UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Medicina Molecolare

CORSO DI DOTTORATO DI RICERCA IN: MEDICINA MOLECOLARE

CURRICULUM : BIOMEDICINE

CICLO XXXV

Presence and role of DNA G-quadruplex structures in the pathogenesis of XDP

Tesi redatta con il contributo finanziario di XDP Collaborative Center

Coordinatore: Ch.mo Prof. Riccardo Manganelli

Supervisore: Ch.ma Prof. Sara Richter

Dottoranda : Giulia Nicoletto

“The most certain way to succeed is always to try just one more time”

Thomas A. Edison

“Disorder”

Joy Division

Index

1. Abstract

XDP is a genetic movement disorder that human males can develop around 40 years of age. Genetic alterations located in the X chromosome are at the base of this disorder and the same haplotype is shared by all probands.[1] An SVA retrotransposon antisense insertion within the intron 32 of *TAF1* gene, which causes lowered mRNA *TAF1* levels, is among the XDP characteristic mutations and is proposed to be crucial for the development of the disease. In fact, removal of the SVA brings *TAF1* levels up to normal. Moreover, a truncated form of *TAF1* comprising an intron which is retained, and truncated exactly at the site of SVA insertion, is present in XDP patients. When the SVA is removed, the truncated *TAF1* levels drop down to healthy cell levels. It is not known how XDP SVA impairs *TAF1* gene transcription. Our hypothesis was that G4s could fold within the SVA retrotransposon slowing down RNA polymerase. We started our investigation with *in vitro* experiments. We first identified putative G4 forming sequences with a G4 predicting tool, and we characterized the highest score sequences by circular dichroism, DMS footprinting and TaqPol STOP assay. Every tested sequence was proved to fold into highly stable, parallel topology G4s. We also studied the interaction of those sequences with different G4 ligands such as BRACO-19 and Quarfloxin. To assess if G4s can form in the double-stranded SVA sequence we set up a PCR STOP assay, in which we amplified the SVA from genome DNA extracted from XDP patient cells and healthy controls. In G4-inducing conditions, SVA amplification was totally impaired, suggesting that G4s were folded and able to block enzyme activity. To identify the SVA domains mainly responsible for this effect, we designed specific primers amplifying SVA domain regions and we proved that the VNTR and Hex domain are the domains that lead to amplification stop. This result was in complete accordance with the initial G4 prediction. To find out if those G4s were present within the SVA also in cells, we set up a BG4-ChIP-seq protocol on a difficult cell line, such as human fibroblasts, that displays 4 time less G4s than the model cell line K-562, that was one of the cell lines used in the development of the published protocol. We found a different G4 landscape between XDP affected cells and healthy control cells, that need to be further investigated. We did not reach a unique mapping alignment in the XDP SVA region, even when performing the more recent and efficient CUT&Tag protocol, that has very little background noise. However, we found coverage for every SVA family, indicating that SVAs (even those different from the SVA present in XDP) display folded G4s, a notion that has never been reported before. We finally proved that the hexanucleotide repeat domain $(CCCTCT)_n$ of the XDP SVA[2] displays folded G4s by BG4-ChIP-qPCR, designing specific Taqman

primers that amplify the last part of the hexameric repeat. To assess the impact of the SVA G4s on *TAF1* transcription, we treated XDP affected and healthy controls with increasing concentrations of G4 ligands for 24 hours. Strikingly, in SVA carrier cells there was an increase in the transcription of the first exons of *TAF1* but not of the other exons. Our hypothesis is that when SVA G4s are stabilized by G4 ligands, the RNA Polymerase that is transcribing *TAF1* stalls on the SVA until G4s are resolved. As a result, premature termination occurs, thus leading to increased truncated *TAF1* with intron 32 retention, less full length transcripts that make the cell induce even more *TAF1* transcription as a negative loop. For this reason, it is important to find a way to destabilize the SVA G4s. Our first attempt was using a new small molecule, PhPc, that was shown to destabilize G4s *in vitro*. Unfortunately, this compound was not able to destabilize the SVA G4s in cells. It did bind to them though, in such a way that led to an effect similar to the stabilizing G4 ligands. In conclusion we proved that G4s can fold within SVA *in vitro* and in cells, and that their being folded has an impact on the transcription of the genes in which SVA is inserted.

2. Introduction

2.1 XDP disease

X-linked Dystonia Parkinsonism (XDP) is a genetic neurodegenerative movement disorder that was discovered in 1975 in the Philippines in Panay island[3](Fig. 1). The estimated prevalence in the Philippines is 1/322,000 and in the Province of Capiz it is at its highest with a prevalence of 1/4,000 in the male population.

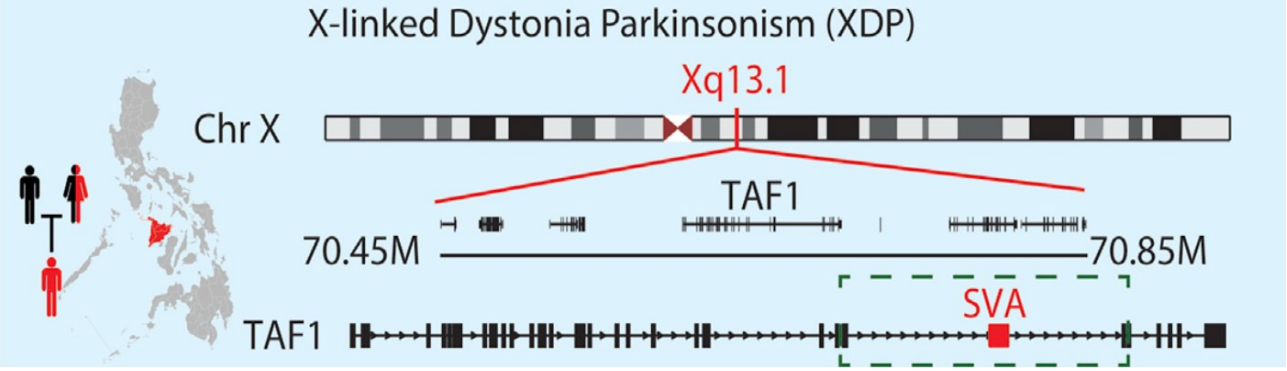


Fig. 1: Genome assembly narrows the causal XDP locus to TAF1. [1]

It is an invalidating disease with an adult age of onset and it is characterized by a first dystonic phase that slightly evolves into parkinsonism[4]. This progression is correlated with the degeneration of dopaminergic neurons in specific brain areas[5]. There is no cure for this disease and current treatments only focus on symptoms, like for other movement disorders[6]. Unlike other neurological diseases, the pathogenesis of which is not known, the cause of XDP is genetic. In fact, all probands share the same genetic aberrant modifications. This typical haplotype is located in the Xq13.1 segment in the X chromosome (Fig. 2).

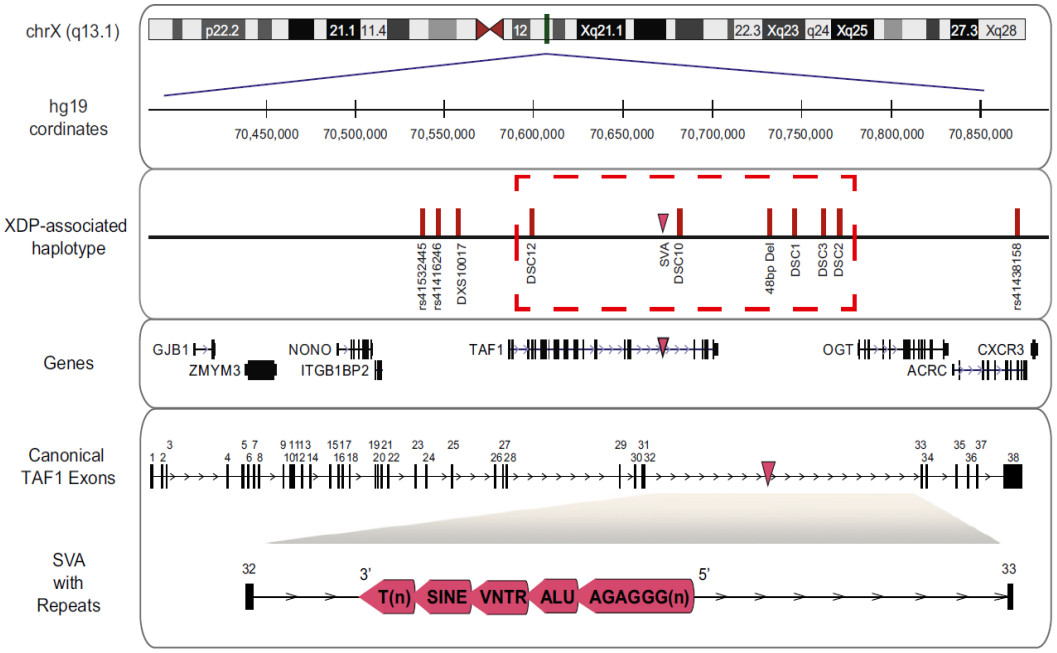


Fig. 2:(Upper) The genomic segment previously associated with XDP on chromosome Xq13.1 with hg19 coordinates, seven known XDP-specific variants that comprise the disease haplotype (boxed region), and flanking markers used to narrow the region. Haplotype variants consist of five single-nucleotide substitutions annotated as DSC-1, 2, 3, 10, and 12; a 48-bp deletion (48 bp Del), and a SVA-type retrotransposon insertion. Eight genes are shown within the broader linkage region, including TAF1.(Lower) Canonical exons of TAF1, the relative position of the SVA inserted antisense to TAF1, and the domain structure of the SVA consisting of (5'–3') a hexameric repeat (CCCTCT) of variable length, an Alu-like domain, a VNTR, a SINE domain, and a poly(A) tail.[2]

For this reason, males develop the disease at around 40 years of age, on the contrary, very few females develop the typical symptoms and the reported cases have higher age of onset, around 80 years of age[7]. The XDP haplotype consists of five single nucleotides variants (DSC), one 48-bp deletion, and an SVA retrotransposon antisense insertion in the *TAF1* gene[8]. *TAF1* is a small protein that is a cofactor essential for RNA Pol II activity. It is also reported as a crucial element in neurodevelopment but the molecular mechanisms remain unknown[9]. The SVA insertion is now considered the most crucial mutation. In fact, it was demonstrated that the excision of the SVA from intron 32 of *TAF1* by Crisp-CAS9 technique in XDP-affected cell lines leads to restoration of normal levels of *TAF1* transcript thus producing a rescued phenotype. Besides, in the rescued clones there is no intron retention close to the SVA insertion, a tract that is instead typical of the XDP disease[1] (Fig. 3).

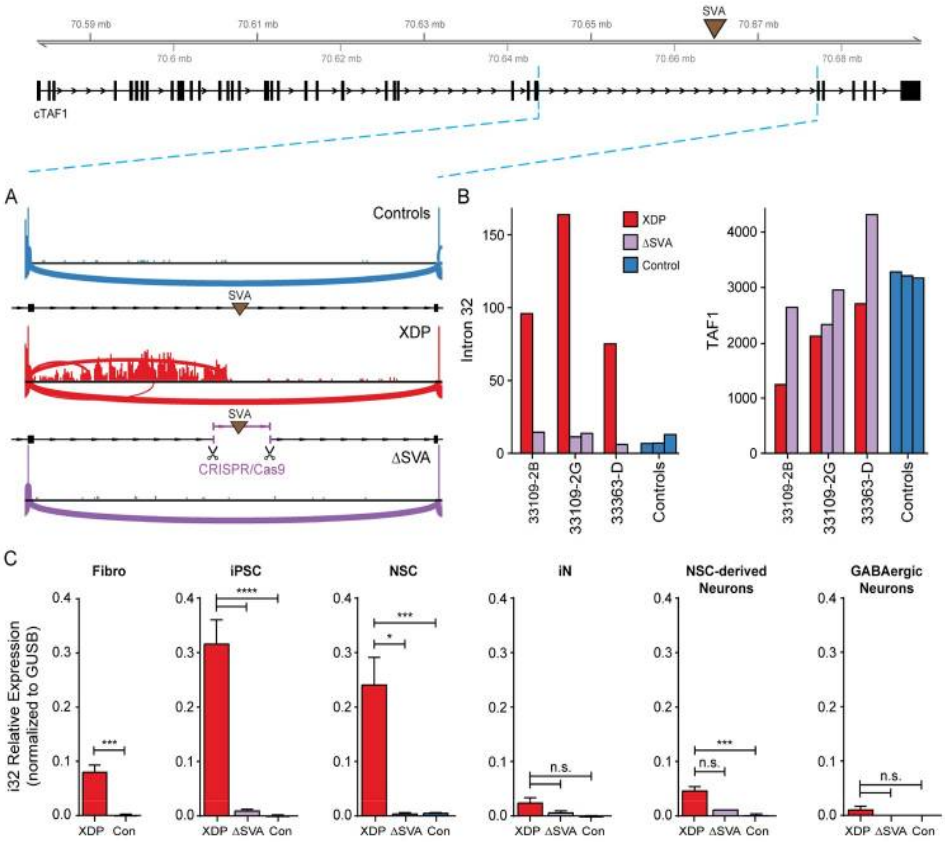


Fig. 3: Excision of the SVA Rescues Aberrant Splicing and Expression in Intron 32 and Expression of TAF1 (A) Sashimi plot depicting coverage and splicing in intron 32 of TAF1 in control, XDP, and SVA-excised (DSVA) proband NSCs. (B) Normalized RNA-seq counts in intron 32 of TAF1 50 to the SVA insertion (left) and TAF1 (right) in proband NSCs, corresponding DSVA clones, and control cells (one clone per individual). (C) Relative expression of intron 32 splice variant in fibroblasts (Fibro), iPSCs, NSCs, iNs, NSC-derived cortical neurons, and GABAergic neurons from XDP, control, and DSVA lines. Graphs represent mean (+ SEM) from clones generated for each cell type. [1]

Focusing on XDP SVA, it belongs to SVA F family which is one of the youngest from an evolutionary point of view. This means that the sequence of those entities did not accumulate many mutations resulting in a highly conservated sequence.[10] Infact the only difference in the SVA sequence among XDP patients is the amplification of the hexameric domain (from 32 to 50 times). This amplification is higher compared to the consensus sequence of SVA F family and positively correlates with the development of more severe symptoms and an earlier development of the disease[2](Fig. 4). More interestingly, amplification of hexameric domain changes in different tissues being more prominent in specific brain areas[11]. Such mosaicism suggests a somatic instability of this amplification that correlates with age onset of the disease, confirming the SVA as the major player in XDP pathology development[12]. Nevertheless it is still not clear how this SVA impairs *TAF1* transcription and why the

effects are more prominent in neurons-like cells[8]. It is more feasible that the toxic effects of SVA insertion are at a transcription level, because if we look at the final protein level TAF1 there in no clear difference between XDP affected and healthy patients[1].

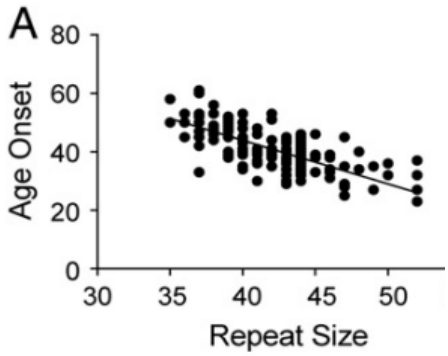


Fig. 4: Length of the hexameric repeat is polymorphic in affected XDP individuals and is inversely correlated with AO based on linear regression analysis. (A) Correlation between repeat length and AO in the entire cohort; n = 140, R2 = 0.507, P = 3.54 × 10−23. [1]

2.2 SINE-VNTR-Alu (SVA) Retrotransposons

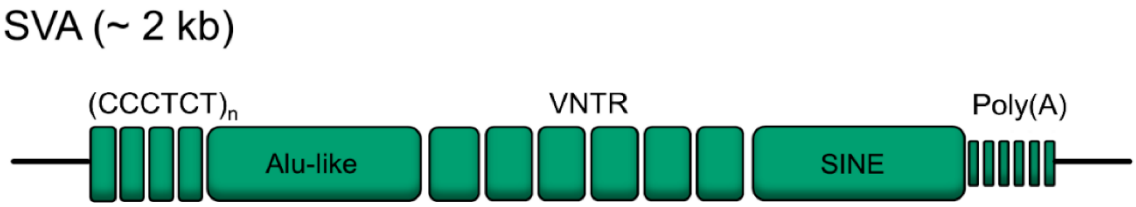


Fig. 5: Canonical SINE-VNTR-Alu (SVA) structure. Canonical SVAs typically contain five distinct regions; a (CCCTCT)n hexamer repeat at the 5' end, an Alu-like domain, a variable number tandem repeat (VNTR), a SINE-derived region, and a poly A tail. [14]

SINE-VNTR-Alu (SVA) are composite non-LTR retrotransposon that are immobilized by LINE-1 protein machinery[13]. 2762 SVA elements were identified in the human genome; they belong to the youngest retrotransposon family, in fact they are found only in primates[14]. SVAs can be divided in 7 subtypes (Fig. 6), named A-F1 in order of age, being SVA-A the oldest and SVA-F1 the youngest. SVA-E, SVA-F and SVA-F1 are found only in humans.[15]

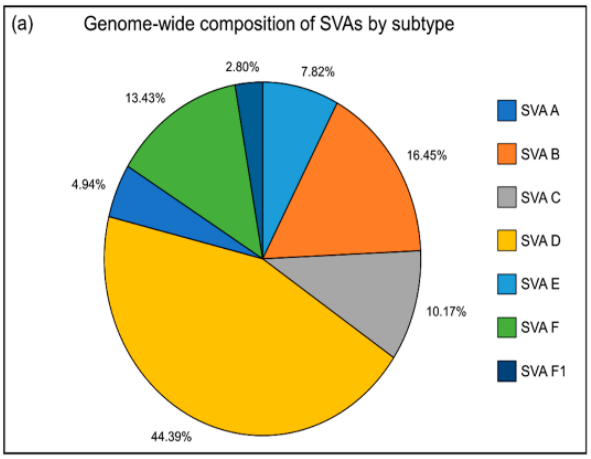


Fig. 6: Composition of SVA subfamilies within 1 Mb zinc finger gene regions compared to the whole genome. (a) Calculating the percentage of each SVA subfamily across the human genome demonstrated that 34.44% of all SVAs were comprised of the evolutionary older A–C subfamilies and are therefore conserved at their respective loci within multiple primate species including humans. The remaining 65.56% is comprised of younger subfamily members D–F1. SVA-D is by far the largest SVA subfamily, comprising 44.39% of all SVA elements in the genome, some of which are human-specific, with others being present in multiple primate species. 21.17% of SVAs (subfamilies E–F1) are entirely human-specific. SVA-B, F, and C each represent over 10% of all SVAs, with a percentage of 16.45, 13.43 and 10.17%, respectively. The remaining subfamilies (SVA-A, E, and F1) make up less than 10% of the total each. [14]

The canonical structure of those entities (Fig. 5) is composed by a simple hexamer repeat of (CCCTCT)n, an Alu-like region of two antisense Alu fragments separated by a region of intervening sequence, one or two variable number tandem repeat (VNTR) regions, a SINE region derived from the 3' LTR of the retroviral HERV-K10 element, and finally a poly-A signal[14].The SVA-F1 subfamily is the only one lacking the hexameric domain. In fact it is substituted by the CpG island containing exon1 of MAST2 gene.[16] With a GC content of around 60%, SVAs have been referred to as "mobile CpG islands"[17]. For this reason, they are considered able to alter the chromatin state of the locus in which they are inserted, thus interfering with transcription processes like other transposable elements do[18]. Indeed, they are not located randomly in the genome but they are found mainly in genes and intergenic regions (Fig.7).

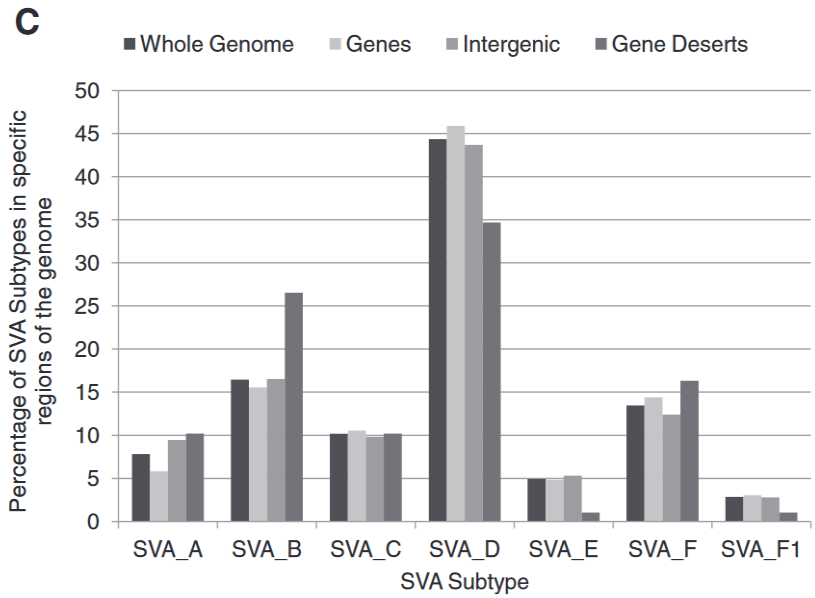


Fig. 7: Distribution of SVAs is associated with genic regions. C) The distribution of SVAs within genes, intergenic regions and gene deserts broken down by subtype and compared to their distribution across the whole human genome. (Genes X2 = 0.71, df = 6, P = 0.99), (Intergenic X2 = 0.47, df = 6, P = 0.99), (Gene deserts X2 = 13.91, df = 6, P < 0.05). [18]

Being very G-rich, SVAs could also adopt G4 structures[19]. Savage et colleagues demonstrated that even if SVAs are only the 0.13% of the human genome, they represent the 2% of putative G4 regions[20]. In particular the hexameric domain is probably the most eligible for G4 formation, in fact if we consider also the size of this domain, the contribution of SVAs to total G4s is not negligible (Fig. 8). In addition, the human-specific SVAs (subtype E, F, F1) have greater potential to adopt the G4 conformation, both in the VNTR and hexameric domain, as they have more GC content compared to the older SVA subtypes (A, B, C, D)[21]. Another interesting detail is that the two SVA G-rich domains are also considered "promoter acting domains"[22], so the possibility

that G4s regulate their expression, as reported for G4s in canonical gene promoters, needs to be taken into consideration.

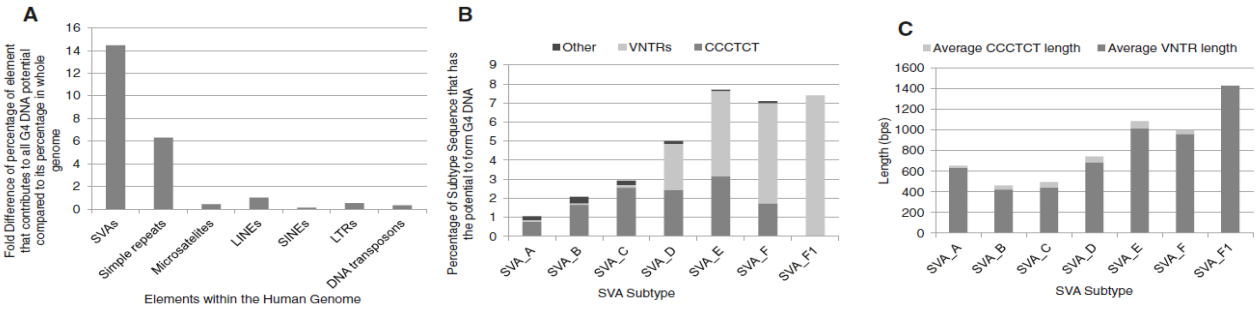


Fig. 8: The primary sequence of SVAs has the potential to form G-quadruplex DNA. A) Potential G4 DNA formation was analysed in silico. The fold difference in the relative contribution of each element to their proportion in the whole human genome was calculated and is displayed. B) The percentage of sequence from each SVA subtype that could potentially form G4 DNA in the human genome according to Quadparser software is shown; it was further sub-divided into the following elements: CCCTCT hexamer repeat, VNTRs and the remainder of the sequence (other). C) Illustrates the relationship between VNTR and hexamer repeat length during evolution of the SVA subtypes. The average lengths are shown in base pairs. [18]

2.3 G-quadruplex structures

G-quadruplexes (G4s) are tetrahedral single-stranded DNA structures that can fold in Guanine-rich regions. The building blocks of G4s are the G-quartets, which are composed of 4 guanines arranged in a plane square and stabilized by Hoogsten hydrogen bonds and monovalent cations such as sodium or potassium. Two or three quartets stack one upon the other stabilized by pi-pi interactions. Even if the G4 canonical motif is fixed (four runs of 2 or 3 guanines), the variety of possible structure topologies is huge, due to strands orientation but also loop length and composition (Fig. 9)[23].

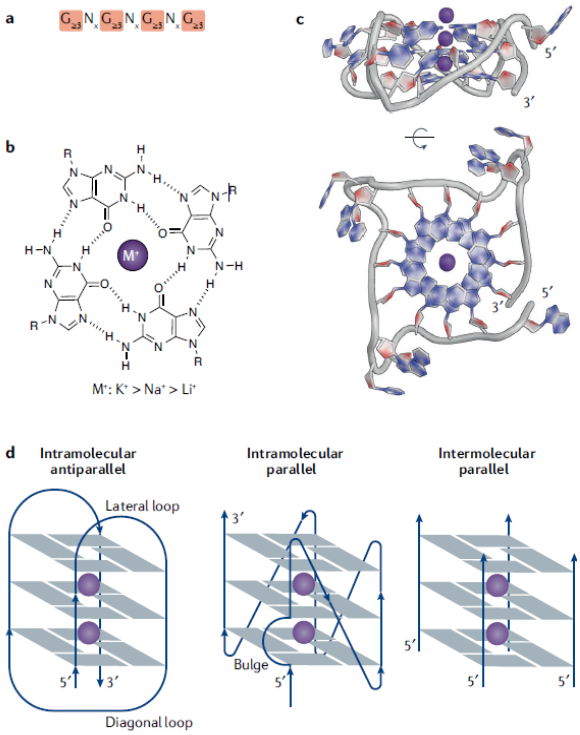


Fig. 9: The structure and topologies of G-quadruplexes. a | The G-quadruplex (G4) consensus sequence. x denotes the number of nucleotides in the loops (see part d). b | A guanine tetrad is stabilized by Hoogsteen base-pairing and by a central cation (M+), with a preference for monovalent cations in the order of potassium (K+) >sodium (Na+) >lithium (Li+). c | X-ray crystal structure of an intramolecular, parallel G4 from a human telomere sequence (PDB: 1KF1)214. d | Schematic representation of some G4 topologies. [22]

G4 distribution across the genome is peculiar, as there is an increase of G-rich sequences in the regulatory regions of the human genome such as promoters, but also protective ones such as telomers[24]. This initially suggested the potential role of G4s in modulation of gene expression[25] and protection of DNA architecture[26]. It was only after the development of a specific antibody, able to

recognize G4s, namely BG4, that G4 biological roles started to be discovered. Biffi and colleagues demonstrated the presence of G4s in living cells. They showed by immunofluorescence that G4 formation is dependent of the cell cycles. They observed increase of foci in the S phase, probably due to the duplication of the DNA that promotes double strand separation. They also saw enrichment of G4 foci after G4 ligands treatment, suggesting once again formation of G4s in cells[27]. They and others were able to use the same antibody in the ChIP-seq technique, with which they mapped G4s in fixed cells finding that G4s are folded predominantly in gene promoter regions. Moreover, it was pointed out that different cell lines have a different *"G4 landscape",* meaning that not every G-rich sequence is always folded, its state varies among different cell types[28],[29]. This evidence suggests an ambivalent role for G4s, where, when embedded in gene promoters they recruit transcription factors to boost transcription[29], when located in the gene bodies they pause transcription (Fig. 10). In fact, G4s can act as steric hindrance elements towards polymerases processing and they need to be resolved to allow full transcription. In the last years many helicases have been found able to specifically resolve G4s[30]. This corroborates the ambivalent nature of G4s that seems to rely more on their location than on their specific topology[31]. For this reason, recently some efforts have been made to develop new tools able to destabilize G4s rather than stabilize them as it was done in the past[32].

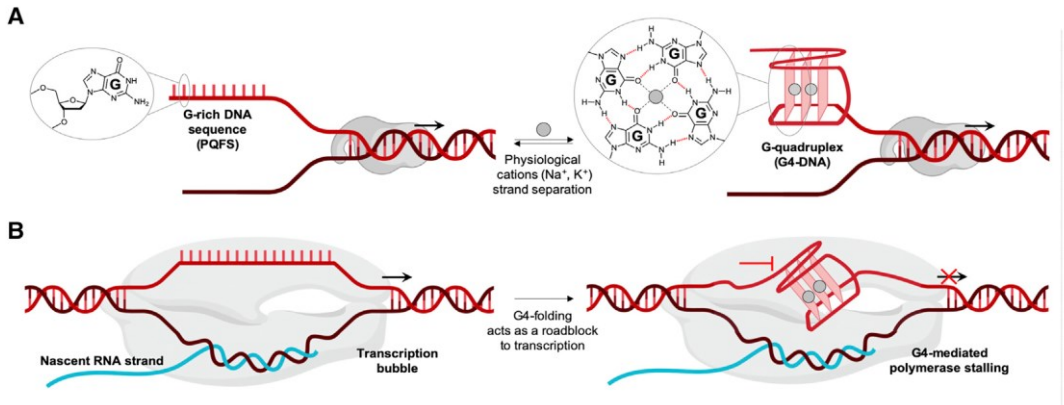


Fig. 10: G-quadruplex-forming sequences and their in vitro characterization Schematic representation of (A) the transient formation of a G-quadruplex-DNA (G4-DNA) when a putative quadruplex-forming sequence (PQFS) is freed from the duplex structure (inserts show detailed chemical structures of guanine (G) and G-quartet); (B) the G4-mediated stalling of the RNA polymerase during transcription. [30]

3. Aim

The project aims at finding if there is a relationship between G4 formation and XDP, focusing on the SVA retrotransposon. Until now there are no data that prove formation of G4s within SVAs in cells, even if some data *in vitro* suggest formation of strong SVA G4 structures[33]. The correlation between severity and early development of XDP in patients with greater hexameric expansions suggests that this domain is crucial in the XDP disease[2]. Our hypothesis is that the negative effects of the SVA insertion in the *TAF1* gene are mediated by formation of G4s forming mainly in the hexameric repeats. To address this hypothesis, we performed both *in vitro* and in cell testing. In particular, *in vitro* we used QGRS tool to predict putative G4 forming sequence within the SVA retrotransposon. We evaluate the topology and the stability of the chosen sequences by Circular Dichroism and DMS footprinting. Moreover we evaluate the ability of two well known G4 ligands, namely BRACO-19 and Quarfloxin to stabilize those G4 structures. We also designed a PCR stop assay to study the ability of those G4s to form within the complete SVA fragments and we validated the presence of two G4 forming domain by NESTED PCR STOP assay. In cells we perform two different G4 mapping technique: G4.ChIP-seq and G4 CUT&TAG that allowed us to identify a different landscape between XDP and healthy fibroblasts. Moreover by G4-ChIP-qPCR we were able to see an enrichment for the hexameric region compared in the XDP patients compared to the healthy controls. By RT-qPCR we measure the levels of expression of SVA-F retrotransposon, that are increase in SVA carrier fibroblasts. We treat the cells with the previously tested G4 ligands and find out that there is an increase in *TAF1* transcription of the first exons but only in XDP paitents. Also there is a strong increase in the 32-intron retention form of *TAF1* after the treatment in particular in NPCs XDP affected cells, suggesting a negative role of SVA G4s in the transcription of *TAF1*. We applied the same approach with a G4 destabilizer Phpc, in order to destabilize the G4s within the SVA but the final result was identical to the other G4 ligands.

4. Methods

4.1 Cell culture

K-562 (CCL-243) were purchased from ATCC. Cells were maintained in IMEM (Gibco) with 10% FBS (Gibco) and 1× penicillin–streptomycin (Sigma). Human fibroblasts (hFib) were purchased from RUCDR infinite biologics (Piscataway Township, NJ). Cells were maintained in DMEM (Gibco) with 15% FBS (Gibco), 1x Non-Essential Amino Acids (Sigma) and 1× penicillin–streptomycin (Sigma). NPCs were shared by Dr D. Christopher Bragg, PhD, from the collaborative center for XDP. NPCs were maintained in Neuronal Progenitors Medium made with DMEM F12 (Gibco, cat no # 11320033), 2% B27 (50 x stock; Gibco, cat no # 17504044), 1% Penicillin/Streptomycin (100x stock; Gibco, cat no # 15140-122), 20ng/Ml EGF (100µg stock; Peprotech, cat no # AF-100-15-100UG), 20ng/ml bFGF (50µg stock; Milipore, cat no # GF003), 5ug/ml heparin (Sigma, cat no # H3149-100KU). NPCs were cultered in Geltrex LDEV-Free hESC-Qualified (cat no # A1413302) coated wells. All cell lines used in this study are resumed in Table 1 in the SI.

4.2 Primers and oligonucleotides

Desalted primers and oligonucleotides were purchased from Sigma Aldrich (Milan, Italy). A detailed list of primers name and sequence is available in the Tables in SI.

4.3 G4 prediction

The presence of putative G4s was assessed by two different computational tools: (i) QGRS mapper[34] to predict the putative G4 forming sequences within XDP SVA; and (ii) Quadparser[24]to predict the putative G4 sequences in the G4-ChIP-seq peaks. QGRS tool was used on line at https://bioinformatics.ramapo.edu/QGRS/ with the following parameters: Max lenght:30 – MinG-GroupSize:3 – Loop size : from 0 To 10. XDP SVA sequence was retrieve from NCBI Genebank AB191243.1[35]. Quadparser script was downloaded from https://github.com/dariober/ as indicated by Puig Lombardi et al[36]., and applied with the regular expressions ([gG]{3,5}\w{1,7}){3,}[gG]{3,5}), in order to allow the matching of loops with length 1–7.

4.4 Circular Dichroism

DNA oligonucleotides were diluted to a final concentration (4 µM) in Lithium Cacodylate buffer (10 mM, pH 7.4, KCl 10-100 mM). All samples were annealed at 95 °C for 5 min and gradually cooled

to room temperature. CD spectra were recorded on a Chirascan-Plus (Applied Photophysics, Leatherhead, UK) equipped with a Peltier temperature controller using a quartz cell of 5 mm optical path length, over a wavelength range of 230–320 nm. For the determination of Tm, spectra were recorded over a temperature range of 20–90 °C, with temperature increase of 5 °C. The reported spectra are baseline-corrected for signal contributions due to the buffer. Observed ellipticities were converted to mean residue ellipticity (θ) = deg × cm2 × dmol−1 (mol ellip). Tm values were calculated according to the van't Hoff equation, applied for a two-state transition from a folded to unfolded state, assuming that the heat capacity of the folded and unfolded states is equal. All oligonucleotides were tested at least twice in independent experiments.

4.5 DMS footprinting

The DNA substrate of interest was gel-purified before use, 5'-end-labeled with [γ-32P]ATP by T4 polynucleotide kinase, purified using MicroSpin G-25 columns (GE Healthcare Europe, Milan, Italy), resuspended in lithium cacodylate buffer 10 mM pH 7.4 with or without KCl 100 mM, heat denatured and folded. Sample solutions were then treated with dimethylsulfate (DMS, 0.5% in ethanol) for 5 min and stopped by addition of 10% glycerol and β-mercaptoethanol. Samples were loaded onto a 16% native polyacrylamide gel and run until the desired resolution was obtained. DNA bands were localized via autoradiography, excised, eluted overnight. The supernatants were recovered, ethanol-precipitated and treated with piperidine 1 M for 30 min at 90 °C. Samples were dried in a speed-vac, washed with water, dried again, and resuspended in formamide gel loading buffer. Reaction products were analyzed on 20% denaturing polyacrylamide gels, visualized by phosphorimaging analysis, and quantified by ImageQuant TL software (GE Healthcare Europe, Milan, Italy).

4.6 TaqPol STOP assay

The DNA primer 5' FAM labelled (final concentration 72 nM) was annealed to the template (final concentration 36 nM) in lithium cacodylate buffer (10 mM, pH 7.4) in the presence or absence of 100 mM KCl by heating at 95°C for 5 min and gradually cooling to room temperature to allow both primer annealing and G4 folding, and incubated overnight. Where indicated, the G4-ligands BRACO-19 and Quarfloxin were added at the concentrations of 2 μM and 5 nM with the template. The primer was subsequently extended on the template strand by adding 2 U/reaction of AmpliTaq Gold DNA polymerase (Applied Biosystem, Carlsbad, CA, USA) at 42°C for 30 min. Reactions were stopped

by sodium acetate precipitation and primer extension products were separated on a 16% denaturing gel, and finally visualized by fluorescence Gel Scanner (Typhoon FLA 9000).

4.7 PCR and NESTED PCR stop assay

Genomic DNA (gDNA) was extracted from 1 million cells using *Gene jet DNA purification kit* (Thermofisher Scientific). 100 ng of gDNA was used for PCR using *PrimeSTAR GLX* enzyme (Takara, Japan) as previously reported[37]. For SVA PCR stop assay, 100 ng of gDNA was folded into G4 adding KCl (0-150 mM) and heating the mixture for 5 min at 95 C and leaving the tubes at RT O/N. The day after we add different G4 ligands to the mix and leave them to equilibrate at RT in the dark for 6 hours. This mix was used as a template for SVA PCR amplification using the same protocol as before. PCR were loaded in an agarose gel 0.8% and run for 1h at 80V., we design specific primers with Primer3plus[38]. 10 µL of PCR samples were loaded in agarose gel 0.8% with NancyDye (Merck) and run at 80 V for 1h. Gels were visualized with Typhoon Gel scanner. Bands were quantified using ImageQuant TL software (GE Healthcare Europe, Milan, Italy). For Nested PCR we use gel purified SVA band using the *Spinnaker kit* (Euroclone). 1 ng of purified SVA band was submitted to PCR using different parameters for each domain. 10 µL of PCR samples were loaded in agarose gel 1.2% with NancyDye (merck) and run at 80 V for 1h. Gels were visualized with Typhoon Gel scanner. Bands were quantified using ImageQuant TL software (GE Healthcare Europe, Milan, Italy).

4.8 SVA-F RT-qPCR

To measure the expression of SVA-F in XDP and ctrl cell family specific primers were designed (Table 3 SI) using Primer3plus[38].SVA-F family consensus sequence, that was used as template, was retrieved from Dfam database[39]. Total RNA was extracted from 500.000 cells with Trizol and further purified with ZymoConcentrator. 1 µg of RNA was retrotranscribed using Random Hexamers using SuperScriptIII (ThermoFisher Scientific). qPCR was performed with a LightCycler (Roche) using SybrGreen Master mix (Applied Biosystems). Row Ct data were normalized using actin as housekeeping gene and -ΔΔCt method for normalization.

4.9 G4-ChIP-qPCR/seq

The G4-ChIP protocol that we used is already reported[40] with few modifications. For nuclei isolation and sheraring buffers we use the buffer composition reported by Schmidt D and collegues[41]. $2x10^6$ of fixed hFib nuclei were sheared at 32x (30 ON/60 OFF), $4x10^6$ NPCs at 40x 30 ON/60 OFF) in 300 µL lisys buffer on a Bioruptor PLUS (Dianogenode). 1 µg of sheared chromatin was incubated with

0.5 µg BG4 antibody (Merck) for 1h at 16 °C. After binding with Anti-FLAG® M2 Magnetperlen M8823 (Merck), beads were wash 5 times with ice cold wash buffer and DNA was decrosslinked and purified as indicated elsewhere[29]. Eluted DNA was submitted to qPCR using Taqman mastermix (thernmofisher Scientific) or to ChIP library preparation using ThruPLEX DNA-Seq Kit (TAKARA) with dual indexes. Libraries were purified with AMPure beads XG (Beckmann), quality check was perform by qPCR to check specific enrichment for G4-positive ctrl and by Bioanalyzer High Sensitivity kit (Agilent) to check library size distribution and adaptor contamination. Libraries were sequenced 50bp PE on a Miniseq500 Illumina platform by LAFUGA (Genecenter, Munich). All bioinformatic analyses were done on a cluster at the Biomedical Center (LMU, Munich). Reads were quality checked with FASTQC[42] and align to hg38 with bowtie2. For XDP patient we generated a custom genome reference inserting the SVA in the referecence genome using reform (https://github.com/gencorefacility/reform). SVA sequence was retrieved from NCBI #AB191243. Alignments were clean and converted to BAMs using Samtools[43]. Bigwig were generated with Bamtools[44]. Peak calling, annotation and repeat analyses was done using Homer[45]. Profile plots were done using deepTools[46] and for peak intersection we used Bedtools[47]. Graphs were done using ggplot2 and VennDiagram on R.

4.10 CUT&Tag

CUT&Tag experiments were performed as described previously[48] with minor modifications. For experiment with nuclei extraction we used a protocol reported elsewhere[49]. 1% BSA (Sigma Aldricht) was used in the antibody buffer, dig-wash buffer and dig-300 buffer to minimize cell clumping. Briefly, 2 × 105 cells were harvested, washed with wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine), and immobilized to concanavalin A–coated beads (Cliniscience) with incubation at room temperature for 10 min. The bead-bound cells were incubated in 200 µl of primary antibody buffer (wash buffer with 1% BSA, 2 mM EDTA and 0.05% digitonin for gentle permeabilization of the plasma and nuclear membrane) with primary antibody 1:25 FLAG-tagged BG4 antibody (Merck, MABE917) or 1:100 IgG rabbit anti-mouse antibody (Sigma, M7023) or 1:50 RNA Pol2 antibody (CellSignaling, 2629) or 1:50 HeK27me (CellSignaling, 9756) dilution at 4°C by rotating overnight. The next day, BG4 antibody-incubated cells were resuspended in 200 µl of dig-wash buffer with 1:100 dilution of mouse anti-FLAG antibody (Sigma, F1804) and incubated at room temperature for 1 h with slow rotation. After BG4 incubation cells were washed with 800 µl of dig-

wash buffer briefly three times to remove unbound antibodies. After they were incubated with 1:100 dilution of rabbit anti-mouse antibody (Sigma, M7023) in 200 µl of dig-wash buffer at room temperature for 1 h with slow rotation, in alternative for rabbit antibody we used a guinea pig anti-rabbit antibody (Antibodies online, ABIN101961). After a brief wash with dig-wash buffer as above, cells were resuspended in 200 µl of dig-300 buffer (20 mM HEPES pH 7.5, 300 mM NaCl and 0.5 mM spermidine, 1% BSA and 0.01% digitonin) with 1:200 dilution of pA-Tn5 adapter complex (CUTANA, Epicypher) and incubated at room temperature for 1 h with slow rotation. pA-Tn5-bound cells were washed with 800 µl of dig-300 buffer three times, followed by tagmentation in 200 µl of tagmentation buffer (dig-300 buffer with 10 mM MgCl2) at 37°C for 1 h. After tagmentation, 15 mM EDTA, 500 µg/ml proteinase K and 0.1% SDS were added and further incubated at 63°C for another 1 h to stop tagmentation and digest protein. Genomic DNA was extracted and purified as reported elsewhere[50]. To generate libraries, purified genomic DNA was amplified with the barcoded i5 primer and barcoded i7 primer[51] using NEBNext Ultra II Q5 Master Mix (NEB, M0544). The library PCR products were cleaned up with Agencourt AMPure XP beads (Beckman Coulter, A63881) and sequenced on an Illumina Nextseq 500 instrument.

4.11 Cytotoxicity assay

Cells were seeded with 80% confluence in a 96-well. The day after they were treated with increasing concentrations of G4 ligands and after 24 hours of incubation ATPlite Luminescence Assay System (Perkin Elmer Italia, Milan, Italy) using manufacturer instructions. Luminescence was acquired by Varioskan (Thermo Fisher Scientific, Waltham, USA)

4.12 G4 ligand treatment RT-qPCR

hFib were seeded in a 6-well, NPCs were seeded in a 24-well to have a 60% confluence. The day after cells were treated with Braco 19 (MedChemExpress) or Quarfloxin (MedChemExpress). After 6h and 24h post treatment hFib were detached using Tripsine 0.05% (Gibco) , NPCs using Accutase (Sigma) and the cell pellet was stored at -80. The day after RNA was extracted with *GeneJet RNA purification kit* (Thermofisher Scientific) and 150 ng were retrotranscribed with *Taqman Retrotranscription kit* (ThermoFisher Scientific) using Random Hexamers. cDNA was diluted and used for realtime qPCR using *Taqman Mastermix*(Applied Biosystem). Row Ct data were normalized using actine as housekeeping gene and -ΔΔCt method for normalization. Data were plotted using GraphPad (Prism 2020).

5. Results

5.1 G4s can form within the XDP SVA *in vitro*

The first question to answer is if G4s can form within the XDP SVA retrotransposon. We first analyzed the SVA retrotransposon sequence with QGRS, a software able to predict potential G4 forming sequences[34]. Using canonical parameters, VNTR and Hexameric domains were identified as G4-rich domains (Fig 11 and SI Fig 1).

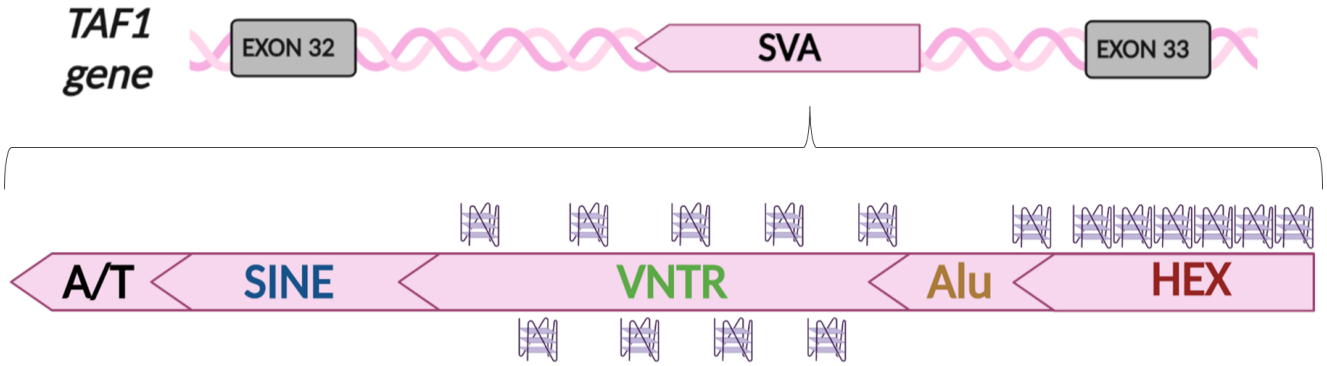


Fig. 11: Graphical presentation of SVA retrotransposon antisense insertion into TAF 1 gene of XDP patients. SVA is represented with its domains in different colours. VNTR and HEX are the ones with higher guanine content, so with the higher probability to form G4s.

The putative G4 forming sequences were submitted to circular dichroism melting experiments[52] in the presence of 100 mM KCl which is essential for G4 formation. Due to the repetitive nature of the SVA sequence, we identified only three different sequences that could form G4 in the forward strand (two sequences in the VNTR domain and one in the hexameric domain, whose minimal G4 forming sequence is formed by four hexanucleotide repeats) and one in the reverse strand (in the VNTR domain). All tested sequences displayed CD signatures of G4 parallel topology, with high melting temperature (Fig 12A, Table1). This suggests that if formed *in vivo*, they would be very stable. We next performed on the same sequences DMS footprinting. This technique identified which Guanines in the sequence are involved in G4 tetrad formation. As reported in Fig. 12B, in all tested sequences protected guanines were identified, confirming formation of G4s.

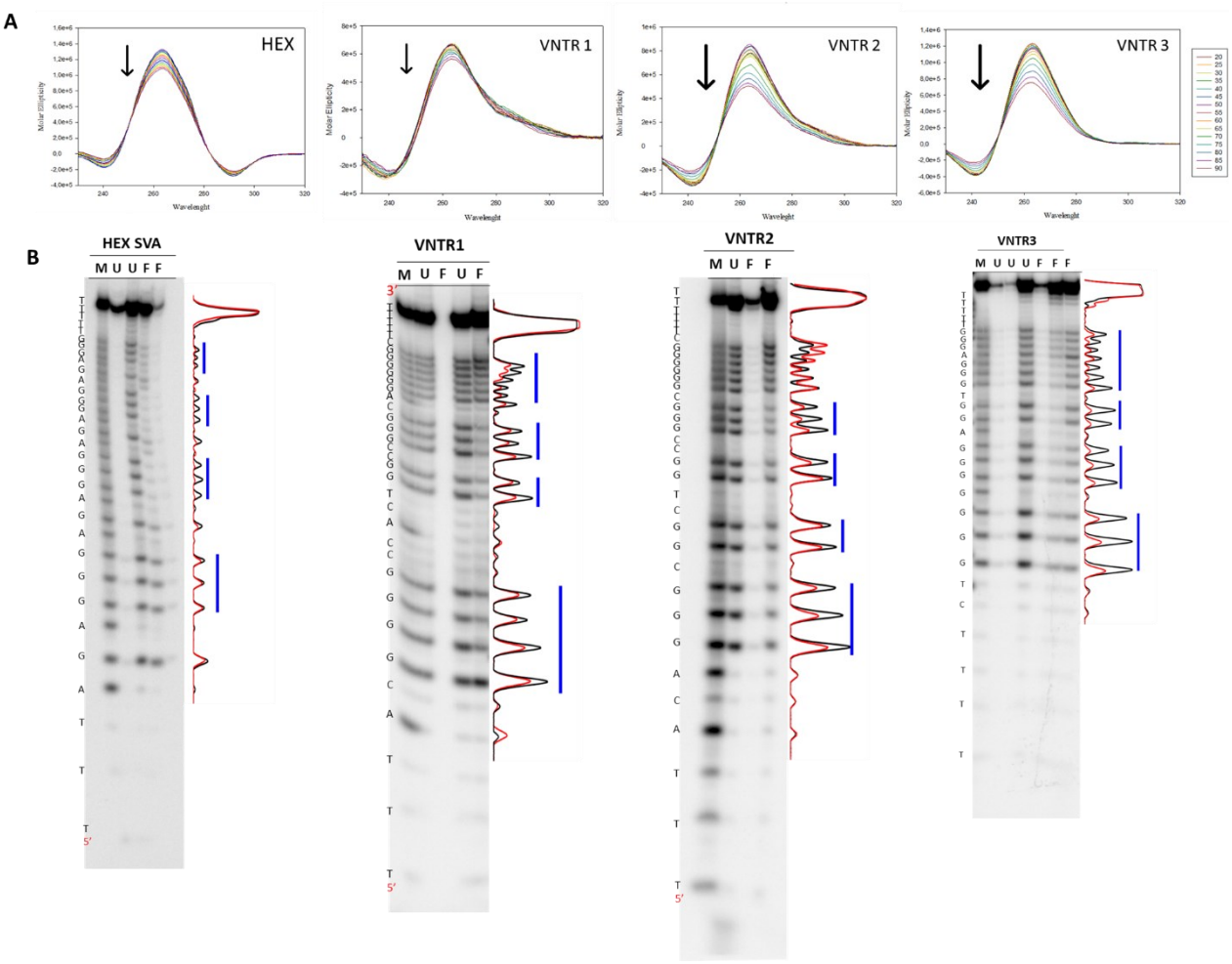
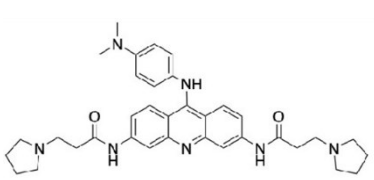
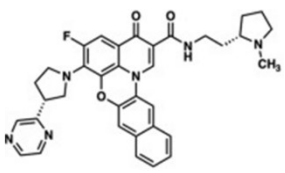


Fig. 12 : In vitro experiments to define G4s within XDP SVA retrotransposon. A. CD melting experiments show a very stable parallel structure with a positive peak at 265 nm. B.For each chosen sequence DMS-footprinting was performed to identify Guanines involved in the formation of the G4 tetrads. M stand for sequence marker, U for unfolded and M for folded. Black profile and ref profile s indicates the grade of protection of each identified guanine. In blue are hightlighted prtected Gs of the folded G4.

In the past years, many small molecules were developed that recognize and stabilize G4 structures *in vitro*: they are known as *G4 ligands*[53]. We tested some of them to check if they were able to



BRACO19



CX-3543

stabilize our sequences. At first we tested two G4 ligands by CD melting experiments: BRACO-19[54] (B-19) and Quarfloxin[55] (QFX) (Fig 13).

In presence of both compounds there is an increase of the melting temperature of the G4 structure of the hexameric domain G4 as shown in the Fig.14A. A similar effect was found also with the other G4 sequences, as shown by the melting temperatures reported in *Table 1*.

Table 1 : Melting temperature of SVA G4s with and without G4 ligands

	Tm (°C)			ΔTm (°C)	
SEQUENCE	**-**	**Braco-19**	**Quarfloxin**	**Braco-19**	**Quarfloxin**
HEX	52.1 ± 0.5	68.2 ± 1.5	>90	16.1	>37.9
VNTR 1	>90	>90	>90	ND	ND
VNTR 2	65.9 ± 2.3	74.6 ± 3.9	77.8 ± 2.1	8.7	11.9
VNTR 3	54.9 ± 1.4	>90	>90	> 35.1	> 35.1

We also performed Taq pol stop assay on the same sequences in the presence or absence of 100 mM KCl and with B19 and QFX. As reported in the Fig. 14B, there was less full-length product in the presence of those compounds and clear stop bands appeared at the first guanines forming the G4 tetrad. These results indicate that the G4 of this sequence can impair Taq activity. Also, the chosen compounds can stabilize G4s even more, thus opening the possibility to use them also in cells to check the effects of G4 stabilization within the SVA.

Table 2: Cell lines used in this study

sample	Onset age	Biopsy age	siblings	hFib (#ID)	NPCs (#ID)
XDP affected	32	35	proband	B2	32517
XDP affected	44	57	proband	G2	34363
XDP affected	38	44	proband	F2	/
XDP at risk NMC	/	7	Son of XDP affected	A2	/
Healthy ctrls	/	42	Son of XDP affected	C2	33113
Healthy ctrls	/	18	Son of XDP affected	E2	/
Healthy ctrls	/	34	Son of XDP affected	D2	33114

To assess if G4s can form also within the complete SVA in double-stranded conditions, we set up a PCR STOP assay[56]. We used as template DNA extracted from XDP-affected or healthy fibroblasts and performed PCR using primers external to the SVA insertion, to obtain amplification in all samples.

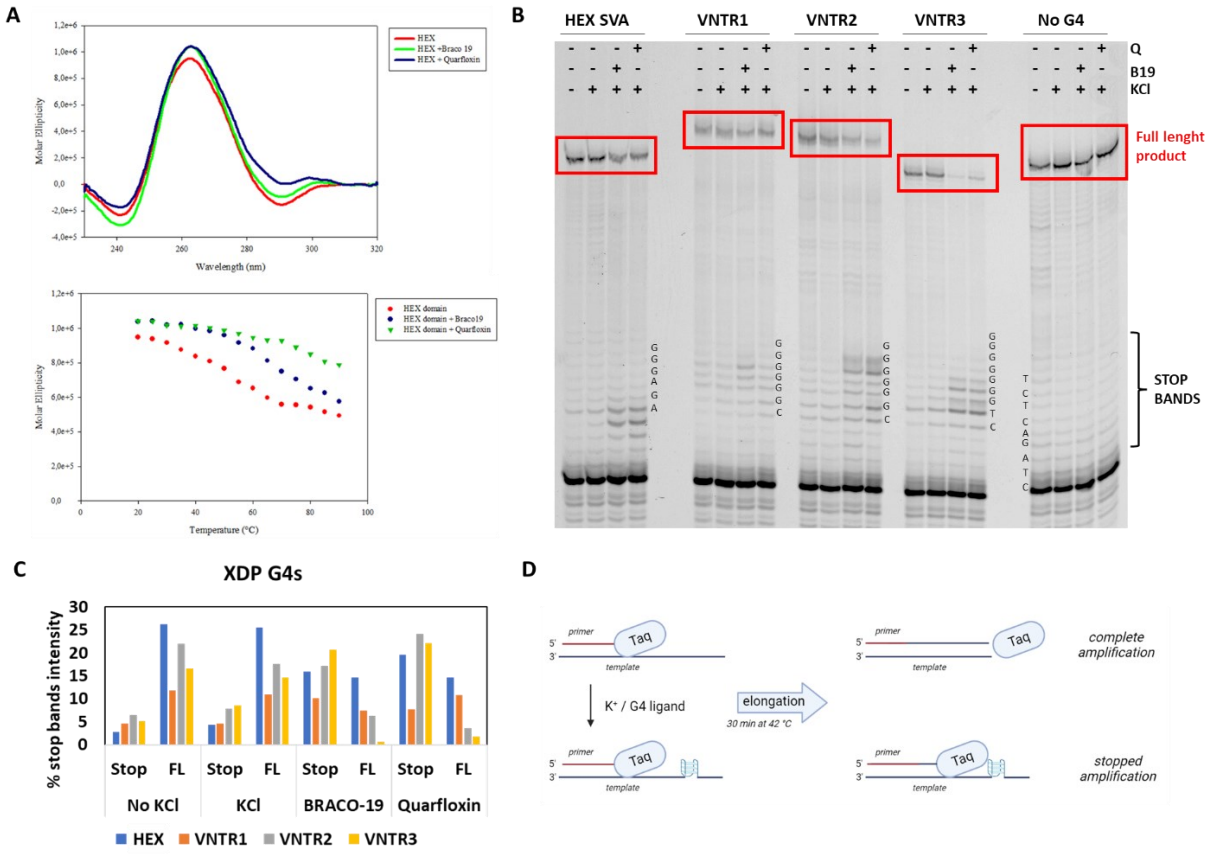


Fig. 14: G4 ligands stabilization of VNTRs and HEX G4s. A. CD spectra show increased stability of G4 hex in the presence of G4 ligands (upper panel). Change in molar elipticity at 265 nm melting at increasing temperature show higher stability in presence of a G4 ligand. B. TaqPol stop assays show less amplificated product and stop bands in the presence of G4 ligands even at a low compound concentration. C. STOP bands and Full lenght product bands quantifications by QuantImage software. D. Graphical scheme of TaqPolStop assay.

The cell lines that we used were hFibs deriving from biopsies of XDP patients (XDP) at different stage of disease and healthy relatives (ctrls) at different ages. We also obtained one cell line coming from a Non Manifesting Carrier (NMC), who is a patient that has inherited the typical haplotype but did not develop the disease at the time of biopsy, due to the young age (7 years old). We also used neuronal progenitor cells that came from the reprogramming of those fibroblasts. A complete table of all the cell lines used are summarized in Table 2, where siblings column indicate the relationship of ctrl cell lines towards the XDP affected cell lines (*probands*). Using genome DNA extracted from these cell lines, we retrieved an amplification band of about 600 bp in healthy patients and a shifted band of around 3200 bp for XDP patients (600+2800 SVA bp), which confirmed SVA insertion in XDP cells (Fig. 15).

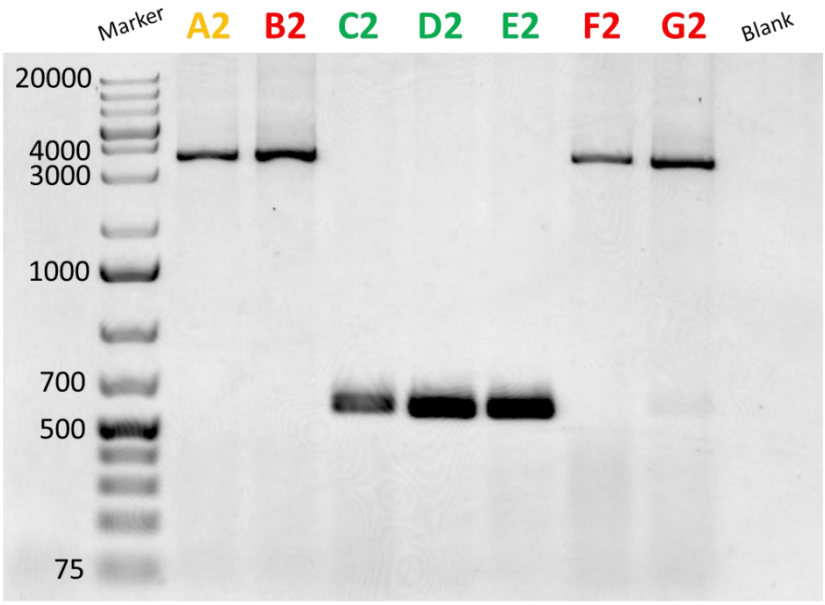


Fig 15: PCR of SVA from genome DNA extracted from fibroblasts derived from different patients. XDP and NMC patients display a shifted band compared to the healthy ctrls because of the SVA presence.

We next used conditions that induce and stabilize G4s in the PCR reaction mix. With increasing concentrations of KCl or G4 ligand, we observed reduction of the band corresponding to the SVA full-length product from XDP-affected patient samples. The decrease of the amplicon was proportional to the increase of KCl or G4 ligand concentration. Intriguingly, at 50 mM KCl SVA amplification was already decreased, suggesting that in cells, where KCl is 150 mM, G4s could have

strong effects. In the same conditions, product amplification was not affected in healthy cells (Fig. 16), as expected since in the ctrl amplicon no G4-forming sequences were retrieved with QGRS.

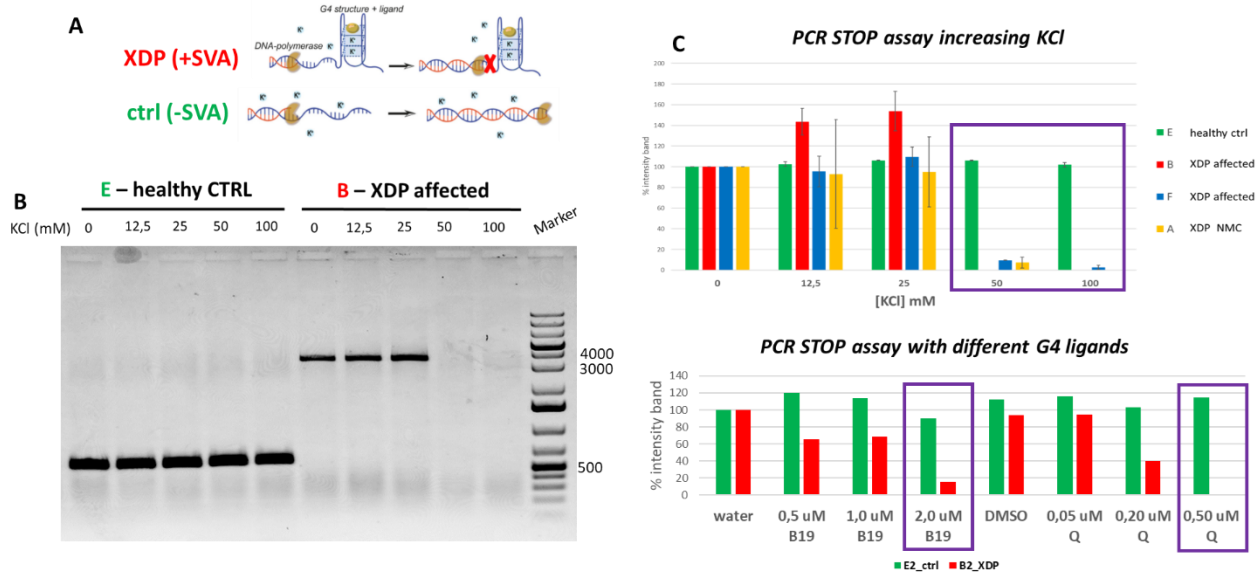


Fig 16: PCR STOP assay. A) schematic representation of the assay. B) Example of PCR stop assay gel with increasing KCl concentrations C) Gel band quantification of different cell lines in different conditions.

Knowing that only two of the SVA domains are G4-rich, we set up a Nested PCR STOP assay. We designed domain-specific primers to amplify each single SVA domain using the gel-purified SVA amplified from the extracted genome DNA (Fig. 17). We could not start from the genome DNA as above because many other SVAs are present in our genome so amplification of the internal domains would have been aspecific.

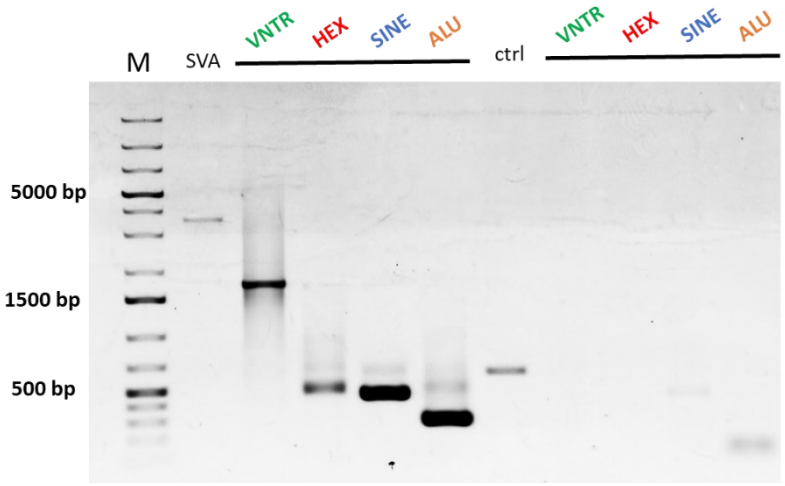


Fig 17: SVA and its domains amplification by PCR. NESTED PCR of each domain of the XDP SVA starting from purified SVA. As expected, no amplification of any domain is detected on samples derived from ctrl patients.

We found that the two G4-rich domains, namely the VNTR and HEX domains, are the only ones where, in G4-inducing conditions such as high KCl concentration or high G4 ligand concentration, PCR amplification is impaired. Amplification of the SINE and the Alu domains, which were predicted not to form G4s at all, was not affected by the same G4-inducing conditions (Fig 18). This indicates that G4s easily form in the VNTR and HEX domains and they are enough stable to impair polymerase progression. We can conclude that SVA *in vitro* adopts parallel G4 structures which are very stable and display high ability to impair Polymerases.

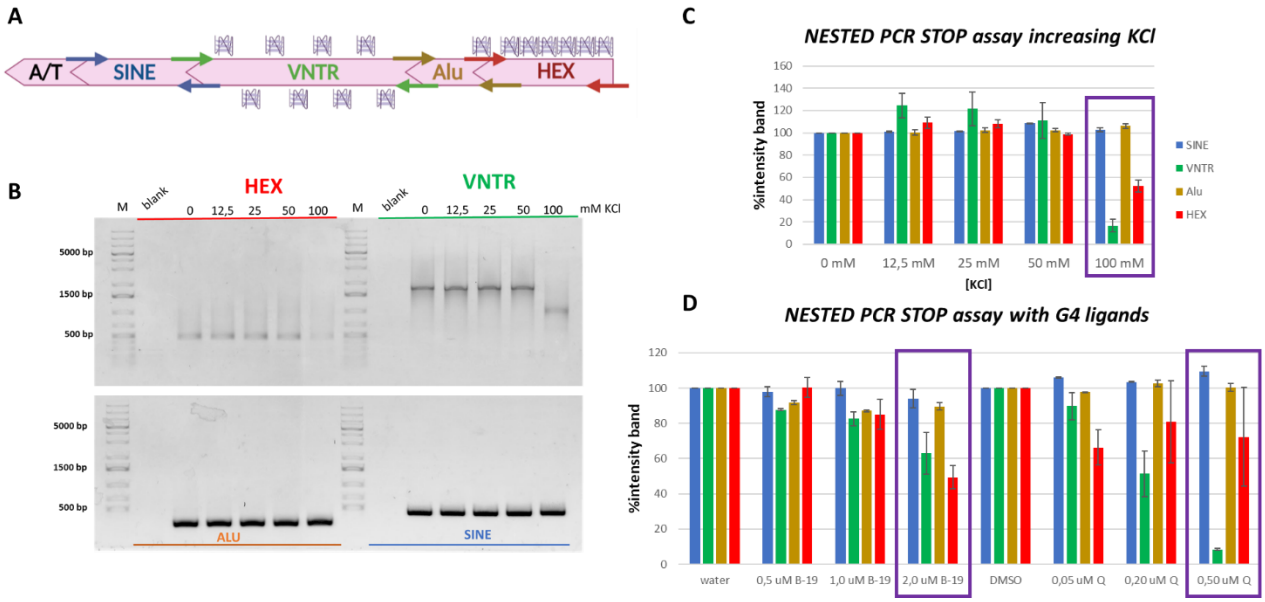


Fig 18: NESTED PCR stop assay. A) schematic representation of the assay. B) example of NESTED PCR stop assay gel with increasing KCl concentrations. C) Gel band quantification with increasing concentration of KCl. D) Gel band quantification of different G4 ligands concentration.

5.2 G4s are present in XDP cells in promoters and in the hex domain

To assess if G4s are present within SVA also in cells, thus in a more complex and physiological context, we first investigated if SVA G4s in cells could trigger transcription as shown for G4s when embedded in gene promoters.[28,29] Normally SVA entities are silenced by DNA methylation and H3K9me3 histone modification, but some remain transcriptionally active in our genome[57,58]. To measure the levels of transcriptionally active SVA-Fs, we performed RT-qPCR of hFib cells using 3 different couples of primers specific for the whole SVA-F family class (Fig.19). We observed an increase of SVA expression compared to ctrl cells, NMC cell displaying an intermediate value of expression. We cannot state that this increase of expression is due only to the presence of the XDP SVA-F. We would need to test more samples to validate this hyphoyesys. Moreover the different levels of expression could be due to a diffetnt coby number of those entites in the tested samples. Nevertheless if those results will be confirmed, they would open the path at the study of new molecular mechanisms that will need to be taken into account when studying XDP pathogenesis.

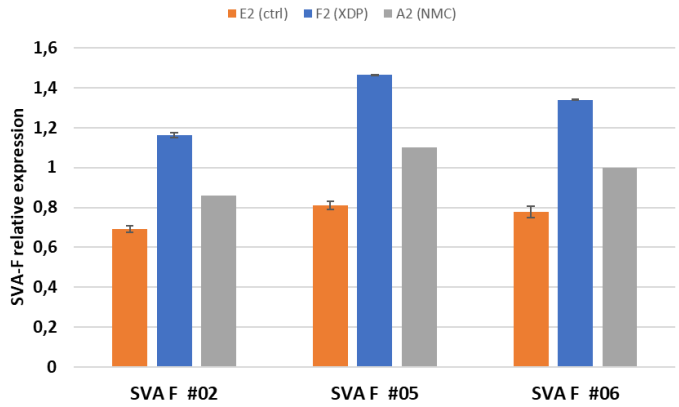


Fig. 19: SVA-F RT-qPCR of hFib cells. Three different couples of primers were designed on the SVA F consensus sequence (#02-05-06)

Next, we set out to find if G4s were folded within the XDP SVA in cells. To do this we performed G4-ChIP[59] (Fig. 20) on both hFib and NPCs. In collaboration with Prof. Gunnar Schotta's lab from LMU (Munich), I was able to adapt the published G4-ChIP protocol[40] to obtain the best results. The protocol needed to be optimized so instead of using primary hFib, we first set it up with a cancer cell model, such as K-562, that had been used by the authors of the first protocol and so the output profiles were deposited. Pooling together 4 immunoprecipitated samples (IPs) we managed to obtain good quality G4-ChIP, as shown in Fig. 21A. There was a very good enrichment over the input for G4 positive regions that were reported in the literature.

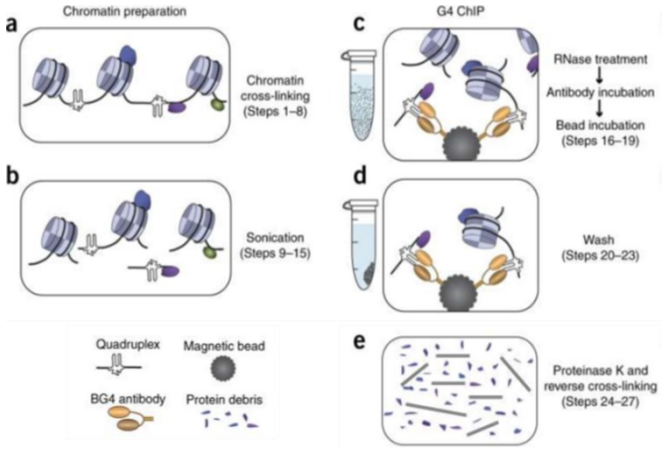


Fig. 20: Overview of the BG4-ChIP protocol. Adapted from Hänsel-Hertsch R, et al., Nat Protoc. 2018 Mar;13(3):551-564.

The negative regions contain no G4s in the amplicon so in theory they should not be enriched. Thus, the amount of negative control enrichment is a measure of the amount of unspecific material that is present in the total DNA recovered after immunoprecipitation. Usually G4-ChIP profiles have high background noise due to unspecific material, so it is fundamental to minimize it as much as possible to obtain good results. We prepared G4 ChIP libraries using a commercially available kit for transcription factor libraries, according to the manufacturer's instruction. We found that the library preparation method the authors proposed usually gave inconsistent results. This was probably due to the low amount of DNA that we immunoprecipitated. So, we opted to use a kit suitable for low input chip that always guaranteed good results. We also checked the quality of the library by bioanalyzer (Fig. 21B) to be sure that the purification did not carry over too much adaptors that could lead to poor sequencing results. Before submitting the libraries to sequencing, we also perform the same qPCR as after immunoprecipitation, to check that the regions that we found positive were also present and enriched in the library sample (Fig. 21C). We find that a fold change of at least 5 times between a positive G4 region compared to a negative region such as ESR1 was mandatory to obtain a good sequencing output. The sequencing profiles were very similar to the ones deposited for the same cell line. We also managed to obtain a good profile with only 10 million reads per sample, instead of the 50 million proposed in the original work, indicating that our optimized protocol was robust enough to be applied also to more difficult cell types (Fig.21D).

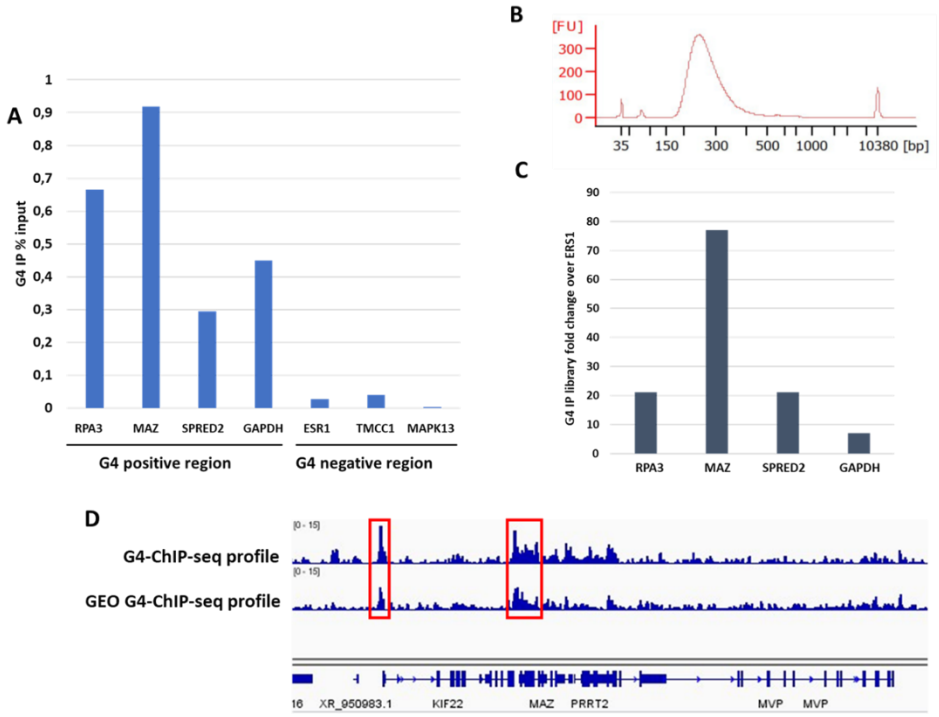
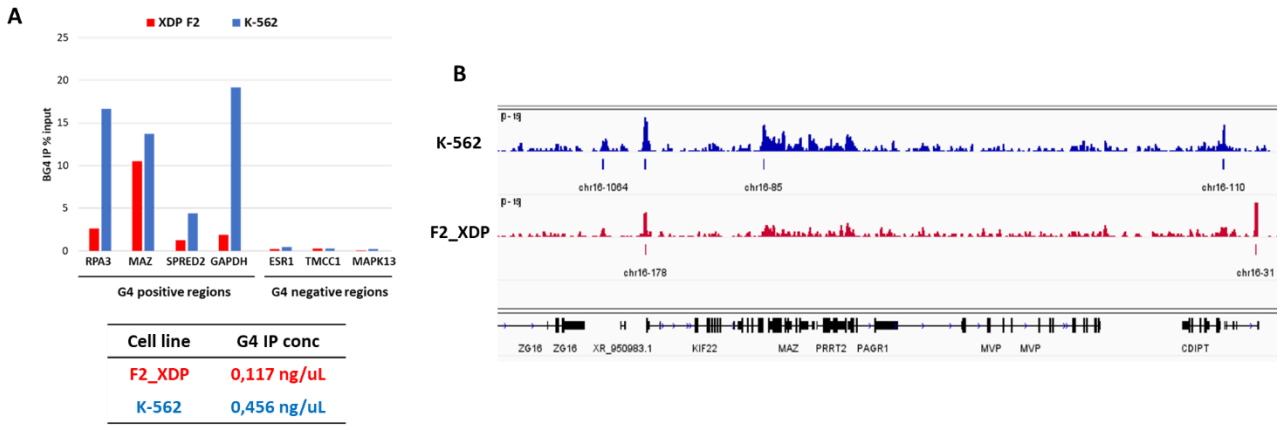


Fig 21. : G4-ChIP on K-562 cells. A) G4-ChIP qPCR of reported G4 positive and negative regions. B-C) Quality check of G4-ChIP libraries. Bioanalyzer profile of the library size distribution, enrichment of positive regions by qPCR. D) G4-ChIP seq profile comparison with the deposited profile on IGV.

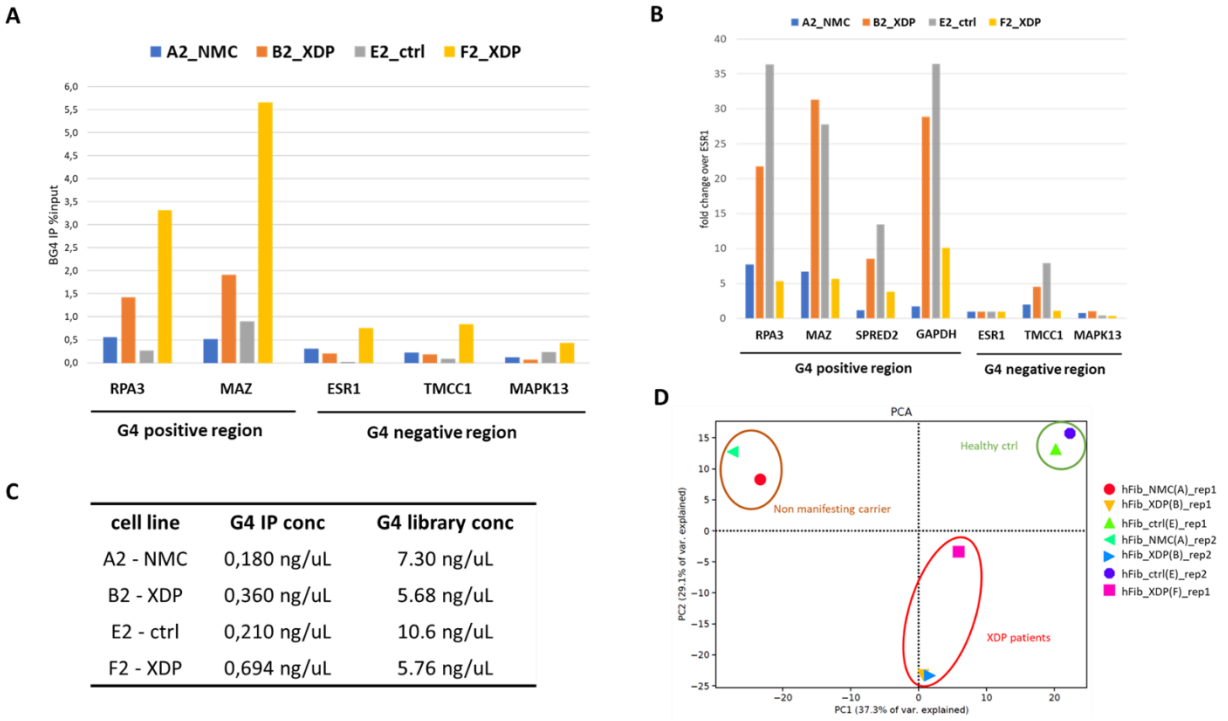
Hence, we applied this protocol to our primary hFib (Fig 22-23). The initial idea was to perform G4-ChIP-seq to find if there was a different G4 landscape between XDP-affected and healthy fibroblasts and to test if there was a G4 positive coverage on the XDP SVA. At first, we worked in parallel with one XDP affected hFib cell line and the K-562 cells. It was not surprising that the amount of DNA isolated from XDP hFib was at least 4- times lower compared to that from K-562. In fact, it is already reported that primary cells display less G4s than cancer cell lines, making the G4-ChIP-seq protocol more difficult to apply in non-cancer cells. It is also important to note that the canonical G4 positive regions were differently enriched in the two cell lines (Fig. 22A), thus confirming once again different G4 landscapes between cells types. This was even more evident in the sequencing profiles where some of the G4 peaks were common while others were cell specific (Fig. 22B).



Cell line	G4 IP conc
F2_XDP	0,117 ng/uL
K-562	0,456 ng/uL

Fig 22. : G4-ChIP on primary cells Vs cancer cells. A) G4-ChIP qPCR of reported G4 positive and negative regions in primary hFib XDP cells (red) and cancer K-562 cells (blue). The table shows the amount of DNA obtained after G4-ChIP for both cell lines. B) IGV screenshot of G4-ChiP-seq profiles in the two different cell lines.

We repeated the experiment in parallel on all available hFibs. The qPCR output was a bit different among the four tested hFibs (Fig.23A-B). In fact, the G4 positive regions were enriched compared to the negative ones, but the F2 XDP sample recovered more unspecific signal, and the amount of the initially recovered DNA was higher than the other samples. This was evident when we checked the quality of the library by qPCR (Fig. 23C).



C

cell line	G4 IP conc	G4 library conc
A2 - NMC	0,180 ng/uL	7.30 ng/uL
B2 - XDP	0,360 ng/uL	5.68 ng/uL
E2 - ctrl	0,210 ng/uL	10.6 ng/uL
F2 - XDP	0,694 ng/uL	5.76 ng/uL

Fig. 23: G4-ChIP on hFib. A) G4-ChIP qPCR of reported G4 positive and negative regions. B) ESR1 negative fold change in G4-ChIP-seq libraries. C) DNA concentration obtained after G4-ChIP and after library preparation. D) PCA of G4-ChIP-seq of hFib duplicates.

The fold change over ESR1 negative region of F2 XDP sample was very low compared to the other three samples. As expected, compared to the other sequencing profiles, this showed more background noise. To assess the degree of similarity among the sequencing profiles we performed PCA analysis (Fig. 23D).

We observed that the biological duplicates clustered together while every cell line clustered separately indicating that their G4-landscapes were essentially different. This gave us enough confidence to continue the analyses. We performed peak calling to identify the G4-positive regions. We identified many G4 positive regions that were common to all hFibs (Fig. 24) and G4 positive regions that were specific, thus confirming a different landscape among the different samples. Also, it was clear from the sequencing profiles that many G4 peaks were present in promoter regions. This was already reported for cancer cells, thus confirming the good quality of our profiles.

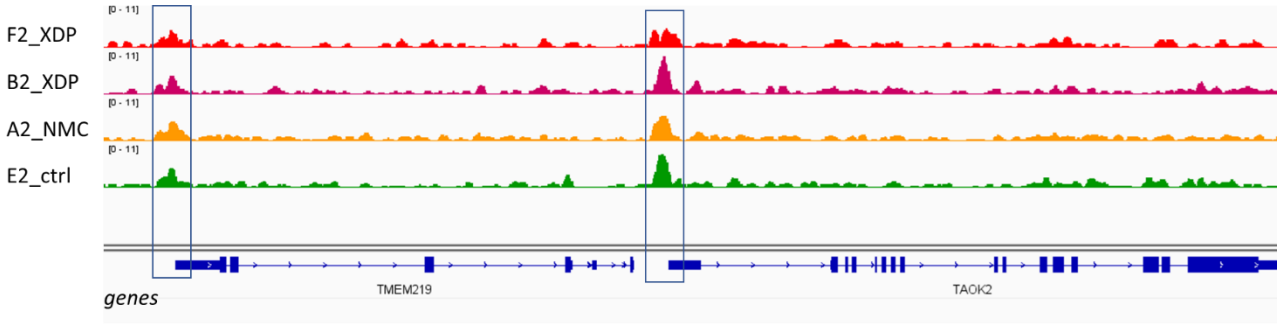


Fig. 24 : G4-ChIP-seq of four hFib cell lines. IGV screenshot of sequencing profiles. In red and dark red XDP affected cell lines, in yellow the NMC and in green the ctrl hfib cell line.

When annotating G4 peaks, the majority were confirmed at gene promoters (Fig 25A). XDP affected cells displayed more peaks than the ctrl ones. The NMC cell line, which presents the SVA insertion but whose patient did not develop the disease yet, displayed several intermediate peaks between the healthy vs XDP-affected conditions (Fig.25B). To check the quality of the identified regions, so to be sure that we immunoprecipitated most G-quadruplex forming regions, we applied two different tools: Quadparser[36] and Homer[60]. Quadparser is a well-known G4 predicting algorithm and in our immunoprecipitated regions it identified around 40% of peaks as canonical G4s (Table 3).

Table 3 : Quadparser G4 prediction on first replicate G4-ChIP-seq hFib

sample	Number of peaks	Canonical G4	%G4
A2 - NMC	10242	4307	42 %
B2 - XDP	9295	3530	38 %
E2 - ctrl	6297	2464	39 %

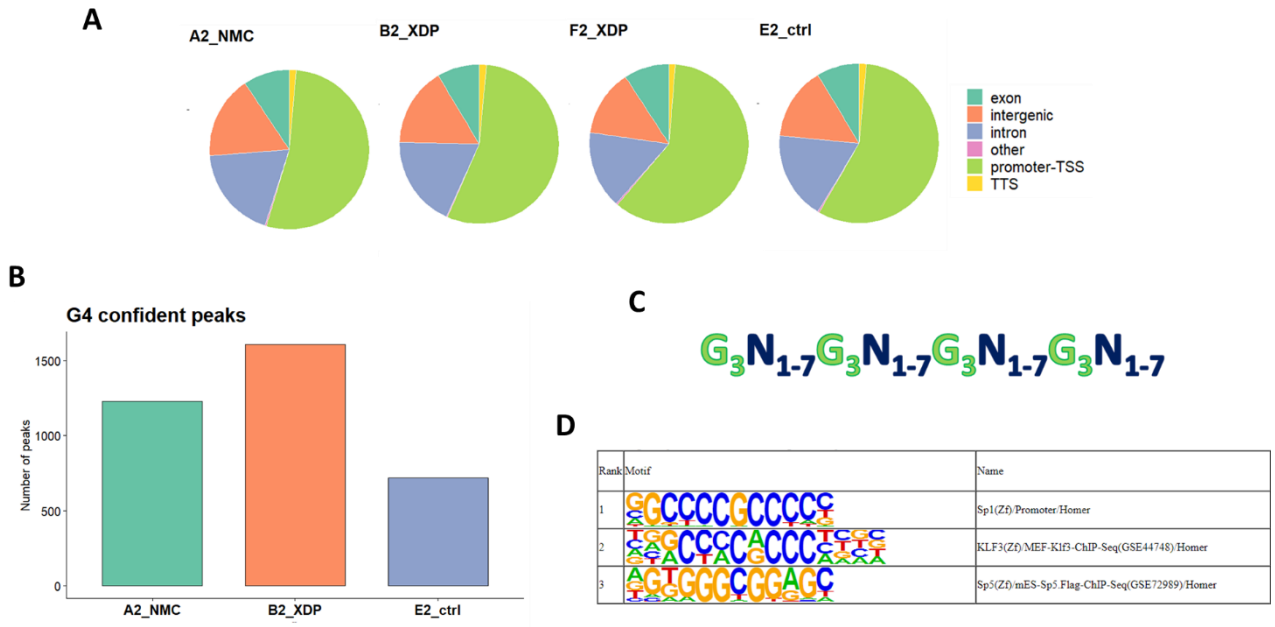


Fig 25: G4-ChIP-seq peaks in hFib. A) Pie charts of G4 peaks annotation in the genome. B) Number of peaks called in hFib cell lines merging duplicates. C) Canonical G-quadruplex forming motif used by Quadparser algorithm . D) HomerFindMotif highest ranked results showing the enriched reported consensu site of the binding protein.

This result could seem low but it is important to underline that the tool only searches for canonical G4 motifs (Fig. 22C), without taking into account all the different topologies of G4s, such as those with bulges or G4s of 2 tetrads. We were also interested in finding if a specific motif was present within the immunoprecipitated sequences, and using HOMER, we identified the SP1 binding motif as the most recurrent motif in the immunoprecipitated sequences (Fig. 25D); SP1 is a transcription factor known to bind G4s[9,17], thus validating once more our results. From this point, we will further analyze those data to find if there are some G4 positive regions specific to XDP affected cells and specific to the controls to find out if there is one of more G4-including gene that is involved in the disease. The primary aim of the G4-ChIP-seq experiment was to find if G4s were present in cells within the XDP SVA retrotransposon, because we know from *in vitro* data that the SVA G4s are very stable. However, due to the highly repetitive sequences that are characteristic of this class of retrotransposons, we could not uniquely align the reads to the XDP SVA sequences, because the reads were too short (50 bp) to be uniquely assigned to one SVA locus rather than another of the same family. When allowing multimapping alignment, some coverage is present within the XDP SVA but the quality of the reads is low, suggesting that they could align to other SVA-F as well (Fig. 26).

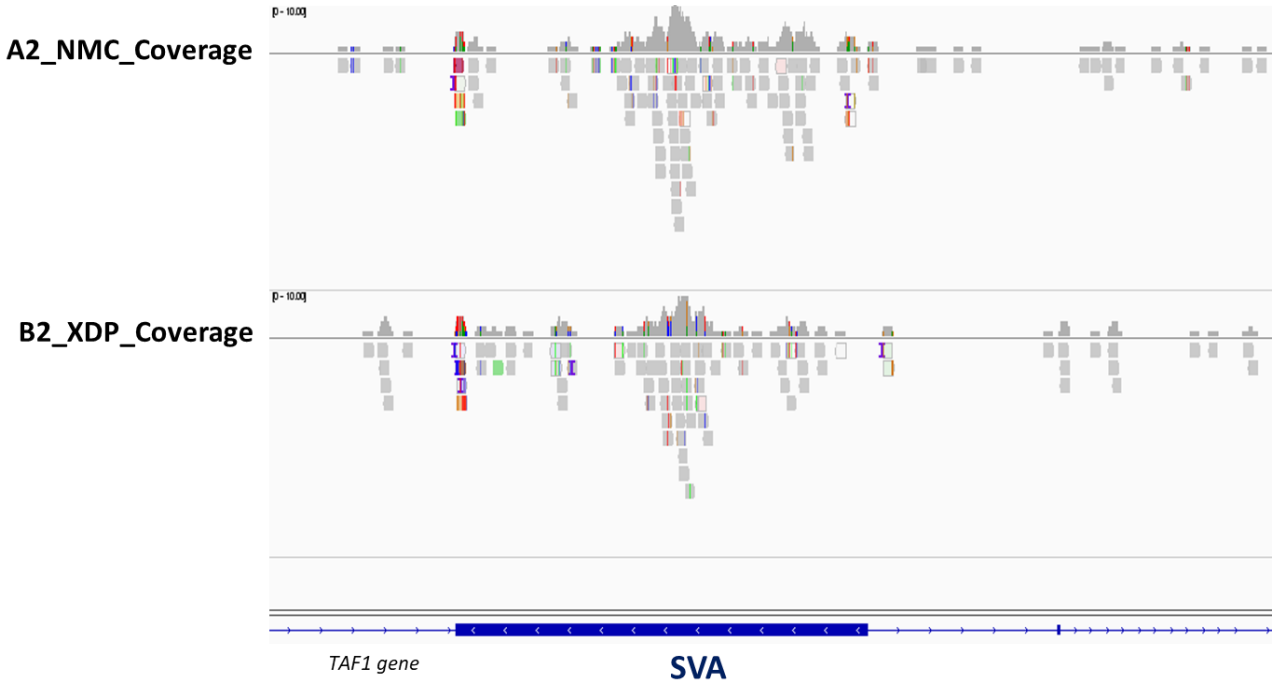


Fig 26: G4-ChIP coverage on XDP SVA. IGV screenshot of BAMs from NMC and XDP G4-ChIP-seq profiles allowing

On the other hand, we could use the SVA consensus sequence to map our IP sequence and check how much coverage we obtained compared to the input. We observed huge enrichment of SVA families in all G4-ChIP samples (Fig. 27A). This means that this kind of retrotransposons has folded G4s within their sequences: this is the first time to our knowledge that this evidence is proved in cells. Next, we compared the SVA and TSS plot profiles (Fig. 27B). TSS plot profiles were enriched and with higher coverage compared to the SVA's. Once again, this is likely due to the higher mappability of TSS regions compared to repetitive elements.

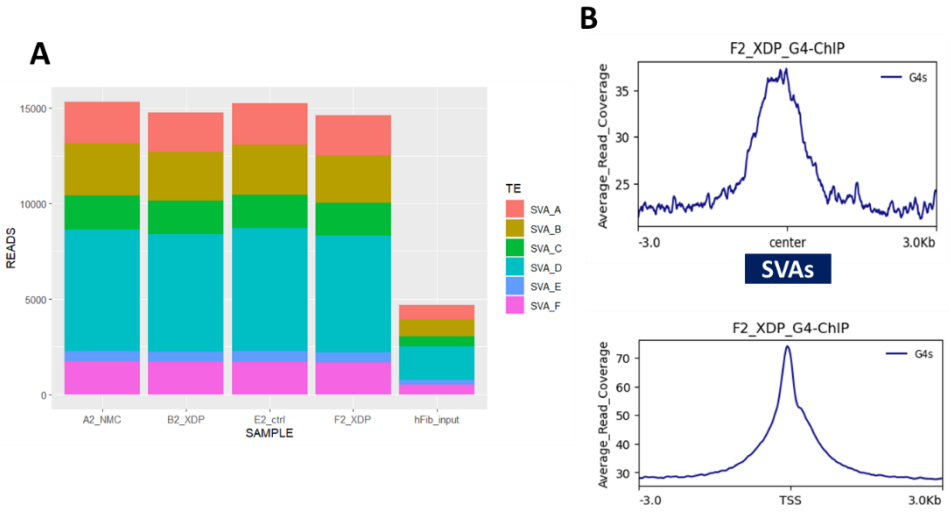


Fig 27: SVAs coverage in G4-ChIP-seq experiments. A. Enrichment of the six SVA families in the G4-ChIP-seq profiles of the tested fibroblast cell lines. B. Profile plot of G4-ChIP coverage of XDP affected fibroblasts (F2): upper coverage on SVA consensus sequences, downer coverage on TSS.

We next decided to test a new emerging method, CUT&Tag, that is supposed to display high quality of the sequencing profiles at much lower sequencing depth compared to the ChIP-seq technique. In the case of G4 mapping, we reasoned that this could increase the quality of our results because, even in our optimized protocol, a lot of background noise was present still.

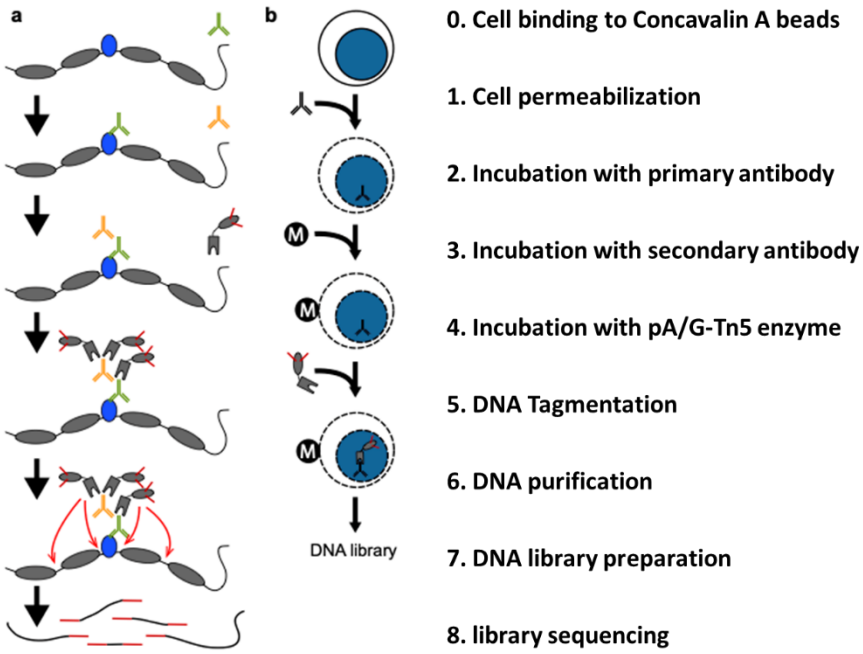


Fig 28: Schematic rapresentation of CUT&TAG protocol. A) infographic of the protocol in the epitope of interest. B) C&T protocol from a cell point of view.[50]

Moreover, this approach can be applied also to live (non-fixed) cells, withdrawing artifacts that can derive from fixation. Essentially cells are gently detached and bound to activated Concavalin-A beads, that by recognizing leptins in the cell membrane, efficiently immobilize the cells on magnetic beads. Cells are then permeabilized and incubated with the antibody of interest. Next, the pA/G-Tn5 enzyme is incubated to be delivered only in the regions where antibodies are. The enzyme is activated in order to obtain DNA tagmentation only in the located epitopes. The tagmented DNA is purified and libraries are prepared to be sequenced (Fig.28). The CUT&Tag technique is very similar to the CUT&RUN protocol[61]: the main difference is the enzyme used for DNA fragmentation, that in in the CUT&Tag is a Tn5, the same enzyme that is used for ATAC-seq experiments. The advantage of using Tn5 is that it is already loaded with adaptors, so the library preparation step is just a simple PCR. On the contrary with MNase used in the CUT&RUN method, the fragments obtained are shorter than the ones obtained with Tn5, so the library preparation step is longer and similar to the

library prep after a ChIP experiment. Tn5 is the enzyme used for ATAC-seq to mark open chromatin regions, for this reason it is fundamental to check that the tagmentation is specific to the epitope of interest and not just random available regions: to do so, a negative control tested with an unspecific antibody, such as IgG, needs necessarily to be included. This kind of negative control is a measure of the aspecificity of the assay and it is always important but even more when mapping epitopes that are normally located in open chromatin regions, such as G4s.

At that time some papers applying this technique to map G4s in cells started to be published. They were a bit different in the protocols compared to the original one, due to BG4 peculiarities. The major differences were that in one case nuclei extraction and sequential permeabilization with TRITON X-100 were performed, so to omit Digitonin to all buffers[49]; in the second case BSA was added to all buffers to reduce BG4 unspecific binding[48]. To be sure that even those little modifications of the protocol lead to the same result we decided to try the two new G4-CUT&Tag protocols in parallel with the original CUT&TAG protocol[62]. We chose our cancer cell line K-562 as cell model because the CUT&Tag protocol had already been tested in this cell line so we could easily check if all the tested protocols were working. Before assaying the protocols with BG4, we tried them with a positive control such as H3K27me, which is a histone marker modification very abundant in cell that is typically a marker of inactive chromatin. In this way we could also indirectly check the level of Tn5 propensity to tagment open chromatin regions. As a negative ctrl, we used a mouse IgG that was supposed to recognize nothing inside the cell. The sequencing profiles obtained were very encouraging. First, the profiles of the negative samples were almost flat with random reads that were not real peaks. The positive control showed specific peaks that identified the same regions in all three sequencing profiles. This means that all protocols worked well, producing the same output and that the DNA that was tagmentated in the positive control sample was specific (Fig. 29).

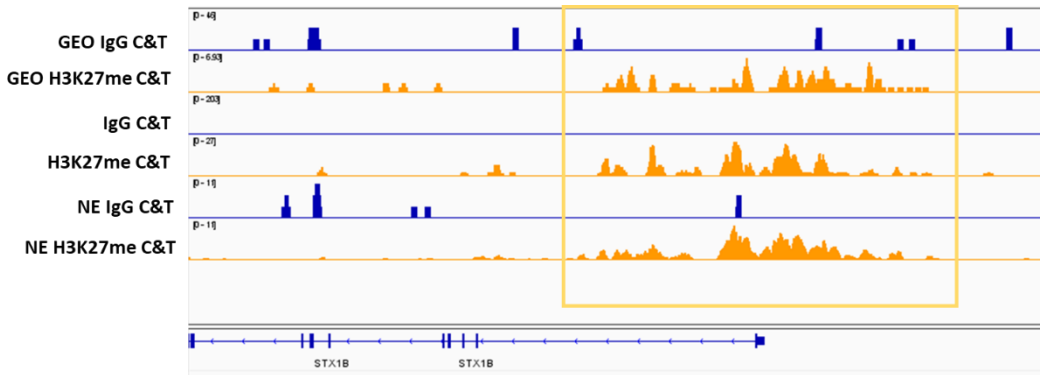


Fig 29: IGV screenshot of CUT&TAG profiles on K-562 cells. In blue IgG negative ctrl samples, in orange H3K27me profiles. GEO profiles are the ones deposited by henickoff and collegues, that were use as a reference. The two profiles in the middle were obtained using the G4 CUT&TAG protocol and the last two with Nuclei Extraction. The yellow box underly the common peaks in the H3K27me.

After we confirmed that both G4-CUT&Tag protocols were robust enough, we tested them with the BG4 antibody (Fig.30). Unfortunately, the G4 sample prepared using the first protocol with nuclei extraction did not work, and the sequencing output was too low to be sequenced. Instead, the other G4 sample prepared with the second protocol, worked very well. With the same sequencing depth, the G4 CUT&Tsg profile displayed the same G4 positive peaks identified also with the G4-ChIP-seq, but with very little background noise. This could help in the identification of G4s in cell with higher confidence in the peak calling step compared to the G4-ChIP. The IgG profile was flat compared to the other profiles. Even if there is a signal near the MAZ promoter, the peak shape is completely different compared to the G4 ones. The experiment worked well also if we look at the positive ctrl. In fact, the G4 signal does not overlap with the signal from H3K27me, and the two signals are mutually exclusive. This evidence confirms what is reported in the literature, i.e. that G4s are formed only in open chromatin region. This also confirms the specificity of Tn5 tagmentation, because if tagmentation were aspecific, we would have obtained identical sequencing profiles for all samples.

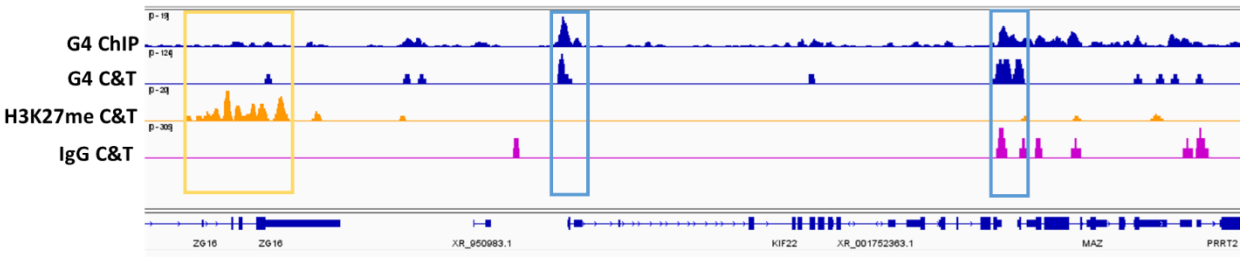


Fig 30: Validation of G4-CUT&TAG protocol on K-562 cells. IGV scrrenshot of sequencing profiles obtained. In blue G4-ChIP and G4-CUT&TAG samples, in orange H3K27me profille (positive ctrl), in pink IgG profile (negative ctrl). Yellow box highlight inactive chromatin region, ligh-blue boxex G4 positive regions.

We next performed the experiment on XDP hFib (Fig.31). Also in this case the experiment worked well, there was a strong correlation between the G4 CUT&Tag sequencing profiles with the G4-ChIP seq profiles obtained in the same cell line. In this case we also use the antibody able to recognize the phosphorylated form of RNA polymerase 2 on Serine 5, which is the form of the polymerase that is transcriptionally active. Noteworthily there is a very good overlap of the promoters that are G4 positive and the active polymerase, thus confirming once again their active role in transcription[29].

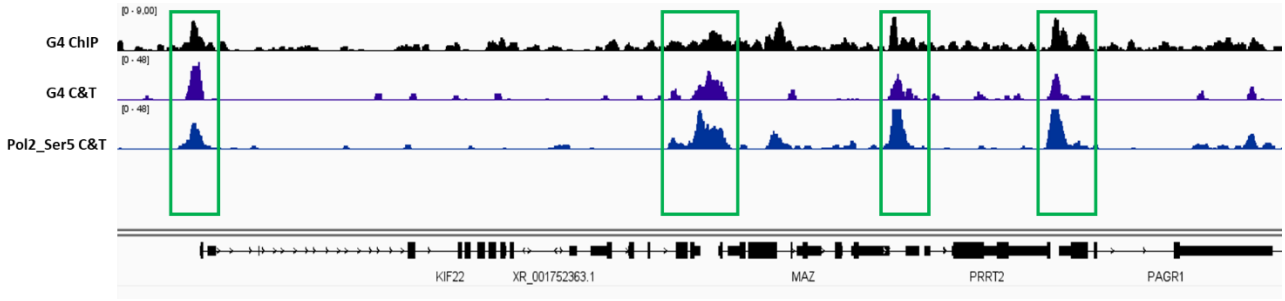


Fig 31: G4-CUT&TAG protocol validation on B2_XDP_hFib. IGV screenshot of sequencing profiles obtained. In black G4-ChIP profile, in violet G4-CUT&TAG profile and in blue Pol2 profile. Green boxes highlight G4 positive regions that correlates also with active Pol2.

Even if we confirmed that CUT&Tag is a powerful method to map G4s in living cells, despite its low background noise, the coverage on SVA regions did not increase. We ascribed this once again to the repetitive nature of this class of retrotransposons that makes them refractory to be uniquely mapped. We calculated the read coverage of the G4-ChIP-seq and G4-CUT&Tag profiles for each SVA family with Homer and we found that CUT&Tag mapped less SVAs compared to ChIP (Fig. 32A)

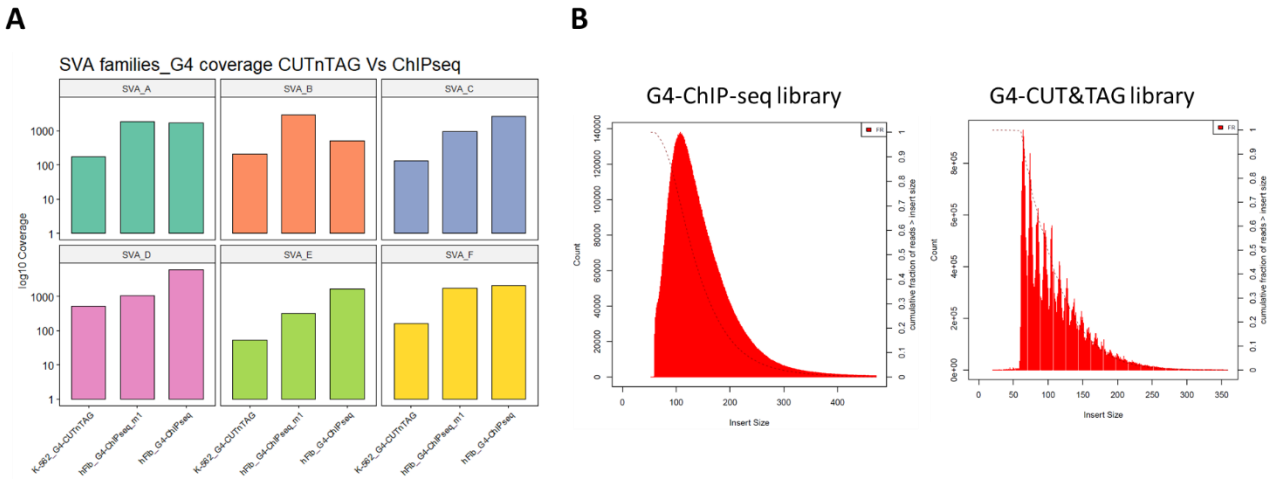


Fig 32: SVA coverage and library insert size comparison between CUT&TAG and ChIP-seq tecnhique. A) SVA family coverage using the two techniques. For ChIp-seq were considered also alignment with and without multimapping. B) Insert size calculation of the libraries prepared in the two different techniques. Plot made with Picard.

The reason is a technical detail that is crucial for uniquely map SVAs and it is very different in the two methods: the insert size in the libraries. We used Picard tool ("Picard Toolkit." 2019. Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/; Broad Institute) to calculate it from each library (Fug. 32B) and the result was very intriguing. The ChIP library displayed an average insert size of 150 bp, that is due to the sonication step that fragments DNA to around 100-

500 bp and to the following immunoprecipitation step and library preparation that adds the adaptors. On the contrary, G4-CUT&Tag libraries have shorter inserts, around 70 bp. If we go back to the protocol steps, adaptors are bound directly to the BG4-recognized epitopes, i.e. small DNA portions of probably 30-40 bp, so it is possible that the resulting libraries contain also smaller DNA fragments. Shorter inserts also mean that if the DNA came from a repetitive element, this will align very poorly to the reference genome, and it will likely map to TEs of different classes.

Even if by sequencing we could not map G4s uniquely to the XDP SVA, we reasoned that performing qPCR with primers specific for the XDP SVA region we could succeed. We managed to do this for the hexameric repeat: we designed a couple of Taqman primers, one complementary to the hexameric repeat, the reverse complementary to the flanking region external to the SVA insertion and a Taqman probe spanning the last nucleotides of the XDP SVA (Figure 33A). We obtained no amplification in the ctrl cell lines, as expected, while amplification was obtained in both XDP fibroblasts and NPCs (Fig 33B-C). This proves that at least the hexameric domain of the XDP SVA is folded into G4 structure in cells. For the VNTR it is impossible to design specific primers because the two flaking domains, namely the SINE and Alu domains, are common to other SVA retrotransposons.

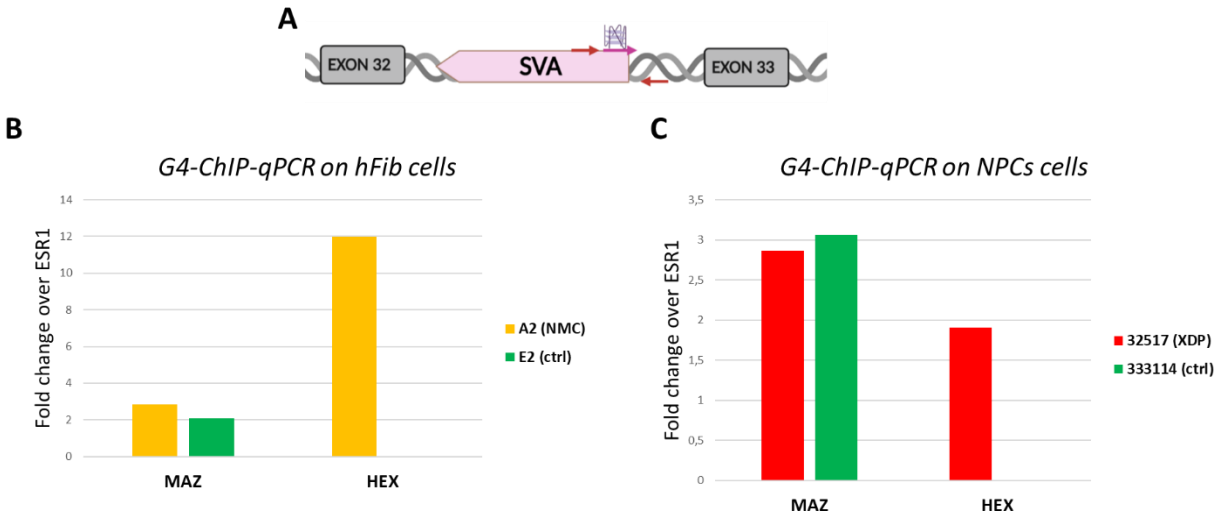


Fig 33: G4-ChIP-qPCR of XDP and ctrl cells. A) Schematic representation of the designed Taqman sistem (primers in red and probe in pink). B-C) Fold enrichment of reported positive G4 region MAZ and Hex over ESR1 (reported G4 negative region). B) on fibroblasts and C) on NPCs.

5.3 G4 ligands induce *TAF1* transcription but only in XDP patients

It is reported in the literature that *TAF1* levels in XDP affected cells are lower than in healthy cells[1]. XDP cells display intron 32 retention and downregulated expression of the last exons (Fig. 34), mostly in neuron-like cells. This could be caused by folded G4s that act as roadblocks leaving RNA pol II stall at intron 32 of *TAF1*. One way to indirectly prove the presence of G4s in the SVA and evaluate the effects of their presence on *TAF1* transcription is by performing RT-qPCR in G4-ligand-treated cells. If SVA G4s are more stabilized by the presence of a G4-ligand, and if the presence of G4s induces a decrease of the last exons of *TAF1* mRNA, RT-qPCR of G4-ligand-treated cells could detect it.

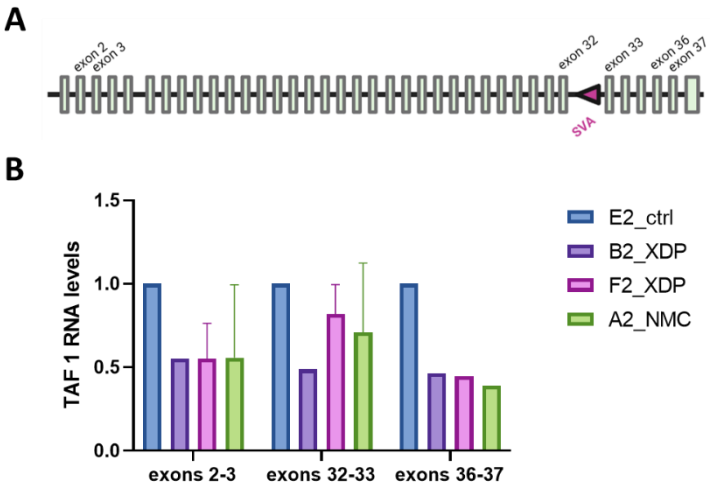


Fig 34: TAF 1 expression in fibroblasts. A) Schematic representation of TAF 1 exons with SVA insertion in intron 32. B) TAF 1 expression of different exons in XDP affected fibroblasts compared to the healthy ctrl.

We first performed cytotoxicity analysis for BRACO-19 and Quarfloxin using ATP-lite kit, to establish the range of subcytotoxic concentrations (Fig.35). We chose two concentrations for each compound (1 and 5 µM for B19, 0.5 and 1.5 µM for QFX) and when the seeded cells reached 60-70% of confluence, we added the compound in fresh medium and recovered RNA after 6 or 24 h of treatment. Being SVA insertion within *TAF1* intron 32, we evaluated the levels of *TAF1* exons by RT-qPCR before (exons 2-3), around (exons 32-33) and after (exons 36-37) the insertion together with intron 32 retention, using a Taqman system that we designed.

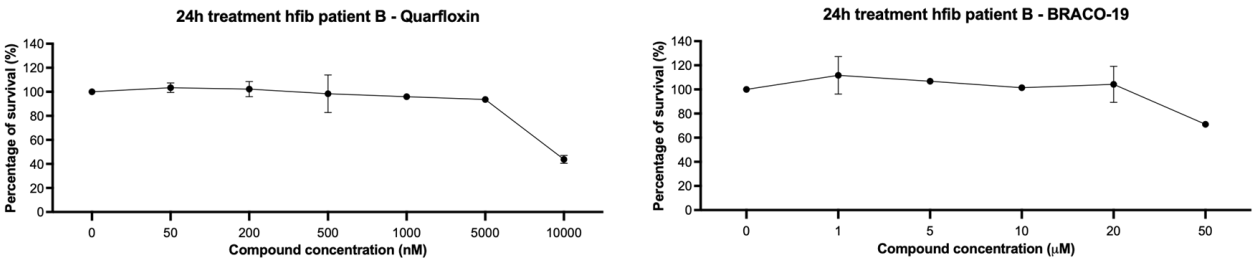


Fig 35: Cytotoxicity assay on hFib XDP to evaluate optimal working concentration. A) Quarfloxin become toxic above 5 uM, B) BRACO 19 become toxic above 5 uM

The treatment induced *TAF1* transcription because we observed an increase in the first exons of *TAF1*, but only in XDP cells. On the other hand, ctrl cells did not show the same behavior. It is interesting to note that this effect was more pronounced in XDP NPCs compared to XDP fibroblast and even if transcription of the first exons were increased, we did not observe the same effect on the exons after the SVA insertion whose levels remained stable or decreased (Fig 36).

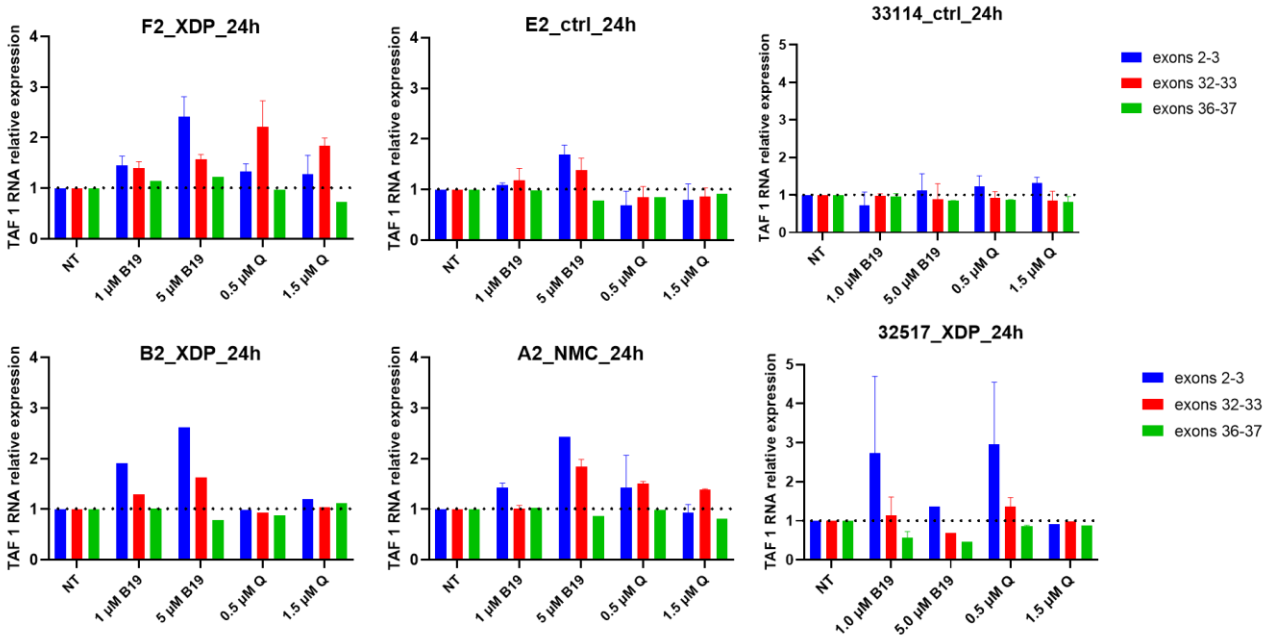


Fig 36: TAF 1 RT-qPCR after 24 h G4 ligand treatment. In different colors the different exons of TAF1 gene. Both ctrl cells hFib and NPCs do not show a clear effect on TAF1 exons transcription. On the other hand B19 induces the transcription on first exons (blue bars) in particular on SVA carrier cells. Quarfloxin show the same effect but only at the lower concentration on XDP NPCs.

Our idea is that G4 ligands stabilizing SVA G4s in XDP cells prevent RNA polymerase to transcribe the full length transcript of *TAF1*, being transcription stalled at intron 32, thus producing a

premature termination and the alternative form with intron retention of *TAF1* transcript. Analyzying the already published RNA-seq data of hFib[1] it is visible that there is retention before the SVA insertion in intron 32. This is peculiar to XDP cells because in normal cells the same pattern in the profile is not found (Fig. 37). This could very likely be the sign of a slowdown of the RNA pol that is actively transcribing *TAF1* in XDP cells.

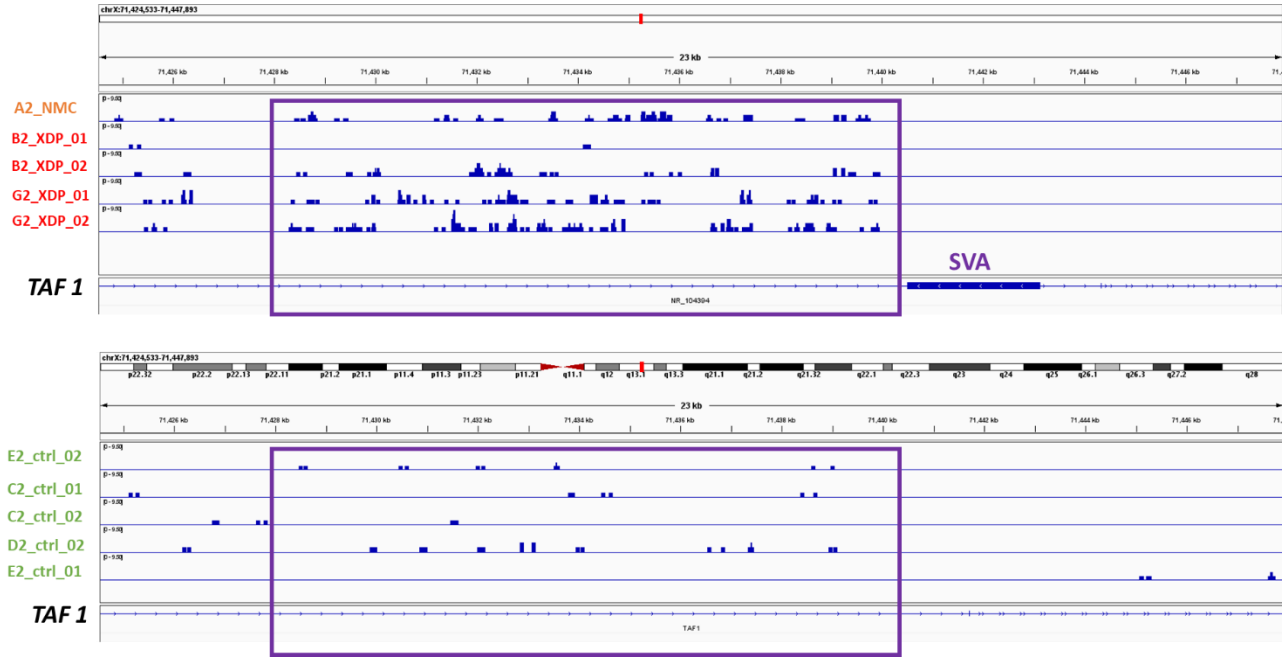


Fig 37: Total RNA seq profile of hFib. Upper panel represent XDP and NMC RNA seq samples were alligned to the cutom genome with XDP SVA insert; down panel only control camples. The violet box indicates the intron retention in the intron 32 of TAF1 gene.

When using a G4 ligand, we observed a prominent increase of intron retention in XDP cells. The increase of intron 32 retention was visible already after 6 h and remained stable after 24 h (Fig. 38). It is notable that Quarfloxin induced more intron retention than BRACO-19, and that the effect is huge in NPCs compared to hFib. We do not have a solid explanation to this behavior, probably NPCs are not able to recover as fast as hFibs from stabilization of XDP SVA G4s. In addition, the helicase panel able to resolve G4s could be different between the two cell types. The increase of *TAF1* first exons expression could also be due to the cells trying to restore the basal level of full length *TAF1* mRNA, which could explain why we observed an increase of the transcription of the first exons.

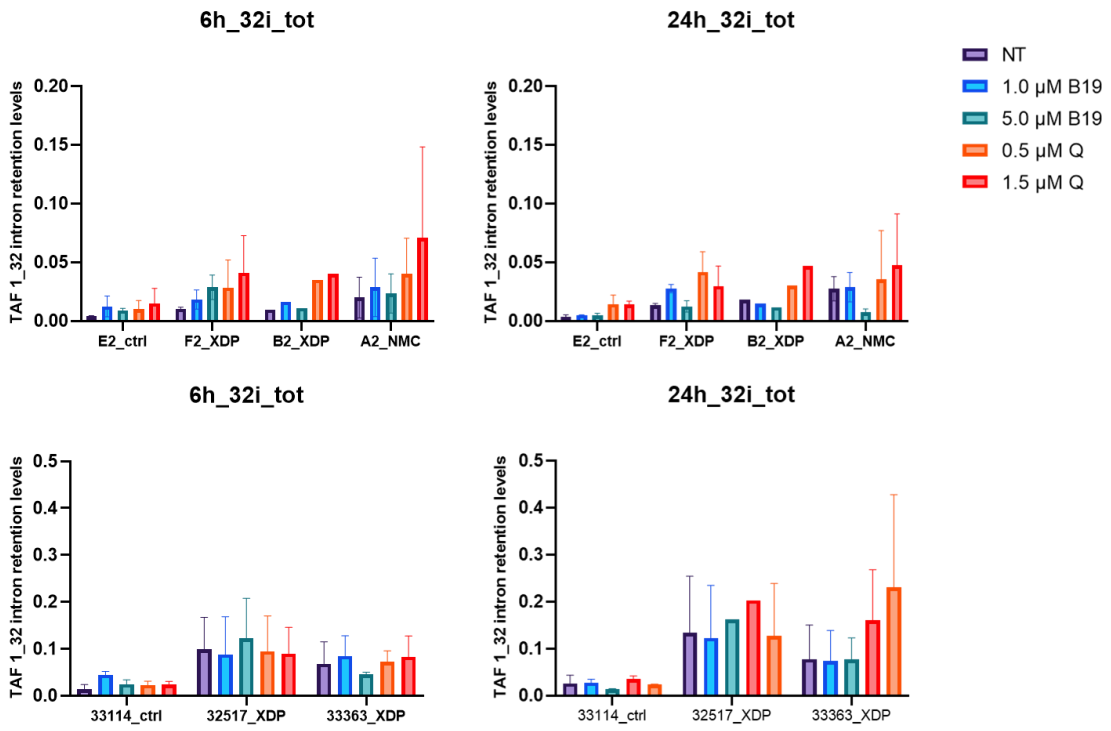


Fig 38: TAF 1 32-intron retention levels increase with G4 ligand treatment. In the upper panel hFib 32 intron retention are reported, in the lower NPCs ones. Quarfloxin increase more 32 intron retention in XDP affected cells, and this effect is more prominent in NPCs compared to hFib.

5.4 Using a small molecule to destabilize G4s within XDP SVA

In the last years, many efforts have been made to find new molecules able to recognize and stabilize G4s[63], in particular as potential antiviral[64] and anticancer[53] innovative treatments. In the case of XDP on the other hand, having a small molecule stabilizing even more the G4s present within the SVA, would not bring benefits but increase damages. XDP SVA G4s anyhow are a good target for XDP treatment: they would need to be unwound. For this reason we decided to try the only molecule shown destabilize G4s: PhPc (Fig 39).

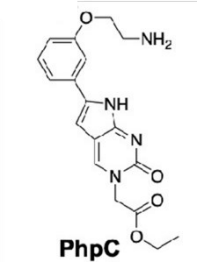


Fig 39: Molecular structure of the G4-destabilizer used in this study: Phpc

This compound can recognize G4 structures and induce their unfolding, but it was never used in cells. Our first approach was to perform circular dichroism experiments to evaluate if the XDP SVA G4s were destabilized by PhPc (Fig. 40).

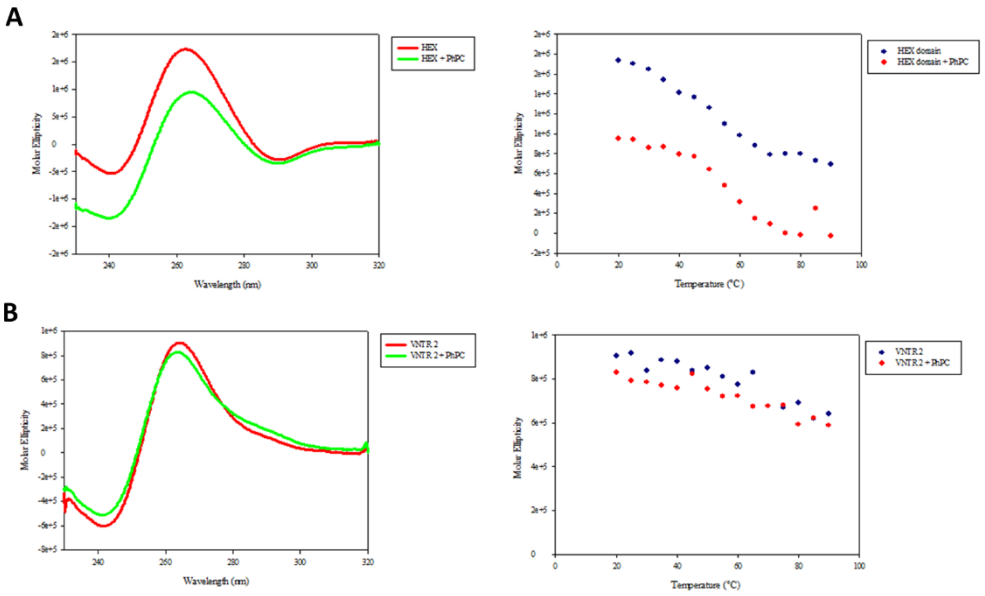


Fig 40: CD melting experiment of two of the G4 forming sequence within SVA. A) CD spectra of Hexameric G4 forming sequence in 100 mM KCl with (green) and without (red) four equivalents of PhPc. Melting curve of the sequence with (red) and without (blue) PhPc. B) CD spectra of VNTR2 G4 forming sequence in 10 mM KCl with (green) and without (red) four equivalent of PhPc. Melting curve of the sequence with (red) and without (blue) PhPc.

The hexameric sequence showed lower melting temperature in the presence of four equivalents of compound, suggesting high destabilization of the G4 structure. In contrast, the VNTR 2 sequence showed very little decrease in the melting temperature, even in low potassium buffer condition, indicating that even if the two G4s have the same topology, for some reason the compound has greater unfolding effect on the hexameric sequence. From our point of view, this was a good result, being the hexameric domain probably more crucial than others in the XDP disease.

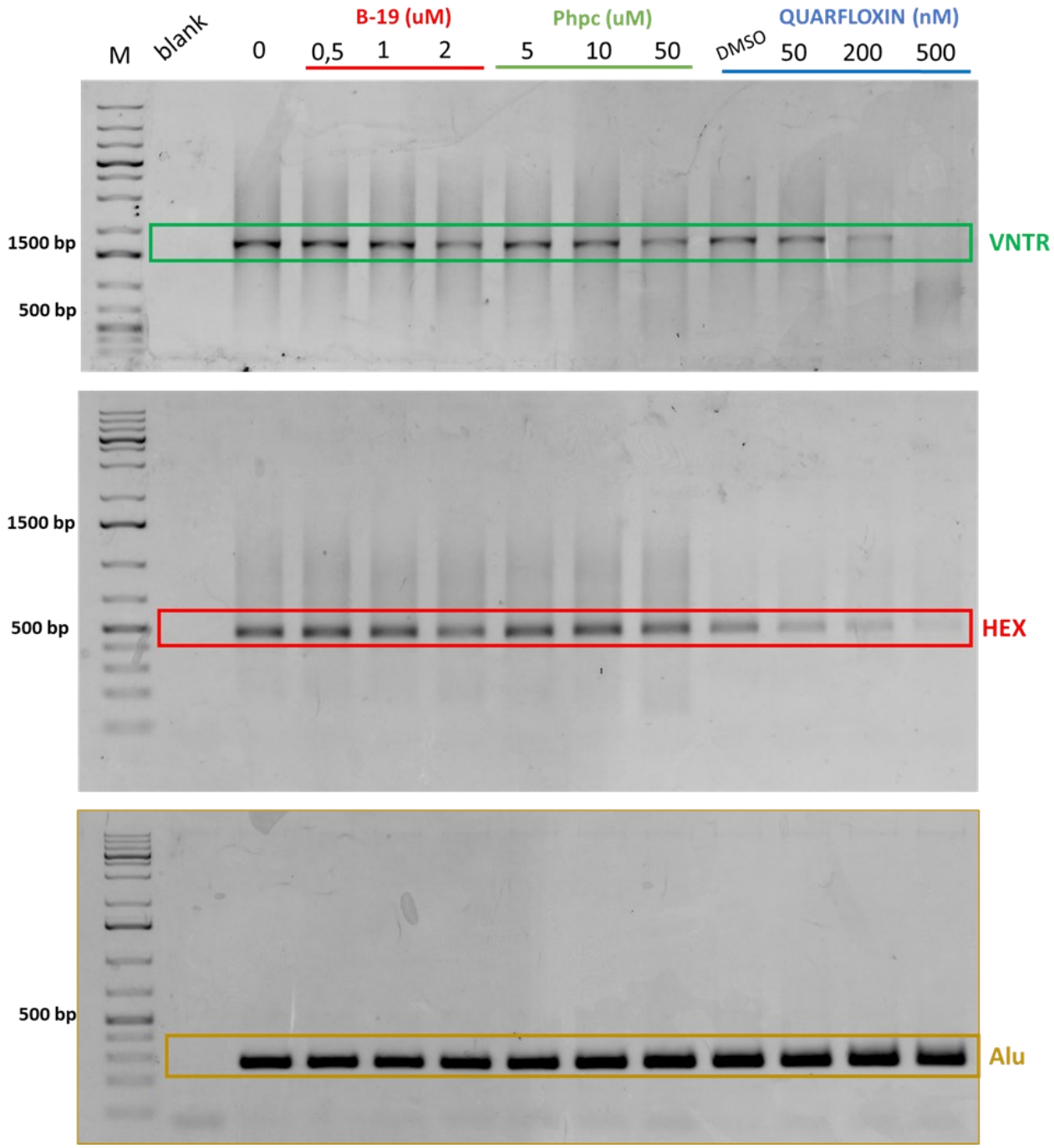


Fig 41: Nested PCR STOP assay with increasing concentration of G4 ligands B-19 and Quarfloxin and G4 destabilizer PhPc. The green box highlight VNTR domain amplification, the red one Hexameric repeat domain, the brown one the Alu domain.

We repeated the Nested PCR STOP assay to assess if PhPc was able to increase the amplification of the VNTR and Hexameric domain (Fig 41). Unfortunately, increasing the concentration of PhPc, we observed less amplification of the SVA G4 domains, as obtained with G4 stabilizing ligands. No effect was evident in Alu amplification, suggesting that there is a preference of the compounds towards G4 regions. It is possible that the G4s are destabilized but the compound remains attached to the DNA, thus impairing the enzyme processivity and resulting in less amplification. From these data we could not define if PhPc was able to destabilize XDP SVA G4s, and thus we treated XDP hFib with 20 uM of PhPC for 24 h and measured *TAF1* levels.

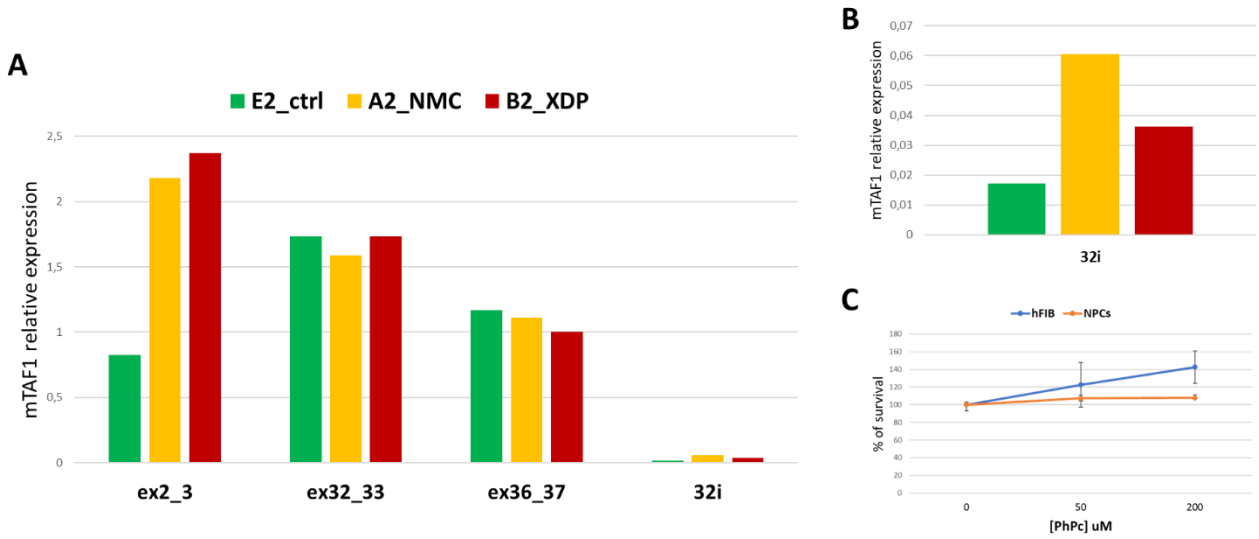


Fig 42: Cell treatment with G4 destabilizer PhPc. A) RT-qPCR on hFib 24h treated with 20uM PhPc. There is an increase of first exons expression but only in SVA carrier cells. The other exons are not affected. B) Zoom in on TAF1 32intron levels. Upon treatment it increases but only in SVA carrier cells, like the other G4ligands treatment. C) Citotoxity assay with PhPc on XDP affected hFib and NPCs.

We thought that even if our *in vitro* data were ambivalent, if the compound was able to destabilize G4s within the XDP SVA in cells, we would see an increase on *TAF1* levels, and a decrease of the *TAF1* intron 32 retention, because the G4s would be less stable and easier to be overcome by RNA Pol II (Fig. 42A-B). Before treating the cells, we performed a citotoxity test (Fig. 42C) and we found that even at very high concentrations, the compound was not cytotoxic in both hFib and NPCs. We would have expected some citotoxity, because PhPc is not specific, so in theory it would destabilize all G4s in the cell genome, generally affecting the transcription process. We next treated hFib with 20 µM for 24 hours. Also in this case we observed the same trend as that of other G4 ligands, i.e. an increase in transcription but only for first exons, and lower expression of

the last exons only in XDP cells. Moreover *TAF1* intron 32 levels were increased compared to ctrl sample. We concluded that even if PhPc was able to destabilize XDP SVA G4s, the final result is the same as any another G4 ligands, probably because the compound still remains attached to the target.

6. Discussion

Our *in vitro* data are clear and neat: G4s form within XDP SVA *in vitro* and they are very stable. The role of G4s in the regulation of SVA transposable elements has never been investigated before in cells. The effects of G4s within XDP SVAs could be manifold. Recently it was demonstrated that G4s positively regulate transcription. In particular, when they are folded in a promoter region, they recruit transcription factors, enhancing gene transcription[29]. From this point of view the hexameric domain could act as promoter for the SVA entity, activating its transcription[65]. To further corroborate this hypothesis a transcriptome study need to be performed to find if there is a correlation between SVA-G4 positive entities and highly transcribed SVAs. Unfortunately being very repetitive the NGS sequencing technology that we used for this study is not enough precise to ascribed the recovered reads to a specific SVA entity. Still the hypothesis G4s as modulator of SVA retrotransposon activity remain intriguing.

We cannot exclude an epigenetic cause of action. Infact SVAs are normally highly silenced from the cells, and they are not located in gene coding region such as TAF1 gene in XDP patients. In this case TAF1 is highly transcribed in cells meaning that the region in which the retrotransposon is insert is very accessible chromatine region. For this reason the typical silencing epigenetic marker of SVAs such as H3K9me3 and DNA methylation could be in conflict with the open chromatin state of TAF1 region. We are planning to investigate also the chromatin state of the region by CUT&Tag experiment mapping the histone modifications in and around TAF1.

Our data suggest that many G4s can fold within the XDP SVA and their G4 structure is very stable in vivo. In cells G4s within XDP SVA are folded in the hexameric domain and this could lead to a stalling of RNA-pol II on *TAF1* 32i, if many G4s are present and are not unwound, they could impair RNA-pol-II activity processivity on the *TAF1* gene[31] (Fig. 43). The presence of an RNA-pol-II enzyme on XDP SVA could also lead to the stalling of RNA-pol-II actively transcribing the *TAF1* gene, leading to a decrease in full length *TAF1* transcript and the increase of alternatively spliced TAF1 transcript with 32-intron retention. Those alternative isoforms of *TAF1* would accumulate and aggregate if it not degraded by cells. Those aberrant aggregates might form stress granules within the nucleus causing cell distress[66], mostly in neuron-like cells. This toxic mechanism is common to other neurodegenerative disease like frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS), in which a long non coding RNA full of G4s, originitating from an intronic G4C2 hexanucleotide repeat expansion (HRE) within C9orf72, was identified as one of the toxic

mechanism that leads to the disease[67]. We think that a similar mechanism could be present also in XDP and we are currently expanding our experiments to deeply understand what roles can be played by G4s in XDP disease and in SVA retrotransposons.

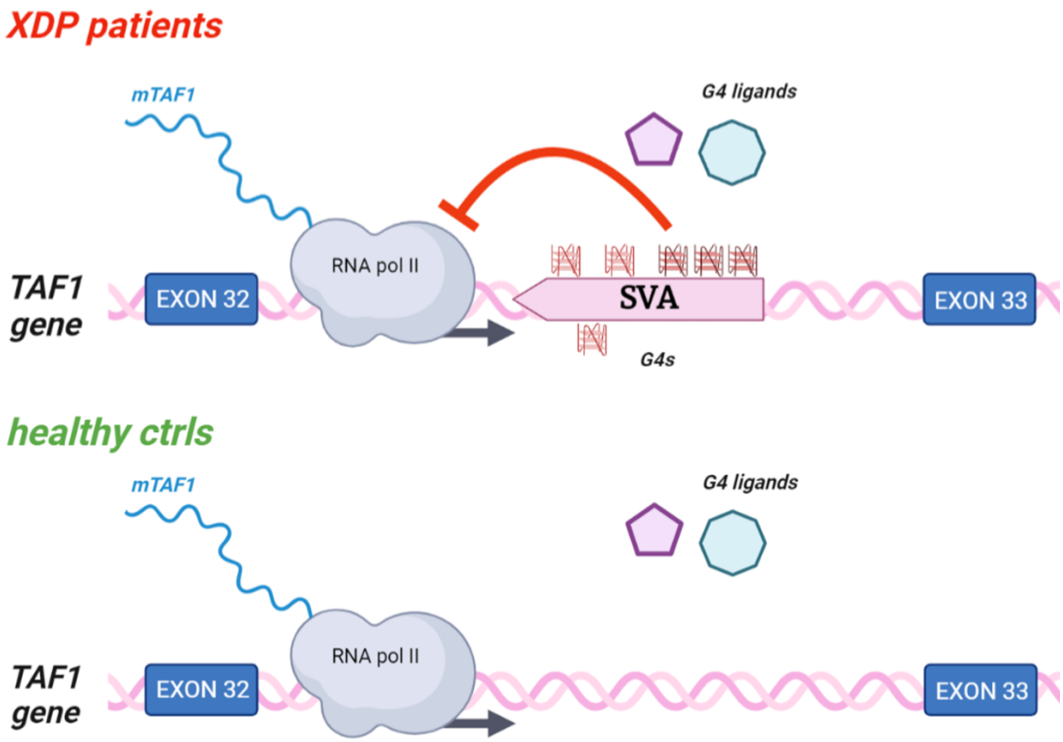


Fig 43: Our proposed model: XDP SVA adopts G4 structures in cells that delay the RNA polymerase actively transcribing TAF1 gene and inducing alternatively sliced forms of mTAF1 that are increased with G4 ligand treatment. In healthy cells that do not display SVA insertion RNA polymerase is not affected even by G4 ligan treatment.

7. Conclusion

Our data confirm the hypothesis that G4s form and are stable within XDP SVA both *in vitro* and in cells. G4s are present in XDP fibroblasts in the SVA hexameric domain and if stabilized they affect *TAF1* transcription. To our knowledge this is the first time that a direct correlation between G4s and SVA is found in living cells. Our results open many new questions that we need to answer. It is fundamental to understand if G4s play an active role in SVA silencing, this could help understand the biology of this class of TE which are poorly studied. Moreover, G4s within SVA can have an impact in nearby gene such in the XDP case. This well correlates with other kind of pathologies in which a TE is involved affecting nearby genes and provides new insights into the study of XDP pathology concerning the SVA retrotransposon. SVA G4s could be new therapeutic targets for a pharmacological treatment of the XDP disease

8. Supplementary information

8.1 XDP SVA sequence with QGRS analyses

AB191243.1:261301-264529 Homo sapiens DNA, transposon SVA sequence, 5'-3'

GTTCCATTGTGTGGTTGTACCAGCGTTTGTTCAAACATTTACTCATTGAAGGACATTCAGATTTTCACTATTTTGCCATTGTGAGTAAAGCTGCTGTGAGCAATCATAGACAG
GGTTTTGGATGAGTTTAATTTTTCATTTCTTTGGGATAAATGCCCAAGAATAGAATGGTATATGTTTAGTTTTACAAGACACGGCACTATTTCATTTTTTTTTTTTTTTTTTTT
TTTTTTTTTATTTTTTTTTTATTTTTTTTTTAATTTATTTTTTTATTGATAATTCTTGGGTGTTTCTCACAGAGGGGGATTTGGCAGGGTCATGGGACAATAGTGGAGGGAAGG
TCAGCAGATAAACAAGTGAACAAAGGTCTCTGGTTTTCCTAGGCAGAGGACCCTGCGGCCTTCCGCAGTGTTTGTGTCCCTGATTACTTGAGATTAGGGATTGGTGATGAC
TCTTAACGAGCATGCTGCCTTCAAGCATCTGTTTAACAAAGCACATCTTGCACCGCCCTTAATCCATTTAACCCTGAGTGGACACAGCACATGTTTCAGAGAGCACAGGGTT
GGGGGTAAGGTCACAGATCAACAGGATCCCAAGACAGAGGAATTTTTCTTAGTGCAGAACAAAATGAAAAGTCTCCCATGTCTACTTCTTTCTACACAGACACGGCAACC
ATCCGATTTCTCAATCTTTTCCCCGCCTTTCCCGCCTTTCTATTCCACAAGGCCGCCATTGTCATCCTGGCCCGTTCTCAATGAGCTGTTGGGCACACCTCCCAGACCGGGTG
GTGGCTGGGCAGAGGCGCCCCTCACCTCCCGGACA**GGGCGGCTGGCCGGGCGGGGGGG**GCTGACCCCCCCCACCTCCCTCCCGGACGGGGCGGCTGGCTGGGCAGA
GGGGCTCCTCACTTCCCAGTAGGGGCAGCCGGGCAGAGGCGCCCCTCACCTCCCGGACGAGGCGGCTGGCCGGGCGTGGGGCTGACACCCCCACCTCCCTCCCGGA
CG**GGGCGGCTGGCCGGGCGGGGGG**CTGACCCCCCCACCTCCCTCCCGGACGAGGCGGCTGGCCGGGCGTGGGGCTGACACCCCCACCTCCCTCCCGGACA**GGGCGGC**
TGGCCGGGCGGGGGGCTGACCCCCCCACCTCCCTCCCGGATGGGGCGGCTGGTCGGGCGGGGGGCCGACCCCCCCACCTCCCTCCCGGACGGGGCGGCTGGCCGGGC
AGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCCGGGCAGAGGCGCCCCTCACCTCCCAGACGGGGCGGCTGGCCGGGCGGAGGGCTGACCCCCCCACCTCCCTCCCG
GACA**GGGCGGCTGGCCGGGCGGGGGGC**TGACCCCCCCACCTCCCTCCCGGACGGGGCGGCTGGCCTGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCCGGGC
AGAGGCGCCCCTCACCTCCCAGACG**GGGCGGCTGGCCGGGCGGGGGGG**CTGACCCCCCCACCTCCCTCCCGGACGGGGCGGCTGGCCAGGCGGGGGGCTGACCCCCCC
ACCTCCCTCCCGGACG**GGGCGGCTGGCCGGGTGGGGGGG**CTGACCCCCCCATCTCCCTCCCGGACGGGGTGGCTGGCCGGGCTGAGGGGCTCCTCACTTCCCAGTAGG
GGCGGCCGGGCAGAGGCGCCCCTCACCTCCCGGACG**GGGCGGCTGGCCGGGCGGGGGGG**CTGACCCCCCCACCTCCCTCCCGGATGGCACGGCTGGCCGGGCGGGGG
GGCTGACCCCCCACCTCCCTCCCGGATGGGGCGGCTGGCCGGGTGGGGGGCTGACCCCCCCCCACCTCCCTCCCGGACGGGGTGGCTGCTGGGCGGAGATGCTCCTCA
CTTCCCAGATGGGGTGGCTGCCGGGCGGAGAGGCTCCTCACTTCTCAGACGGGGCAGCTGCCGGGCGGAGGGGCTCCTCACTTCTCAGACGGGGTGGTTGCCAGGCA
GAGGGTCTCCTCACTTCTCAGACGGGGCGGCCGGGCAGAGACGCTCCTCACCTCCCAGACGGGGTCTCGGCCGGGCAGAGGCGCTCCTCACATCCCAGATGGGGCGGC
GGGGCAGAGGCGCTCCCCACATCTCAGACGATGGGCGGCCGGGCAGAGACGCTCCTCACTTCCTAGATGTGATGGCGGCCGGGAAGAGGTGCTCCTCACTTCCTAGATG
GGATGGCGGCCGGGCGGAGACGCTCCTCACTTTCCAGACTGGGCAGCCAGGCAGAGGGGCTCCTCACATCCCAGACGATGGGCGGCCAGGCAGAGACACTCCTCCCTT
CCCAGACGGGGTGGCGGCCGGGCAGAGGCTGCAATCTCGGCACTTTGGGAGGCCAAGGCAGGCGGCTGGGAGGTGTGGGTTGTAGTGAGCCGAGATCACGCCACTG
CACTCCAGCCTGGGCACCATTGAGCACTGAGTGAACGAGACTCCGTCTGCAATCCCGGCACCTCGGGAGGCCGAGGTTGGCGGATCACTCGCGGTTAGGGGCTGGAGA
CCGGCCCGGCCAACACAGCGAAACCCTGTCTCCACCAAAACCAGTCAGGCGTGGCGGCGCGTGCCTGCAATCGCAGGCACTCGGCAGACTGAGGCAGGAGAATCAGGC
AGGGAGGATGCAGTGAGCCGAGATGGCAGCAGTACAGTCCAGCTTTGGCTCCGCATGAG**AGGGAGACCGTGGGGAGAGGGAGAGGG**AGG**AGAGGGAGAGGGAGA**
GGGAGAGGGAGA**GGGAGAGGGAGAGGGAGAGGG**AGAGGGAGA**GGGACAGGGAGAGGGAGAGGGAGA**GGGAGA**GGGAGAGGGAGAGGGAGAGGGAGA**GCT
ATTTCATTTTTTTTTTTTCCTATTTCATTTTTTTTTTTCCACATCAGATGTGGAAAAAAAATGTACTGGCATTGTAATAAGGCTTGAGGGAGGCACATCTCACACATGAACGT
GAAAACCCAATGTCGTCACACTTACGTCATCATAAGCTTATGAACTACAAAAAGATTGTGCCACACTTTTGGAGTGTCTGTACCATTTTACATTCCCACCAGCAATGTATGTG
ATCCAGTTTCTCTGCAGCATTCACTGTTACCACTTTTTAAAATTTTAGCTGTTCTTCTATTGTGATTTTACTTGGCATTTTTCTTTTTTTAAATTAAAAACATTTTCTTGCTCTCG
GTCTTAAAATTATTTCCTCAAATACATTGTGTTTTTCTAATGGCTAATGATGTTGGGCATCTTTTCATGTG

AB191243.1:261301-264529 Homo sapiens DNA, transposon SVA sequence, 3'-5'

CACATGAAAAGATGCCCAACATCATTAGCCATTAGAAAAACACAATGTATTTGAGGAAATAATTTTAAGACCGAGAGCAAGAAAATGTTTTTAATTTAAAAAAAGAAAAAT
GCCAAGTAAAATCACAATAGAAGAACAGCTAAAATTTTAAAAAGTGGTAACAGTGAATGCTGCAGAGAAACTGGATCACATACATTGCTGGTGGGAATGTAAAATGGTAC
AGACACTCCAAAAGTGTGGCACAATCTTTTTGTAGTTCATAAGCTTATGATGACGTAAGTGTGACGACATTGGGTTTTCACGTTCATGTGTGAGATGTGCCTCCCTCAAGCC
TTATTACAATGCCAGTACATTTTTTTTCCACATCTGATGTGGAAAAAAAAAAAAATGAAATAGGAAAAAAAAAAAATGAAATAGCTCTCCCTCTCCCTCTCCCTCTCCCTCTCC
CTCTCCCTCTCCCTCTCCCTGTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCTCCCTCTCCCTCTCCCCACGGTCTCCCTCTCATG
CGGAGCCAAAGCTGGACTGTACTGCTGCCATCTCGGCTCACTGCATCCTCCCTGCCTGATTCTCCTGCCTCAGTCTGCCGAGTGCCTGCGATTGCAGGCACGCGCCGCCAC
GCCTGACTGGTTTTGGTGGAGACAGGGTTTCGCTGTGTTGGCCGGGCCGGTCTCCAGCCCCTAACCGCGAGTGATCCGCCAACCTCGGCCTCCCGAGGTGCCGGGATTG
CAGACGGAGTCTCGTTCACTCAGTGCTCAATGGTGCCCAGGCTGGAGTGCAGTGGCGTGATCTCGGCTCACTACAACCCACACCTCCCAGCCGCCTGCCTTGGCCTCCCA
AAGTGCCGAGATTGCAGCCTCTGCCCGGCCGCCACCCCGTCTGGGAAGGGAGGAGTGTCTCTGCCTGGCCGCCCATCGTCTGGGATGTGAGGAGCCCCTCTGCCTGGCT
GCCCAGTCTGGAAAGTGAGGAGCGTCTCCGCCCGGCCGCCATCCCATCTAGGAAGTGAGGAGCACCTCTTCCCGGCCGCCATCACATCTAGGAAGTGAGGAGCGTCTCT
GCCCGGCCGCCCATCGTCTGAGATGTGGGGAGCGCCTCTGCCCCGCCGCCCCATCTGGGATGTGAGGAGCGCCTCTGCCCGGCCGAGACCCCGTCTGGGAGGTGAGGA
GCGTCTCTGCCCGGCCGCCCCGTCTGAGAAGTGAGGAGACCCTCTGCCTGGCAACCACCCCGTCTGAGAAGTGAGGAGCCCCTCCGCCCGGCAGCTGCCCCGTCTGAGA
AGTGAGGAGCCTCTCCGCCCGGCAGCCACCCCATCTGGGAAGTGAGGAGCATCTCCGCCCAGCAGCCACCCCGTCCGGGAGGGAGGTGGGGGGGGGTCAGCCCCCCA
CCCGGCCAGCCGCCCCATCCGGGAGGGAGGTGGGGGGTCAGCCCCCCCGCCCGGCCAGCCGTGCCATCCGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCA
GCCGCCCCGTCCGGGAGGTGAGGGGCGCCTCTGCCCGGCCGCCCCTACTGGGAAGTGAGGAGCCCCTCAGCCCGGCCAGCCACCCCGTCCGGGAGGGAGATGGGGG
GGTCAGCCCCCCCACCCGGCCAGCCGCCCCGTCCGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCTGGCCAGCCGCCCCGTCCGGGAGGGAGGTGGGGGGGTCAGC
CCCCCGCCCGGCCAGCCGCCCCGTCTGGGAGGTGAGGGGCGCCTCTGCCCGGCCGCCCCTACTGGGAAGTGAGGAGCCCCTCTGCCAGGCCAGCCGCCCCGTCCGGGA
GGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCTGTCCGGGAGGGAGGTGGGGGGGTCAGCCCTCCGCCCGGCCAGCCGCCCCGTCTGGGAGGTGAGG
GGCGCCTCTGCCCGGCCGCCCCTACTGGGAAGTGAGGAGCCCCTCTGCCCGGCCAGCCGCCCCGTCCGGGAGGGAGGTGGGGGGGTCGGCCCCCCGCCCGACCAGCC
GCCCCATCCGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCTGTCCGGGAGGGAGGTGGGGGTGTCAGCCCCACGCCCGGCCAGCCGCCTCGTCC
GGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCTGTCCGGGAGGGAGGTGGGGGGGTCAGCCCCCCTGCCCGGCCAGTGGCCCCGTCCGGGAG
GTGAGGGGCGCCTCTGCCCGGCTGCCCCTACTGGGAAGTGAGGAGCCCCTCTGCCCAGCCAGCCGCCCCGTCCGGGAGGGAGGTGGGGGGGGTCAGCCCCCCCCGCC
CGGCCAGCCGCCCTGTCCGGGAGGTGAGGGGCGCCTCTGCCCAGCCACCACCCGGTCTGGGAGGTGTGCCCAACAGCTCATTGAGAACGGGCCAGGATGACAATGGCG
GCCTTGTGGAATAGAAAGGCGGGAAAGGCGGGGAAAAGATTGAGAAATCGGATGGTTGCCGTGTCTGTGTAGAAAGAAGTAGACATGGGAGACTTTTCATTTTGTTCT
GCACTAAGAAAAATTCCTCTGTCTTGGGATCCTGTTGATCTGTGACCTTACCCCCAACCCTGTGCTCTCTGAAACATGTGCTGTGTCCACTCAGGGTTAAATGGATTAAGGG
CGGTGCAAGATGTGCTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTTAAGAGTCATCACCAATCCCTAATCTCAAGTAATCAGGGACACAAACACTGCGGAAGGCC
GCAGGGTCCTCTGCCTAGGAAAACCAGAGACCTTTGTTCACTTGTTTATCTGCTGACCTTCCCTCCACTATTGTCCCATGACCCTGCCAAATCCCCCTCTGTGAGAAACACCC
AAGAATTATCAATAAAAAAATAAATTAAAAAAAAAATAAAAAAAAAATAAAAAAAAAAAAAAAAAAAAAAAAAAAAAATGAAATAGTGCCGTGTCTTGTAAAACTAAACA
TATACCATTCTATTCTTGGGCATTTATCCCAAAGAAATGAAAAATTAAACTCATCCAAAACCCTGTCTATGATTGCTCACAGCAGCTTTACTCACAATGGCAAAATAGTGAAAA
TCTGAATGTCCTTCAATGAGTAAATGTTTGAACAAACGCTGGTACAACCACACAATGGAAC

SI Fig 1: XDP SVA retrotransposon sequence. Each color represents a different domain. In bold unique canonical G4 forming motives are identified with QGRS software. Highlighted the characterized sequences VNTR 1 and VNTR 2 and HEX on the forward strand, VNTR 3

8.2 Table 4-8 Oligonucleotides and primers used

Table 4: Oligonucleotides used for the in vitro experiments

name	Sequence (5'-3')
VNTR 1	ACGGGGCCACTGGCCGGGCAGGGGGGCT
VNTR 2	ACAGGGCGGCTGGCCGGGCGGGGGGCT
VNTR 3	TCCGGGAGGGAGGTGGGGGGGTC
HEX	AGAGGGAGAGGGAGAGGGAGAGGG
HEX fooTprinTing	TTTTTAGAGGGAGAGGGAGAGGGAGAGGGTTTTT
VNTR 1 fooTprinTing	TTTTTACGGGGCCACTGGCCGGGCAGGGGGGCTTTTTT
VNTR 2 fooTprinTing	TTTTTACAGGGCGGCTGGCCGGGCGGGGGGCTTTTTT
VNTR 3 fooTprinTing	TTTTTTCCGGGAGGGAGGTGGGGGGGTCTTTTT
VNTR 1 Taq	TTTTTACGGGGCCACTGGCCGGGCAGGGGGGCTTTTTTCTGCATATAAGCAGCTGCTTTTTGCC
VNTR 2 Taq	TTTTTACAGGGCGGCTGGCCGGGCGGGGGGCTTTTTTCTGCATATAAGCAGCTGCTTTTTGCC
VNTR 3 Taq	TTTTTTCCGGGAGGGAGGTGGGGGGGTCTTTTTCTGCATATAAGCAGCTGCTTTTTGCC
HEX_TAQ	TTTTTTAGAGGGAGAGGGAGAGGGAGAGGGTTTTTCTGCATATAAGCAGCTGCTTTTTGCC
No G4 Taq	TTGTCGTTAAAGTCTGACTGCGAGCTCTCAGATCCTGCATATAAGCAGCTGCTTTTTGCC
Primer Taq	CTGCATATAAGCAGCTGCTTTTTGCC

Table 5: PCR and NESTED PCR STOP assay primers

name	Sequence (5'-3')	PCR protocol
SVA - 16153 F	GTTCCATTGTGTGGTTGTACCAGCGTTTGTTC	94 °C 2 min; 30× (98 °C 10 s, 68 °C 3min 30 s); hold at 8 °C
SVA - 19345 R	CACATGAAAAGATGCCCAACATCATTAGCCATTAG	
SVA HEX F	AGCAGTACAGTCCAGCTTTGGC	94 °C 2 min; 30× (98 °C 10 s, 68 °C 20 s); hold at 8 °C
SVA HEX R	CTCAAGCCTTATTACAATGCCAGT	
VNTR 1 for	AATCTTTTCCCCGCCTTTCC	94 °C 2 min; 30× (98 °C 10 s, 68 °C 1 min 45 s); hold at 8 °C
VNTR 1 rev	TCACTACAACCCACACCTCC	
SINE 1 for	ATAGTGGAGGGAAGGTCAGC	94 °C 2 min; 30× (98 °C 10 s, 60 °C 15 s, 68 °C 30s); hold at 8 °C
SINE 1 rev	GTGCCCAACAGCTCATTGAG	
Alu 1 for	GGCACCATTGAGCACTGAG	94 °C 2 min; 30× (98 °C 10 s, 55 °C 15 s, 68 °C 25s); hold at 8 °C
Alu 1 rev	CCACGGTCTCCCTCTCATG	

Table 6: SVA RT-qPCR primer

Name	Sequence (5'-3')
SVA_F_02 f	TTCTCACAGAGGGGGATTTG
SVA_F_02 r	ATCAGGGACACAAACACTGC
SVA_F_05 f	GAGGGAAGGTCAGCAGATAAAC
SVA_F_05 r	TCAAGTAATCAGGGACACAAACA
SVA_F_06 f	GAGATTAGGGATTGGTGATGACTC
SVA_F_06 r	CTGTGTCCACTCAGGGTTAAAT

Table 7: G4-ChIP-qPCR primers

name	Sequence (5’-3’)
HEX_01_F	GAGGGAGAGGGAGA
HEX_01_R	CGTTCATGTGTGAGATG
HEX_01_p	FAM-CCTCAAGCCTTATTACAATGCCAGT-TAMRA
MAZ_F	ACTCAGCGCAGGATTGTAAATA
MAZ_R	CCTCATGCTTCGGCTTCC
MAZ_p	FAM-TGCGTCCTGCAGGCCACCGTCCT-TAMRA
ESR1_F	GAAACAGCCCCAAATCTCAA
ESR1_R	TTGTAGCCAGCAAGCAAATG
ESR1_p	FAM-AGTGGCACCCAGACTTGATGGCCGAC-TAMRA
RPA3_F	CGGAAGTTGACAGATACAGGG
RPA3_R	GATCGCAGAAAGGTAGTCTCAG
SPRED2_F	AACAGGAGGAGGAAGTAGGG
SPRED2_R	TTTCGGTCGCAAGTAGGAAG
GAPDH_F	GCTACTAGCGGTTTTACGGGCG
GAPDH_R	TGCGGCTGACTGTCGAACAGG
TMCC1_F	GTGGTACACTGCCTACAGTATT
TMCC1_R	GTATAACGCCTGGGCTATGT

Table 8: RT-qPCR Taqman primers

name	Sequence (5’-3’)
TAF1_2-3_F	GACTGACGGTGCCTTGGT
TAF1_2-3_R	GTCTGAATAGTCCACAGCATCTTCT
TAF1_2-3_p	FAM-ACCCACCCTTCATCATTT-TAMRA
TAF1_32-33_F	ACCTTATTCTGGCCAACAGTGTT
TAF1_32-33_R	ACAATCTCCTGGGCAGTCTTAGTAT
TAF1_32-33_p	FAM-ACTCTCAGGTCCATTATAC-TAMRA
TAF1_36-37_F	GGAGTGATGAAGAAGGAG
TAF1_36-37_R	GGTTGTTTGGGTCTTATTC
TAF1_36-37_p	FAM-CCACATCAGAGTCACTTCCACT-TAMRA
TAF1_32i_F01	CAAGCACAAGTATCAGAG
TAF1_32i_R01	CGGATTCATGGAAAGAAA
TAF1_32i_p01	FAM-TGAGAAATACCCACCATTATACTTAACAC-TAMRA

9. Acknowledgments

This Phd was very though! So I need to thank all the people that help me scientifically and psicologically.

First of thanks to all the people of the Richter's lab: to my supervisor Prof. Sara Richter for giving me an exciting PhD project and all the XDP girls (Ilaria pICCOLA , Marianna (Lucielle), Manu and Irene) without you this project would not be as great as it is!

I need to thank my supervisor also because she gave me the opportunity to go abroad and to improve myself. Many thanks to Prof Gunnar Schotta to let me stay in his lab a lot more than everyone expected. Thank you Filippo because you are the one who thought me how the work is done in the most efficient and thoughtful way. Thank you for listening to every single question that I have and thank you for never saying that I was not good enough, but encouraging me to improve my techniques and to try another time.

Thanks to all Schotta lab members and the wonderful people of BMC. Thank you Andrea for teaching me the beauty of bioinformatics, for the scripts and for saving my life when I was drunk. Thank you Irina and Angela for partying with me even when party was hard such as during the Covid pandemic. Thank you Eugenio and Sophia for you kind advices. Thank you Isabella and Viola for being the friends that I needed.

And finally a special thanks to Marco. You always have the right words that helped me during this three years. Thank you for all the love and care. This achievement is also yours.

10. References

1. Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172**, 897-909.e21 (2018).
2. Bragg, D. C. *et al.* Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E11020–E11028 (2017).
3. Lee LV, Pascasio FM, Fuentes FD, Viterbo GH. Torsion dystonia in Panay, Philippines. Adv Neurol. 1976;14:137-51. PMID: 941767.
4. Evidente VGH. X-Linked Dystonia-Parkinsonism. 2005 Dec 13 [updated 2018 Feb 15]. In: Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJH, Gripp KW, Amemiya A, editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2022. PMID: 20301662.
5. Kaji R, Goto S, Tamiya G, Ando S, Makino S, Lee LV. Molecular dissection and anatomical basis of dystonia: X-linked recessive dystonia-parkinsonism (DYT3). J Med Invest. 2005 Nov;52 Suppl:280-3. doi: 10.2152/jmi.52.280. PMID: 16366515.
6. Jamora, R. D. G., Diesta, C. C. E., Pasco, P. M. D. & Lee, L. V. Oral pharmacological treatment of X-linked dystonia parkinsonism: successes and failures. *Int. J. Neurosci.* **121 Suppl 1**, 18–21 (2011).
7. Evidente, V. G. H. *et al.* Phenotypic and molecular analyses of X-linked dystonia-parkinsonism ('lubag') in women. *Arch. Neurol.* **61**, 1956–1959 (2004).
8. Ito, N. *et al.* Decreased N-TAF1 expression in X-linked dystonia-parkinsonism patient-specific neural stem cells. *Dis. Model. Mech.* **9**, 451–462 (2016).
9. TAF1, associated with intellectual disability in humans, is essential for embryogenesis and regulates neurodevelopmental processes in zebrafish.
10. Kwon, Y.-J. *et al.* Structure and Expression Analyses of SVA Elements in Relation to Functional Genes. *Genomics Inform.* **11**, 142–148 (2013).
11. Campion, L. N. *et al.* Tissue-specific and repeat length-dependent somatic instability of the X-linked dystonia parkinsonism-associated CCCTCT repeat. *Acta Neuropathol. Commun.* **10**, 49 (2022).

12. Trinh, J. *et al.* Mosaic divergent repeat interruptions in XDP influence repeat stability and disease onset. *Brain J. Neurol.* awac160 (2022) doi:10.1093/brain/awac160.

13. Raiz, J. *et al.* The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **40**, 1666–1683 (2012).

14. Wang, H. *et al.* SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.* **354**, 994–1007 (2005).

15. Gianfrancesco, O. *et al.* The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome. *Int. J. Mol. Sci.* **20**, 5977 (2019).

16. Bantysh, O. B. & Buzdin, A. A. Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. *Biochem. Biokhimiia* **74**, 1393–1399 (2009).

17. Strichman-Almashanu, L. Z. *et al.* A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Res.* **12**, 543–554 (2002).

18. Fasching, L. *et al.* TRIM28 Represses Transcription of Endogenous Retroviruses in Neural Progenitor Cells. *Cell Rep.* **10**, 20–28 (2015).

19. Kejnovsky, E., Tokan, V. & Lexa, M. Transposable elements and G-quadruplexes. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* **23**, 615–623 (2015).

20. Savage, A. L., Bubb, V. J., Breen, G. & Quinn, J. P. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol. Biol.* **13**, 101 (2013).

21. Quinn, J. P. & Bubb, V. J. SVA retrotransposons as modulators of gene expression. *Mob. Genet. Elem.* **4**, e32102 (2014).

22. Savage, A. L. *et al.* An evaluation of a SVA retrotransposon in the FUS promoter as a transcriptional regulator and its association to ALS. *PloS One* **9**, e90833 (2014).

23. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat. Rev. Mol. Cell Biol.* **21**, 459–474 (2020).

24. Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33**, 2908–2916 (2005).

25. Shen, J. *et al.* Promoter G-quadruplex folding precedes transcription and is controlled by chromatin. *Genome Biol.* **22**, 143 (2021).

26. Tan, J. & Lan, L. The DNA secondary structures at telomeres and genome instability. *Cell Biosci.* **10**, 47 (2020).

27. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182–186 (2013).

28. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–1272 (2016).

29. Lago, S. *et al.* Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat. Commun.* **12**, 3885 (2021).

30. Antcliff, A., McCullough, L. D. & Tsvetkov, A. S. G-Quadruplexes and the DNA/RNA helicase DHX36 in health, disease, and aging. *Aging* **13**, 25578–25587 (2021).

31. Lejault, P., Mitteaux, J., Sperti, F. R. & Monchaud, D. How to untie G-quadruplex knots and why? *Cell Chem. Biol.* **28**, 436–455 (2021).

32. Mitteaux, J. *et al.* Identifying G-Quadruplex-DNA-Disrupting Small Molecules. *J. Am. Chem. Soc.* **143**, 12567–12577 (2021).

33. Lexa, M. *et al.* Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics* **15**, 1032 (2014).

34. Kikin, O., D'Antonio, L. & Bagga, P. S. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.* **34**, W676–W682 (2006).

35. Makino, S. *et al.* Reduced Neuron-Specific Expression of the TAF1 Gene Is Associated with X-Linked Dystonia-Parkinsonism. *Am. J. Hum. Genet.* **80**, 393–406 (2007).

36. Puig Lombardi, E. & Londoño-Vallejo, A. A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Res.* **48**, 1–15 (2020).

37. Kawarai, T. *et al.* Application of long-range polymerase chain reaction in the diagnosis of X-linked dystonia-parkinsonism. *Neurogenetics* **14**, 167–169 (2013).

38. Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35**, W71–W74 (2007).

39. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).

40. Hänsel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.* **13**, 551–564 (2018).

41. Schmidt, D. *et al.* ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods San Diego Calif* **48**, 240–248 (2009).

42. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

43. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).

44. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).

45. Homer Software and Data Download. http://homer.ucsd.edu/homer/.

46. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).

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

48. Lyu, J., Shao, R., Kwong Yung, P. Y. & Elsässer, S. J. Genome-wide mapping of G-quadruplex structures with CUT&Tag. *Nucleic Acids Res.* **50**, e13 (2022).

49. Teng, Y.-C. *et al.* ATRX promotes heterochromatin formation to protect cells from G-quadruplex DNA-mediated stress. *Nat. Commun.* **12**, 3887 (2021).

50. Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K. & Henikoff, S. Efficient low-cost chromatin profiling with CUT&Tag. *Nat. Protoc.* **15**, 3264–3283 (2020).

51. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

52. Kejnovská, I., Renčiuk, D., Palacký, J. & Vorlíčková, M. CD Study of the G-Quadruplex Conformation. *Methods Mol. Biol. Clifton NJ* **2035**, 25–44 (2019).

53. Asamitsu, S., Obata, S., Yu, Z., Bando, T. & Sugiyama, H. Recent Progress of Targeted G-Quadruplex-Preferred Ligands Toward Cancer Therapy. *Molecules* **24**, 429 (2019).

54. Burger, A. M. *et al.* The G-quadruplex-interactive molecule BRACO-19 inhibits tumor growth, consistent with telomere targeting and interference with telomerase function. *Cancer Res.* **65**, 1489–1496 (2005).

55. Xu, H. & Hurley, L. H. A first-in-class clinical G-quadruplex-targeting drug. The bench-to-bedside translation of the fluoroquinolone QQ58 to CX-5461 (Pidnarulex). *Bioorg. Med. Chem. Lett.* **77**, 129016 (2022).

56. Jamroskovic, J. *et al.* Identification of putative G-quadruplex DNA structures in S. pombe genome by quantitative PCR stop assay. *DNA Repair* **82**, 102678 (2019).

57. Jacobs, F. M. J. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242–245 (2014).

58. Haring, N. L. *et al.* ZNF91 deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters. *Genome Res.* **31**, 551–563 (2021).

59. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–1272 (2016).

60. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

61. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Protoc.* **13**, 1006–1019 (2018).

62. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).

63. Teng, F.-Y. *et al.* G-quadruplex DNA: a novel target for drug design. *Cell. Mol. Life Sci. CMLS* **78**, 6557–6583 (2021).

64. Ruggiero, E., Zanin, I., Terreri, M. & Richter, S. N. G-Quadruplex Targeting in the Fight against Viruses: An Update. *Int. J. Mol. Sci.* **22**, 10984 (2021).

65. Gianfrancesco, O., Bubb, V. J. & Quinn, J. P. SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides* **64**, 3–7 (2017).

66. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).

67. Wang, X. *et al.* C9orf72 and triplet repeat disorder RNAs: G-quadruplex formation, binding to PRC2 and implications for disease mechanisms. *RNA N. Y. N* **25**, 935–947 (2019).