene conversion in increasing antibody affinity. (Buerstedde et al., 2014)

1.3 Activation induced cytidine deaminase

Activation induced cytidine deaminase is the initiating enzyme in the process of somatic hypermutation, as was already briefly described in the previous chapter. AID deaminates

deoxycytidine residues into deoxyuridine, which leads to mismatched U:G pairs in DNA during the G1 phase of the cell cycle. These pairs will then be recognized and repaired by the DNA repair machinery of the cell. (Di Noia and Neuberger, 2007; Feng et al., 2020; Pilzecker and Jacobs, 2019)

AID is not only important in SHM but works also in other important processes in the germinal centre namely the class switch recombination and gene conversion (Feng et al., 2020; Yeap et al., 2015). Gene conversion is another process that is used by some species to increase the affinity of antibodies and is related to SHM (Arakawa et al., 2002). The mechanism of action of AID in all the processes is the same (Yeap et al., 2015). In CSR, the mutations caused by AID just lead to double strand breaks and the change of the immunoglobulin type from IgM to either IgG, IgE, IgD or IgA (Feng et al., 2020; Qiao et al., 2017). Related to CSR, deficiencies in AID lead to an immunodeficiency called hyper-IgM syndrome which is characterised by very low serum IgG, IgA and IgE and leads to increased susceptibility to infections (Revy et al., 2000).

Some aspects characteristic for functioning of AID have been solved. AID belongs to the APOBEC cytidine deaminase family and is encoded by the *AICDA* gene (Buerstedde et al., 2014; Feng et al., 2020). It acts on single stranded DNA (ssDNA), but even though AID is known to require ssDNA for its function, it has been established that it binds rather structured than completely linear ssDNA (Qiao et al., 2017). Additionally, it has been solved that AID preferentially deaminates deoxycytidines within WRCY motifs where W = A/T, R = A/G and Y = C/T and therefore these are so called hotspots of AID (Feng et al., 2020; Qiao et al., 2017). The frequency of mutations caused by AID in the IgV-regions is around 10^{-3} per base pair per generation which is million times higher than the rate of mutations in other genes of B-cell genome (Odegard and Schatz, 2006; Pilzecker and Jacobs, 2019). Also, the frequency by which AID targets the immunoglobulin switch (IgS) regions in CSR is the same than that of the IgV regions in SHM (Yeap et al., 2015).

For the proper function of AID, its N- and C-terminal regions are important. The N-terminus has a high net positive charge and carries the nuclear localisation signal which makes it possible that AID can work in the nucleus and mutate DNA. Then again, the C-terminus contains a strong nuclear export signal and is especially essential for CSR as cells with mutations to AID C-terminus lose CSR activity but not SHM. On the contrary the opposite is noticeable with the N-terminal region as cells with mutations there lose

SHM activity but not CSR. Thus, it has been proposed that AID interacts with specific cofactors with these two domains, which lead the actions of AID to either somatic hypermutation or class switch recombination. (Muramatsu et al., 2004; Di Noia and Neuberger, 2007)

Posttranslational phosphorylation of AID is also important. It is another factor that has been proposed to affect the choice of action of AID between SHM and CSR. It has been found out that AID is phosphorylated at serine 38, threonine 140 and tyrosine 184. Out of these, mutations at serine 38 decrease levels of CSR and SHM in the cells. Mutations at threonine 140 decrease significantly only SHM levels suggesting that it is a posttranslational modification specific for somatic hypermutation process. (McBride et al., 2008)

As with the total targeting of SHM, it is not completely understood how AID is targeted to the IgV regions. As AID is known to require ssDNA for functioning, it is suggested that the high transcription levels of IgV genes could result in many locations with ssDNA substrates for AID and that could be one reason why AID is acting there. However, as the DIVACs introduced earlier do not function by increasing transcription but still increase AID induced mutations in SHM, there must be some other ways as well than simply increased transcription that work on targeting AID to specific areas. (Kohler et al., 2012; Buerstedde et al., 2014)

1.4 DNA repair mechanisms in SHM

In somatic hypermutation the DNA repair mechanisms that are normally required to work in a high-fidelity fashion work in a very error-prone way resulting in the transitions and transversions that are observed in the SHM process (Liu and Schatz, 2009). Both C-G and A-T base pairs are targeted, and Figure 2 summarises the repair pathways that lead to these mutations.

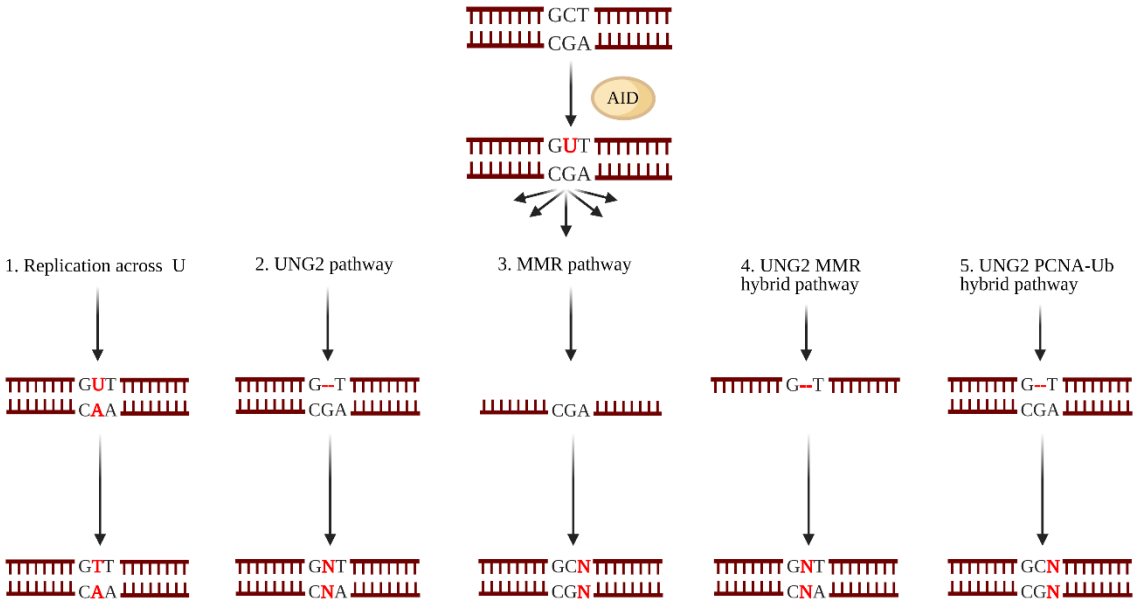


Figure 2. The alternative DNA repair pathways used in somatic hypermutation and the types of mutations they lead to. AID = activation induced cytidine deaminase, UNG= uracil-DNA-glycosylase, MMR = mismatch repair, N = any base

In total five alternative DNA repair pathways have been recognised to lead to the variety of mutations. First possible pathway is followed if the uracil is not processed by base excision repair (BER) or mismatch repair (MMR) and there will simply be transcription over the site. As U is similar to T, it instructs any known DNA polymerase to place an A into the opposite strand leading to a transition from the original C-G pair to T-A base pair. (Pilzecker and Jacobs, 2019)

The second alternative depends on an enzyme called uracil-DNA-glycosylase (UNG) and can lead to both transversions and transitions. If the U is recognised and processed by BER, the enzyme UNG removes it creating an abasic site (Krijger et al., 2009). There are two splice variants of the *Ung* gene out of which UNG2, which is located in the nucleus, is more prominent in SHM (Nilsen et al., 1997). The abasic site gives no instructions on replication on the opposite strand, so both transversions and transitions can result here. Polymerases working to fix these lesions are called translesion synthesis (TLS) polymerases and one important TLS polymerase here is REV1 that can tolerate abasic sites (Pilzecker and Jacobs, 2019). REV1 is especially essential in creating G-C transversions to the site (Krijger et al., 2013).

The U-G mismatch can also be recognised by the mismatch repair protein heterodimer MSH2/MSH6 which expands the mutations to the neighbouring base pairs. After

recognition of the mismatch, exonuclease EXO1 is activated, and it creates a single stranded DNA which will then be patched by DNA polymerase η (Pol η) (Pilzecker and Jacobs, 2019). Pol η is a very error-prone DNA polymerase and causes mutations at a frequency of 10^{-2}. It mutates especially the A-T base pairs and thus expands the mutations to the neighbouring base pairs of the originally deaminated cytosine by AID (Zeng et al., 2001; Wilson et al., 2005). Even though it seems that Pol η can work also independently, for most of the mutations at A-T base pairs the polymerase is activated and recruited to the lesion site by ubiquitinated DNA sliding clamp Proliferating Cell Nuclear Antigen (PCNA-Ub). PCNA-Ub is ubiquitinated at lysine residue 164, which is essential for its function (Krijger et al., 2011).

The UNG2-dependent BER and MMR are not only competitive ways to process the added uracil, but they can work also together. It has even been estimated that about half of the G-C transversions in SHM are due to this hybrid pathway, which forms the fourth possibility for DNA repair in the process of SHM (Krijger et al., 2009; Pilzecker and Jacobs, 2019). However, the system how these transversions are created are not identical between the traditional UNG2 pathway and the UNG2 MMR hybrid pathway. The TLS polymerase REV1 is essential only when the transversions are created downstream of UNG2 processing alone. Then again when UNG2 and MSH2/MSH6 complex work together, the G-C transversions can be created also by action of other polymerases if REV1 is not available (Krijger et al., 2013). To support the idea of the hybrid pathway it has been established that the MMR machinery can interfere with ongoing BER when there are multiple lesion sites in the vicinity of each other (Schanz et al., 2009). It has also been speculated that G-C transversions within AID hotspots depend on UNG2 alone and outside of these hotspots on the hybrid pathway (Pilzecker and Jacobs, 2019). Furthermore, there is also the possibility that the U exists in the single stranded DNA created by EXO1 in MMR and then UNG2 proceeds in removing it as UNG2 is 1.7-fold more reactive on ssDNA when compared to double stranded DNA (Krijger et al., 2009; Pilzecker and Jacobs, 2019).

Lastly, it has also been suggested that UNG2 alone can work together with PCNA-Ub and recruit Pol η to the lesion cite after removing the U and creating an abasic cite. This also expands the mutations to neighbouring A-T base pairs. (Krijger et al., 2009)

It is not completely known why these five alternative pathways work in an error-prone way instead of correctly repairing the lesions. However, there are some theories explaining why the DNA repair mechanisms work in these error-prone fashions during somatic hypermutation and make it possible for B cells to increase the affinity of the antibodies. One reason is in the ubiquitination of PCNA. It has been shown that when only mono-ubiquitinated, it recruits more error-prone DNA polymerases such as Pol η to the lesion site (Fan et al., 2020). Another theory is that some cis-acting elements that take part in targeting of somatic hypermutation itself also recruit machinery for error-prone repair to the Ig genes. However, at this point this has not been much studied as most of the studies have focused on identifying these cis-acting elements (Liu and Schatz, 2009; Odegard and Schatz, 2006). It is also thought that perhaps the N- and C-terminals of AID, that are important for its function, bind specific co-factors that help in recruiting error-prone DNA repair machinery (Liu and Schatz, 2009). But as these are only some of the possible explanations, it is clear that the exact way how specifically error-prone polymerases are recruited to the somatic hypermutation site remains a mystery.

1.5 G-quadruplexes

As it has been mentioned, AID is known to preferentially bind structured ssDNA and one likely secondary DNA structure that AID could bind are the G-quadruplexes (G4) and the field studying these structures is rapidly evolving. (Feng et al., 2020; Qiao et al., 2017; Xu, Y. et al., 2020)

So far, the basic structure of the G-quadruplexes has been solved. They are formed when four guanines interact via Hoogsteen hydrogen bonds and rearrange into co-planar G-quartets (Figure 3). These quartets will then be stacked, and the stacks are stabilised by cations of which potassium (K+) is the most common one as shown in Figure 3 B and C (Sen and Gilbert, 1990). G4s are frequently found in Ig switch regions and telomeres and based on studying their sequences a basic G4 motif of $G_{\geq 2}N_xG_{\geq 2}N_xG_{\geq 2}N_xG_{\geq 2}$ has been solved (Spiegel et al., 2020). This structure consists of stems with a minimum of two consecutive guanines. However, in mammals a stem length of three is more common and more stable. The loops in between the stems (denoted as N_x) can all be of various lengths although it was thought that shorter loops make more stable and faster folding G4s (Dhapola and Chowdhury, 2016; Kikin et al., 2006). However, recently also longer loop

lengths have been found, as well as discontinuations in the G-stems which lead to forming of bulges to the backbone of the G4 structure (Figure 3 C) (Spiegel et al., 2020). Furthermore, it has been proven that a long loop length in the middle of the G4 can actually stabilise the structure if it can form a double helix through complementary base pairing (Nguyen et al., 2020).

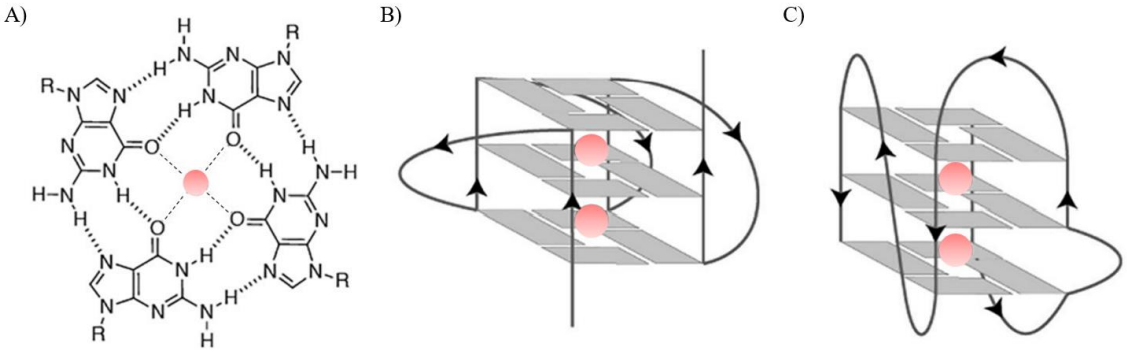


Figure 3. A) Hoogsteen hydrogen bonding between guanines forming a co-planar G-quartet. The central cation (coloured red) is binding to oxygen. B) An example of a possible G4 that can be formed. Central cations stabilizing the structure are shown. In this form the backbone runs parallel on each strand. C) An example of a possible G4 structure showing a bulge on the backbone. Modified from Spiegel J, Adhikari S, Balasubramanian S. The structure and function of DNA G-quadruplexes. Trends in Chemistry. 2020;2(2):123-136.

G-quadruplexes can be found in high numbers in different areas of the genome. Bioinformatic studies have found over 370 000 G4 prone sequences in the human genome (Bedrat et al., 2016). With high-throughput sequencing methods even more have been found, many which were not visible with only computational methods (Chambers et al., 2015). It has also been possible to observe the actual formation of G4 DNA in the nuclei of live cells with fluorescence lifetime imaging microscopy (Summers et al., 2021). Many of the G4 structures in addition to IgS regions and telomeres are found in gene promoter and 5' untranslated regions as well as at splicing sites. These are all functional areas in the genome pointing towards the idea that G4s have biologically essential regulating roles even though their actual function is still unknown (Summers et al., 2021; Chambers et al., 2015).

Additionally, it has been frequently observed that G-quadruplexes are connected to DNA damage (Hänsel-Hertsch et al., 2017). Furthermore, it has been established that G4s can

indeed induce DNA damage through a mechanism involving R-loop formation in cancer cells. R-loops are other noncanonical secondary DNA structures that are associated with genome instability. They form when the DNA duplex is opened and one of the strands forms an RNA:DNA hybrid by annealing to an RNA strand (De Magis et al., 2019). Some cancer tissues such as stomach and liver cancer tissues are also observed to have increased numbers of G4 structures when imaged with a G4 specific antibody BG4 (Biffi et al., 2014). More-over, oncogenes in general are more G-rich than the rest of the genome supporting the fact that they have great potential to form G4s (Duquette et al., 2007).

1.6 G-quadruplexes in AID mediated diversification

As already stated, the Ig constant switch regions, where AID works during CSR, are rich in guanine and have the possibility to form many G4s (Dunnick et al., 1993). Formation of these G4s in switch region sequences following transcription has actually been proven *in vitro* and *in vivo* in *Escherichia coli* by electron and atomic force microscopies. In these experiments the G-rich sequences formed R-loops after unwinding of the DNA double helix during transcription. The formation of these loops was very efficient, one side of the loops composed of a formed G4 structure on the coding strand and the formed structures were stable (Duquette et al., 2004; Neaves et al., 2009). There is also evidence that these G4s within the IgS regions could be one of the nucleic acid secondary structures that AID binds during CSR rather than linear ssDNA (Qiao et al., 2017; Zheng et al., 2015).

It has been shown to be essential for functioning of AID during CSR, that it binds a bifurcated substrate. Folding of a sequence into a G4 structure could form such substrate. In one *in vitro* analysis, it seems that AID rather recognises a single-stranded overhang adjacent to the G4 structure instead of the core structure itself. In the same analysis, it was found out that AID more frequently mutates deoxycytidines close to G4 structures than deoxycytidines on a linear ssDNA composed of the same sequence. In fact, the binding affinity of AID to G4 was about 10 times higher than to a linear substrate with the same sequence. It also seems that AID homologues of the APOBEC family do not have this same preference for G4 and bifurcated substrates over linear substrates pointing towards the idea that G4s are important in AID specific processes. (Qiao et al., 2017)

Furthermore, in the scope of class switch recombination, it has been established that RNA molecules formed from the IgS regions can fold into G4 structures and guide AID to those areas. Especially the RNAs formed from the IgS coding strands were of importance as they are more G-rich (Zheng et al., 2015). Furthermore, if AID of mouse is mutated at glycine 133 so that it cannot bind these G4 structures formed in RNA, CSR is greatly reduced in mouse splenic B cells (Zheng et al., 2015). This same mutation has been also found in humans who have been diagnosed with hyper-IgM syndrome, a genetic Ig class switch deficiency (Mahdaviani et al., 2014). Recent *in vivo* evidence in mice shows that this same mutation in AID leads also to disruption of SHM in addition to CSR (Yewdell et al., 2020). The mutated form of AID is still catalytically active but cannot localise to the correct genomic regions and cannot bind synthetic G4-DNA *in vitro* (Yewdell et al., 2020). In addition to mutating the G4 binding potential of AID, disrupting the formation of these G4s also seemed to reduce CSR (Zheng et al., 2015).

In addition to switch regions, AID has been observed to localise with telomeric sequences *in vitro* (Zheng et al., 2015) and as mentioned, telomeres are also known to be able to form many G4 structures (Spiegel et al., 2020). It even seems that the telomeric sequences can work as off-targets of AID and the DNA repair enzyme UNG protects B-cells from AID-mediated telomere damage (Cortizas et al., 2016; Safavi et al., 2020).

Also, other off-targets of AID, such as oncogenes BCL and c-MYC, have been shown to form G4 structures within their promoter regions (Balasubramanian et al., 2011; Duquette et al., 2007). With c-MYC it has been actually shown *in vitro* that AID localizes to G4 structures within its sequence. The G4 structures there are also formed following transcription and form similar loops that have been observed in switch regions (Duquette et al., 2005). G4s are also greatly associated with genome instability, and where G4s are formed, the normal Watson-Crick base pairing is impaired leaving one strand extendedly open so that AID could work there in addition to binding to the G4s. All this enforces the idea that G4s are associated with AID mediated diversification (Duquette et al., 2007; Spiegel et al., 2020; Qiao et al., 2017).

This all raises a speculation that if G4s could be formed also within Ig variable genes, they could have an effect in targeting AID there during SHM. Although, traditionally it has been thought that the IgV regions do not form G4 structures as they are not G-rich

and therefore would not use similarly DNA secondary structures in recruiting AID during SHM (Pavri, 2017; Qiao et al., 2017). However, this has not been extensively studied.

1.7 G-quadruplexes as possible drug targets

As G-quadruplexes are frequently associated with malignancies and observed also in oncogene promoters, such as c-MYC, VEGF, BCL-2, PDGFA and TERT, there has been various studies regarding their potential on being new drug targets. In addition to G4s in oncogene promoters, G4s in telomeres have been extensively studied as targets for telomerase inhibition in cancer (Asamitsu et al., 2019; Balasubramanian et al., 2011). Indeed, many ligands binding G4 structures have been identified and they can have molecular structures varying for example from anthraquinones to perylenes and porphyrins (Figure 4) (Hurley et al., 2000; Mergny and Hélene, 1998).

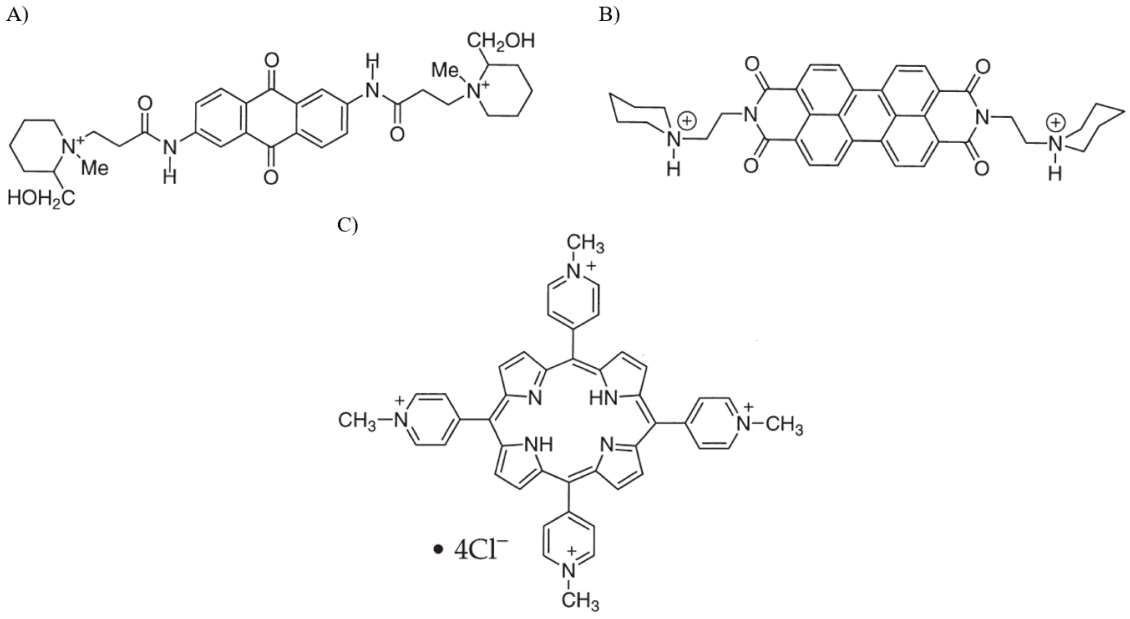


Figure 4. Possible G-quadruplex binding ligands as examples of varying molecular structures they can have. A) anthraquinones B) perylenes C) porphyrins. Modified from Hurley, L.H., R.T. Wheelhouse, D. Sun, S.M. Kerwin, M. Salazar, O.Y. Fedoroff, F.X. Han, H. Han, E. Izbicka, and D.D. Von Hoff. 2000. G-quadruplexes as targets for drug design. Pharmacology and Therapeutics. 85:141-158. doi: 10.1016/S0163-7258(99)00068-6.

The first evidence that G-quadruplex binding ligands can act as telomerase inhibitors comes already from 1990's. The research was then done with an anthraquinone molecule

and has advanced the development of creating other G4 binding small molecules (Sun et al., 1997). The basis of functioning of these molecules is that they stabilise the G4 structures in the telomeres hence inhibiting the normal functioning of telomerase, which eventually leads to shortening of telomeres and death of cancer cells (Mergny and Hélene, 1998). They can also stabilise the G-rich 3´overhangs of telomeres in such form that telomerase cannot attach there anymore (Balasubramanian et al., 2011).

One well known potent telomerase inhibitor is telomestatin which is derived from *Streptomyces anulatus.* It has a very similar structure to the actual planar G-quartet shown in Figure 3 A and it can interact with the G4 structure (Kim et al., 2002). With telomestatin there is *in vitro* and *in vivo* evidence in human acute myelomonocytic leukaemia cell line and mouse xenografts that it can reduce telomerase activity, induce telomere shortening and reduce tumour growth (Tauchi et al., 2006). Similar *in vitro* and *in vivo* effects to telomestatin have been observed with another G-quadruplex interacting molecule called BRACO-19 with the difference that BRACO-19 experiments were carried out to study uterus carcinoma (Burger et al., 2005).

Stabilising the G4 structures in the promoter regions of proto-oncogenes generally leads to inhibition of expression of those genes. Therefore, many ligands for that purpose have also been studied (Asamitsu et al., 2019). Some of these small molecules could even be used specifically for one type of cancer. For instance, ligands targeting especially G4s at c-MYC, KRAS and BCL-2 over G4s at promoter regions of the other proto-oncogenes or telomeres have been developed (Hu et al., 2018; Lavrado et al., 2015; Amato et al., 2018). Similarly to telomestatin and BRACO-19, there is also *in vitro* and *in vivo* evidence of the effect of some of the small molecules targeting G4s at promoter regions (Kendrick et al., 2017).

Despite the frequent studies of G4s as drug targets, none have reached clinical use. Many of the developed small molecules have had unfavourable pharmacological properties and even if some small molecules specific for certain proto-oncogenes have been found, specificity is often an issue (Asamitsu et al., 2019; Balasubramanian et al., 2011). However, still some G-quadruplex targeting molecules, such as quarfloxin which is a fluoroquinolone derivative, have made it to the clinical trial up to Phase II (Xu, H. et al., 2017; Drygin et al., 2009). All in all, G-quadruplexes are a promising drug target and require more research in that field.

1.8 Clinical aspects

Somatic hypermutation and understanding how it and AID are targeted to specific genes is very important in general as aberrant hypermutation is connected to the formation of cancer and B cell malignancies when they act in the wrong place. For instance, proto-oncogenes such as BCL6, c-MYC and PIM1 have been found to be hypermutated by AID which has led to deregulated expression of these genes and B cell malignancies. As described, some of these oncogenes are also rich in guanine and it has been established that G-quadruplexes are formed within their promoters which strengthens the idea that G4s could be related to targeting of somatic hypermutation and AID. (Duquette et al., 2007; Balasubramanian et al., 2011)

In addition to hypermutations in oncogenes, many times also genomic translocations are observed in diseases such as Burkitt's lymphoma, diffuse large B cell lymphoma and multiple myeloma. These translocations often include the immunoglobulin switch regions indicating that AID could be responsible for the lesions leading to these translocations. Moreover, to support this theory, AID has been shown to be responsible for a translocation between c-MYC and the IgH variable region in interleukin 6 transgenic mice. (Dorsett et al., 2007)

1.9 Purpose of the study

The aim of this project is to find out whether somatic hypermutation is targeted near G-quadruplexes in a green fluorescent protein (GFP) reporter construct used and whether a strong G-quadruplex can also increase the mutation frequency. The previous findings in the group suggest that the distribution of the mutations in the process of somatic hypermutation does not correlate well with regions of ssDNA, but rather with the location of potential G-quadruplex forming sequences. The original hypothesis is that G4-structures have an effect in targeting of somatic hypermutation when all the other requirements for mutation targeting are met. The goal is to test this hypothesis.

The question that what makes SHM targeted to Ig genes is important and has been speculated for a long time, so this study also hopes to shed some light in it. This research also focuses more on the possible, less studied connection between SHM and G4s in comparison to CSR and G4s.

2. Results

Somatic hypermutation was studied with GFP loss assay where the decrease in green fluorescence represents mutations caused on green fluorescent protein gene in the somatic hypermutation process. The reporters designed and cloned specifically for this study were transfected into chicken DT40 bursal lymphoma B-cell lines where reporter was integrated in the genome so that somatic hypermutation was targeted on it. In addition to observing the change in green fluorescence from the cells caused by increasing numbers of mutations, the actual location, type and frequency of mutations were also studied by the means of sequencing the reporter constructs after B-cells had grown for one month.

2.1 Analysis of G-quadruplexes with G4Hunter

In the beginning the immunoglobulin variable regions and the GFP gene were studied for possible G-quadruplex forming sequences within them with an online software called G4Hunter (Brázda et al., 2019, http://bioinformatics.ibp.cz/#/). Multiple sequences coding for functional immunoglobulin heavy chain variable regions were studied and always one allele of few genes from every IGHV subgroup were included in the analysis (Table 1).

Only two of the 28 analysed sequences could not form any G4 structures and most of them could form two or more. Positive G4Hunter score indicates that the quadruplex is found in the coding strand and a negative score indicates that it is found in the template strand. In total, 65 % of the quadruplexes found by this analysis are formed in the coding strand.

Table 1. The G-quadruplex analysis of sequences coding for immunoglobulin heavy chain variable regions by G4Hunter.

Gene	Number of G-quadruplexes found in the sequence	Positions of the G-quadruplexes	G4Hunter score of the given quadruplex
IGHV1-2	2	279	1.2
		397	1.2
IGHV1-3	0	-	-

IGHV1-8	2	211	1.2
		329	1.286
IGHV1-18	1	316	1.286
IGHV1-24	4	14	-1.25
		44	-1.2
		133	1.318
		337	1.227
IGHV1-45	4	67	1.136
		107	-1.167
		120	-1.143
		154	1.2
IGHV1-46	1	305	1.2
IGHV1-58	1	216	1.136
IGHV2-5	3	303	1.2
		305	1.167
		565	-0.844
IGHV2-26	1	41	-1.2
IGHV2-70	0	-	-
IGHV2-70-D	4	148	1.083
		175	-1.231
		746	-0.806
		811	-0.867
IGHV3-7	5	79	-1.037
		184	1.16
		343	1.065
		458	1.088
		704	1.13
IGHV3-9	5	27	-1.12
		34	-1.25
		65	-1.217
		284	0.943
		401	1.227
IGHV3-11	2	206	0.943
		316	1.059
IGHV3-13	4	1	1.115
		166	0.943
		189	1.25
		508	1.583
IGHV3-15	4	166	0.889
		188	1.273
		276	1.118
		466	1.136

IGHV3-20	2	174	1.267
		285	1.091
IGHV3-21	2	174	1.034
		283	1.061
IGHV4-4	8	350	1.143
		355	1.19
		393	-0.909
		522	1.083
		562	0.893
		572	1.238
		575	1.2
		603	0.786
IGHV4-28	6	380	1.2
		414	1.238
		419	1.238
		459	-1.2
		461	-1.2
		586	1.083
IGHV4-30-2	2	173	-1.2
		175	-1.2
IGHV4-30-4	3	312	-1.2
		314	-1.2
		441	1.182
IGHV4-31	2	199	-1.2
		201	-1.2
IGHV5-10-1	1	139	0.933
IGHV5-51	6	74	-1.2
		97	-1.333
		175	-1.167
		434	1.136
		611	-0.889
		620	-1.348
IGHV6-1	1	69	1.182
IGHV7-4-1	1	223	1.286

When the G4Hunter algorithm was validated, it was noted that the average G4Hunter score for known G4 forming sequences was 1.64 ± 0.46 but some G4 structures had also lower scores (Bedrat et al., 2016). 53 % of the G4 forming sequences found in this IGHV analysis scored propensity scores that fit in this previously defined average value. This

warrants further experimentation to find out if G4s are formed in IgH variable regions *in vivo* and whether they could play a role in SHM.

After analysing the IGHV sequences and establishing that they have the possibility to form G4 structures, the GFP gene used as a reporter gene in this study was also analysed with G4Hunter. Based on previous studies of the group, it was already established that the GFP sequence can potentially form G4 structures so now the exact locations of G4 forming sequences with the same settings used for analysing IGHV were found. In total, 10 sequences were found likely to form G4s in the GFP coding gene and they have similar G4Hunter propensity scores to the G4 forming sequences found from the IgV regions. Most of the sequences overlap together to form 4 areas that are likely to form quadruplex structures (Table 2). All these areas were found in the template strand of the gene and Figure 5 visualises how they are positioned along the GFP gene.

Table 2. Possible G-quadruplex forming sequences found from GFP when analysed with G4Hunter

G-quadruplex	Position (bp from the start of GFP)	Sequence (coding strand)	G4Hunter score
1	159	CTGCCCGTGCCCTGGCCCACCCTCG	-1.2
2	168	CCCTGGCCCACCCTCGTGAC	-1.2
3	173	GCCCACCCTCGTGACCACCCTG	-1.227
4	178	CCCTCGTGACCACCCTGACC	-1.2
5	212	CTTCAGCCGCTACCCCGACC	-1.2
6	214	TCAGCCGCTACCCCGACCACA	-1.143
7	544	ACTACCAGCAGAACACCCCCATC	-1.13
8	558	ACCCCCATCGGCGACGGCCCCG	-1.318
9	616	CCGCCCTGAGCAAAGACCCCA	-1.238
10	619	CCCTGAGCAAAGACCCCAAC	-1.2

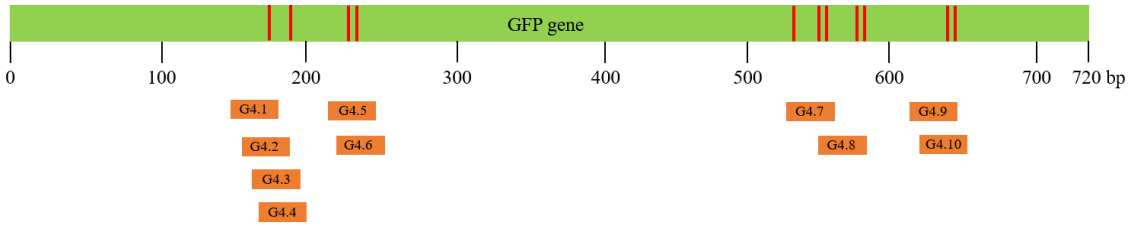


Figure 5. The positions of possible G-quadruplex forming sequences in GFP gene. The G-quadruplex structures have been solved with G4Hunter online software. The red lines in the GFP denote the positions of mutations designed to diminish the G4 forming potential of the sequence. GFP = green fluorescent protein, G4 = G-quadruplex, bp = base pair

2.2 Different reporter constructs modelling somatic hypermutation with GFP loss

For this study different reporter constructs based on green fluorescent protein for modelling somatic hypermutation were designed. The reporters were based on already existing GFP2 and GFP4 reporters described by Blagodatski *et al.* in 2009 and Buerstedde *et al.* in 2014 respectively. The basic reporters also included DIVAC sequences being DIVAC 2-3 in GFP2 reporter and a human Ig lambda enhancer core (Igλ) sequence in GFP4 (Figure 6). These acted also as positive controls in the study and are further referred as GFP2 2-3 and GFP4 Igλ reporters. Corresponding negative controls being the original ones described in the aforementioned papers contained no DIVAC sequences and are denoted only as GFP2 and GFP4 here.

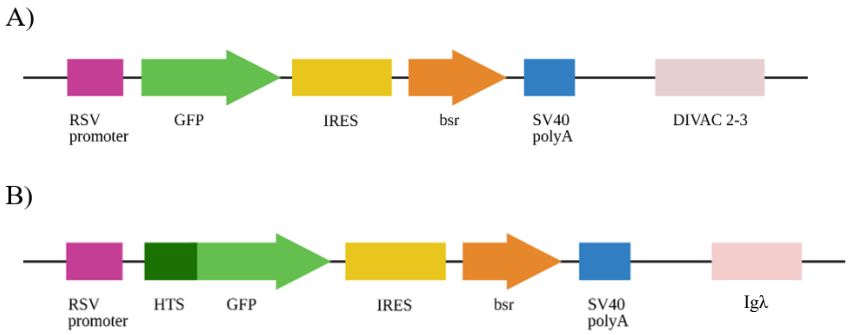


Figure 6. The reporters used as bases for the reporters cloned for this study. A) GFP2 2-3 reporter, which works as a positive control. The negative control GFP2 is otherwise the exact same reporter construct but lacks the DIVAC 2-3. B) GFP4 Igλ reporter, which works as the other positive control. The negative control GFP4 is otherwise the exact same reporter construct but lacks the Igλ DIVAC.

For studying the effects of removing G4 structures from the reporter construct on targeting of somatic hypermutation, a so called GFP2 2-3 G4- reporter was cloned (Figure 7). In that reporter, the possible G-quadruplex forming sequences analysed by G4Hunter (Table 2 and Figure 5) were modified so that their G4 forming potential was removed. Point mutations were introduced to the sequence at those points and the mutated sequence was again analysed with the G4Hunter to make sure that the G4 forming potential of the sequence was diminished with the designed mutations. In total 11 mutations were designed to make the required changes and after the changes the program could not give any G4Hunter scores to the GFP sequence anymore. These mutations and their exact positions are listed in Table 3. The locations of the mutations in relation to the identified possible G4 forming sequences are also shown in Figure 5.

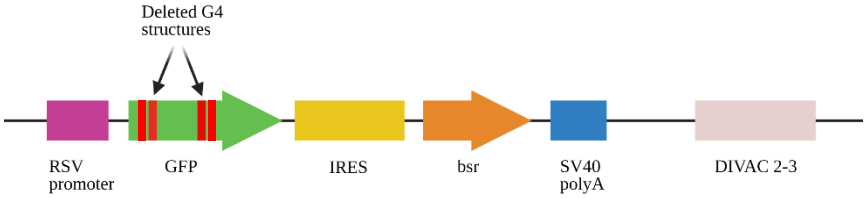


Figure 7. The GFP2 2-3 G4- reporter where possible G-quadruplex forming sequences are deleted. G4 = G-quadruplex.

Table 3. The point mutations introduced to the GFP gene to remove the possible G4 forming sequences

Mutated amino acid (position as counted from the first amino acid of GFP)	**Original codon**	**New codon**
Threonine (60)	ACC	ACA
Threonine (63)	ACC	ACA
Tyrosine (75)	TAC	TAT
Proline (76)	CCC	CCT
Tyrosine (183)	TAC	TAT
Threonine (187)	ACC	ACA
Proline (188)	CCC	CCT
Glycine (192)	GGC	GGA

Proline (193)	CCC	CCT
Aspartic acid (211)	GAC	GAT
Proline (212)	CCC	CCT

Also based on the GFP2 reporter, a tandem reporter with a combination of GFP and DsRed was cloned. Two versions of this so called GFP2+DsRed tandem reporter were created, GFP2+DsRed and GFP2+DsRed G4+, of which the latter contains a strong G4 structure in between of the two fluorescent proteins (Figure 8).

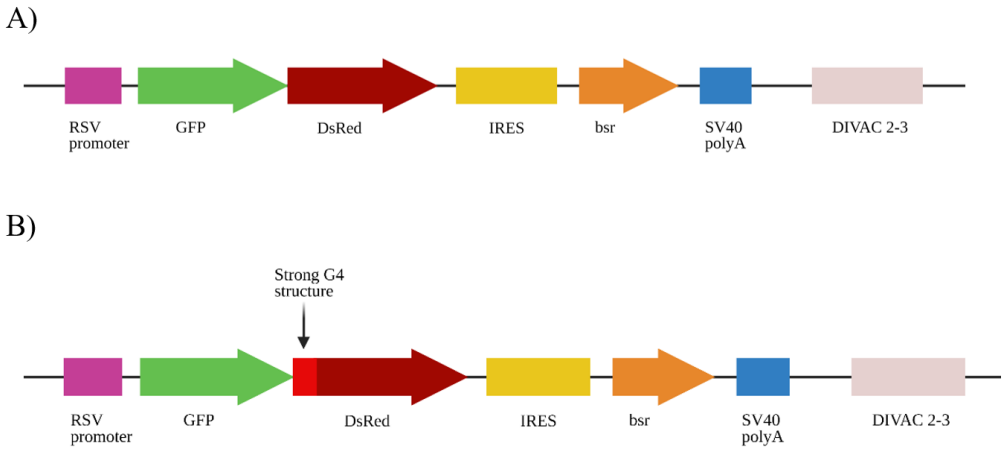


Figure 8. GFP2+DsRed tandem reporters. A) The basic tandem reporter with no modifications to G-quadruplex structures. B) The GFP2+DsRed G4+ tandem reporter with a strong G-quadruplex between the fluorescent proteins. G4 = G-quadruplex

Additionally, to further study the effects of adding a strong G4 structure to the reporter on the targeting of somatic hypermutation, the same strong G4 structure as in the GFP2+DsRed G4+ reporter was added in the middle of the hypermutation targeting sequence (HTS) of GFP4 Igλ creating a so called GFP4 Igλ G4+ reporter (Figure 9).

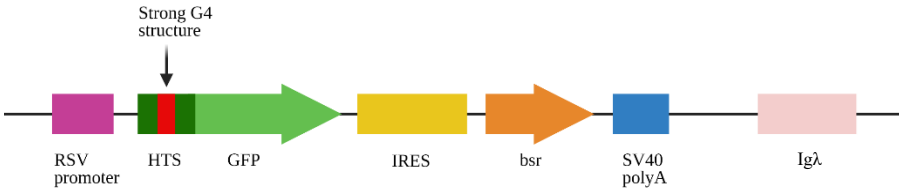


Figure 9. GFP4 Igλ G4+ reporter where a strong G-quadruplex structure is added in the hypermutations targeting sequence. G4 = G-quadruplex

To ensure the success of cloning these reporters, they were sequenced, and it was proven that they contain the desired changes. The complete reporters were then integrated into the genome of DT40 chicken bursal lymphoma B-cell lines. The GFP2 based reporters were integrated into a place where the IgL locus of the cells had been deleted and the GFP4 based reporters into a place where locus containing the gene for AID had been deleted. The reporters where then used to study the effects of G quadruplexes in targeting of somatic hypermutation on the green fluorescent protein in the settings of the GFP loss assay used.

2.3 Removing the potential G quadruplex forming sequences of GFP gene decreases GFP loss significantly in GFP2 2-3 reporter

The correctly targeted DT40 cells that had taken in the cloned GFP reporters were subcloned. The targeting efficacy varied for each reporter (Table 4) and always the best growing clones were chosen for subcloning. 14 days after subcloning, the cells were measured for their green fluorescence with fluorescence activated cell sorting (FACS).

Table 4. Targeting efficacy of transfection of GFP2 reporters.

Reporter	**Number of primary clones**	**Number of correctly targeted clones**	**Targeting efficacy (%)**
GFP2	49	3	6
GFP2 2-3	40	9	23
GFP2 2-3 G4-	34	4	12

The difference in GFP loss between GFP2 2-3 G4- and GFP2 2-3 reporters is statistically significant so that removing the potential G4 forming sequences of the GFP gene reduces the median GFP loss from 3.5 % to 1.8 % ($p = 0.0096$) (Figure 10). This suggests that removing G4 structures reduces somatic hypermutation.

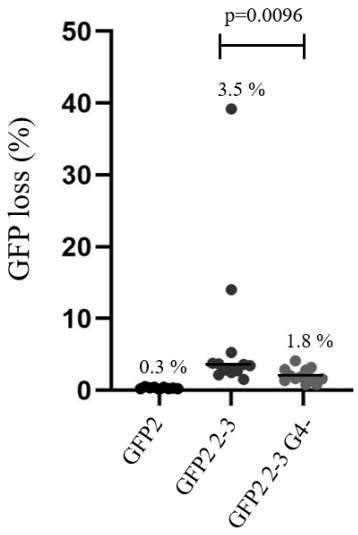


Figure 10. Percentage GFP loss of GFP2 based reporters. All the subclones are plotted in the graph and the line represents the median. The actual value for median is shown above each reporter and the p-value is determined by Mann-Whitney test.

In general, the percentage loss of green fluorescence in the GFP2 2-3 reporters was quite small even in the positive control as it has been observed in the previous studies that the GFP loss in the same reporter should be at least 4 % or more (Blagodatski et al., 2009; Kohler et al., 2012). Figure 11 shows one of the 12 subclones from each control and GFP2 2-3 G4- reporter types measured by FACS. The figure represents the median of the subclones and also shows how the percentage of green fluorescent negative cells decreases when the G4 structures are removed and moves closer to the negative control, where almost no GFP loss is expected.

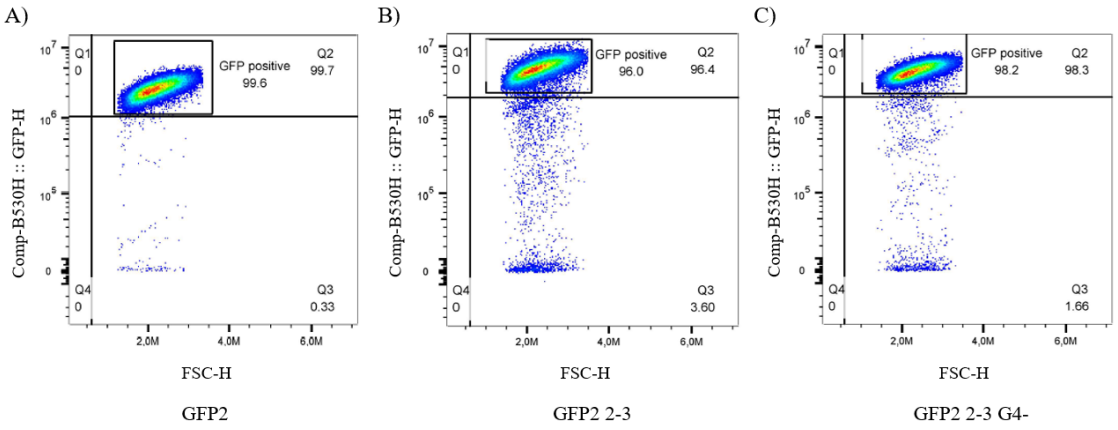


Figure 11. Examples of GFP loss flow cytometry results from subclones of different GFP2 reporters. A) negative control, B) positive control, C) GFP2 reporter where G4 structures have been deleted

2.4 Adding a strong G4 structure only slightly increases GFP loss in GFP4 Igλ reporters

Like the GFP2 reporters, the correctly targeted DT40 cells containing the GFP4 reporters were subcloned. The targeting efficacy varied also for each GFP4 based reporter (Table 5) and were generally worse than the targeting efficacy of the GFP2 reporters. This time 12 days after subcloning, the cells were measured for their green fluorescence by FACS.

Table 5. Targeting efficacy of transfection of GFP4 reporters.

Reporter	Number of primary clones	Number of correctly targeted clones	Targeting efficacy (%)
GFP4	63	9	14
GFP4 Igλ	71	5	7
GFP4 Igλ G4+	54	2	4

Unlike a clear statistically significant effect of removing G4 structures, adding one in the HTS sequence of the GFP4 Igλ reporter did not result in a statistically significant change in GFP loss after growing the subclones for 12 days. However, a slight increase from a median GFP loss of 12.9 in control to 14.9 in the reporter with an added G4 occurred (p=0.9682) (Figure 12). So, the amount of SHM happening was only slightly increased.

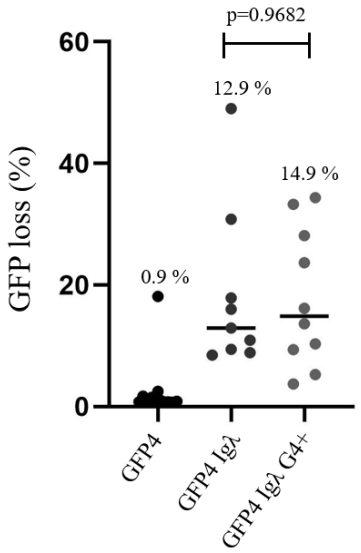


Figure 12. Percentage GFP loss of GFP4 reporters. All the subclones are plotted in the graph and the line represents the median. The actual value for median is shown above each reporter and the p-value is determined by Mann-Whitney test.

In general, the intensity of fluorescence from GFP4 reporters is noticeably lower than that from GFP2 reporters and the GFP positive and negative populations are not as far apart from each other. Despite of that they were still well distinguishable and Figure 13 representing the median GFP losses shows that the positive control here has already higher GFP loss than with GFP2 reporters. This goes well with previous studies denoting that the GFP4 assay is more sensitive than GFP2 assay and therefore resulted in bigger median GFP loss values (Buerstedde et al., 2014).

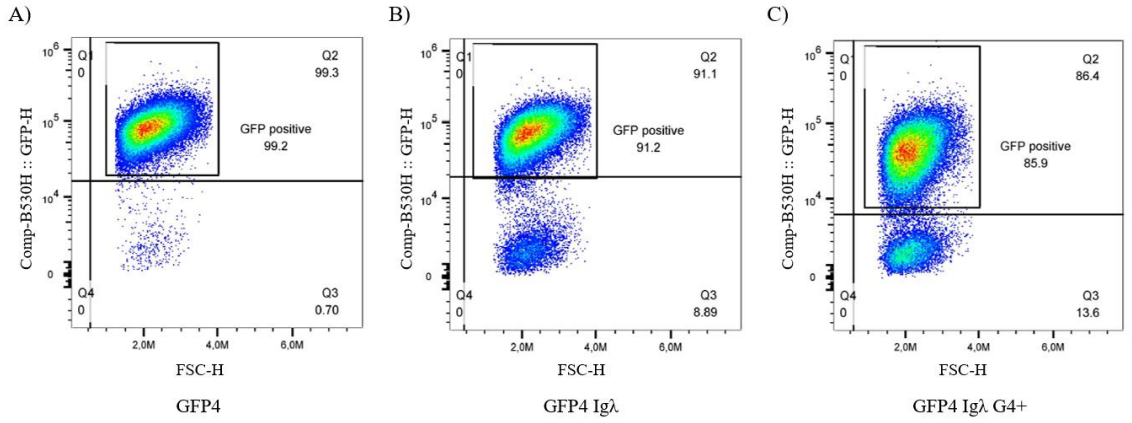


Figure 13. Examples of GFP loss flow cytometry results from subclones of different GFP4 reporters. A) negative control, B) positive control, C) GFP4 reporter where a strong G4 has been added in the hypermutation targeting sequence.

2.5 Adding a G4 in GFP2+DsRed tandem reporter decreases GFP loss

Lastly, adding a strong G4 in the tandem reporter was also considered. The targeting efficacies of transfecting cells with these reporters were the highest of all (Table 6) and also here the best growing clones were chosen for subcloning. These tandem reporters were treated as the other GFP2 based reporters and measured at same timepoints.

Table 6. Targeting efficacy of transfection of GFP2+DsRed tandem reporters.

Reporter	**Number of primary clones**	**Number of correctly targeted clones**	**Targeting efficacy (%)**
GFP2+DsRed	35	9	26
GFP2+DsRed G4+	25	11	44

Unlike adding a strong G4 in the GFP4 Igλ reporter, adding one in the tandem reporter did not increase GFP loss. Instead adding the quadruplex structure decreased the median GFP loss from 8.3 to 3.7 ($p=0.0317$) resulting in a statistically significant change (Figure 14). There is no clear explanation why adding a G4 structure in the tandem reporter resulted in an opposite outcome than when adding a G4 in the GFP4 Igλ reporter.

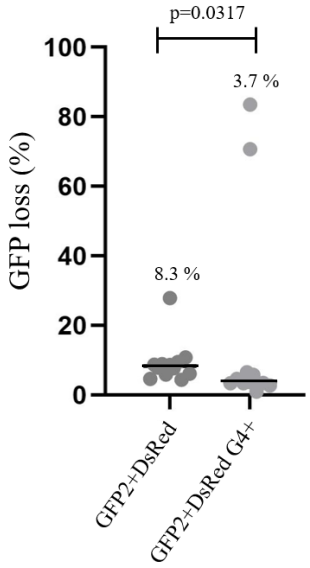


Figure 14. Percentage GFP loss of GFP2+DsRed tandem reporters. All the subclones are plotted in the graph and the line represents the median. The actual value for median is shown above each reporter and the p-value is determined by Mann-Whitney test.

In general, the intensity of the green fluorescence coming from the cells containing the tandem reporters was between that of GFP2 2-3 and GFP4 Igλ reporters and Figure 15 represents the median GFP loss observable. Here it also seems that adding the G4 structure to the reporter decreases the intensity of fluorescence as well.

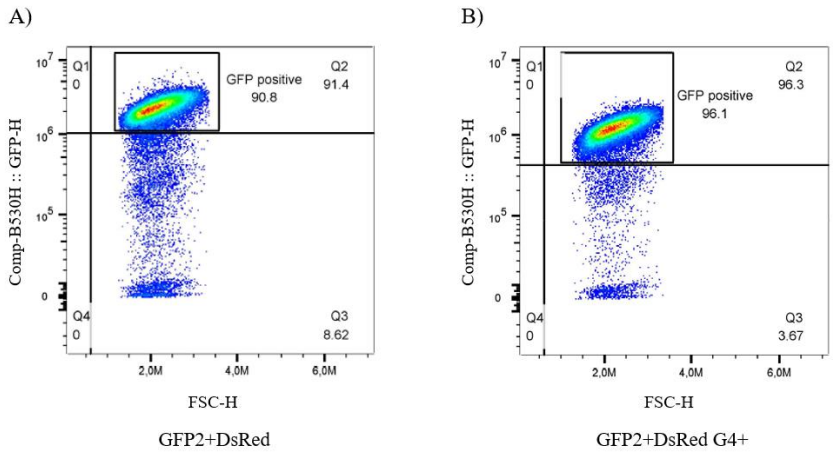


Figure 15. Examples of GFP loss flow cytometry results from subclones of GFP2+DsRed tandem reporters. A) Basic tandem reporter B) Tandem reporter with a strong G4 between the two fluorescent proteins.

As there was no control with only DsRed as a fluorescent protein, the flow cytometry results of loss of red fluorescence were not analysed due to lack of compensations possibilities. By studying the spectra of GFP and DsRed and taking into account the wavelengths used in the measurements, it was noted that there is some leakage of fluorescence from GFP to DsRed channel of the NovoCyte apparatus used for FACS and therefore cannot be analysed reliably.

2.6 Sequencing gives more accurate information on how removing and adding of G4s affect the targeting of somatic hypermutation

After subcloning the primary clones for GFP loss assay, they were further grown up to one month in total. The GFP loss at that point was again measured (Figure 16) and the GFP reporters were sequenced from the DT40 cell genomes. At the moment of extracting the genomic DNA from the cells, the GFP loss was markedly increased from the GFP loss of the subclones as expected as more time had elapsed. Cells containing the GFP2 2-3 reporter were also measured for their fluorescence, but they were not sequenced as the reporter construct has been already sequenced (Tarsalainen and Alinikula, unpublished observations) and the existing data was used for comparing the results from GFP2 2-3 G4- reporter to them.

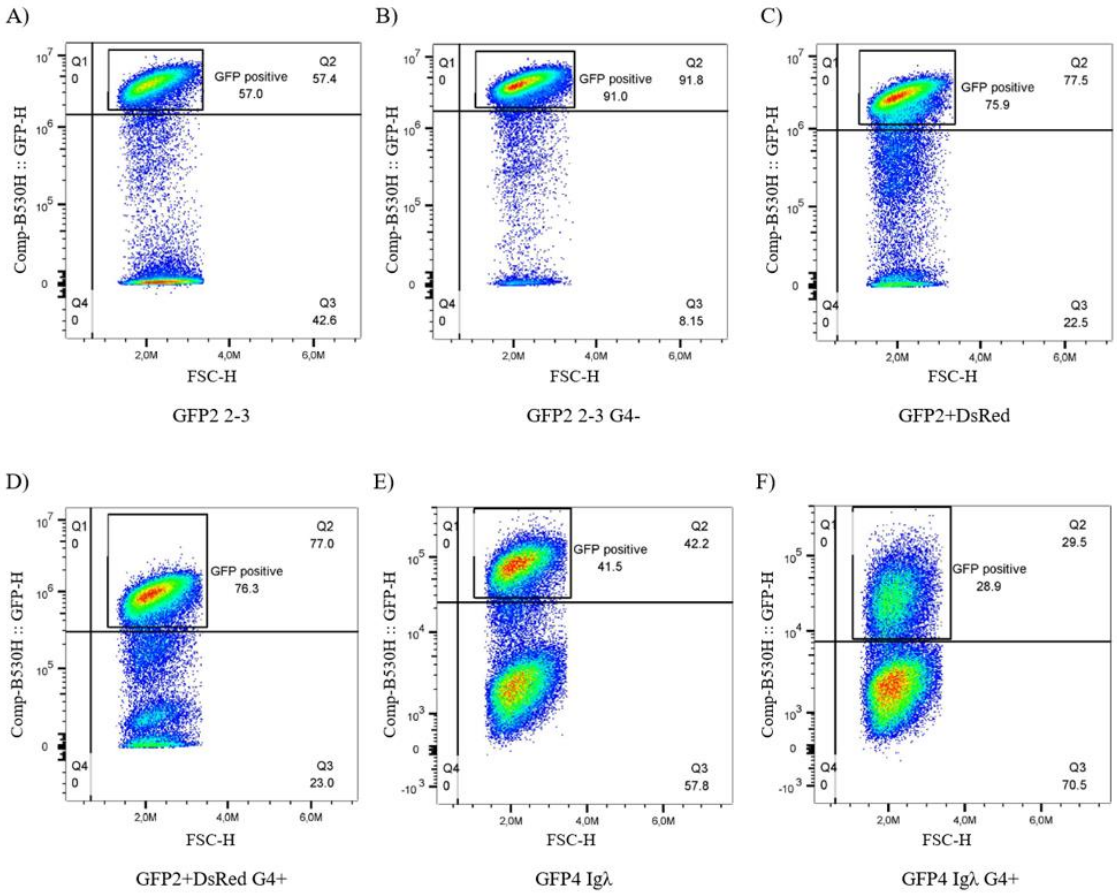


Figure 16. GFP loss of primary clones right before sequencing to see how the accumulation of mutations had advanced from the subclone GFP loss analysis. A) GFP2 positive control, B) GFP2 reporter where G4 structures had been deleted, C) GFP2-DsRed tandem construct, D) GFP2-DsRed tandem construct with strong G4 structure in between the genes of the two fluorescent proteins, E) GFP4 positive control, F) GFP4 where a strong G4 structure has been added in the hypermutation targeting sequence.

As already seen from Figure 16, the situation at the moment of sequencing showed that there is less GFP loss in GFP2 2-3 G4- than in GFP2 2-3 reporter suggesting that there would be fewer mutations. Also, especially in the GFP4 Igλ reporters, adding a G4 has noticeably increased the GFP loss.

From the sequenced reporters, the number and type of mutations were recorded, and the mutation frequency calculated (Table 7). The sequencing results fit together with the FACS results as the mutation frequency is lower in GFP2 2-3 G4- reporter when compared to GFP2 2-3 reporter and again higher in GFP4 Igλ G4+ reporter when compared to GFP4 Igλ reporter. Also, in the tandem reporter, adding a G4 has slightly increased the mutation frequency.

Table 7. The number, frequency and types of mutations found in the reporters after one month of growing. GFP2 2-3 had been sequenced already before and that data was analysed from suitable parts to compare GFP2 2-3 G4- reporter to it.

	GFP2 2-3	GFP2 2-3 G4-	GFP2+ DsRed	GFP2+ DsRed G4+	GFP4 Igλ	GFP4 Igλ G4+
Number of sequences	94	56	53	52	39	48
Number of bp sequenced	70876	42224	55438	56108	38844	49392
Number of mutations	53	25	61	80	135	288
Mutation frequency	0,000748	0,000592	0,0011	0,001426	0,003475	0,005831
Mutations/ 1000 bp	0,747785	0,59208	1,100328	1,425822	3,47544	5,830904
Transitions	18	5	21	17	118	267
Transversions	31	10	25	28	10	13
Insertions	4	18	6	23	19	27
Deletions	0	8	9	12	13	6
Mutations/ seq	0,56383	0,446429	1,150943	1,538462	3,461538	6

Furthermore, the mutations were positioned along the GFP constructs with transcription start site (TSS) as starting point. For the Figures 17-19 the number of mutations has been divided into 50 base pair bins and normalized to the number of sequences obtained for analysis. The figures clearly point out the differences in the mutation distributions along the different reporters.

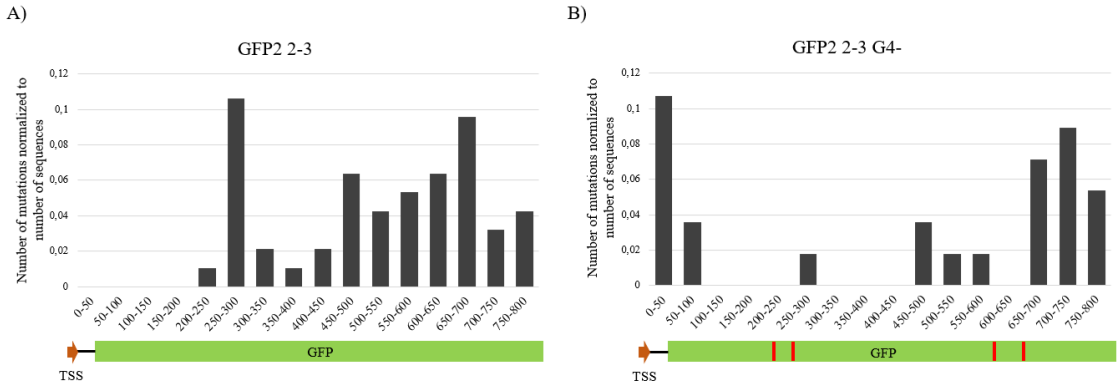


Figure 17. The distribution of mutations in A) GFP2 2-3 and B) GFP2 2-3 G4- reporters. The positions from where the G4 structures were deleted in GFP2 2-3 G4- reporter are shown as red lines in the GFP gene. TSS = transcription start site

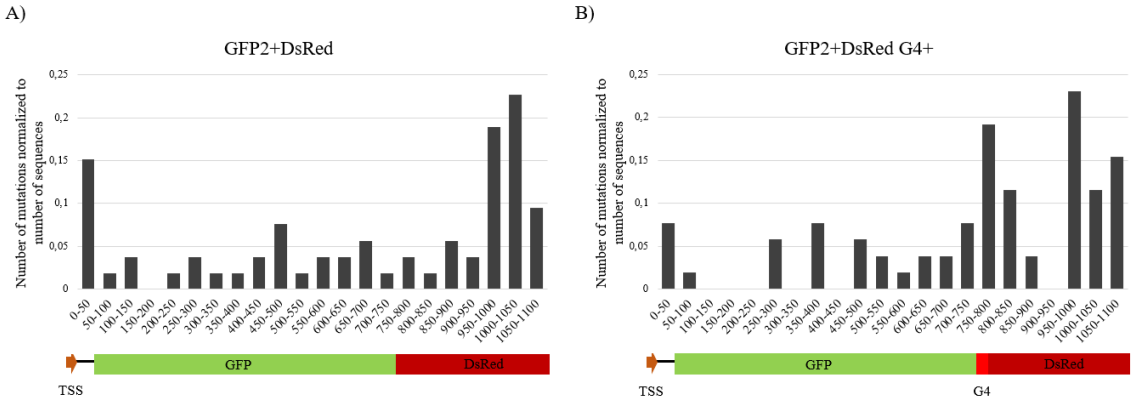


Figure 18. The distribution of mutations in A) GFP2+DsRed and B) GFP2+DsRed G4+ reporters. The position of the added strong G4 structure in the second reporter is marked with red in between the GFP and DsRed genes. TSS = transcription start site, G4 = G-quadruplex

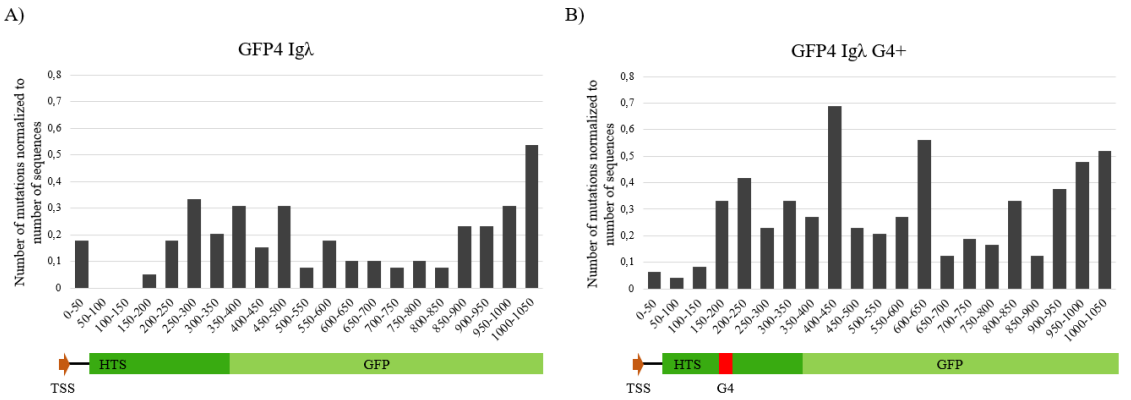


Figure 19. The distribution of mutations in A) GFP4 Igλ and B) GFP4 Igλ G4+ reporters. The position of the added G4 is marked with red in the second reporter. TSS = transcription start site, G4 = G-quadruplex, HTS = hypermutation targeting sequence

As seen from Figure 17 the most common mutation sites in GFP2 2-3 are located close to the points where the possible G4 forming sequences were. However, the distribution of the mutations has clearly shifted in the GFP2 G4- reporter and it seems that removing the G4s has not only decreased the mutation frequency, but it has also clearly reduced the number of mutations especially at those points where the G4s used to be.

Figure 18 moves on showing the effects of adding a strong G4 to the distribution of the mutations along the reporters. Here it also seems that adding a G4 has not only an effect in increasing the mutation frequency, but it also “attracts” mutations. This is seen as a peak in number of mutations around the G4 structure in GFP2+DsRed G4+ reporter in comparison to the same location from TSS in GFP2+DsRed reporter.

The GFP4 reporters in Figure 19 continue showing the effects of adding a strong G4 structure to the locations of the mutations. Here a similar effect is observable as in Figure 18 with GFP2+DsRed tandem reporters. A slight shift in the mutation distribution towards the G4 structure is seen in GFP4 Igλ G4+ reporter when compared to the same area in GFP4 Igλ reporter.

2.7 Studying the strand bias of mutations in GFP4 reporters

The GFP4 based reporters were integrated into the genome of DT40 cells that were devoid of the UNG enzyme. In UNG deficient cells, most of the mutations are G to A and C to T transitions as the mutations site is mainly repaired by transcription over the site. (Saribasak et al., 2006) Therefore, they can be used to assess the strand of origin of the formed mutations.

This is also well represented in this analysis as around 80 % of all the mutations observed are transitions (Table 7). The reporters were sequenced only in the direction of the coding strand and it can be deducted that if the observed mutation was G to A, the original mutation was in the template strand and if C to T, the mutation in the coding strand. As seen in Table 8, the percentage of G to A transitions slightly increases from 43 to 51 % ($p=0.7548$) in the GFP4 Igλ G4+ reporter when compared to the reporter without strong G4 structure. However, this is not statistically significant and there is close to no difference in the portion of C to T transitions. The biggest change is observable in the amount of insertions and deletions, but from there, there is no reliable evidence that on which strand the original mutation has occurred.

Table 8. Types of mutations in GFP4 Igλ reporters

Reporter	G-A transitions (%)	C-T transitions (%)	Transversions (%)	Deletions (%)	Insertions (%)
GFP4 Igλ	43.0	35.6	4.4	8.1	8.1
GFP4 Igλ G4+	51.0	36.8	4.5	5.2	2.1

3. Discussion

The results of this study suggest that G-quadruplexes have an effect in targeting of somatic hypermutation. Removing the possible G4 structures in the used GFP reporters significantly reduced GFP loss as well as mutation frequency and there were fewer mutations at the positions from where the G4s had been removed. Adding a strong G4 did not lead to a statistically significant increase in GFP loss, but still increased the mutation frequency and caused an increase in the number of mutations around the position where the G4 was added when compared to the same position in the control reporter. This goes well together with the previous observations in CSR that AID preferably binds and mutates cytosines close to G4 structures (Qiao et al., 2017). Previous observations in mice and with mouse B cells also suggested that mutating AID so that it cannot bind G4s, diminishes CSR and SHM (Yewdell et al., 2020; Zheng et al., 2015). CSR was also diminished if the folding of the G4 structures was disrupted (Zheng et al., 2015). In light of this experiment, it also seems that removing the G4 structures from the target gene also decreases SHM without mutations in AID.

As the median GFP loss of the measured GFP4 Igλ G4+ subclones was slightly increased, it raises a question whether increasing the sample size from 12 could have resulted in a statistically significant result if the difference in GFP loss had persisted. Also, as the mutation frequency in the GFP4 Igλ G4+ is almost twice as much as in GFP4 Igλ reporter it lets us expect that there should have been more GFP loss observed already with this sample size. It can be speculated that the reason for the statistically unsignificant change in GFP loss could lie also in the possibility that the mutations happening in GFP4 Igλ G4+ have not led to stop codons as often or have changed the codon into another one that still codes for the same amino acid as the original one and thus do not change the amino acid sequence of GFP.

What causes the decrease in GFP loss when a strong G4 structure is added to the GFP2+DsRed tandem reporter is not known. Perhaps also for this reporter the study could be repeated with a larger sample size. However, also in this reporter the mutation frequency was increased, and the locations of mutation shifted towards the G4 structure. Therefore, it seems that the effects of adding a G4 could be more in increasing the mutations then decreasing them even though the GFP loss was reduced in this case.

Another interesting aspect that comes to the G-quadruplex structures removed in this study from the GFP2 2-3 reporter is that they are at this point only theoretical G4 structures as the sequences forming them have been predicted by an online software but have not been actually observed to form the quadruplex structure in GFP gene. On the contrary, the strong G4 structure used has been studied before and is known to form a very stable G-quadruplex (Nguyen et al., 2020).

The effects of the strong G4 were noted also in the transcription of GFP as the transcription of the fluorescent protein was always lower in the reporter containing the quadruplex structure when compared to the same report without it. This suggests that the DNA secondary structure really is there as it is known that it can hinder the functioning of RNA polymerase and hence lead to lower transcription (Nayun Kim, 2019; Broxson et al., 2011; Duquette et al., 2004).

On the other hand, removing the G4 forming potential of the GFP gene did not affect transcription and the transcription levels of GFP in GFP2 2-3 G4- reporter remained the same compared to GFP2 2-3 reporter. As adding a G4 structure decreases transcription, it could have been assumed that removing them would respectively increase it by removing hinderance from transcription. This leads to the question that could the decrease in GFP loss and mutation frequency and the change in mutation distribution be a result of something else than removing actual G4 structures. However, important to note here as well is that the possible removed G4s were in the template strand and the G4 was added to the coding strand. The articles that have discussed the effects of G4s in transcription have described G-quadruplexes especially in the coding strand (Nayun Kim, 2019; Broxson et al., 2011; Duquette et al., 2004) and therefore it cannot be expected that G4s in template strand would work in the same way and have a similar effect in transcription.

As disclosed before, WRCY hotspots where W = A/T, R = A/G and Y = C/T have been recognized to be the preferable locations for mutations caused by AID (Feng et al., 2020; Qiao et al., 2017). These hotspots were not taken into account when designing the mutations in GFP gene to remove the G4 forming potential of the gene. By afterwards analysis of the changed bases in GFP gene together with the hotspots, it was noted that 2 out of the 11 mutations created when cloning the GFP2 2-3 G4- reporter changed such hotspots from TACC to TATC. In theory this could affect at least to some extent the number of mutations close to those hotspots. However, these were only two hotspots

disrupted from the many that can be found from GFP gene. Additionally, in a previous study from Qiao *et al.* in 2017, it has been established that in CSR, the G4 structure could possibly even override the hotspot preference of AID so perhaps the same could apply here in SHM. Therefore, disruption of these two hotspots is most likely not the only reason for the significant change observed.

What is also many times discussed when it comes to the mutations in somatic hypermutations process, is the strand bias of the mutations. It is known that both DNA strands, the coding and the template strand, are targeted by AID and some bias over deaminated cytosines in the coding strand over the template strand has been observed (Sohail et al., 2003; Saribasak et al., 2011). As the GFP4 reporters where transfected into cells deficient of UNG2, most of the mutations were only G to A and C to T transitions as the repair of the uracil in the DNA was mainly conducted by replication over the lesion site (Saribasak et al., 2006).

As seen from Table 8 there is a slight increase in G to A transitions in the GFP4 Igλ G4+ reporter when compared to the GFP4 Igλ. So, a slight increase in the mutations happened in the template strand when the G4 structure was added to the coding strand. However, this increase was not statistically significant and the portion of mutations in the coding strand itself did not decrease and the change was mainly dependent in the decrease of insertions and deletions that happened less in the GFP4 Igλ G4+ than in the GFP4 Igλ reporter. Of the deletions and insertions, it is still difficult to say on which strand the original mutation had been. From the deletions it can be always speculated that if the deleted base from the coding strand is C, most likely it could be the C that was originally deaminated. However, this would not change the numbers drastically and the result would still be inconclusive.

From the GFP2 reporters it is difficult to say that on which strand the mutations have originally happened as the DNA repair machinery in them is intact and all the possible mutations with any method described in SHM (Figure 2) can occur. Also, the results even from the GFP4 Igλ reporter are not in line with the previously obtained data in other studies where more mutations happen in the coding strand (Sohail et al., 2003; Saribasak et al., 2011). Also, the results from Qiao *et al.* suggest that the ideal deamination site for AID would be at the third position from a guanine stem in G4, but still on the same strand with the structure and the results here oppose this observation as well. However, the

difference between the percentage of mutations on each strand is small and the sample size very limited so a larger analysis with perhaps more high-throughput sequencing could be helpful in getting clearer results. Therefore, in the scope of this thesis work a clear statement on the effect of G4s on the strand bias cannot be given.

Also, for future studies it would be interesting to test if adding two or more strong G4 structures could result in a stronger effect in the number and locations of mutations. In GFP2 2-3 G4- reporter, there are several possible G4 structures close to each other that are disrupted, and the results of that reporter compared to the control are more significant. Therefore, maybe adding two G-quadruplexes next to each other would also result in more GFP loss in GFP4 Igλ G4+ reporter.

One big question still remaining is that are these findings done within the GFP gene actually relevant in the immunoglobulin variable genes in B cells. As pointed out in the introduction, it is well known that the switch regions can form G-quadruplexes but evidence from the variable regions remains sparse (Dunnick et al., 1993; Pavri, 2017; Qiao et al., 2017). However, when studying some of the human IgV regions with the G4Hunter program in this study, it seems that the sequences have a potential to form G4s as well (Table 1) even though they have not been extensively studied. This suggests that the role of G4s in AID mediated diversification could extent from the more studied CSR to SHM as well. Still to validate this in the future, it would be interesting to see if the G4s within IgV regions could be imaged similarly with electron and atomic force microscopies like within the switch regions (Neaves et al., 2009; Duquette et al., 2004).

An interesting idea is also that if G-quadruplexes do not have an important role in targeting of somatic hypermutation to the IgV areas, maybe their relevance is in the off targeting of somatic hypermutation as it is known that oncogenes are rich in their guanine content, their promoters are observed to form G4s and G4 structures are observed in cancerous tissue (Balasubramanian et al., 2011; Biffi et al., 2014).

As a conclusion, this thesis suggests that G-quadruplexes can play a role in determining the exact locations of mutations when other requirements for targeting of somatic hypermutation are met. Even though the G4s seem to have an effect, they most likely cannot override the other targeting mechanisms. Meaning that there still must be transcription going on and stalling of polymerase so that there are ssDNA substrates available for AID and mutation enhancers such as DIVACs are needed to possibly recruit

all the machinery needed for SHM. Also, the exact mechanism how G4s work in targeting of the mutations is still unknown. Perhaps they can stabilise the single-strandedness of the opposite DNA strand so that AID can more easily target deoxycytidines in it, or it can form a good, bifurcated structure for AID to bind and mutate deoxycytidines around it. The G4s also lowered GFP transcription and can therefore possibly contribute to polymerase stalling, but with the limited data from this study, it is impossible to provide a detailed description of the actual mechanism. So, there is still a lot of work to do for the future. Additionally, the exact role of G4s in the targeting of SHM to IgV regions when compared to their role in the off targeting of SHM requires more looking into.

4. Materials and methods

4.1 Analysing the G-quadruplex structures in immunoglobulin variable genes

The sequences coding for immunoglobulin heavy chain variable regions were studied with an online program called G4Hunter (version 3.3.3-3-g32cee6f, http://bioinformatics.ibp.cz/#/). It is an available to all platform that can be used to analyse possible G4s formed by sequences. The parameters for the search, the window and the threshold, were chosen to be 20 and 1.2 respectively (Bedrat et al., 2016; Brázda et al., 2019). The sequences for the IgV regions were obtained from the IMGT Database (Giudicelli et al., 2005). In total 28 sequences were analysed comprising of few functional sequences from each IGHV subgroup (Table 9).

Table 9. The genes for immunoglobulin variable regions used for G-quadruplex analysis and IMGT accession numbers for them.

Gene	**IMGT accession number**
IGHV1-2	X07448
IGHV1-3	X62109
IGHV1-8	M99637
IGHV1-18	M99641
IGHV1-24	M99642
IGHV1-45	X92209

IGHV1-46	X92343
IGHV1-58	M29809
IGHV2-5	X62111
IGHV2-26	M99648
IGHV2-70	L21969
IGHV2-70-D	KC713935
IGHV3-7	M99649
IGHV3-9	M99651
IGHV3-11	M99652
IGHV3-13	X92217
IGHV3-15	X92216
IGHV3-20	M99657
IGHV3-21	M99658
IGHV4-4	X62112
IGHV4-28	X56358
IGHV4-30-2	L10089
IGHV4-30-4	Z14238
IGHV4-31	L10098
IGHV5-10-1	X92227
IGHV5-51	M99686
IGHV6-1	X92224
IGHV7-4-1	L10057

4.2 Designing the GFP reporters used

The reporters that were designed for this study were based on already existing GFP2 and GFP4 reporters described in articles by Blagodatski *et al.* in 2009 and Buerstedde *et al.* in 2014 respectively (Figure 6). The original GFP reporters will serve as negative controls.

The GFP2 reporter includes a DIVAC sequence called DIVAC 2-3 and will be therefore called GFP2 2-3. The DIVAC 2-3 is a part of the chicken immunoglobulin lambda enhancer sequence and is known to lead to high levels of GFP loss (Kohler et al., 2012).

In the reporter it is inserted immediately downstream of the polyA-site sequence of the GFP -ires-Bsr-transcription unit. Additionally, mouse R1 intronic sequence is added right upstream of the Rous sarcoma virus promoter included in the basic reporter (Alinikula et al., Manuscript).

The GFP4 reporter includes a DIVAC sequence called Igλ and is therefore referred to as GFP4 Igλ reporter. The Igλ is the human immunoglobulin lambda enhancer sequence which has been shown to act as a diversification activator in GFP4 reporter assay (Buerstedde et al., 2014). The Igλ has been inserted into the reporter downstream of the GFP gene at BamHI restriction cite.

4.2.1 GFP2 2-3 G4- reporter

This reporter was used for studying how removing G4s from GFP coding sequence affects the targeting of somatic hypermutation on it.

For studying the G4 structures formed naturally by GFP gene, the same G4Hunter online program was used as for analysing the G4 structures in IgV genes. The parameters for analysis were also the same. Mutations were designed to those sequences with the help of SnapGene (version 3.3.4) so that they could not form G4 structures anymore.

4.2.2 GFP2+DsRed reporter

Gene for DsRed from another vector (pX458-DsRed, Addgene) was taken and inserted into GFP2 2-3 reporter construct after the GFP coding sequence with only a sequence for a T2A self-cleaving peptide between them.

4.2.3 GFP2+DsRed G4+ reporter

The GFP2+DsRed reporter was further modified by adding a strong G4 structure directly after GFP gene before the T2A sequence. The sequence for G-quadruplex used here has been studied before by Nguyen et al. in 2020. It has a sequence 5′-TTGGGTGGGTTTCGCGCAGCGTTTGGGTGGGTT-3′, and it was added to the

coding strand. It is known to quickly form a stable G4 structure with a duplex stem-loop (Nguyen et al., 2020).

4.2.4 GFP4 G4+ reporter

This reporter is based on a GFP4 reporter. There preceding the GFP coding sequence is a hypermutation targeting sequence and in this study, the strong G4 was created in the middle of this HTS, on the coding strand, after tryptophan 32. The sequence used for G4 is the same as in GFP2+DsRed G4+ reporter.

4.3 Cloning the GFP reporters

The designed GFP reporters were cloned with a process called In-Fusion Cloning. The pieces of the reporters for the cloning process were created by polymerase chain reactions (PCR).

4.2.1 Polymerase chain reaction

The primers for the PCR phase were designed with the help of SnapGene (version 3.3.4) so that the desired mutations or the beginning/end of the desired insert are contained at the 5’ end of the primer. Full list of used primers is found in Appendix 1.

All PCR reactions were carried out with the Q5® High-Fidelity DNA Polymerase from New England Biolabs (NEB) and the recommended reaction setup given in the official protocol for this enzyme by NEB was used (see Appendix 2 for the setups used). For longer PCR reactions (PCR product > 13 000 bp) the amount of template used was 500 pg-5 ng and for smaller inserts (PCR product 50-800 bp) it was 100 pg.

After the PCR, the samples were run on an agarose gel. 2.5 % agarose was used for small DNA fragments and 0.6 % for long. The desired bands were then cut out from the gel and the DNA was extracted and purified with NucleoSpin Gel and PCR clean-up mini kit (REF 740609.50) from Macherey-Nagel by following the instructions from the manufacturer.

4.3.2 In-Fusion reactions

To combine all the parts of the reporter constructs created by PCR, In-Fusion reactions were carried out. The In-Fusion® HD Cloning Kit from Takara Bio was used and the manual given by them was followed for optimal reactions.

4.3.3 Miniprep

To produce the plasmids containing the GFP reporters created by In-Fusion reactions, the plasmids were transformed into Stellar™ Competent Cells (Clontech Laboratories, Inc. Takara Bio Company). For each transformation, 30 µl of competent cells in a 1.5 ml Eppendorf tube were used and always 2.5 µl of In-Fusion reaction was added on them. The mixtures were incubated for 30 minutes on ice, heat shocked for exactly 60 seconds at 42 °C and incubated for further 2 minutes on ice. The volume of the cell suspension was increased to 500 µl with SOC-medium that was pre-warmed to 37 °C and incubated for one hour at 37 °C by shaking at 225 rpm. An appropriate amount of the suspension was then plated on agar plates containing ampicillin (100 µg/ml). The plates were let to incubate at 37 °C overnight and afterwards in room temperature for 6 hours. Well grown bacterial colonies were chosen and suspended in 3 ml LB-medium containing ampicillin (100 µg/ml). This bacterial culture was then grown overnight at 37 °C.

The plasmids were then purified from the bacteria with GeneJET Plasmid Miniprep Kit (#K0503) from ThermoFisher Scientific. The purification was carried out as instructed by the manufacturer of the kit.

To ensure that the plasmids produced and purified were correct and contained the desired changes, samples from them were prepared for sequencing with Mix2Seq Overnight kit (Eurofins Genomics) and sent to Eurofins in Germany where the sequencing took place. After receiving the results, the sequenced sections were aligned with the desired plasmid with the help of SnapGene (version 3.3.4) to see which minipreps contained the correct mutations and insertions.

4.3.4 Maxiprep

To produce needed amounts of plasmids containing the reporter constructs for transfection, maxipreps were made from the minipreps proven correct by sequencing.

Roughly 3 ng of plasmid DNA and 25 µl of Stellar Competent Cells were used for each transformation. The cells and DNA were mixed in an Eppendorf tube and incubated 30 min on ice. Next the cells were heat shocked for 60 seconds at 42 °C and incubated for further 2 minutes on ice. The volume of the cell suspension was increased to 500 µl with SOC-medium pre-warmed to 37 °C and incubated for an hour at 37 °C by shaking at 225 rpm. After the hour, an appropriate amount of this cell suspension was plated on an agar plate containing ampicillin (100 µg/ml). The plates were let to grow overnight at 37 °C.

From every plate one bacterial colony was chosen and transferred into 2 ml LB-medium containing ampicillin (100 µg/ml) and this pre-culture was incubated for 7 hours at 37 °C by shaking at 225 rpm. 1 ml of this pre-culture was then transferred to 150 ml fresh LB-medium with ampicillin (100 µg/ml) and returned to 37 °C and incubated overnight by shaking at 225 rpm.

The plasmids were then purified with GenElute™ HP Endotoxin-Free Plasmid Maxiprep Kit from Sigma-Aldrich (Catalogue number NA0410) by following the official instructions.

Complete maps of the final plasmids can be found from Appendix 3.

4.4 GFP loss assay

4.4.1 Linearization and precipitation of the reporter plasmids

To prepare the DNA needed for transfection, the plasmids containing the reporter constructs were linearized with NotI restriction enzyme. Always 50 ng of the plasmid DNA was linearized for one transfection. 2 µl of restriction enzyme was used per digestion reaction and the reaction was let to process overnight at 37 °C.

On the next day, 1 µl of the linearized plasmid was run on a 1 % agarose gel to ensure a complete linearization. When the result was satisfactory, the linearized plasmid DNA was precipitated so that the final volume of the DNA was 300 µl in suitable buffer for transfection.

The precipitation was started by adding 500 µl of ice cold 100 % EtOH on the DNA. The mixture was vortexed well and the precipitate was centrifuged down with a speed of 15 000 g for 5 minutes. The supernatant was removed, 500 µl of cold 70 % EtOH was added on the pellet and the centrifugation was repeated for three minutes. The supernatant was again carefully completely removed, and the pellet was air dried at room temperature for about 10 minutes until it was completely dry.

The dried pellet was eluted in 50 µl TE-buffer and the volume was increased to 300 µl with PBS.

4.4.2 Cell culture and transfection

Two cell lines were used in the experiments, namely DT40 pseudoV^{-}IgL$^{-/Puro}$AID^{R2} (Blagodatski et al., 2009) and DT40 UNG$^{-/-}$AID$^{R/Puro}$ (Buerstedde et al., 2014). They are both chicken DT40 bursal lymphoma cell lines and are grown at 40 °C with 5 % CO_2 in medium consisting of RPMI-1640 with 10 % FBS, 1 % NCS, 1 % penicillin-streptomycin (10 000 U/ml) 1 % Glutamax (Gibco) and 0.1 % β-mercaptoethanol (DT40-medium). The cells were grown till the cell viability was 90 % or higher and density of viable cells was between 0.5 and 0.9 million cells per millilitre. The reporters based on GFP2 were transfected into DT40 pseudoV^{-}IgL$^{-/Puro}$AID^{R2} cell line and those based on GFP4 into DT40 UNG$^{-/-}$AID$^{R/Puro}$ cell line.

12 million cells were harvested for one transfection by centrifugation at 200 g for 5 minutes. The correct number of cells was resuspended in 400 µl PBS and transferred into a pre-chilled 0.4 cm cuvette. 300 µl of plasmid DNA containing 50 µg of DNA was added on the cells in the cuvette and incubated for 10 minutes on ice. The cells were electroporated with MicroPulser™ Electroporation Apparatus (Bio-Rad) with a voltage of 0.7 kV, resistance of the circuit 200 Ω and the capacitance of the apparatus being 25 µF. After electroporation, the cells were moved back on ice for five minutes and then transferred into 10 ml of pre-warmed DT40 medium. The cells were divided on a 96-well plate so that each well contained 100 µl of the transfected cells.

The cells were grown overnight at 40 °C and on the next day, blasticidin selection was added on the cells to make sure that only the cells that have taken in the plasmid are growing. For the blasticidin selection, 100 µl of medium containing 30 µg/µl blasticidin

was added on each well with the transfected cells. Here the medium used was conditioned medium with a consistency of 40 % old media harvested from growing cells, 5 % FBS and 55 % fresh DT40-medium. After the selection, the cells were let to grow at 40 °C for 9 days.

To prepare the conditioned medium, the old medium was harvested from the growing cells by centrifugation at 500 g for 5 minutes so that all the cells were pelleted to the bottom. The medium was collected and filtered through a 0.2/0.22 µm filter and was then ready to use.

On days 7-9 after transfection, the primary clones were picked from the wells and transferred to clean wells on a new 96-well plate containing 200 µl of freshly made conditioned medium consisting of 25 % old medium harvested from growing cells, 70 % fresh DT40 medium and 5 % FBS. On the next day, 100 µl of the cells were moved into a neighbouring empty well and replaced with 100 µl of fresh medium on the original wells. Into the new wells 100 µl of medium with puromycin (2 µg/ml) was added. The plates were placed back at 40 °C over night.

The results of the puromycin selection were studied on the next day, when correctly targeted cells died in the wells containing puromycin. The correctly targeted clones were then moved into bigger volume and split every 1 to 2 days according to their growth.

When the viability of the cells was over 90 %, part of them were saved by freezing and storing them in liquid nitrogen and part were further subcloned. For subcloning, 96-well plates with 300, 600 and 900 cells were created and let to grow for 7-8 days at 40 °C. After this time, 12 clones were picked and transferred into 200 µl of fresh DT40-medium and let to grow.

4.4.3 Flow cytometry

For DT40 pseudoV^{-}IgL$^{-/Puro}$AID^{R2} cells containing the GFP2 reporters, 14 days after subcloning the 12 subclones picked were studied with flow cytometry for their degree of GFP loss. For DT40 UNG$^{-/-}$AID$^{R/Puro}$ cells, the measurements were done 12 days after subcloning.

For the fluorescence activated cell sorting (FACS) NovoCyte Flow Cytometer was used. The laser used was 488 nm, filter for GFP 530/30 and for DsRed 585/40 and always 50 000 live cells were recorded.

Appropriate gates were adjusted for each subclone measured.

4.5 Sequencing of the fluorescent proteins

For sequencing, the primary clones subcloned for GFP loss assay were further grown up to four weeks. Then the genomic DNA was extracted from the cells with Quick-DNA™ Miniprep Kit from Zymo Research (Catalogue number D3024). 3-5 x 10^6 cells per each clone were used for the gDNA extraction and the protocol for cell suspensions given by the manufacturer was followed.

From the extracted gDNA, the reporter sequences were amplified by PCR. Always 400 ng of DNA was used as a template and like when cloning the reporters, all PCR reactions were carried out with the Q5® High-Fidelity DNA Polymerase and the recommended reaction setup was used (Appendix 2). To ensure correct PCR products, 3 µl of each product was run on a 1.5 % agarose gel with 120 V for 1 hour and 15 minutes.

The correct PCR products where then cloned into pCR™4Blunt-TOPO® plasmid vectors (Zero Blunt® TOPO® PCR Cloning Kit for Sequencing from Invitrogen, catalogue number 450159). For each reaction 2 µl of PCR product was used and the reaction was set up according to instructions given in the official protocol of the kit. The incubation time of the reaction was increased from 5 minutes to 20 before transforming the clones into Stellar™ Competent Cells.

For the transformation 2 µl of TOPO® Cloning reaction and 40 µl competent cells were used. The transformation reactions were incubated for 30 minutes on ice, heat shocked for exactly 60 seconds at 42 °C and incubated for further 2 minutes on ice. Then the volume of the cell suspension was increased to 500 µl with SOC-medium that was pre-warmed to 37 °C and incubated for one hour at 37 °C by shaking at 225 rpm. After that, appropriate amounts of bacteria were plated on agar plates containing kanamycin (50 µg/ml) and let to incubate at 37 °C overnight.

On the next day, 57 bacterial colonies from each transformation reaction were inoculated in separate wells on PlateSeq Kit Clone agar plates from Eurofins Genomics. These plates were let to incubate further at 37 °C degrees overnight and on the next day they were sent to Eurofins to be sequenced.

4.6 Statistical analysis

The data received from NovoCyte was firstly analysed by FlowJo-Win64-10.7.1 to obtain the mean GFP loss (%) for each subclone from every reporter construct. The means of the experimental reporter constructs where then compared with GraphPad Prism (version 9.0.2) to their positive controls. P-values for the differences were calculated with Mann-Whitney test. From the analysis of the data, subclones with GFP loss greater than 95 % were excluded.

Sequencing analysis was done with SnapGene (version 3.3.4) where the sequenced data was aligned with the original reporters and the number, type and location of mutations were collected manually. These were then used to calculate the frequency of mutations and analyse the mutation distribution along the reporter genes with the help of Microsoft Excel. P-value for studying the strand bias of the mutations was calculated with Fisher's exact test by GraphPad Prism.

5. Acknowledgments

This thesis was conducted at the Institute of Biomedicine at the University of Turku and many of the illustration in this thesis were created with BioRender.

Firstly, I would like to thank my supervisor and the principal investigator of the Somatic hypermutation research group, adjunct professor Jukka Alinikula for taking me as part of his group and making this study possible. I would also like to thank him for being a supportive and easily approachable supervisor and being enthusiastic throughout the whole project.

I am also thankful for the rest of the group members Anni Soikkeli, Alina Tarsalainen and Ann Sofie Wierda for helping me a lot in laboratory with the practical part of my thesis and for creating a fun and relaxed working environment.

Lastly, I would also like to thank my family and friends for being supportive throughout my studies and the thesis process.

6. List of abbreviations

AID	activation-induced cytidine deaminase
BER	base excision repair
Bsr	blasticidin resistance gene
CSR	class switch recombination
DIVAC	diversification activator
FACS	fluorescence-activated cell sorting
G4	G-quadruplex
GC	germinal centre
HSC	haematopoietic stem cell
HTS	hypermutation targeting sequence
Ig	immunoglobulin
IgH	immunoglobulin heavy chain
IgL	immunoglobulin light chain
IgS	immunoglobulin switch region
IgV	immunoglobulin variable region
IRES	internal ribosome entry site
MMR	mismatch repair
MZ	marginal zone
PCNA	proliferating cell nuclear antigen
RSV	Rous sarcoma virus
SHM	somatic hypermutation
ssDNA	single-stranded DNA
TSS	transcription start site
UNG	uracil-DNA-glycosylase

7. References

Alinikula, J., Y. Maman, A. Tarsalainen, F. Meng, M. Kyläniemi, A. Soikkeli, P. Budzynska, J.J. McDonald, F. Senigl, F. Alt, and D.G. Schatz. Manuscript. Immunoglobulin enhancers modulate RNA polymerase progression through a somatic hypermutation target gene.

Amato, J., A. Pagano, D. Capasso, S. Di Gaetano, M. Giustiniano, E. Novellino, A. Randazzo, and B. Pagano. 2018. Targeting the BCL2 Gene Promoter G-Quadruplex with a New Class of Furopyridazinone-Based Molecules. *ChemMedChem.* 13:406-410. doi: 10.1002/cmdc.201700749.

Arakawa, H., J. Hauschild, and J. Buerstedde. 2002. Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. *Science.* 295. doi: 10.1126/science.1067308.

Asamitsu, S., S. Obata, Z. Yu, T. Bando, and H. Sugiyama. 2019. Recent Progress of Targeted G-Quadruplex-Preferred Ligands Toward Cancer Therapy. *Molecules.* 24. doi: 10.3390/molecules24030429.

Balasubramanian, S., L.H. Hurley, and S. Neidle. 2011. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat Rev Drug Discov.* 10:261-275. doi: 10.1038/nrd3428.

Bedrat, A., L. Lacroix, and J. Mergny. 2016. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Research.* 44:1746-1759. doi: 10.1093/nar/gkw006.

Biffi, G., D. Tannahill, J. Miller, W.J. Howat, and S. Balasubramanian. 2014. Elevated Levels of G-Quadruplex Formation in Human Stomach and Liver Cancer Tissues. *PloS One.* 9:e102711. doi: 10.1371/journal.pone.0102711.

Blagodatski, A., V. Batrak, S. Schmidl, U. Schoetz, R.B. Caldwell, H. Arakawa, and J. Buerstedde. 2009. A cis-Acting Diversification Activator Both Necessary and Sufficient for AID-Mediated Hypermutation. *PLoS Genetics.* 5:e1000332. doi: 10.1371/journal.pgen.1000332.

Brázda, V., J. Kolomazník, J. Lýsek, M. Bartas, M. Fojta, J. Šťastný, and J. Mergny. 2019. G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics.* 35:3493-3495. doi: 10.1093/bioinformatics/btz087.

Briney, B.S., J.R. Willis, and J. Crowe J E. 2012. Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes and Immunity.* 13:523-529. doi: 10.1038/gene.2012.28.

Broxson, C., J. Beckett, and S. Tornaletti. 2011. Transcription arrest by a G quadruplex forming-trinucleotide repeat sequence from the human c-myb gene. *Biochemistry.* 50:4162-4172.

Buerstedde, J., J. Alinikula, H. Arakawa, J.J. McDonald, and D.G. Schatz. 2014. Targeting Of Somatic Hypermutation By immunoglobulin Enhancer And Enhancer-Like Sequences. *PLoS Biology.* 12:e1001831. doi: 10.1371/journal.pbio.1001831.

Burger, A.M., F. Dai, C.M. Schultes, A.P. Reszka, M.J. Moore, J.A. Double, and S. Neidle. 2005. The G-Quadruplex-Interactive Molecule BRACO-19 Inhibits Tumor Growth, Consistent with Telomere Targeting and Interference with Telomerase Function. *Cancer Research.* 65:1489-1496. doi: 10.1158/0008-5472.CAN-04-2910.

Cerutti, A., M. Cols, and I. Puga. 2013. Marginal zone B cells: virtues of innate-like antibody-producing lymphocytes. *Nature Reviews. Immunology.* 13:118-132. doi: 10.1038/nri3383.

Chambers, V.S., G. Marsico, J.M. Boutell, M. Di Antonio, G.P. Smith, and S. Balasubramanian. 2015. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature Biotechnology.* 33:877-881. doi: 10.1038/nbt.3295.

Cortizas, E.M., A. Zahn, S. Safavi, J.A. Reed, F. Vega, J.M. Di Noia, and R.E. Verdun. 2016. UNG protects B cells from AID-induced telomere loss. *Journal of Experimental Medicine.* 213:2459–2472. doi: 10.1084/jem.20160635.

De Magis, A., S.G. Manzo, M. Russo, J. Marinello, R. Morigi, O. Sordet, and G. Capranico. 2019. DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proceedings of the National Academy of Sciences - PNAS.* 116:816-825. doi: 10.1073/pnas.1810409116.

Dhapola, P., and S. Chowdhury. 2016. QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Research.* 44:W277-W283. doi: 10.1093/nar/gkw425.

Di Noia, J.M., and M.S. Neuberger. 2007. Molecular Mechanisms of Antibody Somatic Hypermutation. *Annual Review of Biochemistry.* 76:1-22. doi: 10.1146/annurev.biochem.76.061705.090740.

Dorsett, Y., D.F. Robbiani, M. Jankovic, B. Reina-San-Martin, T.R. Eisenreich, and M.C. Nussenzweig. 2007. A role for AID in chromosome translocations between c-myc and the IgH variable region. *The Journal of Experimental Medicine.* 204:2225-2232. doi: 10.1084/jem.20070884.

Drygin, D., A. Siddiqui-Jain, S. O'Brien, M. Schwaebe, A. Lin, J. Bliesath, C.B. Ho, C. Proffitt, K. Trent, J.P. Whitten, J.K.C. Lim, D. Von Hoff, K. Anderes, and W.G. Rice. 2009. Anticancer Activity of CX-3543: A Direct Inhibitor of rRNA Biogenesis. *Cancer Research.* 69:7653-7661. doi: 10.1158/0008-5472.CAN-09-1304.

Dunnick, W., G.Z. Hertz, L. Scappino, and C. Gritzmacher. 1993. DNA sequences at immunoglobulin switch region recombination sites. *Nucleic Acids Research.* 21:365-372. doi: 10.1093/nar/21.3.365.

Duquette, M.L., P. Handa, J.A. Vincent, A.F. Taylor, and N. Maizels. 2004. Intracellular transcription of G-rich DNAs induces formation of G-loops, novel

structures containing G4 DNA. *Genes & Development.* 18:1618-1629. doi: 10.1101/gad.1200804.

Duquette, M.L., M.D. Huber, and N. Maizels. 2007. G-Rich Proto-Oncogenes Are Targeted for Genomic Instability in B-Cell Lymphomas. *Cancer Research.* 67:2586-2594. doi: 10.1158/0008-5472.CAN-06-2419.

Duquette, M.L., P. Pham, M.F. Goodman, and N. Maizels. 2005. AID binds to transcription-induced structures in c-MYC that map to regions associated with translocation and hypermutation. *Oncogene.* 24:5791-5798. doi: 10.1038/sj.onc.1208746.

Fan, L., T. Bi, L. Wang, and W. Xiao. 2020. DNA-damage tolerance through PCNA ubiquitination and sumoylation. *Biochemical Journal.* 477:2655-2677. doi: 10.1042/bcj20190579.

Feng, Y., N. Seija, J.M. D:i Noia, and A. Martin. 2020. AID in Antibody Diversification: There and Back Again. *Trends in Immunology.* 41:586-600.

Giudicelli, V., D. Chaume, and M. Lefranc. 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Research.* 33:D256-D261. doi: 10.1093/nar/gki010.

Hänsel-Hertsch, R., M. Di Antonio, and S. Balasubramanian. 2017. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nature Reviews Molecular Cell Biology.* 18:279-284. doi: 10.1038/nrm.2017.3.

Hu, M., Y. Wang, Z. Yu, L. Hu, T. Ou, S. Chen, Z. Huang, and J. Tan. 2018. Discovery of a New Four-Leaf Clover-Like Ligand as a Potent c-MYC Transcription Inhibitor Specifically Targeting the Promoter G-Quadruplex. *J. Med. Chem.* 61:2447-2459. doi: 10.1021/acs.jmedchem.7b01697.

Hurley, L.H., R.T. Wheelhouse, D. Sun, S.M. Kerwin, M. Salazar, O.Y. Fedoroff, F.X. Han, H. Han, E. Izbicka, and D.D. Von Hoff. 2000. G-quadruplexes as targets for drug design. *Pharmacology and Therapeutics.* 85:141-158. doi: 10.1016/S0163-7258(99)00068-6.

Kendrick, S., A. Muranyi, V. Gokhale, L.H. Hurley, and L.M. Rimsza. 2017. Simultaneous Drug Targeting of the Promoter MYC G-Quadruplex and BCL2 i-Motif in Diffuse Large B-Cell Lymphoma Delays Tumor Growth. *J. Med. Chem.* 60:6587-6597. doi: 10.1021/acs.jmedchem.7b00298.

Kikin, O., L. D'Antonio, and P.S. Bagga. 2006. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Research.* 34:W676-W682. doi: 10.1093/nar/gkl253.

Kim, M., H. Vankayalapati, K. Shin-ya, K. Wierzba, and L.H. Hurley. 2002. Telomestatin, a Potent Telomerase Inhibitor That Interacts Quite Specifically with the Human Telomeric Intramolecular G-Quadruplex. *Journal of the American Chemical Society.* 124:2098-2099. doi: 10.1021/ja017308q.

Kohler, K.M., J.J. Mcdonald, J.L. Duke, H. Arakawa, S. Tan, S.H. Kleinstein, J. Buerstedde, and D.G. Schatz. 2012. Identification of Core DNA Elements that Target Somatic Hypermutation1. *Journal of Immunology (Baltimore, Md. : 1950).* 189:5314-5326. doi: 10.4049/jimmunol.1202082.

Krijger, P.H.L., P. Langerak, van den Berk, Paul C M, and H. Jacobs. 2009. Dependence of nucleotide substitutions on Ung2, Msh2, and PCNA-Ub during somatic hypermutation. *The Journal of Experimental Medicine.* 206:2603-2611. doi: 10.1084/jem.20091707.

Krijger, P.H.L., A. Tsaalbi-Shtylik, N. Wit, den Berk, Paul Cornelius Maria, N. Wind, and H. Jacobs. 2013. Rev1 is essential in generating G to C transversions downstream of the Ung2 pathway but not the Msh2+Ung2 hybrid pathway. *European Journal of Immunology.* 43:2765-2770. doi: 10.1002/eji.201243191.

Krijger, P.H.L., van den Berk, Paul C. M, N. Wit, P. Langerak, J.G. Jansen, C. Reynaud, N. de Wind, and H. Jacobs. 2011. PCNA ubiquitination-independent activation of polymerase η during somatic hypermutation and DNA damage tolerance. *DNA Repair.* 10:1051-1059. doi: 10.1016/j.dnarep.2011.08.005.

Lavrado, J., H. Brito, P.M. Borralho, S.A. Ohnmacht, N. Kim, C. Leitão, S. Pisco, M. Gunaratnam, C.M.P. Rodrigues, R. Moreira, S. Neidle, and A. Paulo. 2015. KRAS oncogene repression in colon cancer cell lines by G-quadruplex binding indolo[3,2-c]quinolines. *Scientific Reports.* 5:9696. doi: 10.1038/srep09696.

Lebien, T.W., and T.F. Tedder. 2008. B lymphocytes: how they develop and function. *Blood.* 112:1570-1580. doi: 10.1182/blood2008-02-078071.

Liu, M., and D.G. Schatz. 2009. Balancing AID and DNA repair during somatic hypermutation. *Trends in Immunology.* 30:173-181. doi: 10.1016/j.it.2009.01.007.

Mahdaviani, S.A., A. Hirbod-Mobarakeh, N. Wang, A. Aghamohammadi, L. Hammarström, M.R. Masjedi, Q. Pan-Hammarström, and N. Rezaei. 2014. Novel mutation of the activation-induced cytidine deaminase gene in a Tajik family: special review on hyper-immunoglobulin M syndrome. *Expert Review of Clinical Immunology.* 8:539-546. doi: 10.1586/eci.12.46.

McBride, K.M., A. Gazumyan, E.M. Woo, T.A. Schwickert, B.T. Chait, and M.C. Nussenzweig. 2008. Regulation of class switch recombination and somatic mutation by AID phosphorylation. *The Journal of Experimental Medicine.* 205:2585-2594. doi: 10.1084/jem.20081319.

Melchers, F. 2015. Checkpoints that control B cell development. *The Journal of Clinical Investigation.* 125:2203-2210. doi: 10.1172/JCI78083.

Mergny, J., and C. Hélene. 1998. G-quadruplex DNA: A target for drug design. *Nature Medicine.* 4:1366-1367.

Muramatsu, M., H. Nagaoka, Y. Sakakibara, R. Shinkura, S. Ito, N.A. Begum, H. Hijikata, K. Kinoshita, and T. Honjo. 2004. Separate domains of AID are required for

somatic hypermutation and class-switch recombination. *Nature Immunology.* 5:707-712. doi: 10.1038/ni1086.

Nayun Kim. 2019. The Interplay between G-quadruplex and Transcription. *Current Medicinal Chemistry.* 26:2898-2917. doi: 10.2174/0929867325666171229132619.

Neaves, K.J., J.L. Huppert, R.M. Henderson, and J.M. Edwardson. 2009. Direct visualization of G-quadruplexes in DNA using atomic force microscopy. *Nucleic Acids Research.* 37:6269-6275. doi: 10.1093/nar/gkp679.

Neuberger, M.S., and C. Milstein. 1995. Somatic Hypermutation. *Current Opinion in Immunology.* 7:248-254.

Nguyen, T.Q.N., K.W. Lim, and A.T. Phan. 2020. Folding Kinetics of G-Quadruplexes: Duplex Stem Loops Drive and Accelerate G-Quadruplex Folding. *The Journal of Physical Chemistry. B.* 124:5122-5130. doi: 10.1021/acs.jpcb.0c02548.

Nilsen, H., M. Otterlei, T. Haug, K. Solum, T.A. Nagelhus, F. Skorpen, and H.E. Krokan. 1997. Nuclear and mitochondrial uracil-DNA glycosylases are generated by alternative splicing and transcription from different positions in the UNG gene. *Nucleic Acids Research.* 25:750-755. doi: 10.1093/nar/25.4.750.

Odegard, V.H., and D.G. Schatz. 2006. Targeting of somatic hypermutation. *Nature Reviews. Immunology.* 6:573-583. doi: 10.1038/nri1896.

Pavri, R. 2017. R Loops in the Regulation of Antibody Gene Diversification. *Genes.* 8:154. doi: 10.3390/genes8060154.

Pieper, K., MSc, B. Grimbacher MD, and H. Eibel PhD. 2013. B-cell biology and development. *Journal of Allergy and Clinical Immunology.* 131:959-971. doi: 10.1016/j.jaci.2013.01.046.

Pilzecker, B., and H. Jacobs. 2019. Mutating for Good: DNA Damage Responses During Somatic Hypermutation. *Frontiers in Immunology.* 10. doi: 10.3389/fimmu.2019.00438.

Qiao, Q., L. Wang, F. Meng, J.K. Hwang, F.W. Alt, and H. Wu. 2017. AID Recognizes Structured DNA for Class Switch Recombination. *Molecular Cell.* 67:361-373.e4. doi: 10.1016/j.molcel.2017.06.034.

Revy, P., T. Muto, Y. Levy, F. Dé, R. Geissmann, A. Plebani, O. Sanal, N. Catalan, M. Forveille, R. Mi Dufourcq-Lagelouse, A. Gennery, I. Tezcan, F. Ersoy, H. Kayserili, A.G. Ugazio, N. Brousse, M. Muramatsu, L.D. Notarangelo, K. Kinoshita, T. Honjo, A. Fischer, and A. Durandy. 2000. Activation-Induced Cytidine Deaminase (AID) Deficiency Causes the Autosomal Recessive Form of the Hyper-IgM Syndrome (HIGM2) . *Cell.* 102:565-575. doi: 10.1016/s0092-8674(00)00079-9.

Safavi, S., A. Larouche, A. Zahn, A. Patenaude, D. Domanska, K. Dionne, T. Rognes, F. Dingler, S. Kang, Y. Liu, N. Johnson, J. Hébert, R.E. Verdun, C.A. Rada, F. Vega, H. Nilsen, and J.M. Di Noia. 2020. The uracil-DNA glycosylase UNG protects the fitness

of normal and cancer B cells expressing AID. *NAR Cancer.* 2:zcaa019. doi: 10.1093/narcan/zcaa019.

Sagaert, X., B. Sprangers, and C. De Wolf-Peeters. 2007. The dynamics of the B follicle: understanding the normal counterpart of B-cell-derived malignancies. *Leukemia.* 21:1378-1386. doi: 10.1038/sj.leu.2404737.

Saribasak, H., N.N. Saribasak, F.M. Ipek, J.W. Ellwart, H. Arakawa, and J. Buerstedde. 2006. Uracil DNA Glycosylase Disruption Blocks Ig Gene Conversion and Induces Transition Mutations. *The Journal of Immunology (1950).* 176:365-371. doi: 10.4049/jimmunol.176.1.365.

Saribasak, H., R. Woodgate, W. Yang, R.L. McClure, D.M. Wilson, P.J. Gearhart, H.S. Gramlich, S.A. Martomo, A. Vaisman, D.G. Schatz, and R.W. Maul. 2011. Uracil residues dependent on the deaminase AID in immunoglobulin gene variable and switch regions. *Nature Immunology.* 12:70-76. doi: 10.1038/ni.1970.

Schanz, S., D. Castor, F. Fischer, and J. Jiricny. 2009. Interference of mismatch and base excision repair during the processing of adjacent U/G mispairs may play a key role in somatic hypermutation. *Proc Natl Acad Sci U S A.* 106:5593–5598. doi: 10.1073/pnas.0901726106.

Sen, D., and W. Gilbert. 1990. A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature (London).* 344:410-414. doi: 10.1038/344410a0.

Sohail, A., J. Klapacz, M. Samaranayake, A. Ullah, and A.S. Bhagwat. 2003. Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. *Nucleic Acids Research.* 31:2990-2994. doi: 10.1093/nar/gkg464.

Spiegel, J., S. Adhikari, and S. Balasubramanian. 2020. The Structure and Function of DNA G-Quadruplexes. *Trends in Chemistry.* 2:123-136. doi: 10.1016/j.trechm.2019.07.002.

Summers, P.A., B.W. Lewis, J. Gonzalez-Garcia, R.M. Porreca, A.H.M. Lim, P. Cadinu, N. Martin-Pintado, D.J. Mann, J.B. Edel, J.B. Vannier, M.K. Kuimova, and R. Vilar. 2021. Visualising G-quadruplex DNA dynamics in live cells by fluorescence lifetime imaging microscopy. *Nat Commun.* 12. doi: 10.1038/s41467-020-20414-7.

Sun, D., B. Thompson, B.E. Cathers, M. Salazar, S.M. Kerwin, J.O. Trent, T.C. Jenkins, S. Neidle, and L.H. Hurley. 1997. Inhibition of Human Telomerase by a G-Quadruplex-Interactive Compound. *Journal of Medicinal Chemistry.* 40:2113-2116. doi: 10.1021/jm970199z.

Tauchi, T., K. Shin-Ya, G. Sashida, M. Sumi, S. Okabe, J.H. Ohyashiki, and K. Ohyashiki. 2006. Telomerase inhibition with a novel G-quadruplex-interactive agent, telomestatin: In vitro and in vivo studies in acute leukemia. *Oncogene.* 25:5719-5725. doi: 10.1038/sj.onc.1209577.

Wilson, T.M., A. Vaisman, S.A. Martomo, P. Sullivan, L. Lan, F. Hanaoka, A. Yasui, R. Woodgate, and P.J. Gearhart. 2005. MSH2–MSH6 stimulates DNA polymerase η,

suggesting a role for A:T mutations in antibody genes. *The Journal of Experimental Medicine.* 201:637-645. doi: 10.1084/jem.20042066.

Winkler, T.H., and I. Mårtensson. 2018. The Role of the Pre-B Cell Receptor in B Cell Development, Repertoire Selection, and Tolerance. *Frontiers in Immunology.* 9:2423. doi: 10.3389/fimmu.2018.02423.

Xu, H., M. Di Antonio, S. Mckinney, V. Mathew, B. Ho, N.J. O'neil, N.D. Santos, J. Silvester, V. Wei, J. Garcia, F. Kabeer, D. Lai, P. Soriano, J. Banáth, D.S. Chiu, D. Yap, D.D. Le, F.B. Ye, A. Zhang, K. Thu, J. Soong, S. Lin, A.H.C. Tsai, T. Osako, T. Algara, D.N. Saunders, J. Wong, J. Xian, M.B. Bally, J.D. Brenton, G.W. Brown, S.P. Shah, D. Cescon, T.W. Mak, C. Caldas, P.C. Stirling, P. Hieter, S. Balasubramanian, and S. Aparicio. 2017. CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nat Commun.* 8. doi: 10.1038/ncomms14432.

Xu, Y., P. Jenjaroenpun, T. Wongsurawat, S.D. Byrum, V. Shponka, D. Tannahill, E.A. Chavez, S.S. Hung, C. Steidl, S. Balasubramanian, L.M. Rimsza, and S. Kendrick. 2020. Activation-induced cytidine deaminase localizes to G-quadruplex motifs at mutation hotspots in lymphoma. *NAR Cancer.* 2:1-14. doi: 10.1093/narcan/zcaa029.

Yeap, L., J.K. Hwang, Z. Du, R.M. Meyers, F. Meng, A. Jakubauskaitė, M. Liu, V. Mani, D. Neuberg, T.B. Kepler, J.H. Wang, and F.W. Alt. 2015. Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. *Cell (Cambridge).* 163:1124-1137. doi: 10.1016/j.cell.2015.10.042.

Yewdell, W.T., Y. Kim, P. Chowdhury, C.M. Lau, R.M. Smolkin, K.T. Belcheva, K.C. Fernandez, M. Cols, W. Yen, B. Vaidyanathan, D. Angeletti, A.B. McDermott, J.W. Yewdell, J.C. Sun, and J. Chaudhuri. 2020. A Hyper-IgM Syndrome Mutation in Activation-Induced Cytidine Deaminase Disrupts G-Quadruplex Binding and Genome-wide Chromatin Localization. *Immunity (Cambridge, Mass.).* 53:952-970.e11. doi: 10.1016/j.immuni.2020.10.003.

Zeng, X., D.B. Winter, C. Kasmer, K.H. Kraemer, A.R. Lehmann, and P.J. Gearhart. 2001. DNA polymerase η is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. *Nature Immunology.* 2:537-541. doi: 10.1038/88740.

Zheng, S., B. Vuong, B. Vaidyanathan, J. Lin, F. Huang, and J. Chaudhuri. 2015. Non-coding RNA Generated following Lariat Debranching Mediates Targeting of AID to DNA. *Cell (Cambridge).* 161:762-773. doi: 10.1016/j.cell.2015.03.020.

Appendix 1

Complete list of primers used

Complete list of primers used during the project	
Primer name	Primer sequence (5' --> 3')
Deleting G-quadruplexes from GFP2 2-3	
GFP2_T60T63_F	CACACTCGTGACAACCCTGACCTACGGCGTGCAGTGC
GFP2_T60T63_R	GGGTTGTCACGAGTGTGGGCCAGGGCACGGGCAGCTT
GFP2_Y75P76_F	CGCTATCCTGACCACATGAAGCAGCACGACTTCTTCAAG
GFP2_Y75P76_R	GTGGTCAGGATAGCGGCTGAAGCACTGCACGCCGTA
GFP2_T187P188_F	AACACACCTATCGGCGACGGACCTGTGCTGCTGCCCGACAACCACTAC
GFP2_T187P188_R	GCCGATAGGTGTGTTCTGCTGATAGTGGTCGGCGAGCTGCAC
GFP2_D211P212_F	AAAGATCCTAACGAGAAGCGCGATCACATGGTCCTG
GFP2_D211P212_R	CTCGTTAGGATCTTTGCTCAGGGCGGACTGGGT
Cloning a strong G4 in the middle of HTS in GFP4 Igλ	
GFP4_G4_F	TTTCGCGCAGCGTTTGGGTGGGTTCAGCCTTGGCAGCAACGA
GFP4_G4_R	AAACGCTGCGCGAAACCCACCCAACCATTGCTGCCACTGTAACC
Amplifying T2A + DsRED insert from pX458-DsRed vector	
T2A_DsRED_F	GGCAGTGGAGAGGGCAGA
T2A_DsRED_R	CTACAGGAACAGGTGGTGGC
Cloning GFP2+DsRed reporters	
GFP2_T2A_G4_F	TTTCGCGCAGCGTTTGGGTGGGTTGGCAGTGGAGAGGGCAGA
GFP2_BB_G4_R	AAACGCTGCGCGAAACCCACCCAAGGACTTGTACAGCTCGTCCAT
GFP2_BB_DsRED_F	CACCTGTTCCTGTAGCGGGAGATCACCCCTCT
GFP2_BB_T2A_R	GCCCTCTCCACTGCCGGACTTGTACAGCTCGTCCAT
Primers for sequencing	
JA568	CCCAACGAGAAGCGCGATCA
JA705	CAAGGGCGAGGAGCTGTTCA
JA698	CTCGATACAATAAACGCCATTTGACCATT
JA699	GGGTGATCTCCCGCTAGGACTTGTA
JA701	TAGGACTTGTACAGCTCGTCCAT
0986	CTACAGGAACAGGTGGTGGCG

Appendix 2

Polymerase chain reaction setups used

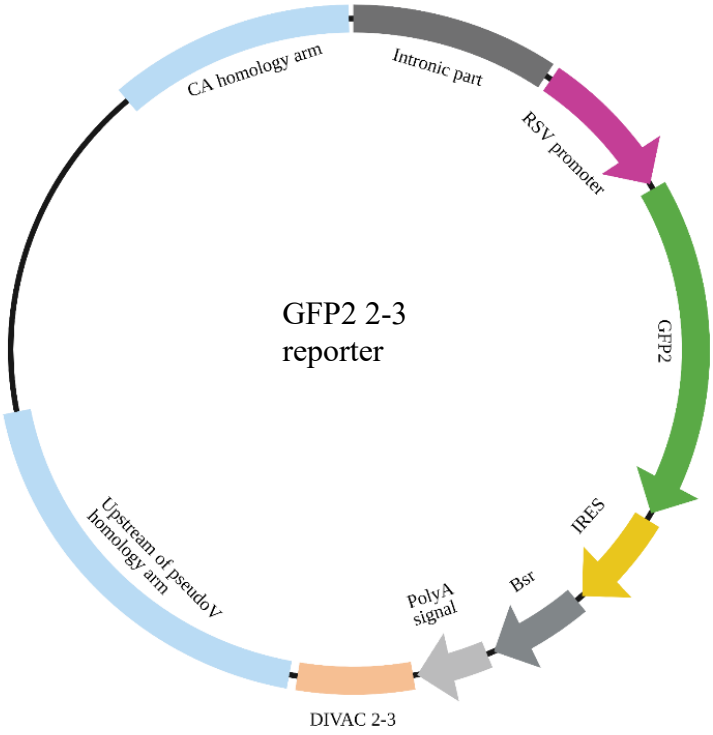
Long DNA fragments		
Temperature (°C)	Time	Cycles
98	30 s	1
98	5 s	32
68/72	20 s	
72	14 min	
72	2 min	1
10	hold	∞

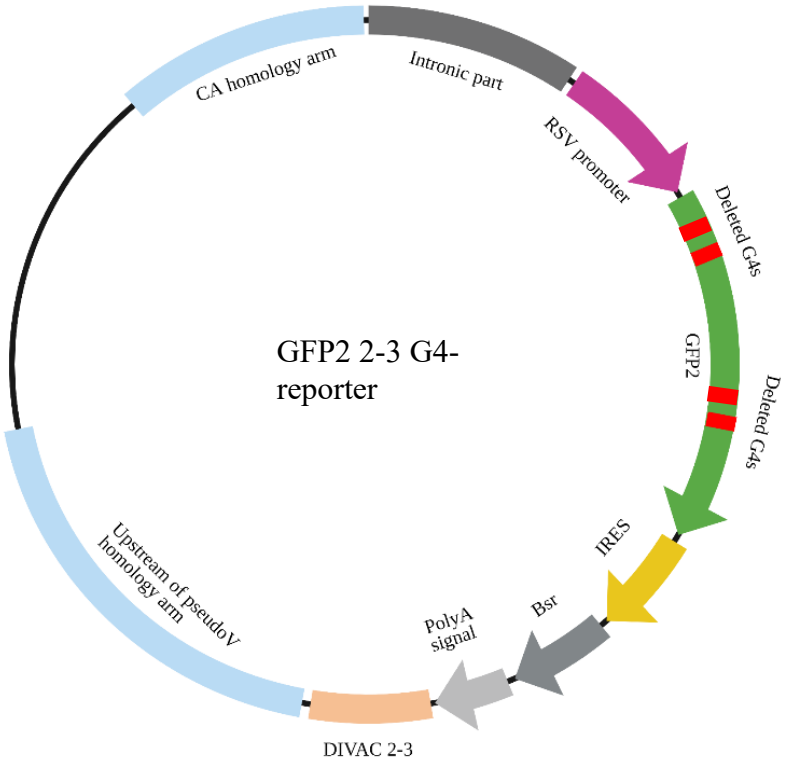
Short DNA fragments		
Temperature (°C)	Time	Cycles
98	30 s	1
98	5 s	32
68/72	20 s	
72	15 s	
72	2 min	1
10	hold	∞

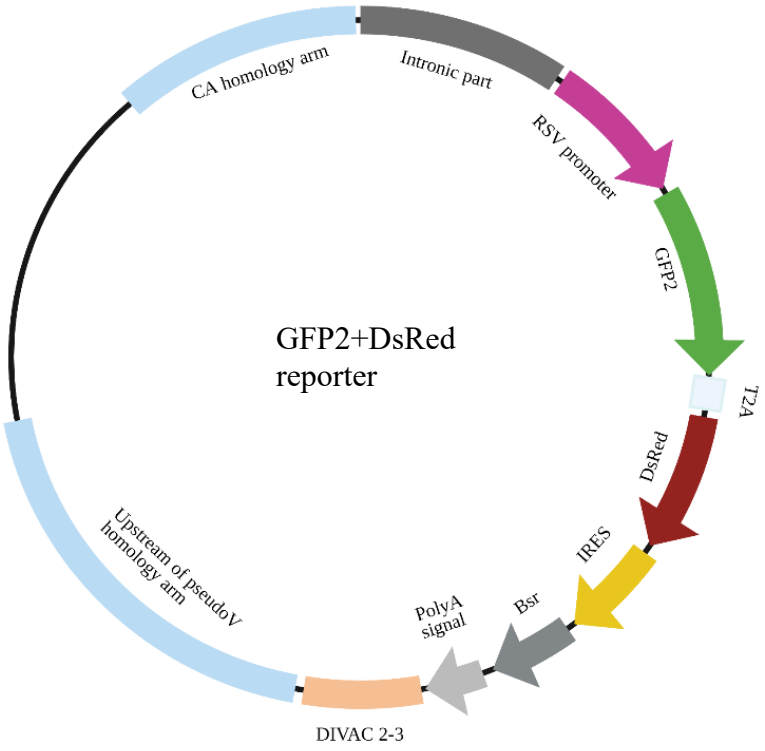
DT40 genomic DNA		
Temperature (°C)	Time	Cycles
98	30 s	1
98	5 s	32
68	20 s	
72	45 s	
72	2 min	1
10	hold	∞

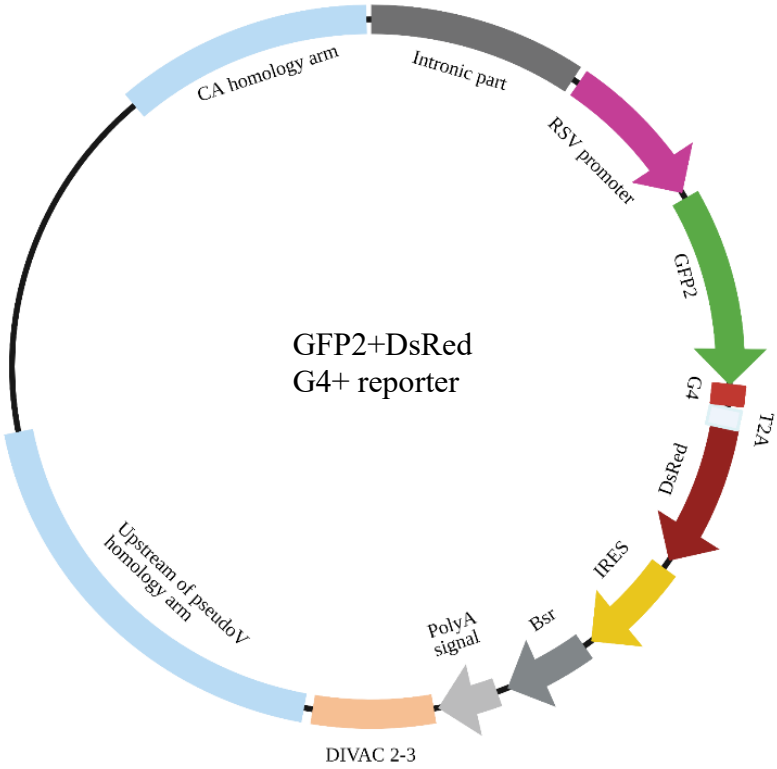
Appendix 3

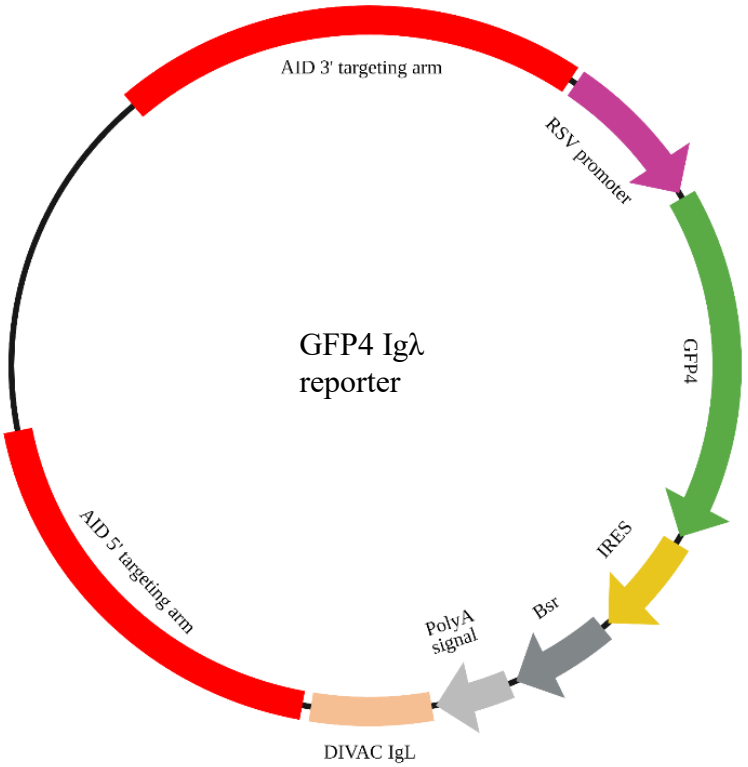
Complete plasmids containing the reporters

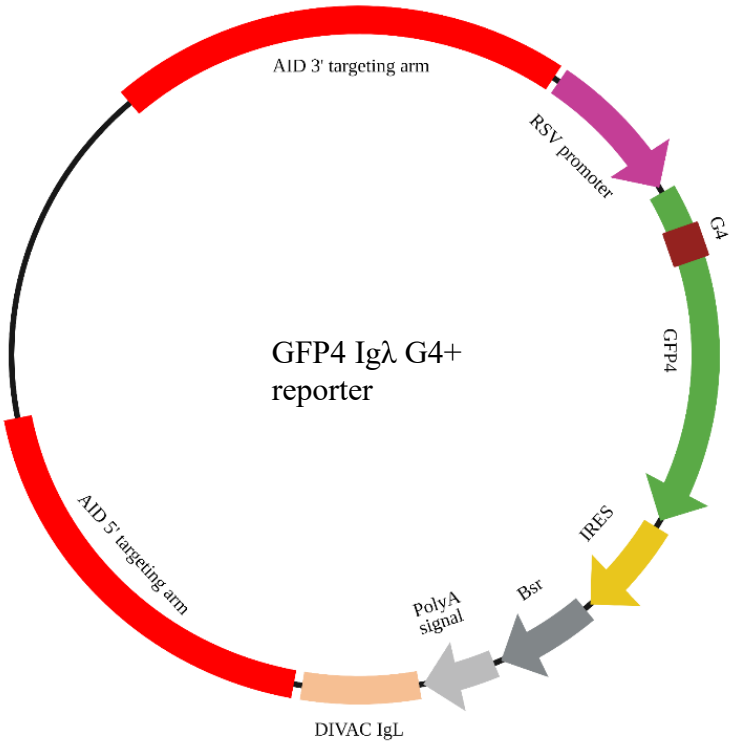





CA homology arm
Intronic part
RSV promoter
GFP2
T2A
DsRed
IRES
Bsr
PolyA signal
Upstream of pseudoV homology arm
DIVAC 2-3
GFP2+DsRed reporter


CA homology arm
Intronic part
RSV promoter
GFP2
G4
T2A
DsRed
IRES
Bsr
PolyA signal
Upstream of pseudoV homology arm
DIVAC 2-3
GFP2+DsRed G4+ reporter


AID 3' targeting arm
RSV promoter
GFP4
IRES
Bsr
PolyA signal
DIVAC IgL
AID 5' targeting arm
GFP4 Igλ reporter


AID 3' targeting arm
RSV promoter
G4
GFP4
IRES
Bsr
PolyA signal
DIVAC IgL
AID 5' targeting arm
GFP4 Igλ G4+ reporter