TITLE

G-quadruplex landscape and its regulation revealed by a new antibody capture method

Authors

Subhamoy Datta[1], Manthan Patel[2], Chakkarai Sathyaseelan[3], Divyesh Patel[4], Thenmalarchelvi Rathinavelan[3], Umashankar Singh[1].

1. HoMeCell Lab, Discipline of Biological Engineering, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat 382355, India
2. Centre for Genomics and Child Health, Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK
3. Department of Biotechnology, Indian Institute of Technology Hyderabad, Kandi Campus, Telangana 502285, India
4. Research Programs Unit, Applied Tumor Genomics Program, Faculty of Medicine, University of Helsinki, Biomedicum, Helsinki 00290, Finland

ABSTRACT

Our understanding of DNA G-quadruplexes (G4s) from in vitro studies has been complemented by genome-wide G4 landscapes from cultured cells. Conventionally, formation of G4s is accepted to depend on G-repeats such that they form tetrads. However, genome-wide G4s characterized through high-throughput sequencing suggests that these structures form at a large number of regions with no such canonical G4-forming signatures. New functions, such as regulation of gene expression regulation and replication have been attributed to the genomic G4s. Many G4-binding proteins have been described, but their roles in formation and maintenance of G4s remain unknown. The representation of G4s not bound to proteins in these genomic landscapes remains unaddressed. We have developed a simple variation of the conventional G4-chromatin immunoprecipitation (G4-ChIP) assay to capture G4s in the cells using the 1H6 antibody. A comparison with the conventional G4-ChIP and in vitro DNA pull-down assays shows that our method minimizes false G4 detection. We use this method along with a transfected or in vitro-added control DNA (containing canonical G4-forming sequences) to study the G4 regulatory effects of the CGG repeat-binding protein CGGBP1. We present a reliable G4 landscape and apply the new protocol to discover a functionally relevant regulation of G4s by CGGBP1. Our results show that sharp interstrand transitions of G/C-skew, rich in binding sites for CTCF and CGGBP1, form G4s. Thus, our method demonstrates a sequence property-specific constraints of the nuclear environment that mitigates G4 formation.

KEYWORDS: G-quadruplex, ChIP, *AbC* G4-ChIP, CGGBP1, CTCF, G/C-skew

INTRODUCTION

The containment and maintenance of the genomic DNA inside the human cellular nuclei is a tightly regulated process. Higher order chromatin conformation and chromosomal territories exemplify the manifestation of an orderly containment of the genomic DNA at a broad scale [1,2]. These arrangements selectively allow regulatory interactions between genomic regions separated by long distances or located in trans. It is well established that this orderly arrangement of the genome depends on interactions between DNA and chromatin regulatory proteins, such as CTCF [3] and its co-operators such as CGGBP1 [4–6]. The complexity of such

regulatory proteins and their cognate binding sites in the genome has increased in the course of evolution. Essentially, a significant amount of cellular resources are directed to ensure that the DNA does not adopt conformations randomly [7,8].

The innate nature of cross-strand annealing combined with the presence of repetitive sequences in the human genome makes the DNA prone to adopting multiple conformations randomly. This problem is further exacerbated by the sheer size of the human genome, the restricted nuclear volume and the need to reset the genomic DNA arrangement in the nuclei through replication and transcription [9,10]. Obviously, this challenge could not be addressed just by the regulators of higher order chromatin regulators with sequence-specific DNA-binding. The highly conserved histone proteins, related protein machinery dedicated to packaging DNA into chromatin and single strand DNA-binding proteins exhibit little DNA sequence preference [11]. These mechanisms disallow the existence of nascent DNA, which would be prone to adopting multiple conformations as allowed biochemically and thermodynamically in the nuclear environment. Collectively, these mechanisms constitute constraints of the nuclear environment on the property of genomic DNA to adopt secondary structures.

The DNA in the nucleus is however under a state of flux through various processes that involve strand separation and new strand synthesis. In addition, certain functional regions of the genome, such as transcription enhancers and promoter elements interact facultatively with transcription factors and may evade dense packaging with histone and related proteins in a cell type and developmental state-specific manner. The DNA in these regions, often associated with G/C richness [12,13], becomes amenable to adopt secondary structures. One such secondary structure of DNA that has attracted wide interest is the DNA G4s [8]. The G4s form due to a preference for Hoogsteen base pairing over Watson-Crick base pairing in certain sequence contexts and involves four strands of DNA. The G4s could form from multiple parallel or antiparallel conformations of either a single DNA strand, two complementary strands or even different DNA molecules. The variety of G4s discovered so far indicates that the structure formed is the lowest energy state conformation possible in a certain biochemical context; a process solely dependent on the chemical properties of the DNA, its nucleotide sequence, strandedness and the biochemical environment [14–16]. Much of our understanding of the nature of G4s come from *in vitro* characterization of DNA secondary structures [17,18]. A commonly accepted signature sequence for G4s formation is a tandem repeat of [G_3-$N_{(1-7)}$] four times on a single strand. However, in a complex mix of DNA, such as the human genome, such single units of [G_3-$N_{(1-7)}$] from different molecules could collaborate and form intermolecular G4s [17,18]. It is obvious that formation of secondary structures such as G4s in cellular nuclei would be enhanced if the DNA escapes the constraints of the nuclear environment. Determination of the G4s landscape of the human genome has the potential to elucidate their biological significance.

Adoption of secondary structures by the DNA is a hindrance to genomic DNA usage and maintenance. Hairpin loops and G4s formed by tandem repeats cause polymerase stalling and replication errors [19]. Repetitive sequences prone to formation of secondary quadruplex structures depend on helicases [20] for efficient replication and repair. These energy-dependent secondary structure resolution mechanisms collaborate with the constraints of the nuclear environment to mitigate secondary structure formation. More recently, regulatory functions have been attributed to quadruplex structures [8,11,21,22]. Studies indicate that

regulation of expression of some genes depend on quadruplex formation in their *cis* regulatry regions [23–25]. It remains unclear if the G4s formation in *cis* regulatory regions is a consequence of nucleosome-depletion and relaxation of constraints of the nuclear environment, or, if there are dedicated mechanisms to ensure formation and stabilization of such regulatory quadruplexes. An increasing number of studies have reported an array of G4s-binding proteins [26–30]. However, most of these studies do not distinguish between proteins binding to sequences capable of forming G4s or actual G4s structures. The consequences of protein-G4s interactions remain unknown for many such candidate quadruplex binding proteins. There is evidence that many helicases with RRH domains bind to the G4s and likely resolve the secondary folding of the DNA [31,32]. Hundreds of other proteins have been identified as associated with G4s with no indications of any common quadruplex-specific binding domain [26], such as the RGG domain which has a putative quadruplex binding activity.

While there is overwhelming evidence that G4s affect DNA-protein interactions thereby affecting the epigenome, there is no evidence for a targeted G4s-stabilizing mechanism in the human cells. Most of these G4s-binding proteins seem to have affinity for generally GC-rich DNA and nucleosome depleted regions suggesting that their binding could exert constraints of the nuclear environment on G4s formation. The G4s landscape determination with knockdown of these G4s-binders (and putative G4 regulators) will elucidate the role of these proteins in quadruplex establishment in cells. A comprehensive determination of the G4s landscape and their regulation by quadruplex-binding proteins shall address the non-canonical G4s as well.

Valuable studies have described the genome-wide G4s landscapes in human cells. The canonical [G_3-$N_{(1-7)}$]xn signature is only mildly enriched in these quadruplex pulldown-sequencing assays. Currently we do not know if the large amount of DNA pulled down by antibodies specific for G4s that does not show any [G_3-$N_{(1-7)}$]xn signature is an artifact or not. Intriguingly, there is an unexplained bias for G/C-rich open chromatin and nucleosome depleted regions in G4s landscape datasets and the same G/C-rich regions are predominantly the binding sites for the G4s-binding proteins (unpublished observations derived from data published vide [33]). In addition, tetrads formed by non-G nucleotides also exist interspersed between G tetrads with some even stabilizing the G4s to varying degrees [34–37]. It remains unknown if the G/C-rich and non-G/C-rich G4s are regulated by same or distinct mechanisms. Nevertheless, it remains established that although G/C richness is not an essential feature of quadruplex forming regions, the G/C-rich sequences readily form stable quadruplexes *in vitro* and to a certain extent in the cellular nuclei. Our understanding of the DNA sequence properties associated with G4s formation *in cellulo* depends on the global G4s profiles, mostly generated using the G4-ChIP protocol [38].
Formaldehyde crosslinking, as used in conventional chromatin assay protocols has been adapted in G4-ChIP as well. Although several G4s-binding proteins have been described, there are no known dedicated G4s-binding domains which stabilize the quadruplexes common to them. Formaldehyde crosslinking can efficiently capture only those G4s which are stably bound to proteins. This creates a detection bias against the G4s not bound to proteins. That formaldehyde is a passive DNA denaturing agent [39,40] suggests that protein unbound G4s would be underrepresented in the G4-ChIP seq. On the other hand, an overrepresentation of the [G_3-$N_{(1-7)}$] sequences in the G4-ChIP-seq could be due to the rapid adoption of quadruplex structures by [G_3-$N_{(1-7)}$] sequences post fixation. The melting of DNA

by formaldehyde would render it amenable to forming quadruplexes post fixation in the lysates or during any step before the anti-G4s antibody is added. The antibody would capture the *in vitro* formed quadruplexes and these would be falsely represented in the ChIP-seq results as bonafide quadruplexes *in cellulo*. Such a potential misrepresentation of G4s landscape in the conventional ChIP-seq experiments contaminates our understanding of *in cellulo* G4s landscape with *in vitro* formed quadruplexes.

Thus, multiple interrelated challenges hinder our interpretation of the G4s landscape and its regulation. Does formaldehyde fixation truly capture the *in cellulo* G4s only? Do post-fixation formed quadruplexes *in vitro* get misrepresented as *in cellulo* G4s? Does G4 formation/stabilization depend on some regulatory protein(s)? Are there constraints of the nuclear environment exerted by G4s-binding proteins on quadruplex formation?

In this work, we shed light on these questions. We have developed a G4s ChIP-sequencing protocol using which we stabilize *in cellulo* quadruplexes and eliminate any quadruplexes formed post-fixation *in vitro*. We use this protocol with a control DNA, having the [G_3-$N_{(1-7)}$] signature sequence, to establish quantitatively the extent of constraints of the nuclear environment on quadruplex formation by [G_3-$N_{(1-7)}$] sequences *in cellulo*. Finally, we apply this protocol in cells depleted for CGGBP1, a protein capable of binding to G/C-rich sequences. We describe the most native *in cellulo* G4s landscape and show that CGGBP1 exerts selective constraints of the nuclear environment on G4s formation on sequences with G/C-skew. Enhanced G4s formation upon CGGBP1 depletion appears to enhance CTCF occupancy at non-canonical CGGBP1-regulated CTCF-binding sites.

RESULTS

1. G4 landscapes determined by G4-ChIP from cells or in vitro G4 immunoprecipitations show unexpected similarity

We performed *in vitro* G4 DNA-immunoprecipitations (DNA-IPs) and the conventional G4 chromatin immunoprecipitation [38] using the 1H6 antibody [41,42]. G4 DNA-IP on sonicated HEK293T genomic DNA stored in Tris buffer (pH=8.0) for 7 days at 4˚C was followed by sequencing. The DNA in this sample (called G4-DNA-IP^{nat}) was amenable to form secondary structures. The long-term stable G4s in the sample were precipitated without any disruption. In parallel, the same sample was heat-denatured and snap-chilled to melt away the secondary structures before the immunoprecipitation (called G4-DNA-IP^{denat}). We also performed G4 chromatin immunoprecipitation in HEK293T cells (called G4-ChIP) using a previously described protocol [38]. The immunoprecipitated DNA, along with the respective inputs from the three samples were also sequenced (Table S1). The samples G4-DNA-IP^{nat} and G4-DNA-IP^{denat} shared a common input.

A comparison of these G4 DNA-IPs and G4-ChIP profiles showed that there was a very weak correlation between the G4-DNA-IP^{nat} and its respective input (Fig S1). G4-DNA-IP^{nat} co-clustered with G4-ChIP and the input for G4-ChIP (Fig S1). The G4-DNA-IP^{denat} was distinct from other samples and clustered with the common *in vitro* G4 DNA-IP input (Fig S1). Considering that the inputs were sample-specific non-enriched representations of the genomic DNA fragments, we could conclude that G4-ChIP and G4-DNA-IP^{nat} showed enrichments for G4s with some similarities and were thus co-clustered. It seemed that a

small fraction of G4s captured in the G4-DNA-IP^{nat} were also partially represented in the G4-ChIP.

The 1H6 antibody can capture secondary DNA structures without the canonical [G_3-$N_{(1-7)}$] signature sequences, especially upon denaturation [37]. Quality filtered reads were searched for the presence of G4s-forming [G_3-$N_{(1-7)}$] sequences using *allquads* [18]. As expected, we observed that the [G_3-$N_{(1-7)}$] sequences were enriched in the G4-DNA-IP^{nat} sample whereas the G4-DNA-IP^{denat} sample showed such sequences only at a level expected in the input (Table S2). The G4-ChIP sample, that showed co-clustering with the non-denatured DNA-IP sample (Fig S1), also showed an enrichment for [G_3-$N_{(1-7)}$] sequences compared to the respective input. Similar enrichment of the [G_3-$N_{(1-7)}$] sequences were found in independently reported G4-ChIP using BG4 antibody in HaCaT and human keratinocytes cells (Table S3). The results suggested that some [G_3-$N_{(1-7)}$]-rich sequences, which form quadruplexes *in vitro*, were also unexpectedly represented in the G4-ChIP.

To clearly establish the G4s landscape under constraints of the nuclear environment, it was important to first generate the quadruplex landscape in cells with minimum *in vitro* artifacts. We worked on eliminating the possibility that some sequences, which do not exist as quadruplexes in live cells, adopt quadruplex structures during various stages of the G4-ChIP protocol post-fixation. Such an *in vitro* formation of G4s during the G4-ChIP protocol can generate some artificial similarity between the *in vitro* DNA-IPs and G4-ChIP datasets.

2. A method for capture of G4s genome-wide without in vitro interference

The G4-ChIP protocol relies on fixation of G4s-protein complexes by formaldehyde followed by lysis of the cellular nuclei, sonication and target precipitation using an antibody of interest. The capture of the G4s by this method depends on (i) an assumed interaction of the quadruplexes with proteins required for formaldehyde fixation, and (ii) that the binding of the anti-G4s antibody is not sterically hindered by proteins covalently fixed onto the quadruplexes. Together these factors are expected to create false negatives in the G4 landscape identified by G4-ChIP (Fig 1A). Additionally, the melting effect of formaldehyde on nascent DNA segments [39,40] of the sonicated chromatin creates single-stranded DNA capable of forming false positive G4s. To overcome these challenges, we developed a variation of the G4-ChIP to minimize these false positives and false negatives.

HEK293T cells were pretreated with a buffer containing Triton-X 100 for permeabilization and KCl for maintaining/stabilizing the G4s in their native states (Fig 1B and C). These permeabilized cells, having a dynamic chromatin, were bathed with 1H6 antibody for 2h at room temperature. The antibody-bound G4s were fixed with formaldehyde ensuring that the quadruplexes were captured as antibody-G4s complexes thereby minimizing false negatives. Also, no further addition of the antibody ensured that any quadruplexes formed post-fixation were not captured. We call this new protocol *A*nti*b*ody *C*apture G4-ChIP (*AbC* G4-ChIP).

The DNA captured using *AbC* G4-ChIP was expected to enrich only the nuclear G4s. This sample (called CT^{nuc}) was sequenced (Table S4) and compared against the *in vitro* G4 DNA-IPs (Table S2). The G4-ChIP and G4-DNA-IP^{nat} co-clustered with each other whereas the *AbC* G4-ChIP samples formed an independent cluster (Fig 1D). We observed that the *AbC*

G4-ChIP returns a G4 landscape most different from either of the *in vitro* G4 DNA-IPs (heat denatured or not) and the G4-ChIP. These results suggest that the G4-ChIP captures G4s from the cellular nuclei but with some *in vitro* artifacts. The *AbC* G4-ChIP minimizes this artifact.

A measure of this artifact was derived by analyzing G4 signals captured in these assays at randomly drawn genomic regions (details in the methods). Unlike the regions captured in these assays using the 1H6 antibody (where G4 signals were expected), these randomly selected regions were not expected to be rich in G4 signals. As expected, the G4 signals in these random genomic regions were very low; across all the samples, about 78-83% of the regions contained just 0.001% of all the G4 signals captured (Fig S2). However, strikingly, the largest deviation from low or no signals in these random regions were observed in G4-ChIP and G4-DNA-IP^{nat} (Fig S2). Thus, the formation of G4s at random regions gives rise to artifacts similarly in the G4 DNA-IP^{nat} and G4-ChIP. This formation of G4s at random regions in G4-ChIP likely happens during the antibody incubation step.

Peaks were called on the read sets of all the samples using respective inputs as backgrounds. The peaks that could be called in the four datasets were very different, indicating that the reads were differently distributed (Table S5). In G4-ChIP and *in vitro* G4 DNA-IP^{nat}, the G4 signals were spread out over a larger number of regions and consequently these samples had the lowest numbers of peaks (Table S5). The CT^{nuc} captured a larger number of high frequency G4 signals resulting in a larger number of peaks in CT^{nuc}. Apparently, the highest discretization of G4-signals was observed in *in vitro* G4 DNA-IP^{denat} (Table S5). The concentration of G4 signals resulting in peaks was however due to different reasons. As indicated by the widely different levels of G4s-forming sequence signature in the peak sequences (Table S5), the association of the signature sequence with G4s was predominantly observed *in vitro*. Although G4-ChIP as well the *AbC* G4-ChIP captured G4s without such a dependence on the signature sequences, the *AbC* G4-ChIP was more efficient at capturing discrete G4s signals with the least dependence on the signature sequences.

The peaks were also compared by Jaccard indices. These indices recapitulated the similarity between G4-ChIP and *in vitro* G4 DNA-IP^{nat}. As evident for the CT^{nuc} sample, the *AbC* G4-ChIP minimized this similarity with the *in vitro* G4 DNA-IP^{nat} sample (Fig S3). These results suggested that enriched G4 regions discovered using the G4-ChIP consisted of the endogenous G4 landscape as well as some *in vitro* artifacts. The *AbC* G4-ChIP protocol retains the genuine G4 profile from G4-ChIP and minimizes false positive representation from the *in vitro* artifacts.

We next analyzed the presence of known G4s-forming sequence properties in these peaks. Using *allquads* we identified that only 10% of the CT^{nuc} peaks contained the G3-N1-7-G3 sequences. In comparison, 13% of the G4-ChIP contained this signature sequence (Table S6). The *in vitro* G4 DNA-IPs had the maximum G3-N1-7-G3 contents, upto 22% (Table S6). We also compared the representation of sequence tags in peaks rich or poor in GGG/CCC triplets, a component of the G3-N1-7-G3 signature. Although the G3-N1-7-G3 content was only mildly different between CT^{nuc} and G4-ChIP, we found that the GGG/CCC-containing regions were represented by a higher read density in the latter. This comparison was performed by calculating the GGG/CCC contents per peak weighted by the read-count

(product of total GGG/CCC content and the sequence tag density per peak). Consistent with the *allquads* analysis, we found that the CT^{nuc} peaks were significantly poor in GGG/CCC triplet content than the *in vitro* G4 DNA-IPs and the G4-ChIP (p-values <0.0001 for two-way comparisons) (Fig 1E). In a recapitulation of the similarities between the G4-ChIP and *in vitro* G4 DNA-IP^{nat}, the weighted GGG/CCC contents were only marginally different (p-value 0.0323).

These results showed that the G4-ChIP captures the G4s with some similarities to the *in vitro* G4 DNA-IPs. The G4s landscape determined by the *AbC* G4-ChIP minimizes similarities with the *in vitro* G4s landscape. To test if *AbC* G4-ChIP actually eliminates *in vitro* folded false positive quadruplexes, we needed to apply additional controls to differentiate the *in cellulo* and *in vitro* formed G4s with confidence. As described below, we generated a pair of control DNA with or without the canonical G4s-forming signature and characterized their quadruplex forming properties. Subsequently these DNA were used in the new protocol as a control to distinguish G4s formed only under constraints of the nuclear environment from those formed only *in vitro*.

3. An application of AbC G4-ChIP identifies CGGBP1 as an agent of constraints of the nuclear environment on G4 formation

We generated a G4-control DNA to compare how well the *AbC* G4-ChIP captured G4s from cellular nuclei versus *in vitro*. This G4-control DNA contained three repeats of the canonical [G_3-N_5-G_3-N_5-G_3-N_5-G_3] signature sequence (Fig 2A). Quadruplex forming property of the control DNA was established using CD spectroscopy. The unit sequence of the control DNA (5´-GGG TCT AAG GGC TCG AGG GTC TGC GGG-3') showed an expected positive peak around 210 nm in the CD spectra (Fig S4A). The negative peak at around 240 nm and the positive peaks at 260 nm and 290 nm further confirmed the formation of hybrid G4s [43,44]. These signature peaks were absent in a control WC sequence (Fig S4B), indicating that the hybrid quadruplex formation was specific to Scheme-DG and not observed in the Scheme-WC. The hypochromic thermal melting curve observed for the Scheme-DG was independent of the KCl concentration indicating higher-order G4s formations [45,46] (Fig S4B). Electrophoretic mobility shift assays further indicated that the hybrid G4s conformations preferred by the Scheme-DG were intramolecular (Lanes 2-6 in Fig S4C) as it moved faster than the $d(T)_{27}$ (Lane 1 in Fig S4C). We could conclude from these CD and EMSA experiments that the control DNA could form intramolecular hybrid G4s which can have any of the quadruplexes folds depicted in Fig S4D.

The control DNA could be used in two different ways: (i) added to the sonicated lysate to represent the *in vitro* false-positive quadruplexes formed without any constraints of the nuclear environment, or (ii) transfected into the cells so that the quadruplexes formation on the same sequence would be subjected to constraints of the nuclear environment (Fig 2A). The differential capture of the control DNA in the two samples would reveal the extent of constraints of the nuclear environment on G4s formation on sequences with canonical G4s forming signatures. We applied the *AbC* G4-ChIP with these control DNA for functional studies on a candidate G4s-regulatory protein CGGBP1.

CGGBP1 is a GC-rich DNA triplet repeat-binding protein. It exhibits a heightened occupancy at putatively G4s forming GGG triplets in growth-stimulated cells (as described elsewhere [47]

and Fig S5). We hypothesized that if the possibility of constraints of the nuclear environment by CGGBP1 is true then the depletion of CGGBP1 would affect the G4 landscape. *AbC* G4-ChIP with control DNA was employed to determine the G4s landscape in HEK293T cells with or without CGGBP1 knockdown.

CGGBP1 was knocked down in HEK293T cells as described elsewhere ([5] and Fig S6). The non-targeting shRNA sample was called CT and the CGGBP1-depleted sample was called KD. The CT and KD samples were transfected with the control DNA and subjected to *AbC* G4-ChIP (Fig 2A). To rule out any transfection efficiency bias between CT and KD samples due to secondary structure formation in the control DNA, we co-transfected an equimolar amount of a carrier plasmid with no detectable G4s-forming sequence (*allquads* output, not shown). The recovery of the control DNA, as well as the carrier plasmid from the inputs, was verified by quantitative PCR (Fig S7). There was no significant difference in the transfection efficiency between the CT and KD. In these samples, the control DNA was expected to be captured with 1H6 antibody only if it formed quadruplexes in the nuclei, presumably under constraints of the nuclear environment. These samples were called CT^{tr} and KD^{tr}. In parallel, CT and KD were also processed for *AbC* G4-ChIP without any transfections of control DNA (called CT^{ut} and KD^{ut} respectively). These samples were processed till nuclear lysis without any control DNA. The control DNA was folded *in vitro* in presence of KCl, incubated with 1H6 antibody, formaldehyde-fixed and then added to the CT^{ut} and KD^{ut} lysates (Fig 2A). The CT^{tr}, KD^{tr}, CT^{ut} and KD^{ut} samples were processed identically for the rest of the *AbC* G4-ChIP protocol and sequenced. In addition, we also sequenced the *AbC* G4-ChIP pulldown from a KD sample without any transfection or *in vitro* addition of control DNA to capture only the nuclear G4s. This sample (called KD^{nuc}) was comparable against the CT^{nuc} sample described above (Fig 1). The sequencing data are summarized in table S7.

First, we analyzed the control DNA pulldown in CT^{tr}, KD^{tr}, CT^{ut} and KD^{ut} datasets. There were no differences in the transfection of the control DNA in the six samples as established by quantitative PCR (fig S7). The coverages of sequencing reads on the control DNA in the input samples were low, with an enrichment in the *AbC* G4-ChIP samples. The number of reads mapping to the control DNA was increased upon CGGBP1 knockdown in KD^{tr} as compared to CT^{tr} (Table S8). The read counts mapping to the control DNA in CT^{ut} and KD^{ut} remained unaffected. In CT^{tr} the control DNA was captured as only 6 reads, which increased to 390 reads in KD^{tr}. Linked to the control DNA sequences, the vector backbone also showed similar variations (not shown). It could be concluded from these results that upon transfection, quadruplex formation on the control DNA was under an inhibitory influence of CGGBP1. Depletion of CGGBP1 relieved this inhibition on the control DNA, including the linked vector backbone. The capture of G4s similarly between CT^{ut} and KD^{ut} samples and very strongly only in KD^{tr} also established that the *AbC* G4-ChIP captures the G4s efficiently and specifically in a manner that is quantitatively verifiable. Thus, the differences in G4s prevalence on the control DNA between CT^{tr} and KD^{tr} could represent the influence of CGGBP1 on G4s formation throughout the genome in the constraints of the nuclear environment. We analyzed the genome-wide G4s landscape captured in the three CT and KD samples and asked if, just like on the control DNA, the constraints of the nuclear environment by CGGBP1 is observed on the genomic sequences as well.

G4 enrichment in the six samples was confirmed using peak calling and by calculating the read density within the peaks as compared to the flanking regions (Fig 2B-D). We observed

a generally weaker enrichment at the peak centers in KD samples as compared to the respective CT samples (Fig 2B-D). This less concentration of G4s at the peak regions suggested that interspersed *de novo* quadruplexes were formed and captured at non-peak regions upon CGGBP1 depletion. Unsupervised clustering of the pooled CT and KD datasets and the respective inputs showed that CGGBP1 depletion affected the G4s landscape drastically. CGGBP1 depletion accounted for the largest variance along with two of the highest-ranking components (Fig 2E).

To compare the G4s profiles of CT and KD, we first calculated signals in 0.2kb bins genome-wide, selected bins with ≥ 3 (removal of no signal biases) and ≤100 reads (removal of PCR artifacts) in all the samples separately. These G4s signals were then corrected for regional representational biases through a normalization based on signals from the same regions in the respective inputs. Paired t-tests were formed independently for each pair of 873929 genomic bins from CT and KD (n=3 pairs; CT^{nuc}–KD^{nuc}, CT^{tr}–KD^{tr} and CT^{ut}–KD^{ut}). We discovered 12319 differentially captured regions (DCRs) with a significant difference in G4s between CT and KD samples (p value <0.01, Q:1%: two-stage linear step-up procedure of Benjamini, Krieger and Yekutieli; pair-wise p and Q values for each bin in Table S9). A correlation analysis between pooled G4s signals in CT and KD (Fig 3A) showed two major categories of 0.2kb bins; most bins showed no difference between CT and KD (non-DCRs), and the DCRs with a conspicuous increase in the signal in KD over CT (Fig 3A). These were apparently the regions where CGGBP1 exerted constraints of the nuclear environment on G4s formation and depletion of CGGBP1 led to an increase in G4s signals. Conspicuously, there were a negligible number of regions with a decrease in G4s formation in KD as compared to CT (Fig 3A and Fig S8). We further analyzed the properties of these differentially captured regions (DCRs) where G4s-formation is regulated by CGGBP1.

Signals from the three pairs of CT and KD in the flanks of these DCRs showed that they were rich in G4s with an increase upon CGGBP1 depletion (Fig 3B and C). The proximity of the DCRs with functional regions was calculated using UCSC annotations for genes and CpG islands, FANTOM annotations for transcription start sites and enhancers and ChIP-seq data for markers of replication origins in human cells. There were no obvious spatial associations of these DCRs with any of these functionally annotated landmarks (not shown).

If the DCRs discovered by the application of the *AbC G4-ChIP* were indeed biologically relevant, they would exhibit features predictable by the known functions of CGGBP1. We analyzed the properties of these regions with reference to the features of DNA sequences associated with regulation by CGGBP1: interspersed repeats, GC-richness, G/C-skew and CTCF-binding sites [47–50]. The DCRs where G4s formation was mitigated by CGGBP1 were rich in GC content (Fig 4A). Although CGGBP1 binds to interspersed repeats and regulates their transcription as well as cytosine methylation levels, the DCRs showed no differences in the repeat contents (not shown). Interestingly, these regions were found to be rich in CGGBP1-binding sites (Fig 4B), even if these CGGBP1 binding sites were characterized in a different cell line elsewhere [47]. No such enrichment of CGGBP1 binding sites non-DCRs (Fig 4B).

G/C-skew is a property of DNA regions regulated by CGGBP1 [49]. This includes non-canonical CTCF-binding sites described in HEK293T cells at which CTCF occupancy

depends on CGGBP1 [5]. The DCRs were also found to be rich in CTCF-binding sites (Fig 4C). A G/C-skew calculation showed that these regions, including 1kb flanking regions, had an unexpectedly high G/C-skew (Fig 4D). Most interestingly, the G/C-skew was asymmetrically distributed across the DCRs (Fig 4E). The upstream and downstream 0.1kb regions in the 1kb flanks segregated into separate clusters (Fig 4F) due to unexpectedly high transitions of G/C-skew at DCRs.

Collectively, our findings showed that using *AbC* G4-ChIP we could identify G4s genome-wide. We could apply this method to CT and KD HEK293T cells and discover regions where mitigation of G4s formation by CGGBP1 coincides with unexpected G/C-skew and CGGBP1-regulated CTCF occupancy.

DISCUSSION

G4s are some of the most well-studied DNA secondary structures. Decades of *in vitro* studies have defined the chemical environment (presence of KCl) in which the quadruplexes can form on some canonical sequences. Very important and lasting conclusions about the nature of bonding leading to formation of G-tetrads and resultant formation of quadruplexes have been generated using *in vitro* assays. A combination of G triplets separated by spacers have long been considered as model sequences for G4s formation. Many computational tools rely on such sequences to report potential quadruplex formation. However, two major advances have challenged the restriction of quadruplexes to such canonical sequences. First, antibodies such as 1H6 antibody that detect a variety of T and G combinations and second, characterization of G4s genome-wide using anti-G4s antibodies. The G4s landscapes generated by using BG4 antibody do not show any restriction of quadruplexes to the canonical sequences. In fact, these studies underscore the complexity of DNA secondary structure landscape in cellular nuclei. Centrally enriched sequence patterns in G4-ChIP sequencing datasets are likely to represent the properties required for quadruplex formation. However such a signature sequence for G4s remains elusive. This is reaffirmed by the findings reported elsewhere as well as by running the G4-ChIP sequencing data through programmes such as *allquads*.

The disconnect between our knowledge from *in vitro* studies on G4s and the complex G4s landscapes discovered in cells also leads to questions central to G4s biology: Are there sequence properties other than $[G_3\text{-}N_{(1\text{-}7)}]xn$ that form the quadruplexes as well? Are there different mechanisms to regulate G4s formation at different types of sequences? Are there protein regulators of G4s formation? What are these proteins? Do they work in a sequence-dependent manner? Some of these questions can be answered by restricting the sequence analysis of G4s landscapes with or without perturbation of putative G4s regulatory genes. Despite our knowledge of multiple G4s-binding proteins, such targeted studies to dissect subtypes of quadruplex-forming sequences based on their dependence on putative regulator genes have not been reported. However, identification of such sequence properties at which quadruplex formation is regulated by specific proteins relies on capturing G4s in the native cellular environment with minimal *in vitro* artifacts. The conventional ChIP protocol has been applied to G4s landscape determination in keratinocytes and HaCaT cells [33,38] to generate valuable knowledge in the field. However, the efficiency of this protocol in representing the G4 landscape depends on crosslinking of G4s unbound to proteins, no steric hindrance to

BG4 antibody when proteins get crosslinked on/in the vicinity of the quadruplexes, and formation/deformation of quadruplexes post-lysis.

Our study attempts to provide some advance in this direction by (i) developing a protocol that captures G4s as natively as possible in cells and (ii) applying the protocol to study the potential regulation of G4s by the human protein CGGBP1. Our findings using the *AbC* G4-ChIP and control DNA show that G4 formation *in vitro* is different from the G4 landscape in the cellular nuclei. While the *in vitro* G4 formation is unconstrained, the nuclear environment presentes conditions which mitigate G4 formation. The *AbC* G4-ChIP captures the cellular G4s landscape with a discernible reduction in *in vitro* artifacts as compared to the conventional G4-ChIP protocol. The application of the *AbC* G4-ChIP protocol to a CGGBP1 knockdown system also allows us to query the quadruplex formation and changes therein show overlaps with the sequences previously reported to be regulated by CGGBP1. This validatory detection of differential G4s formation at sequences with properties previously reported for regulation by CGGBP1 shows that the *AbC* G4-ChIP protocol is reliable and reports the quadruplexes reliably such that the data are interpretable with reference to previously reported independent experiments. The *AbC* G4-ChIP avoids the capture of the G4s formed *in vitro* during the incubation steps, such as the incubation with the antibody. Thus, the G4s landscape captured by *AbC* G4-ChIP is a better representative of the actual G4s landscape in the cellular nuclei. We also discover a generic and widespread sequence property of interstrand G/C-skew that makes the DNA prone to G4s formation.

Our analysis reveals that the *allquads* G4s signature is strongly associated with *in vitro* formation of these structures. The G4s formation in the cellular nuclei is relatively less dependent on the $[G_3\text{-}N_{(1\text{-}7)}]xn$ signature.

G4s formation in double-stranded genomic DNA is widespread as indicated by the spread of the sequence tags. The reads in the *in vitro* G4 DNA-IP sequence data were highly dispersed and thus showed a low level of discretization of G4s signal resulting in fewer peaks. The discretization of G4s formation is strikingly increased when the DNA is rendered single stranded due to heat denaturation and is captured as a much larger number of peaks. These findings suggest that in double-stranded genomic DNA, the G4s are formed at a background rate genome-wide and that the concentration of G4s formation at hotspot regions is prevented by the double stranded nature of the DNA. Between the double stranded and single stranded genomic DNA, such a difference in the G4s landscape can be attributed only to the dependence of quadruplex formation on DNA sequence itself and the strandedness. These data from the DNA *in vitro* represent the inherent nature of DNA to fold into G4s when no regulatory proteins are around.

We observe that the constraints of the nuclear environment discretize as well as restrict G4s formation. This is reflected in the number of peaks which is larger in CT^{nuc} than that in G4-DNA-IP^{nat} but less than that in G4-DNA-IP^{denat}. In the absence of any heat denaturation, the G4-DNA-IP^{nat} retains its stable long-term G4s landscape with no significant *de novo* formation of G4s at random regions during the antibody incubation. The nature of the constraints of the nuclear environment on G4s formation is expected to be widespread, given the hindrances to DNA processes imposed by random formation of secondary structures. Clearly, there is a larger repertoire of helicases that resolve these secondary structures, including the G4s than the possibility of some proteins with dedicated quadruplex

stabilizing functions. The growing number of G4s interacting proteins do not indicate any such quadruplex stabilizing function although there is a paucity of targeted studies that test this possibility.

Our study reports such a targeted test of function for CGGBP1 on G4s formation. CGGBP1 is not a major G4s-binding protein. Similarly, it is not established for the many G4s-binding proteins if their binding is to the actual quadruplex structure or to the immediate vicinity of the structure. This argument is supported by the absence of any quadruplex-binding domain common to these proteins [26]. An additional set of proteins important in this context are those which can bind to the DNA sequences capable of forming G4s but not to the quadruplex structures per se. The known DNA-binding properties of CGGBP1 indicate that it is capable of binding to G/C-rich sequences with G-repeats [26,47,48,50]. These sequences are strong candidates for G4s formation. The binding to quadruplex-forming sequences but not the quadruplexes per se is logically a feature of proteins, which through binding to the DNA, would preclude quadruplex formation. These proteins, such as CGGBP1, would be able to exert constraints of the nuclear environment on quadruplex formation. Our results show that CGGBP1 indeed binds to sequences capable of G4s formation as well as exert constraints of the nuclear environment on quadruplex formation. The DCRs, with their preference for CTCF occupancy, underscore potential functions of G4s-forming regions as chromatin boundary elements. The preponderance of G4s in the heterochromatin [42] and mitigation of G4 formation at CTCF-binding sites suggest that heterochromatinization, G4s formation and regulation by CGGBP1 are interlinked phenomena. Interestingly, our results highlight a property of quadruplex-forming DNA sequences which is much less appreciated in G4s biology; G/C-skew. The *AbC* G4-ChIP presents a powerful technique to decipher the cellular G4s landscape and its regulation.

METHODS

G4-control DNA synthesis and cloning

We have designed and procured a set of oligonucleotides (Table S10) to generate the G4-control DNA by PCR amplification. The synthesized G4-control DNA was gel-purified and cloned into pGEM-T Easy vector (Promega) as per the manufacturer's protocol. The carrier DNA was also synthesized using the same methodology and has the same order of base composition, however, the GGG in [G_3-N_5-G_3-N_5-G_3-N_5-G_3] were mutated to ATG. The carrier DNA was also cloned into pGEM-T Easy vector using the same method (see below). The sequences of the control and carrier DNA are provided below:

Sequence of the G4-Control DNA:

5'-
TAATACGACTCACTATAGGGCACCCCTACATCGCAGCGGTCTTTCGGGCTAACGGGAT
CATGGGACTCAGGGGACAGCCTCAAGCAACATGCCCAGTCCTGACCTTCAATAAGGAA
GCAAACTGGGAAGGAGGGTGTCAGGGATAAAGGGGAGTCCTAGTCAAGGTGTCGGAT
GTCCTAAGACTTATGATCATTTTCTTAGGGTCTAAGGGCTCGAGGGTCTGCGGGTCGGT
TTCCTTCTAGAATTAGTATCTTCTATAGTGTCACCTAAAT-3'

Sequence of the Carrier DNA:

5'-
TAATACGACTCACTATAGGGCACCCCTACATCGCAGCGGTCTTTCATGCTAACATGATC
ATATGACTCAATGGACAGCCTCAAGCAACATGCCCAGTCCTGACCTTCAATAAGGAAGC
AAACTATGAAGGAATGTGTCAATGATAAAATGGAGTCCTAGTCAAGGTGTCGGATGTCC

TAAGACTTATGATCATTTTCTTAATGTCTAAATGCTCGAATGTCTGCATGTCGGTTTCCTT
CTAGAATTAGTATCTTCTATAGTGTCACCTAAAT-3'

Biophysical characterization of control DNA sequence: Sample preparation

Commercially synthesized HPLC grade single stranded 5'-GGGTCTAAGGGCTCGAGGGTCTGCGGG-3' (oligo1) and 5'-CCCGCAGACCCTCGAGCCCTTAGACCC-3' oligodeoxyribonucleotide (oligo2) sequences were purchased from Bioserv Biotechnologies, India. The oligodeoxyribonucleotides were diluted to a final concentration of 40μM in 50mM Tris-HCl (pH 7.4) buffer with a different KCl concentrations (0.05M, 0.1M, 0.15M, 0.5M and 1M) for oligo1 and 0.05M KCl concentration for oligo2. Subsequently, the DNA samples (oligo1 alone (Scheme-DG)) and oligos 1 and 2 together (Scheme-WC) were heated at 95°C for 5 minutes and cooled down to room temperature for 3 hours. The secondary structural conformational preferences of the DNA oligonucleotides were explored by acquiring the CD spectra immediately after the sample preparation.

Circular Dichroism (CD) spectroscopy

The CD spectroscopic study was carried out in JASCO-1500 spectrophotometer at 25°C and the spectra were processed using spectral manager software (www.jascoinc.com). The spectral data were collected in triplicate in the 190-320nm wavelength region with 1mm pathlength and 1nm bandwidth. The baseline correction was done using the Tris-HCl buffer in the presence of an appropriate KCl concentration. Note that only 0.05M KCl was used for Scheme-WC unlike the Scheme-DG. The triplicate average of the CD spectra was considered for the analysis. Further, the thermal melting was carried out for the Scheme-DG by heating the sample with a temperature in the range of 15 to 95°C at an interval of 1°C with a 30sec incubation time before the acquisition of data at each temperature. The triplicate averaged thermal denaturation curve was used for the analysis.

Electrophoretic mobility shift assay

Electrophoretic mobility shift assay (EMSA) was carried out using 14% native polyacrylamide gel which is pretreated with ethidium bromide (EtBr). The native polyacrylamide gel was prepared using 1X TAE buffer (pH 7). Subsequently, 25μM concentration of the Scheme-DG sample in 0.05M Tris-HCl (pH 7.4) buffer with varying concentrations of KCl (0.05M to 1M) was mixed with 50% of glycerol and loaded into the well. The electrophoresis was carried out using 1X TAE running buffer at 60 V, 4°C for 2 hours. After electrophoresis, the gel was photographed under UV light using SmartView Pro 1100 Imager System, UVCI-1100 from Major science.

Modeling of DNA conformation

Based on the information obtained from CD and EMSA the secondary structural preference of Scheme-DG was modeled using the 3D-NuS web server [51]. Pymol 2.1.1 software package (https://pymol.org/2/) was used for the visualization of the generated models.

Cell culture and transfection of control DNA

HEK293T cells were transduced with a control (CT) or CGGBP1-targeting (KD) shRNA as described elsewhere [5]. These cells were maintained in DMEM (SH30243.01, HyClone) supplemented with 10% heat-inactivated fetal bovine serum (FBS) (10270106, Invitrogen) and selected with 0.3 μg/ml of the final concentration of puromycin (Invitrogen). For the equality of transfection of the control DNA, the cloned-in carrier DNA in the same vector was cotransfected in CT and KD using the calcium-phosphate transfection method. 1.25 μg of each of control DNA and carrier DNA (cloned in pGEM-T Easy vector) and 20 μl of 2 M $CaCl_2$ mixed. A final volume of 200 μl is attained with sterile autoclaved water. This mix was added dropwise in 200μl of 2X HEPES buffered saline (Invitrogen) with gentle agitation on a

vortex mixer to form calcium phosphate-DNA precipitate. The precipitate was incubated at room temperature for 20 minutes and added drop-by-drop on approx 2 X 10^6 cells per 10 cm dish. On the next day, the cells were processed for *AbC* G4-ChIP.

In vitro G4 DNA-IP

1.25 µg of each control DNA and the carrier DNA were mixed in a 1.5 ml centrifuge tube at room temperature and the volume was made up to 100 µl using deionized autoclaved water. 22.5 µl of 1M KCl was added to the final concentration of 150 mM in a final volume of 135 µl. This mix was then heated in a dry bath at 95°C for 5 minutes. It was cooled down at room temperature and continued for further 30 minutes to allow the formation of G4s. Following this incubation, 15 µl of FBS was added to the final concentration of 10% v/v. The formed G4s were allowed to be recognized by 1 µl (1 µg) of Anti-G4 1H6 antibody. The mixture was flicked a couple of times and incubated at 4°C for 1 hour. The antibody-G4s interaction was crosslinked using 4 µl of 37% formaldehyde to the final concentration of 1% at room temperature for 10 minutes with a gentle invert mixing. 15 µl of 1.25 M glycine was added to the final concentration of 125 mM and gently mixed to quench the reaction for 5 minutes. These antibody-crosslinked G4s were included in the cell lysates for samples CT^{nuc} and KD^{nuc} and sonicated as mentioned in *AbC* G4-ChIP.

AbC G4-ChIP

G4s in CT^{nuc}, KD^{nuc}, CT^{tr}, KD^{tr}, CT^{ut} and KD^{ut} were captured using the *AbC* G4-ChIP protocol. The culture medium was removed from two 10 cm dishes with or without transfection for each sample. The cells were gently washed with ice-cold 1X Ca^{2+}-free PBS (Cytiva, SH30256.01) twice. After the last wash, the cells were treated with 1ml of trypsin per dish and incubated at 37°C for 10 minutes. Following the trypsinization, 5 ml of serum-containing DMEM was added in each dish and the cells were declumped by vigorous pipetting. Then the cells were collected in a 50 ml screw-capped centrifuge tube by passing the cell suspension through a 40 µm cell strainer (Corning). The volume was adjusted to a total of 30 ml with serum-free DMEM. The cells were crosslinked with 4 ml of 37% formaldehyde stock solution and incubated for 5 minutes at room temperature to fix DNA-protein interactions. The suspension of cells was quickly quenched with 4 ml of 1.25 M Glycine along with 4 ml of 1M KCl. The suspension was quickly spun down and resuspended in 5 ml of Tris-KCl buffer (50 mM Tris, pH 7.5; 150 mM KCl). The cells were washed once more in the Tris-KCl buffer, spun down and the supernatant was removed. The pelleted cells were resuspended and permeabilized in 15 ml of Tris-KCl buffer supplemented with 0.5% of Triton X-100 for 10 minutes at room temperature with gentle invert mixing. Following the incubation, the suspension was spun down and resuspended in 15ml of Tris-KCl buffer supplemented with 0.01% v/v of Triton X-100 (called, TKT buffer) to maintain the permeability of the cell membrane. To the suspension, 10% v/v FBS was immediately added and incubated at room temperature for 30 minutes (for the first 15 minutes with intermittent invert mixing and stationary till the end of the 30th minute). The suspension was spun down and all but 5 ml of the supernatant was gently removed without disturbing the pellet. The pellet was resuspended gently with the help of a pipette avoiding mechanical rupturing of the cells and the suspension of cells was then transferred into a fresh 15 ml screw-capped tube. This suspension of cells was incubated with Anti-G4 1H6 antibodies (Sigma; MABE1126, 5 µg of antibody/100 µl of cell pellet) for 2 h at room temperature. Following the incubation, the suspension was spun down and the pellet was gently washed once with the TKT buffer and resuspended in the same. The antibody-DNA G4s interactions were then crosslinked with

1% final concentration of formaldehyde and incubated at room temperature for 10 minutes with gentle invert mixing. 125 mM of the final concentration of Glycine was added to quench formaldehyde and the mix was incubated at room temperature for 5 minutes. The cells were pelleted down and lysed in SDS lysis buffer (50 mM Tris-HCl pH 8.0; 1% SDS and 25 mM EDTA) supplemented with 1X v/v Halt Protease and Phosphatase inhibitor (Thermo Scientific; 87786) twice the volume of cell pellet. The cells were lysed on ice for 30 minutes with intermittent tapping of the tube. The lysate was equally distributed into 1.5 ml centrifuge tubes (no more than 500 µl/tube) and sonicated in Diagenode Bioruptor for 30 sec ON - 30 sec OFF for 20 cycles at constant preset maximum power. The sonicated fraction was gently spun down and kept separately at 4°C. To increase the recovery of sonicated fraction, the pellet was resuspended in an equal volume of ChIP dilution buffer (0.01% SDS; 15 Triton X-100; 1.2 mM EDTA; 16.7 mM Tris-HCl pH 8.0 and 167 mM NaCl) as the volume of sonicated fraction retained earlier. At this stage, the antibody-crosslinked control DNA (for samples CT^{ut} and KD^{ut}) was included and sonicated for 30sec ON - 30 sec OFF for 10 cycles at constant power. The respective sonicated fractions were pooled and centrifuged at 14,000 rpm for 30 minutes at 4°C. Meanwhile, Protein G Sepharose® 4 Fast Flow (GE; 17061801) beads were washed thrice gently with 1X PBS. The clear sonicated fraction was transferred into a 15ml tube and diluted with ChIP dilution buffer to ten times of the volume of SDS lysis buffer used. 10% of this diluted fraction was kept as input. In the rest, 10% v/v slurry of beads was added and tumbled at low rpm for 2 hours at 4°C. The beads were gently collected and washed with PBST (0.1% Triton X-100 in 1X PBS) five times. Finally, the beads were vigorously vortexed in the presence of an elution buffer (1% SDS; 0.1 M sodium bicarbonate) for 30 minutes. The elution was pooled in 500 µl volume by collecting 250 µl of eluted fraction after 15 minutes and another 250µl at the end of the 30th minute. The eluted fractions and the inputs were subjected to reverse-crosslinking by adding 20 µl of 5 M NaCl and incubating overnight at 65°C. The reverse-crosslinked DNA was digested with Proteinase K (2 µl of 10 mg/ml stock) in presence of 10 µl of 0.5 M EDTA and 20 µl of 1M Tris-HCl pH 6.8 at 42°C for 1 hour. The DNA was purified using the ethanol precipitation method and dissolved in deionized autoclaved water.

Sequence alignment and annotation

The CT^{nuc}, KD^{nuc} and $Input^{nuc}$ were sequenced on the IonTorrent S5 sequencer while CT^{tr}, KD^{tr}, CT^{ut}, KD^{ut} and respective inputs were sequenced on MinION from the Oxford Nanotechnologies sequencing platform. The G4-ChIP, *in vitro* G4 DNA-IPs and the inputs were sequenced on MinION as well. The base-calling of the reads on MinION was carried out by *guppy_basecaller* in real-time. For reads sequenced on MinION, the poor-quality reads were further subjected to basecalling by locally installed GPU-based *guppy_basecaller*. The sequenced reads were subjected to removal of the sequencing adapter using Porechop using *default* options. The repeat-unmasked hg38 genome (excluding unannotated and mitochondrial chromosomes and chromosome Y) with or without the control DNA was indexed by using *minimap2 -x map-ont* option. The filtered reads were aligned using *minimap2* on the indexed genome, however, the samples, CT^{nuc}, KD^{nuc} and $Input^{nuc}$, were aligned on the indexed genome without the control DNA. Only uniquely aligned reads were fetched for further analyses. A detailed description of the analyses is mentioned in the Appendix-I. The sequence data are deposited to NCBI GEO vide ID GSE202456.

Generation of correlation matrices

The human genome was binned into 0.2kb bins. Coverage of the uniquely aligned reads on those genomic bins were calculated using *bedtools coverage* with the option *-counts*. Any genomic bins with read coverage less than 3 and more than 100 were excluded. The bins satisfying this criterion were used to fetch the reads from all the samples and the inputs. The reads were then converted into *bam* format using *bedtools bedtobam* function. The bam files were then indexed using *samtools* and converted to *bigWig* signal using *deepTools bamCoverage* with the bin size of 200. These *bigWig* signals were used to compute the correlation using *deepTools multiBigwigSummary*. The heatmaps of correlation matrices were generated using *deepTools plotCorrelation* with the options *-c spearman -p heatmap --skipZeros --removeOutliers*.

G/C-skew calculation

The human genome hg38 was binned in 100bp regions using *bedtools makewindows* and fasta sequences were fetched using *bedtools getfasta* from the unmasked hg38. For every bin, the presence of “G”s and “C”s were calculated separately using *emboss fuzznuc*. Any corresponding bin having *zero* count was eliminated from the G/C-skew calculation. Once all the bins with “G” and “C” counts were available, G/C-skew was calculated using the formula [(G-C)/(G+C)] for each 100 bp genomic bin. This information was used to generate bigWig signal using UCSC *bedGraphToBigWig* for generating heatmap.

Discovery of Differentially Captured Regions (DCRs)

The coverage of uniquely aligned sequenced reads were computed for all *AbC* G4-ChIP samples in the 0.2 kb genomic bins. The coverage counts were normalized using sequencing depth in the corresponding CT or KD samples. Those 0.2 kb genomic bins were chosen wherein the raw coverage counts follow the criterion of ≥ 3 and ≤ 100 for all three sets of *AbC* G4-ChIP samples. Then the corrected coverage counts were used to calculate the “diff/sum” ratios [(KD-CT)/KD+CT)] for each pair of the *AbC* G4-ChIP samples. The same exercise was performed for the inputs, however, no limit to coverage counts was applied. Thus, the diff/sum ratios for the three *AbC* G4-ChIP samples and three inputs were calculated for each bin individually and a separate one-tailed T-Test was applied for each bin with n=3 for CT as well as KD. The 0.2 kb genomic bins with p-value < 0.01 (true discoveries established through q-value < 0.01) were selected and denoted as differentially captured regions (DCRs).

GGG/CCC weighted content calculation

GGG/CCC weighted content was calculated on all peak datasets. Fasta sequences for the peaks were fetched from the repeat-unmasked hg38 followed by “GGG” and “CCC” patterns were searched using *emboss fuzznuc* with every 1-base succession for each peak. Any peaks with zero count of either “GGG” or “CCC” patterns were discarded for GGG/CCC weighted content calculation. The weighted GGG/CCC content of a peak was calculated by taking the product of the sum of “GGG” and “CCC” counts and the coverage of reads on the given peak. The GGG/CCC weighted content was plotted using GraphPad Prism9.

Peak calling

Broad peaks were called on the datasets against the corresponding inputs as background noise using *MACS2 callpeak* function. For all *AbC* G4-ChIP samples and *in vitro* G4 DNA-IP^{denat}, a band width of 3000 (*--bw 3000*) was used to call the broad peaks with the options: --

nolambda --broad --broad-cutoff 0.01. However, for G4-ChIP and *in vitro* G4 DNA-IP^{denat} samples broad peaks were called at default band with value (300). The details of the discovered peaks are listed in table S5.

Jaccard Index calculation

Jaccard indices on the peaks were calculated using *bedtools jaccard* option regardless of strandedness. The indices between samples are mentioned in figure S3.

DECLARATIONS

The authors declare no conflict of interest.

REFERENCES

1. Wolffe AP, Guschin D. Review: chromatin structural features and targets that regulate transcription. J Struct Biol 2000; 129:102–22.

2. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome [Internet]. Nature Reviews Genetics2019; 20:207–20. Available from: http://dx.doi.org/10.1038/s41576-018-0089-8

3. Xiang J-F, Corces VG. Regulation of 3D chromatin organization by CTCF. Curr Opin Genet Dev 2021; 67:33–40.

4. Patel D, Patel M, Datta S, Singh U. CGGBP1-dependent CTCF-binding sites restrict ectopic transcription. Cell Cycle 2021; 20:2387–401.

5. Patel D, Patel M, Datta S, Singh U. CGGBP1 regulates CTCF occupancy at repeats. Epigenetics Chromatin 2019; 12:57.

6. Patel M, Patel D, Datta S, Singh U. CGGBP1-regulated cytosine methylation at CTCF-binding motifs resists stochasticity. BMC Genet 2020; 21:84.

7. Bochman ML, Paeschke K, Zakian VA. DNA secondary structures: stability and function of G-quadruplex structures. Nat Rev Genet 2012; 13:770–80.

8. Spiegel J, Adhikari S, Balasubramanian S. The Structure and Function of DNA G-Quadruplexes. Trends Chem 2020; 2:123–36.

9. Groth A, Rocha W, Verreault A, Almouzni G. Chromatin challenges during DNA replication and repair. Cell 2007; 128:721–33.

10. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell 2007; 128:707–19.

11. Robinson J, Raguseo F, Nuccio SP, Liano D, Di Antonio M. DNA G-quadruplex structures: more than simple roadblocks to transcription? Nucleic Acids Res 2021; 49:8419–31.

12. West AG, Fraser P. Remote control of gene transcription. Hum Mol Genet 2005; 14 Spec No 1:R101–11.

13. Hapgood JP, Riedemann J, Scherer SD. Regulation of gene expression by GC-rich DNA cis-elements. Cell Biol Int 2001; 25:17–31.

14. Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. Nat Rev Mol Cell Biol 2017; 18:279–84.

15. Nikolova EN, Kim E, Wise AA, O'Brien PJ, Andricioaei I, Al-Hashimi HM. Transient Hoogsteen base pairs in canonical duplex DNA. Nature 2011; 470:498–502.

16. Alvey HS, Gottardo FL, Nikolova EN, Al-Hashimi HM. Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics. Nat Commun 2014; 5:4786.

17. Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology. Nucleic Acids Res 2015; 43:8627–37.

18. Kudlicki AS. G-Quadruplexes Involving Both Strands of Genomic DNA Are Highly Abundant and Colocalize with Functional Sites in the Human Genome. PLoS One 2016; 11:e0146174.

19. Murat P, Guilbaud G, Sale JE. DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. Genome Biol 2020; 21:209.

20. Mendoza O, Bourdoncle A, Boulé J-B, Brosh RM Jr, Mergny J-L. G-quadruplexes and helicases. Nucleic Acids Res 2016; 44:1989–2006.

21. Frasson I, Pirota V, Richter SN, Doria F. Multimeric G-quadruplexes: A review on their biological roles and targeting. Int J Biol Macromol 2022; 204:89–102.

22. 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 2020; 21:459–74.

23. Gong J-Y, Wen C-J, Tang M-L, Duan R-F, Chen J-N, Zhang J-Y, Zheng K-W, He Y, Hao Y-H, Yu Q, et al. G-quadruplex structural variations in human genome associated with single-nucleotide variations and their impact on gene activity. Proc Natl Acad Sci U S A [Internet] 2021; 118. Available from: http://dx.doi.org/10.1073/pnas.2013230118

24. David AP, Margarit E, Domizi P, Banchio C, Armas P, Calcaterra NB. G-quadruplexes as novel cis-elements controlling transcription during embryonic development. Nucleic Acids Res 2016; 44:4163–73.

25. Lago S, Nadai M, Cernilogar FM, Kazerani M, Domíniguez Moreno H, Schotta G, Richter SN. Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. Nat Commun 2021; 12:3885.

26. Zhang X, Spiegel J, Martínez Cuesta S, Adhikari S, Balasubramanian S. Chemical profiling of DNA G-quadruplex-interacting proteins in live cells. Nat Chem 2021; 13:626–33.

27. Brázda V, Hároníková L, Liao JCC, Fojta M. DNA and RNA quadruplex-binding proteins. Int J Mol Sci 2014; 15:17493–517.

28. Huang Z-L, Dai J, Luo W-H, Wang X-G, Tan J-H, Chen S-B, Huang Z-S. Identification of G-Quadruplex-Binding Protein from the Exploration of RGG Motif/G-Quadruplex Interactions. J Am Chem Soc 2018; 140:17945–55.

29. Williams P, Li L, Dong X, Wang Y. Identification of SLIRP as a G Quadruplex-Binding Protein. J Am Chem Soc 2017; 139:12426–9.

30. Ribeyre C, Lopes J, Boulé J-B, Piazza A, Guédin A, Zakian VA, Mergny J-L, Nicolas A. The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. PLoS Genet 2009; 5:e1000475.

31. Yan KK-P, Obi I, Sabouri N. The RGG domain in the C-terminus of the DEAD box helicases Dbp2 and Ded1 is necessary for G-quadruplex destabilization. Nucleic Acids Res 2021; 49:8339–54.

32. Wallgren M, Mohammad JB, Yan K-P, Pourbozorgi-Langroudi P, Ebrahimi M, Sabouri N. G-rich telomeric and ribosomal DNA sequences from the fission yeast genome form stable G-quadruplex DNA structures in vitro and are unwound by the Pfh1 DNA helicase. Nucleic Acids Res 2016; 44:6213–31.

33. Hänsel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, et al. G-quadruplex structures mark human regulatory chromatin. Nat Genet 2016; 48:1267–72.

34. Kuryavyi V, Patel DJ. Solution structure of a unique G-quadruplex scaffold adopted by a guanosine-rich human intronic sequence. Structure 2010; 18:73–82.

35. Lightfoot HL, Hagen T, Tatum NJ, Hall J. The diverse structural landscape of quadruplexes. FEBS Lett 2019; 593:2083–102.

36. Liu H, Wang R, Yu X, Shen F, Lan W, Haruehanroengra P, Yao Q, Zhang J, Chen Y, Li S, et al. High-resolution DNA quadruplex structure containing all the A-, G-, C-, T-tetrads. Nucleic Acids Res 2018; 46:11627–38.

37. Kazemier HG, Paeschke K, Lansdorp PM. Guanine quadruplex monoclonal antibody 1H6 cross-reacts with restrained thymidine-rich single stranded DNA. Nucleic Acids Res 2017; 45:5913–9.

38. 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 2018; 13:551–64.

39. Traganos F, Darzyndiewicz Z, Sharpless T, Melamed MR. Denaturation of deoxyribonucleic acid in situ effect of formaldehyde. J Histochem Cytochem 1975; 23:431–8.

40. Hoffman EA, Frey BL, Smith LM, Auble DT. Formaldehyde crosslinking: a tool for the study of chromatin complexes. J Biol Chem 2015; 290:26404–11.

41. Henderson A, Wu Y, Huang YC, Chavez EA, Platt J, Johnson FB, Brosh RM Jr, Sen D, Lansdorp PM. Detection of G-quadruplex DNA in mammalian cells. Nucleic Acids Res 2017; 45:6252.

42. Hoffmann RF, Moshkin YM, Mouton S, Grzeschik NA, Kalicharan RD, Kuipers J, Wolters AHG, Nishida K, Romashchenko AV, Postberg J, et al. Guanine quadruplex structures localize to heterochromatin. Nucleic Acids Res 2017; 45:6253.

43. Sathyaseelan C, Vijayakumar V, Rathinavelan T. CD-NuSS: A Web Server for the Automated Secondary Structural Characterization of the Nucleic Acids from Circular Dichroism Spectra Using Extreme Gradient Boosting Decision-Tree, Neural Network and Kohonen Algorithms. J Mol Biol 2021; 433:166629.

44. del Villar-Guerra R, del Villar-Guerra R, Trent JO, Chaires JB. G-Quadruplex Secondary

Structure Obtained from Circular Dichroism Spectroscopy [Internet]. Angewandte Chemie2018; 130:7289–93. Available from: http://dx.doi.org/10.1002/ange.201709184

45. Ajjugal Y, Kolimi N, Rathinavelan T. Secondary structural choice of DNA and RNA associated with CGG/CCG trinucleotide repeat expansion rationalizes the RNA misprocessing in FXTAS. Sci Rep 2021; 11:8163.

46. Zhou B, Geng Y, Liu C, Miao H, Ren Y, Xu N, Shi X, You Y, Lee T, Zhu G. Characterizations of distinct parallel and antiparallel G-quadruplexes formed by two-repeat ALS and FTD related GGGGCC sequence. Sci Rep 2018; 8:2366.

47. Agarwal P, Enroth S, Teichmann M, Jernberg Wiklund H, Smit A, Westermark B, Singh U. Growth signals employ CGGBP1 to suppress transcription of Alu-SINEs. Cell Cycle 2016; 15:1558–71.

48. Singh U, Westermark B. CGGBP1--an indispensable protein with ubiquitous cytoprotective functions. Ups J Med Sci 2015; 120:219–32.

49. Patel D, Patel M, Westermark B, Singh U. Dynamic bimodal changes in CpG and non-CpG methylation genome-wide upon CGGBP1 loss-of-function. BMC Res Notes 2018; 11:419.

50. Müller-Hartmann H, Deissler H, Naumann F, Schmitz B, Schröer J, Doerfler W. The human 20-kDa 5’-(CGG)(n)-3'-binding protein is targeted to the nucleus and affects the activity of the FMR1 promoter. J Biol Chem 2000; 275:6447–52.

51. Patro LPP, Kumar A, Kolimi N, Rathinavelan T. 3D-NuS: A Web Server for Automated Modeling and Visualization of Non-Canonical 3-Dimensional Nucleic Acid Structures. J Mol Biol 2017; 429:2438–48.

Legends

Fig 1: A comparative analysis of G4s landscape in cultured cells using AbC G4-ChIP: A: A schematic of the conventional G4-ChIP highlighting the possibilities of false positives and negatives. B: A schematic of the AbC G4-ChIP protocol highlighting the key differences from the conventional G4-ChIP and minimization of false positives and false negatives. C: Character legend for A and B. D: Correlation matrix and clustering for the G4 landscapes reported by various methods. The absolutely *in vitro* G4s (*in vitro* G4 DNA-IP^{nat} and *in vitro* G4 DNA-IP^{denat}) show similarities with the conventional G4-ChIP, whereas the AbC G4-ChIP eliminates this similarity. Notably, all ChIP samples cluster with the respective inputs except for the *in vitro* G4 DNA-IP^{nat} which co-clusters with the G4-ChIP. The numbers in the insets are spearman coefficients. E: Violin plots of read representation-weighted GGG/CCC counts in peaks (product of GGG/CCC count and the number of reads mapping to each peak). The Y-axis shows the median and IQR on a log10 scale. Remarkably, CT^{nuc} has the lowest GGG/CCC representation in peaks. Peak counts and properties are presented in table S5.

Fig 2: Validation of the AbC G4 ChIP protocol using a control DNA and its application to a CGGBP1 knockdown system: A: A schematic of the control DNA (sequence with G-triplets highlighted in bold) cloned in pGEMT vector and its usage in AbC G4-ChIP. The left flank shows the protocol in which the control DNA (along with the carrier DNA) is transfected (samples CT^{tr} and KD^{tr}). The right flank shows the samples CT^{ut} and KD^{ut} in which the preformed G4s on the control DNA are added to the lysates. As indicated, the effects of the intranuclear environments of live cells on G4s would be seen only in CT^{tr} and KD^{tr} but not in CT^{ut} and KD^{ut}. B-D: Heatmaps and signal plots showing G4s enrichment at peak regions as compared to the neighboring 5kb flanks. The signals were consistently weaker at the peaks in KD samples as compared to the respective CT samples. E: A PCA plot of the combined signals of the three CT and KD experiments show that CGGBP1 depletion remarkably alters the G4s landscape. The values used for PCA loading were derived from summation of signals in synchronous 0.2kb genomic bins after removal of bins with signals < 3 and > 100.

Fig 3: Distribution of G4s signals as captured in pooled CT and KD samples. A: The correlation shows the abundance of G4s signals in the 0.2 kb bins genome-wide (n = 873929). Most 0.2 kb bins show no quantitative difference in the signal (r = 0.26) as represented by the cluster of non-DCRs (dark blue dots). However, a small fraction of the 0.2 kb genomic bins (~0.01%) show stronger presence of G4s signals in KD than in CT with a negative correlation (r = -0.16) represented by the DCRs (orange dots). The Y-axis represents input-normalized signal for KD and the X-axis represents input-normalized signal for CT per 0.2 kb bins. The DCRs and non-DCRs are negatively correlated. The correlation values for DCRs (yellow dots) and non-DCRs (blue dots) are mentioned in the inset. B-C: The pooled G4s signal in CT and KD show a qualitative difference on the 10 kb flanks from the center the DCRs. In KD, the center of the DCRs carries higher enrichment of the signal than in CT. In general, the neighbouring regions of the DCRs show weaker G4s signal in CT compared to KD. The heatmaps were generated using *deepTools plotHeatmap* with the adjusted Y-axis values for the signal to plot the profile (mean±SE).

Fig 4: Features of the DCRs. A: The DCRs show that these CGGBP1-regulated genomic regions have stronger GC content than their immediate flanks of 2 kb. The GC content signal is plotted as mean±SE and the Y-axis represents the signal values. B: The signal from

genome-wide occupancy of CGGBP1 under growth-stimulation (GSE53571) has been plotted on the 1 kb flanks from the center of DCRs and non-DCRs. The DCRs (green) show enrichment of CGGBP1 occupancy in their centers as compared to the non-DCRs (dark red) with a background level of the signal devoid of any specific enrichment. The lines were drawn as mean±SE and the Y-axis shows the signal values. C: The non-canonical CGGBP1-dependent CTCF binding sites show that upon knockdown of CGGBP1, the occupancy of CTCF on DCRs has heightened occupancy (dark red) than seen at normal levels of endogenous CGGBP1 (green). The signal is plotted in 10 kb flanks from the center of the DCRs with the lines representing the mean±SE of the signal and the Y-axis representing the signal values. D: The DCRs show a quantitative stronger G/C-skew as measured in their 1 kb flanks compared to the same number of randomly expected genomic regions sampled 10 times. The randomly expected genomic regions show a similar difference in G/C skew on their 1 kb flanks as compared to the DCRs. E: The heatmap shows a qualitative and asymmetric distribution of G/C-skew on the 1 kb flanks of the DCRs. The DCRs are clustered to show stronger changes in G/C-skew in the flanks (cluster 2 and cluster 3) as well as minimum to no change in the G/C-skew (cluster 1). The lines represent the mean signal of G/C-skew calculated for the DCRs. F: The G/C-skew has been binned into 0.1 kb bins in 1 kb flanks of DCRs and forms distinct clusters. The upstream and downstream bins form separate clusters wherein the respective bins cluster differently with the proximal bins. The yellow line represents the length of DCRs and the numbers represent the distance (0.1 kb) from the center of the DCRs either upstream (blue) or downstream (red). The Y-axis shows an arbitrary distance from the 0.1 kb bins. The dashed lines indicate the end of the 0.1 kb bins. The 0.1 kb bins are not drawn on scale. The cluster has been generated in R using the hierarchical method “manhattan”.

Fig S1: Spearman correlation matrix and clustering of G4s profile for the G4-ChIP and *in vitro* G4 DNA-IPs. The *in vitro* G4 DNA-IP^{denat} forms the most distant cluster with similarity to the *in vitro* input. The *in vitro* G4 DNA-IP^{nat} shows similar correlation with G4-ChIP and its input in a separate cluster. The numbers indicate the correlation coefficients for pairs of samples.

Fig S2: Discrete G4s signals are captured by *AbC* G4-ChIP. Using 123190 (10 rounds of randomization matching 12319 DCRs in each round) randomly selected regions genome-wide, the G4 signals at random regions were calculated for the four samples as indicated. The CT^{nuc}, in which *AbC* G4-ChIP was applied showed a very restricted low level detection of G4 signals at these random regions, similar to the *in vitro*^{denat}. The G4-ChIP and in vitro nat on the other hand showed a relatively high prevalence of G4 capture at random regions. These comparisons showed that the discrete and specific G4s are detected by application of the *AbC* G4-ChIP protocol and it minimizes *in vitro* G4s captured at random regions as artifacts.

Fig S3: Jaccard indices of G4s peaks identified in CT^{nuc}, G4-ChIP, *in vitro* G4 DNA-IP^{denat} and *in vitro* G4 DNA-IP^{nat}. The numbers indicate the proximity of occurrences of G4s peaks genome-wide. The *in vitro* G4 DNA-IP^{denat} represents a very different profile of G4s peaks than in the rest of the samples. The profile of the G4s peaks in G4-ChIP showed close resemblance to the *in vitro* G4 DNA-IP^{denat} and CT^{nuc}. The lines with the bidirectional arrows indicate the pairs of samples within which the jaccard indices were calculated. The numbers on the bottom of the lines are not on scale.

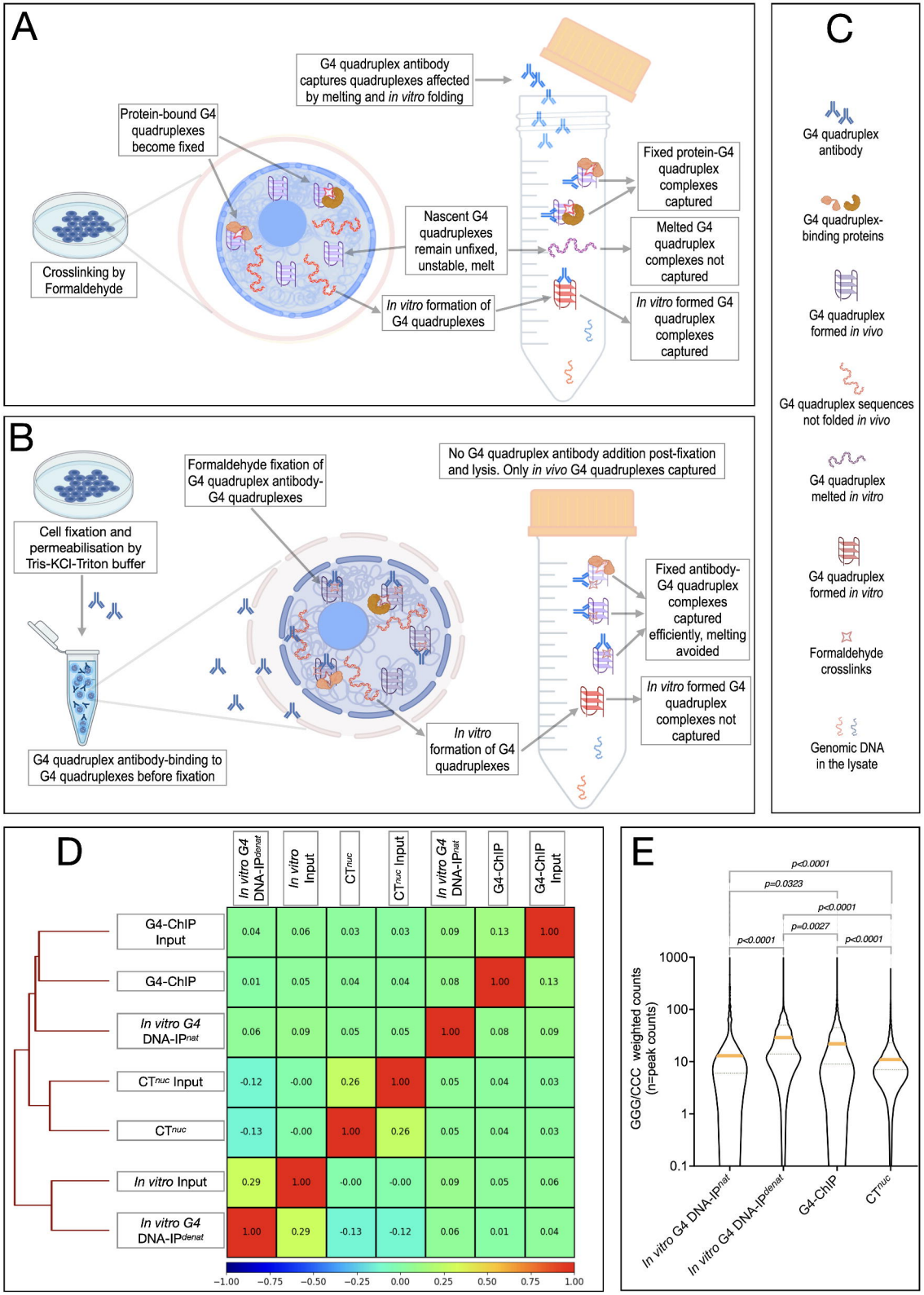
Fig S4: CD wavelength scan, CD thermal melt experiment and EMSA showing the preference for intramolecular hybrid G4s by control DNA unit sequence (5´-GGG TCT AAG GGC TCG AGG GTC TGC GGG-3´). A: CD spectra showing the preference for the hybrid G4s conformation by the Scheme-DG in the presence of 50mM Tris-HCl (pH 7.4) buffer irrespective of different KCl concentrations (0.05M, 0.1M, 0.15M, 0.5M and 1M). Note the absence of such a conformational preference by the corresponding WC duplex (Scheme-WC, colored in peach). B: The hypochromic thermal melting pattern seen for the Scheme-DG at different KCl concentrations (i) 0.05M, (ii) 0.1M, (iii) 0.15M, (iv) 0.5M and (v) 1M indicates the higher order G4s formation. C: EMSA corresponding to the Scheme-DG indicates the intramolecular G4s formation at different KCl concentrations as it moves faster (Lanes 2-6) than the control $d(T)_{27}$ (Lane 1). D: Possible intramolecular hybrid G4s conformation preferred by the Scheme-DG which are modeled using the 3D-NuS web server. The guanines of the G-quartet are colored green and the loop residues are shown in pink.

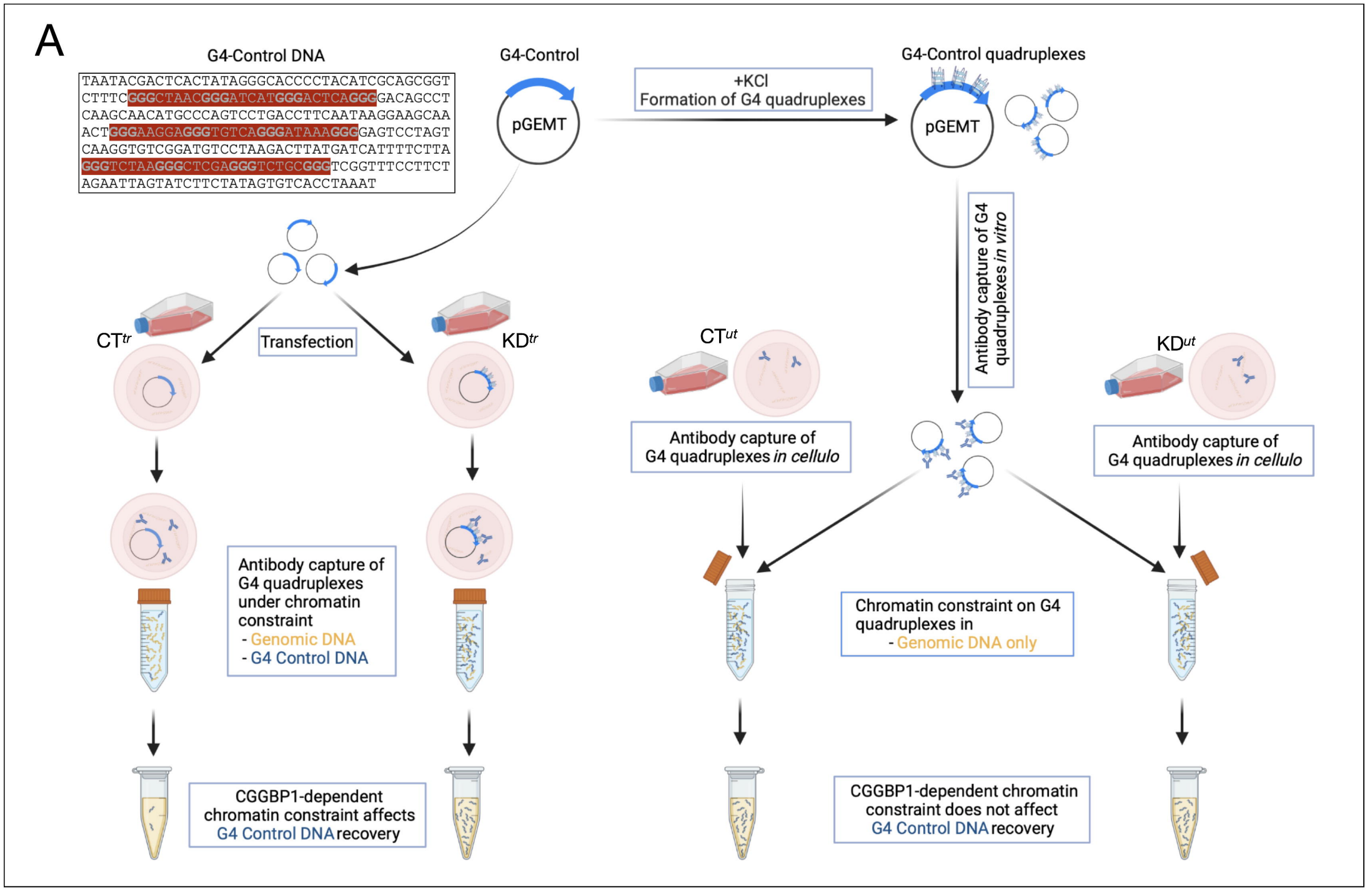
Fig S5: GGG/CCC content on CGGBP1 peak datasets (GSE53571). GGG/CCC content calculated in two different CGGBP1 peak datasets (growth-stimulated and starved) shows that the CGGBP1 occupancy is facilitated on genomic regions with higher GGG/CCC content upon growth stimulation. The difference of the CGGBP1 occupancy in these two datasets are denoted by a strong p-value. The orange lines indicate the median values presented by the Y-axis on a log10 scale.

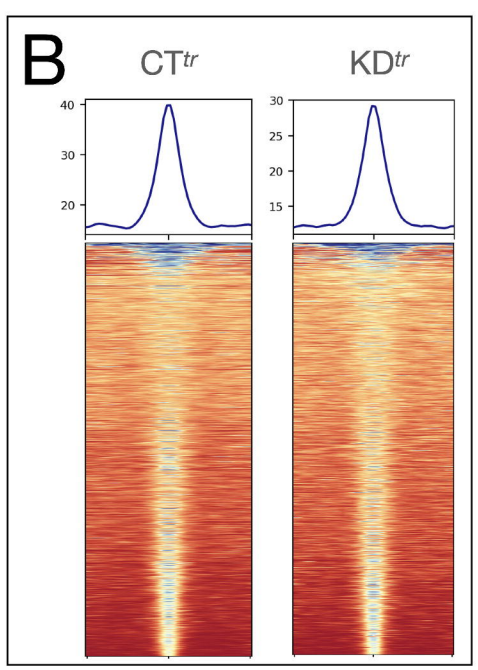
Fig S6: Western blot showing the knockdown of endogenous levels of CGGBP1. The levels of CGGBP1 and GAPDH are shown in the lower and upper panel respectively. The level of knockdown of CGGBP1 is normalized to the levels of GAPDH. Approximately, 99% knockdown of endogenous levels of CGGBP1 was observed in KD.

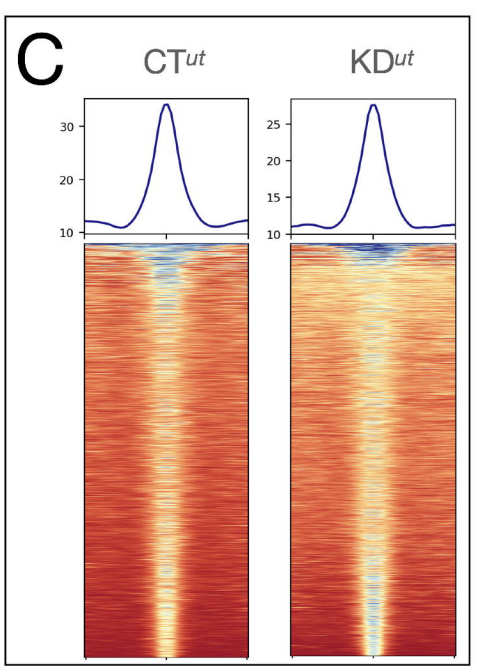
Fig S7: Recovery of control and carrier DNA in the inputs of CT^{ut}, KD^{ut}, CT^{tr} and KD^{tr}. The qPCR quantification confirms the bias-free presence of the control and control DNA between CT and KD as quantified in their input samples. Notably, the *in vitro* addition of the control and carrier DNA in the untransfected samples yielded a higher amount of the control DNA than that can be quantified in the transfected samples. The Y-axis represents the concentration of the control and control DNA on a log10 scale.

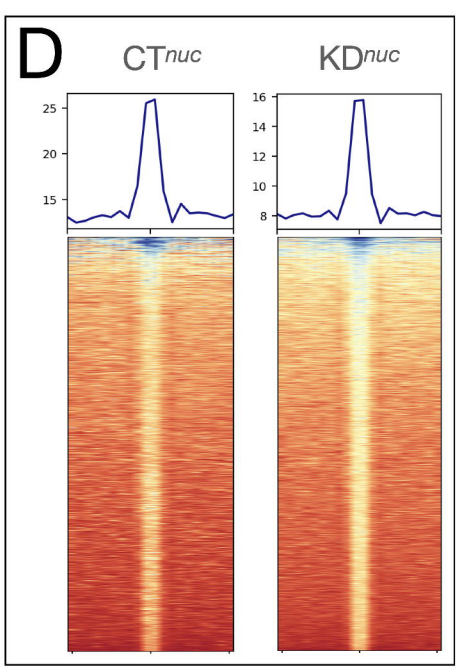
Fig S8: A volcano plot of 0.2kb genomic bins compared pairwise in triplicates between CT and KD. The X-axis represents $\log_2$ fold change calculated as KD/CT and the corresponding p-values on a $-\log_{10}$ scale are plotted on the Y-axis. The violin plots represent the density of the data points along the two axes. The red-green partition marks the non-DCRs and DCRs respectively. The green data points correspond to the blue data points in fig 3A.

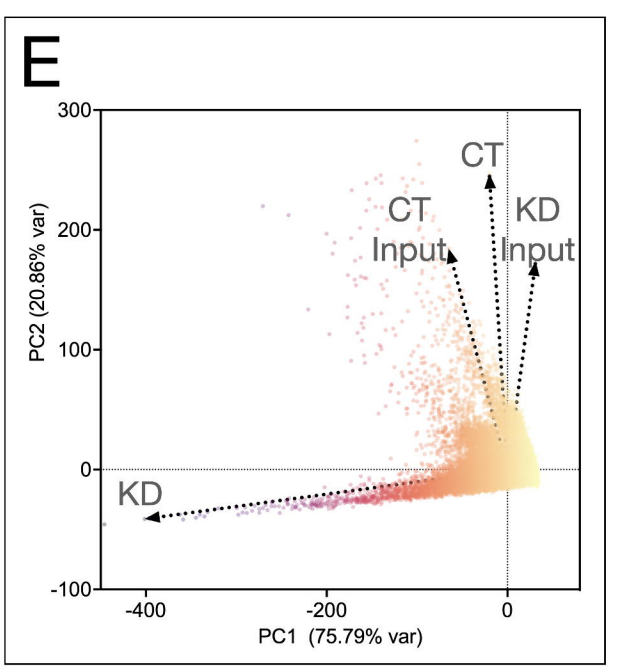

A
G4 quadruplex antibody captures quadruplexes affected by melting and in vitro folding
Protein-bound G4 quadruplexes become fixed
Crosslinking by Formaldehyde
Nascent G4 quadruplexes remain unfixed, unstable, melt
In vitro formation of G4 quadruplexes
Fixed protein-G4 quadruplex complexes captured
Melted G4 quadruplex complexes not captured
In vitro formed G4 quadruplex complexes captured
B
Cell fixation and permeabilisation by Tris-KCl-Triton buffer
G4 quadruplex antibody-binding to G4 quadruplexes before fixation
Formaldehyde fixation of G4 quadruplex antibody-G4 quadruplexes
No G4 quadruplex antibody addition post-fixation and lysis. Only in vivo G4 quadruplexes captured
In vitro formation of G4 quadruplexes
Fixed antibody-G4 quadruplex complexes captured efficiently, melting avoided
In vitro formed G4 quadruplex complexes not captured
C
G4 quadruplex antibody
G4 quadruplex-binding proteins
G4 quadruplex formed in vivo
G4 quadruplex sequences not folded in vivo
G4 quadruplex melted in vitro
G4 quadruplex formed in vitro
Formaldehyde crosslinks
Genomic DNA in the lysate
D
In vitro G4 DNA-IPdenat
In vitro Input
CTnuc
CTnuc Input
In vitro G4 DNA-IPnat
G4-ChIP
G4-ChIP Input
G4-ChIP Input 0.04 0.06 0.03 0.03 0.09 0.13 1.00
G4-ChIP 0.01 0.05 0.04 0.04 0.08 1.00 0.13
In vitro G4 DNA-IPnat 0.06 0.09 0.05 0.05 1.00 0.08 0.09
CTnuc Input -0.12 -0.00 0.26 1.00 0.05 0.04 0.03
CTnuc -0.13 -0.00 1.00 0.26 0.05 0.04 0.03
In vitro Input 0.29 1.00 -0.00 -0.00 0.09 0.05 0.06
In vitro G4 DNA-IPdenat 1.00 0.29 -0.13 -0.12 0.06 0.01 0.04
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
E
p<0.0001
p=0.0323
p<0.0001
p=0.0027
p<0.0001
p<0.0001
GGG/CCC weighted counts (n=peak counts)
1000
100
10
1
0.1
In vitro G4 DNA-IPnat
In vitro G4 DNA-IPdenat
G4-ChIP
CTnuc

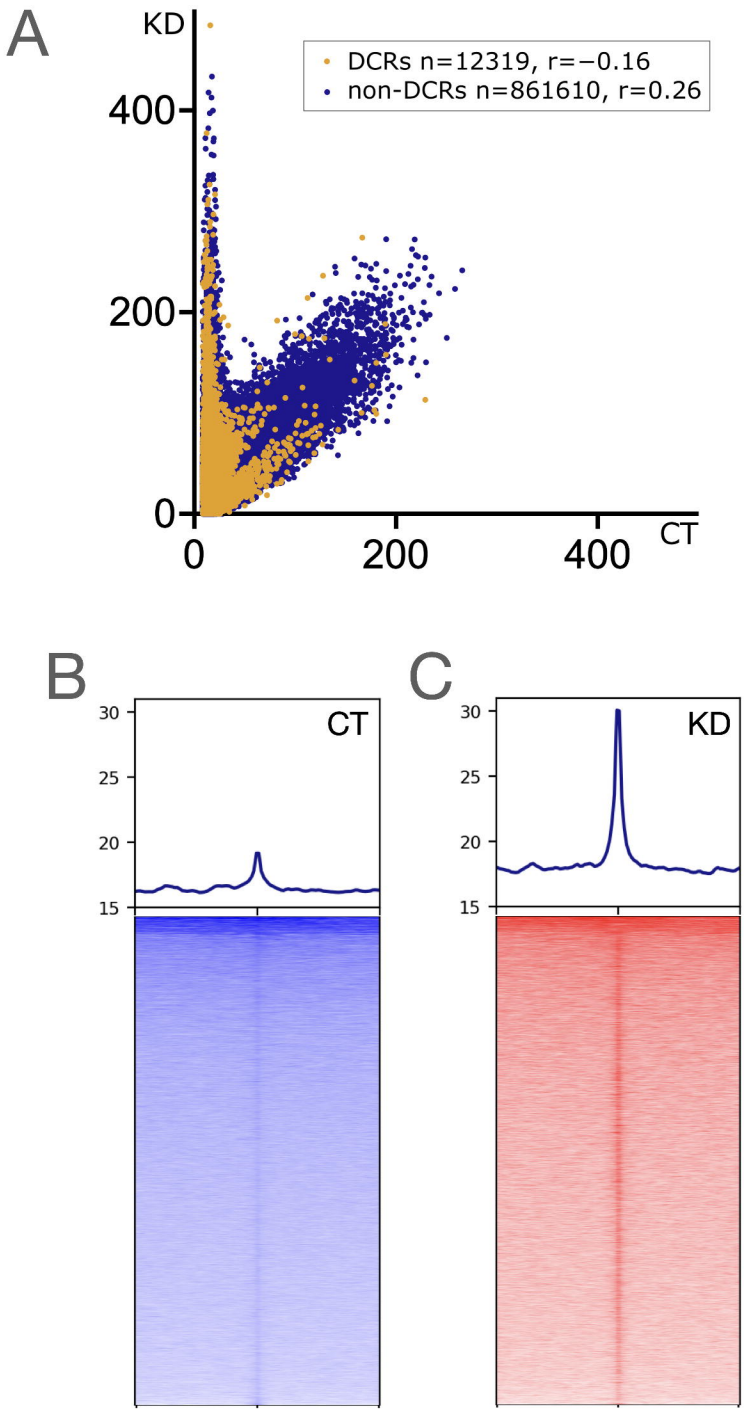

A
G4-Control DNA
TAATACGACTCACTATAGGGCACCCCTACATCGCAGCGGT
CTTTCGGGCTAACGGGATCATGGGACTCAGGGGACAGCCT
CAAGCAACATGCCCAGTCCTGACCTTCAATAAGGAAGCAA
ACTGGGAAGGAGGGTGTCAGGGATAAAGGGGAGTCCTAGT
CAAGGTGTCGGATGTCCTAAGACTTATGATCATTTTCTTA
GGGTCTAAGGGCTCGAGGGTCTGCGGGTCGGTTTCCTTCT
AGAATTAGTATCTTCTATAGTGTCACCTAAAT
G4-Control
pGEMT
+KCl
Formation of G4 quadruplexes
G4-Control quadruplexes
pGEMT
Antibody capture of G4 quadruplexes in vitro
Transfection
CTtr
KDtr
CTut
KDut
Antibody capture of G4 quadruplexes in cellulo
Antibody capture of G4 quadruplexes in cellulo
Antibody capture of G4 quadruplexes under chromatin constraint
- Genomic DNA
- G4 Control DNA
Chromatin constraint on G4 quadruplexes in
- Genomic DNA only
CGGBP1-dependent chromatin constraint affects G4 Control DNA recovery
CGGBP1-dependent chromatin constraint does not affect G4 Control DNA recovery

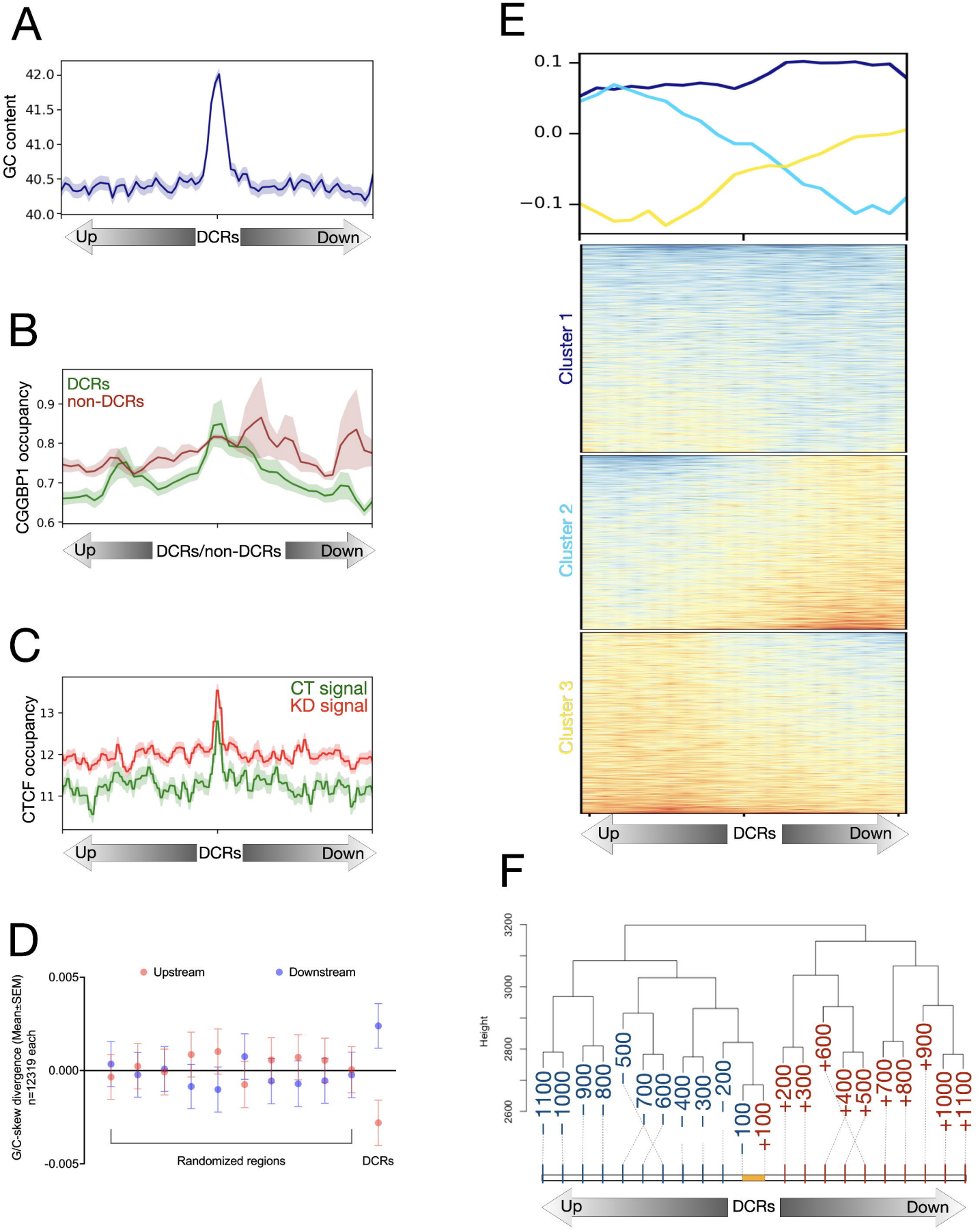

B
CTtr
KDtr


C
CTut
KDut


D
CTnuc
KDnuc


E
CT
CT Input
KD Input
KD
PC2 (20.86% var)
PC1 (75.79% var)
300
200
100
0
-100
-400
-200
0


A
KD
DCRs n=12319, r=−0.16
non-DCRs n=861610, r=0.26
400
200
0
0
200
400
CT
B
CT
30
25
20
15
C
KD
30
25
20
15


A
GC content
42.0
41.5
41.0
40.5
40.0
Up
DCRs
Down
B
CGGBP1 occupancy
DCRs
non-DCRs
0.9
0.8
0.7
0.6
Up
DCRs/non-DCRs
Down
C
CTCF occupancy
CT signal
KD signal
13
12
11
Up
DCRs
Down
D
G/C-skew divergence (Mean±SEM)
n=12319 each
Upstream
Downstream
0.005
0.000
-0.005
Randomized regions
DCRs
E
0.1
0.0
−0.1
Cluster 1
Cluster 2
Cluster 3
Up
DCRs
Down
F
Height
3200
3000
2800
2600
−1100
−1000
−900
−800
−500
−700
−600
−400
−300
−200
−100
+100
+200
+300
+600
+400
+500
+700
+800
+900
+1000
+1100
Up
DCRs
Down