CGGBP1-dependent CTCF-binding sites restrict ectopic transcription from SV40 early promoter

Divyesh Patel, Manthan Patel, Subhamoy Datta, Umashankar Singh
HoMeCell Lab, Biological Engineering, Indian Institute of Technology Gandhinagar, 382355, India.

ABSTRACT
Binding sites of the chromatin regulator protein CTCF function as important landmarks in the human genome. The recently characterized CTCF-binding sites at LINE-1 repeats depend on another repeat-regulatory protein CGGBP1. These CGGBP1-dependent CTCF-binding sites serve as potential barrier elements for epigenetic marks such as H3K9me3. Such CTCF-binding sites are associated with asymmetric H3K9me3 levels as well as RNA levels in their flanks. The functions of these CGGBP1-dependent CTCF-binding sites remain unknown. By performing targeted studies on candidate CGGBP1-dependent CTCF-binding sites cloned in an SV40 promoter-enhancer episomal system we show that these regions act as inhibitors of ectopic transcription from SV40 promoter. CGGBP1-dependent CTCF-binding sites that recapitulate their genomic function of loss of CTCF binding upon CGGBP1 depletion and H3K9me3 asymmetry in immediate flanks are also the ones which show the strongest inhibition of ectopic transcription. By performing a series of strand-specific reverse transcription PCRs we demonstrate that this ectopic transcription results in synthesis of RNA from the SV40 promoter in a direction opposite to the downstream reporter gene in a strand specific manner. The unleashing of the bidirectionality of the SV40 promoter activity and a breach of the transcription termination sequence required for the upstream transcription seems to depend on depletion of CGGBP1 and loss of CTCF binding proximal to the SV40 promoter. These findings suggest a role of CGGBP1-dependent binding sites in restricting ectopic transcription.

KEYWORDS
CT (Control non-targeting shRNA); KD (anti-CGGBP1 shRNA); SV40 (Simian Virus 40); CGGBP1-dependent CTCF-binding sites, H3K9me3; RNA Polymerase 2; ChIP (Chromatin immunoprecipitation); transcription.

INTRODUCTION

CTCF binding sites serve as insulator and chromatin barrier elements (Ong and Corces, 2014; Phillips and Corces, 2009). CTCF binds to insulators to organize genome in chromatin regulatory domains. Regulatory elements within one chromatin regulatory domain interact more frequently than regulatory elements across the domain (Arzate-Mejía et al., 2018; Nichols and Corces, 2015; Van Bortle et al., 2012). CTCF along with cohesin by partitioning the genome into regulatory domains establishes a cell-type specific gene expression (Hou et al., 2010; Ren and Zhao, 2019). The two loci pioneering the characterisation of the insulator function of CTCF were beta-globin locus control region and Igf2-H19 locus (Farrell et al., 2002; Han et al., 2008; Kurukuti et al., 2006; Ulaner et al., 2003; Valadez-Graham, 2004). CTCF-binding sites, along with those of other associated chromatin regulatory proteins, serve as the boundaries of the chromatin loops and higher order chromatin structures such as Topologically Associated Domains (TADs) (Dekker and Mirny, 2016; Giorgetti et al., 2016; Narendra et al., 2016). Thus, allowing the enhancer-promoter communication within TAD and preventing the inter-TAD interaction of the gene-regulatory elements (Dekker and Mirny,

2016; Galupa and Crocker, 2020; Ghirlando and Felsenfeld, 2016). Further, CTCF acts as a barrier element by marking the boundaries of heterochromatin domains that are enriched with repressive histone marks such as constitutive silencing H3K9me3 (may bound by HP1) and temporary silencing marks H3K27me3 (may bound by polycomb repressors) (Barkess and West, 2012; Kim et al., 2011; Lu et al., 2016). CTCF prevents the spread of heterochromatin into the gene-rich euchromatin (Barkess and West, 2012; Cuddapah et al., 2009).
Recently described CGGBP1-dependent CTCF-binding sites function as chromatin barrier elements (Patel et al., 2019). They restrict H3K9me3 signal spread and function as boundaries of H3K9me3-rich and H3K9me3 depleted regions (Patel et al., 2019). This property of CGGBP1-dependent CTCF-binding sites seems to not affect H3K27me3 and H3K4me3 levels (Patel et al., 2019). H3K9me3 is a major gene silencing epigenetic mark which is also employed for chromatin compaction and prevention of noisy transcription. This epigenetic mark is predominant in the Giemsa-positive gene-poor regions and hence aids in condensation of such DNA and rendering it unavailable for active transcription (Becker et al., 2016; Ninova et al., 2019). H3K9me3 plays a prominent role in repeat silencing and retrotransposon inactivation (Bulut-Karslioglu et al., 2014; van Kruijsbergen et al., 2017). The different kinds of repeats marked by H3K9me3 includes satellite repeats, tandem repeats and DNA transposons (Bulut-Karslioglu et al., 2014; Iglesias and Moazed, 2017). H3K9me3 plays a vital role in maintaining genomic integrity by preventing the genomic integration of the LTR and non-LTR retrotransposons, specifically interspersed repeats such as Alu-SINEs and L1-LINEs (Bulut-Karslioglu et al., 2014; van Kruijsbergen et al., 2017). Interestingly, the CGGBP1-dependent CTCF-binding sites are also mostly L1 repeat-rich and CTCF motif-poor. CTCF occupancy at these sites depends on CGGBP1 levels (Patel et al., 2019). This regulation of CTCF binding to repeat-derived binding sites depends on CGGBP1 levels but does not seem to require formation of a protein-protein complex between the two proteins. A comparison of CTCF occupancy between three different CGGBP1 levels (normal, depleted and overexpressed) suggests that there is a cooperative facilitation of CTCF-repeat binding by CGGBP1 (Patel et al., 2019). However, since CGGBP1 itself is a repeat binding protein, this cooperativity of CGGBP1-CTCF binding is lost upon CGGBP1 overexpression just the way it is lost upon CGGBP1 depletion (Patel et al., 2019). The repeat-origins of CTCF-binding sites are established. In the primates, Alu-SINEs have diverged into a large number of CTCF-binding sites and they occur in millions of locations in the human genome (Schmidt et al., 2012). However, CGGBP1-dependent CTCF-binding sites seem to have preference for L1-LINEs. Remarkably, the CGGBP1-dependent CTCF-binding sites in L1 repeats concentrate on motif-like subsequences that are common between Alus and L1 elements (Patel et al., 2019). Thus, even with disparate evolutionary origins, L1 and Alu repeat function as sites where the CGGBP1-CTCF axis operates to regulate the patterns of H3K9me3 patterns. The functional significance of the CGGBP1-dependent CTCF-binding sites however remain unclear. This is especially interesting given the evolutionary unrelatedness of CTCF and CGGBP1. The former is conserved in vertebrates whereas the latter is present only in amniotes. One significance of the CGGBP1 regulation of CTCF binding is highlighted by our recent work that CGGBP1 levels regulate cytosine methylation at CTCF-binding motifs in a non-stochastic manner (Patel et al., 2018, 2020).

The proteins with which CTCF and CGGBP1 form complexes shed some light on the possible functions of CGGBP1-dependent CTCF-binding sites and the mechanisms through which they are regulated. CGGBP1 itself is a component of the enhancer-binding proteins complexes containing YY1 and CTCF (Weintraub et al., 2017). The histone

methyltransferases SUV4 and SUV39 family member enzymes form complexes with cohesin ring family members and thus associate with CTCF (Hahn et al., 2013). HMT SUV39H2 is also associated with CGGBP1 (Singh and Westermark, 2015; Singh et al., 2011). Although a fraction of CTCF and CGGBP1 do co-immunoprecipitate with each other, such indirect interactions seem to direct the coregulation between CTCF and CGGBP1 at CGGBP1-dependent CTCF-binding sites (Patel et al., 2019). Nucleophosmin forms complexes with both CTCF and CGGBP1 (Hein et al., 2015; Yusufzai et al., 2004). Interestingly, boundaries of the L1-rich lamina associated domains show contrasting levels of CTCF occupancy outside and within the LAD. This LAD boundary specific pattern also depends on the functions of CGGBP1. However, the CTCF-CGGBP1 complexes detected *in situ* do not localize to the nuclear periphery (Patel et al., 2019).

CGGBP1 has been implicated in regulation of transcription of interspersed repeats. It is also required for normal RNA Polymerase 2 activity. Upon growth stimulation of normal human fibroblasts, the transcript elongation by RNA Polymerase 2 seems to depend on the levels of CGGBP1 (Agarwal et al., 2016; Cardiello et al., 2014; Ichiyanagi, 2014; Patel et al., 2019; Singh and Westermark, 2015). Circumstantial evidence suggests that this H3K9me3 asymmetry in the flanks of CGGBP1-dependent CTCF-binding sites may lead to asymmetrical RNA abundance in 10 kb long flanking regions (Patel et al., 2019). CTCF binding sites exert a transcription repressive effect in *cis*. CTCF was first reported as a transcriptional repressor of the chicken c-myc gene (Holwerda and de Laat, 2013; Lobanenkov et al., 1990). Further studies have found transcriptional activator activity of the CTCF (Klenova et al., 1993; Lobanenkov et al., 1990). The regulatory function of CTCF binding sites is determined by its interacting proteins, the location of CTCF binding sites relative to the transcription site of a gene (Holwerda and de Laat, 2013; Lobanenkov et al., 1990; Nishana et al., 2020).

Thus, there seems to be a functional link between CGGBP1-dependent CTCF-binding sites, regulation of specific histone modifications including H3K9me3 and RNA Polymerase 2 activity. Here we have investigated the effects CGGBP1 levels exerts on how the candidate CGGBP1-dependent CTCF-binding sites affect histone modifications, transcript levels and RNA Polymerase 2 occupancy. By using an episomal system that drives transcription from SV40 promoter under the control of SV40 enhancer , we show that CGGBP1 depletion promotes ectopic transcription activity from SV40 promoter. We report that the early transcriptional activity downstream of the SV40 promoter remains unaffected by CGGBP1 depletion. However, the upstream late transcription, that normally depends on the SV40 enhancer, is induced upon CGGBP1 depletion. This generates an ectopic strand-specific transcription from the SV40 promoter. By using different CTCF-binding sites juxtaposed with the SV40 promoter we have found that this effect of CGGBP1 depletion is linked to the presence of CGGBP1-dependent CTCF-binding sites. These results shed light on the mechanisms of action of CGGBP1-dependent CTCF-binding sites and highlight the role of CGGBP1 in transcription regulation that can bypass the enhancer-dependence of promoters.

RESULTS AND DISCUSSION

Genomic CGGBP1-dependent CTCF-binding sites and characterization of LoB5, a candidate CGGBP1-dependent CTCF-binding site

The locations of the 879 CGGBP1-dependent CTCF-binding sites were analyzed with respect to the known TSSs (Fantom database). These CGGBP1-dependent CTCF-binding sites were located at long distances from the TSSs (185 kb ± 388.021 kb) in the gene-poor regions and did not seem to be involved in *cis*-regulation of gene expression. We could detect transcripts generated from some of these TSS pairs of the same genes (such that the TSS pairs were located on either side of the CGGBP1-dependent CTCF-binding site) which were the closest to the CGGBP1-dependent CTCF-binding sites (Table 1). However, there was no consistent effect of CGGBP1 depletion on transcript levels of these genes (Figure 1A). A survey of transcript levels derived from five such TSSs revealed that although some pairs of TSSs do display a differential activity of the two TSSs in CT that is lost in KD, such an effect was not observed consistently in the selected TSS pairs (Figure 1A). For some genes (NRXN2 and OPRL) the transcript level differences were in line with the RNA Polymerase occupancy observed at the same regions suggesting that to some extent the transcript levels reflect the activity of RNA Polymerase 2 at these TSSs (Figure 1B).
One of the known functions of these CTCF-binding sites is the maintenance of differential levels of H3K9me3 in the flanks such that the differences depend on the levels of CGGBP1. The *cis* length range in which H3K9me3 asymmetry has been studied earlier is 10 kb. It was not clear however if these TSSs were under the influence of H3K9me3 silencing or not. We selected some CGGBP1-dependent CTCF-binding sites with highest reported asymmetries in the cumulative H3K9me3 signals in their 10 kb flanks and verified two important parameters at them for further studies: (i) CGGBP1-dependence of CTCF occupancy at them, and (ii) H3K9me3 asymmetry in their immediate flanks. Three of these regions which we pursued are called LoB3, LoB5 and GoB4. The genomic locations and contexts of these CGGBP1-dependent CTCF-binding sites are shown in figure 2 (Fig 2, A to C). We identified one region, called LoB5, where we could consistently perform specific PCR amplification. At LoB5, CTCF occupancy was lost upon CGGBP1 depletion (Fig 3A). The other two regions LoB3 and GoB4 showed a weaker loss and gain of CTCF binding respectively upon CGGBP1 depletion (Fig 3, B and C). Also, the H3K9me3 levels were asymmetric in LoB5 flanks only in the presence of CGGBP1 and was lost upon CGGBP1 depletion (Fig 3D). The immediate flank transcription at LoB5 also exhibited expected asymmetry in CT as well as KD (not shown). Although GoB4 and LoB3 showed some CGGBP1-dependence of CTCF binding, they did not recapitulate the expected H3K9me3 levels in flanks (Fig 3, E and F). Thus, LoB5 presented us with a region where the CTCF and H3K9me3 ChIP-seq data could be independently verified. These findings suggested that LoB5 could be used as a model region to explore the role of CTCF-CGGBP1 axis at such candidate barrier elements.
We cloned LoB5 in an episomal vector system pGL3-Control. The LoB5 element was inserted upstream of the SV40 promoter in *Kpn*I-*Xho*I sites (Fig 4A; the figure also shows the locations of the various regions and primer binding sites used further on in various experiments). Unlike the endogenous LoB5 locus that is distant from TSSs, in this construct the LoB5 element was artificially juxtaposed against an RNA Polymerase 2 promoter. We tested if the LoB5 element in the episomal system retained its properties of CGGBP1-dependent CTCF binding and H3K9me3 asymmetry in immediate flanks. ChIP-qPCRs revealed that LoB5 is a potent CTCF-binding site in CT. Upon CGGBP1 knockdown CTCF-binding at episomal LoB5 was lost (Fig 4B). This was verified using primer pairs that amplified the endogenous LoB5 and episomal LoB5 exclusively or commonly. Similarly, constructs containing LoB3 and GoB4 were also subjected to CTCF ChIP-qPCRs in CT and KD but the expected loss of binding of CTCF upon CGGBP1 depletion were not observed on those clones (Fig 4, C and D). The H3K9me3 levels in the upstream and downstream regions of episomal LoB5 showed the same effect of CGGBP1 depletion as was observed for the genomic endogenous LoB5 locus (Fig 4E). Again for LoB3 and GoB4 constructs, the

H3K9me3 levels in the flanks of the inserts did not mimic the expected genomic H3K9me3 patterns (Fig 4, F and G). We thus focussed on LoB5 construct as an episomal system to study the functions of LoB5 as a model CGGBP1-dependent CTCF-binding site.

Regulation of LoB5-SV40 promoter activity by CGGBP1

We next tested if CGGBP1 levels affected the LoB5-SV40 promoter activity. In this episomal system the SV40 promoter activity is driven by an upstream enhancer located ~2 kb away (Fig 4A) (Benoist and Chambon, 1981; Kadesch and Berg, 1986; Kelly and Wildeman, 1991; Sassone-Corsi et al., 1984; Shaw et al., 1985). Compared to the empty vector (SV40 promoter), the LoB5-SV40 promoter did not show any significant difference in the promoter activity as measured by Firefly Luciferase activity (Fig 5). In these experiments we used Renilla Luciferase as an internal control for normalization of systemic variables (Fig 5).
The SV40 promoter is a bidirectional promoter (Byrne et al., 1983; Hertz and Mertz, 1988). Its basal downstream activity is driven by the 21 bp element that contains the CAAAT box and is proximal to the TSS (Byrne et al., 1983; Hertz and Mertz, 1988). The upstream transcriptional activity is dependent on the 70 bp repeat elements located distal to the TSS (Wasylyk et al., 1983). We first measured the transcript levels of the Luciferase gene. Luciferase transcript levels were not changed by CGGBP1 depletion (Fig 6A). Since CGGBP1 depletion also caused a loss of CTCF binding, these findings reinforced that a CGGBP1, and potentially CTCF as well, do not act as *cis* regulators of SV40 promoter activity. The upstream enhancer drives the SV40 promoter (Hertz and Mertz, 1988) and insertion of foreign DNA upstream of the SV40 promoter in pGL3-Control has been used as a tool to identify potential insulator sequences which can block the communication between enhancer and the promoter. However, these findings also suggested that there is no insulator-like activity of the LoB5 element in the episomal system.
Next, we measured transcript levels from various other regions of the episome. We found that just like the Luciferase gene, at the LoB5-SV40 promoter, AmpR gene, Ori and f1Ori there were no significant changes in transcript levels in KD compared to CT (Fig 6B). However, the SV40 enhancer showed a mild decrease (Fig 6B).
We further investigated if CGGBP1 depletion affected the occupancy of RNA Polymerase 2 in all these regions. Unexpectedly, the RNA Polymerase 2 occupancy was reduced in KD as compared to CT at all the regions except the LoB5-SV40 promoter and the Luciferase gene (Fig 6C). We then calculated a ratio of transcript abundance and RNA Polymerase 2 occupancy to gauge the transcript productivity of RNA Polymerase 2 presence on the DNA. We found that f1Ori had the highest transcript abundance to RNA Polymerase 2 occupancy ratio followed by the SV40 enhancer (derived from data shown in Fig 6, B and C; not shown). This indicated that upon CGGBP1 depletion, transcription activity at f1Ori increases. At the same time, the coupling between SV40 enhancer and the LoB5-SV40 promoter, that is expected to drive the promoter activity, was not retained in KD. Upon CGGBP1 knockdown, the increase in the transcriptionally active RNA Polymerase 2 presence at SV40 enhancer did not produce a concomitant increase in Luciferase gene transcript levels. However, highly similar increase in RNA Polymerase 2 occupancy and transcript abundance at f1Ori and SV40 Enhancer suggested that upon CGGBP1 depletion, the SV40 Enhancer and LoB5-SV40 promoter couple to drive transcription upstream towards the f1Ori.

Restriction of bidirectional transcription from LoB5-SV40 promoter by CGGBP1

For the possibility of transcription of f1Ori by LoB5-SV40 promoter to materialize, the RNA Polymerase 2 activity from SV40 promoter must overcome the transcription pause signal located between the MCS and the f1Ori (Fig 4A). Thereby, transcripts must be generated that span the region from LoB5-SV40 promoter to the f1Ori.

PCRs on randomly primed cDNA using primers located in f1Ori and SV40 revealed that transcripts spanning f1Ori and SV40 promoter are formed for LoB3, LoB5 as well as GoB4 (Fig 7A). The amplification was observed from DNase-digested RNA templates but not observed upon digestion with RNaseH or RNaseI (not shown). Interestingly, the CGGBP1-dependence of this SV40 promoter-f1Ori transcript was observed very strongly for LoB5, weakly for GoB4 and not at all for LoB3 (Fig 7A). The same length of SV40 promoter-f1Ori transcript was obtained using oligo-dT for cDNA synthesis followed by PCR using primers located in f1Ori and SV40 Promoter (Fig 7B). These results suggested that the SV40 promoter-f1Ori transcript is polyadenylated at a location 3´ to the f1Ori. The depletion of CGGBP1 thus allows RNA Polymerase 2 to breach the transcription pause and polyA sites located between SV40 promoter and f1Ori (Fig 4A). As a control, we tested if the polyA signal and RNA Polymerase 2 pause site between the 3' end of the Luciferase gene and SV40 enhancer is also breached. Using strand-specific cDNA synthesized using P5 or P10 (Fig 4A), we were not able to amplify any products using one primer in SV40 enhancer and another in SV40 promoter (not shown). Thus, the directionality of SV40 promoter towards Luciferase gene was not altered but ectopic transcription from SV40 promoter through LoB5 towards f1Ori occurred upon CGGBP1 depletion.
To further characterize the direction of transcription of the SV40-LoB5-f1Ori we performed strand-specific PCRs. cDNA were prepared from the terminal primers located in f1Ori or SV40 promoter and the PCRs were performed using both these primers. The entire ~ 2 kb long product was amplified from cDNA synthesized using f1Ori forward primer only (Fig 7B). The cDNA generated using the SV40 promoter reverse primer did not give rise to any product (Fig 7B). The f1Ori forward primed cDNA showed stronger amplification of the SV40-LoB5-f1Ori product in KD and only minimal amplification was detected in CT (Fig 7B). The 1.97 kb long PCR product was sequence-verified using two opposite outgoing primers in the LoB5 (data not shown). Further, by using the cDNA generated using the f1Ori terminal primer, we performed PCRs for terminal fragments in the 2kb SV40-LoB5-f1Ori transcript (Fig 7C). The f1Ori levels were much less compared to the levels at Sv40 promoter (Fig 7C). These findings indicated that the SV40 promoter-f1Ori transcript synthesis begins at the SV40 promoter and due to incomplete synthesis of transcripts and truccations before f1Ori, the levels of the 3´ end of the transcript are lower than those at the 5´end.
Together, these findings showed that upon CGGBP1 depletion the SV40 promoter activity is driven to synthesize the SV40-LoB5-f1Ori transcript that is in the opposite direction and on the strand complementary to the Luciferase gene sense strand. Interestingly, depletion of CGGBP1 allowed this atypical transcription to occur despite the presence of the TTS that normally restricts transcription. However, this ability of RNA Polymerase 2 to breach the pause site was observed only for the SV40-upstream region that contained LoB5. For the downstream TTS located between the Luciferase gene and SV40 enhancer no ectopic transcription was observed.
The LoB5 element recapitulates some of its key endogenous properties to the episomal system. The LoB5 belongs to a set of regions that exhibit CGGBP1-dependent CTCF binding and are rich in L1 repeats. The L1 repeats are also strong binding sites for SV40-LargeT antigen. This raised the possibility that the SV40 T-antigen binding sites in the L1-rich CGGBP1-dependent CTCF-binding sites genome-wide could exhibit ectopic transcription in the absence of CGGBP1.

The SV40-LargeT antigen binding sites on L1 repeats are rich in SP1 binding motifs and SP1 binding on SV40 promoter is required for bidirectional transcription (Gidoni et al., 1985; Gruss et al., 1988). In order to study this further, we are generating a transcriptomic landscape that will reveal the scale of ectopic strand-specific non-coding transcription at

CGGBP1-dependent CTCF-binding sites and its possible inhibition by CGGBP1. To establish strand-specificity, we are standardizing a new protocol to identify strand-specific transcripts from gene-poor regions. More work in this direction will reveal how CTCF-binding cooperated with other proteins, such as CGGBP1, to shape the non-coding transcriptome.

Table and Figure Legends

Table 1: The selected alternative TSS pairs of genes with least distances to the CGGBP1-regulated CTCF-binding sites. The table shows the gene symbols and the locations of the TSSs from the nearest CGGBP1-regulated CTCF-binding sites.

Figure 1: RNA levels and RNA Polymerase 2 occupancy at the TSSs of selected genes tabulated in Table 1. **A:** The RNA levels at the TSSs of the genes are calculated by double delta Ct method. Beta actin (ACTB) has been used as a quantitation control in all the PCRs. The Y axis shows the ddCt values on a log2 scale. The location of the TSSs relative to the nearest CGGBP1-regulated CTCF-binding sites are indicated by suffixes “upstream” or “downstream”. **B:** RNA Polymerase 2 occupancy at the same TSSs as shown in **A** shows that apart from NRXN2, for none of the TSSs showed a correspondence between RNA levels and RNA Polymerase 2 occupancy. For PBX1 the Ct values obtained were too low to be used reliably for ddCt analysis and were eliminated. The levels of amplification from the input DNA was used as a quantitation control against respective ChIPs. The Y axis shows the ddCT values on a log2 scale.

Figure 2: Characterisation of CGGBP1-regulated CTCF binding sites for repeats and CTCF-motif. **A-C:** Genome browser views showing the distribution of CTCF ChIP-seq reads, predicted CTCF binding motifs by MEME-suite FIMO and CTCFBSDB and permissive TSS. Repeat tracks show distribution of the LINE-1, LINE-2, Alu, MIR-SINEs, Simple repeats and DNA-repeat elements. **A:** The genome browser view on chromosome 1 (~Chr1:16976 kb - 16996 kb) shows loss of CTCF ChIP-signal at LoB5 upon CGGBP1 depletion. Alu-SINEs and MIR-SINEs are clustered in immediate flanks of the LoB5 CTCF binding sites. The nearest permissive TSS is approximately 8 kb upstream from LoB5 site. **B:** The genome browser view on chromosome 16 (~Chr16:12564 kb - 12584 kb) represents GoB4 CTCF binding site. CTCF ChIP-signal is potentiated in KD at GoB4 site. Predicted CTCF motifs (EMBL_M1 and MIT_LM23) are present in the GoB4 CTCF binding site. Interspersed repeats (LINE-1, LINE-2, Alu and MIR-SINEs) and DNA-elements are clustered downstream of the GoB4 CTCF binding site. **C:** The genome browser view on chromosome 2 (~Chr2:128362 kb - 128382 kb) represents LoB3 CTCF binding site. CTCF ChIP-signal is reduced at LoB3 CTCF binding site. SINEs (Alu and MIR-SINEs) and DNA repeats are enriched in the upstream and downstream of the LoB3 CTCF binding site respectively.

Figure 3: LoB5 shows expected loss of CTCF occupancy and H3K9me3 asymmetry upon CGGBP1 depletion. **A-C:** LoB5, GoB4 and LoB3 sites were cloned in pGL3-control vector. Cloned plasmid were transfected in HEK293T cells with normal and depleted levels of CGGBP1 separately. Total CTCF occupancy at CTCF binding sites (Endogenous and episomal) was compared between CT and KD for LoB5, GoB4 and LOB3. The Ct value for CTCF enrichment was normalised with input and CTCF enrichment are plotted in arbitrary units. Statistical significance was determined by using the unpaired t-test (p-value = 0.05). CTCF binding is strongly reduced at the LoB5 site in KD **(A)**. Depletion of CGGBP1 leads to decreased occupancy of CTCF at GoB4 site **(B)** and increased occupancy at LoB3 site **(C)**. **D-F:** H3K9me3 levels in the immediate flanks of the endogenous CTCF binding sites were

compared between CT and KD. The Ct value for H3K9me3 enrichment was normalised with input and H3K9me3 enrichment is represented in arbitrary units (y-axis). Statistical significance was determined by using the unpaired t-test (p-value = 0.05). H3K9me3 levels show significant asymmetry in the immediate flanks of the endogenous LoB5 CTCF binding site in CT (p value < 0.05). The decrease in H3K9me3 levels in the downstream region causes loss of H3K9me3 asymmetry in KD **(D)**. A non-significant asymmetry in H3K9me3 levels is observed in the immediate downstream flanks of GoB4 CTCF binding site in CT. However, H3K9me3 levels show mild but significant asymmetry in KD **(E)**. The immediate flanks of the LoB3 sites maintained similar levels of H3K9me3 in CT and KD **(F)**.

Figure 4: Characterisation of CGGBP1-regulated CTCF binding sites in episomal system. **A:** Schematic representation of the pGL3-control episomal vector sysyem. Candidate LoB and GoB sites along with approximately 250 bp flanking sequences were cloned upstream of the SV40 promoter. The schematic shows LoB5 CTCF binding site along with primers used to determine CTCF binding, transcript levels, RNA-Polymerase II, H3K9me3 and H3K27me3 levels at different regions of the episome. **B-D:** CTCF occupancy at episomal LoB and GoB sites were compared between CT and KD by using primer P3 (cloned CTCF sites forward) and P6 (SV40 promoter reverse). The Ct value for CTCF enrichment was normalised with input and CTCF enrichment are plotted in arbitrary units. Statistical significance was determined by using the unpaired t-test (p-value = 0.05). Episomal LoB5 CTCF binding has shown a strong decrease in CTCF binding in KD **(B)**. CTCF occupancy at the episomal GoB4 site increases significantly **(C)**, similarly, LoB3 site in the episome show significant increase in CTCF binding upon CGGBP1 depletion **(D)**. **E-G:** H3K9me3 levels at f1Ori and SV40 promoter located upstream and downstream respectively of the episomal CTCF binding sites were compared between CT and KD. The Ct value for H3K9me3 enrichment was normalised with input and H3K9me3 levels are represented as arbitrary units (y-axis). Statistical significance was determined by using the unpaired t-test (p-value = 0.05). H3K9me3 levels are comparable in upstream and downstream of the episomal LoB5 site in CT and KD **(E)**. The upstream (f1Ori) of the episomal GoB4 CTCF binding site display comparatively reduced levels of the H3K9me3 than SV40 promoter in CT. The asymmetric distribution of the H3K9me3 is potentiated in KD **(F)**. Comparable levels of H3K9me3 are observed at f1Ori and SV40 promoter in CT. A strong decrease in H3K9me3 levels at f1Ori increases asymmetry of H3K9me3 levels in immediate flanks of episomal LoB3 site in KD.

Figure 5: Luciferase reporter gene assays. For the three episomal constructs LoB3, LoB5 and GoB4, the Luciferase reporter gene activity was determined to measure the *cis*-regulatory effect of SV40 promoter (justaposed with LoB3, LoB5 or GoB4 respectively). The luminescence values for each sample (X axis) is reported on the Y axis (log 10 scale). Values are from separate experiments and account for multiple systemic experimental variations. Values shown are mean and standard deviations. The suffixes FL, RL and FL/RL stand for luminescence values of Firefly Luciferase, Renilla Luciferase and ratio of Firefly Luciferase to Renilla Luciferase respectively. Renilla Luciferase values serve as an internal control.

Figure 6: Regulation of the LoB5-SV40 promoter activity by CGGBP1. **A:** pGL3-LoB5 construct was transfected in HEK293T cells with normal and depleted levels of CGGBP1. Transcript levels were compared at different episomal regions between CT and KD. SV40 Promoter and FF luciferase show a non-significant increase in transcript levels upon removal of CGGBP1. Similarly, immediate upstream located f1Ori and AmpR regions also show a non-significant increase in transcript levels in the absence of CGGBP1. However, SV40

enhancer exhibits a strong decline in transcriptional activity, while Ori remains immune to any significant transcriptional changes upon CGGBP1 depletion. **B:** RNA Polymerase II occupancy was compared across the episomal landscape between CT and KD. In agreement with transcript levels, the SV40 promoter and FF luciferase do not show significant changes in RNA polymerase II levels. However, f1Ori and AmpR, the immediate upstream regions of the episomal LoB5 site, show significant decrease in RNA Polymerase II levels in KD. Similarly, RNA Polymerase II levels at SV40 enhancer and Ori portray significant reduction in KD.

Figure 7: CGGBP1 regulates atypical promoter activity of the SV40 promoter in pGL3-Control vector system. **A:** Randomly primed cDNA from LoB3, LoB5 and GoB4 pGL3-control construct transfected CT and KD cells were used to detect transcripts spanning CGGBP1-regulated CTCF binding sites by using primer P1 (f1Ori Forward) and P6 (SV40 Promoter Reverse). The strongest difference in transcript levels was observed in LoB5 followed by GoB4. **B:** RNA was isolated from pGL3-control LoB5 construct transfected cells. To study atypical promoter activity of the LoB5-SV40 promoter strand-specific cDNA were synthesised from DNaseI digested RNA by using promoter P1 (f1Ori Forward) and P6 (SV40 Promoter Reverse) primers. The full-length 1.97 kb PCR product was detected using oligo-dT primed cDNA as well as P1 primed cDNA but not P6 primed cDNA. For oligo-dT as well as P1 primed cDNA, the levels of this transcript were higher in KD than in CT. Using P6 primed cDNA multiple non-specific fragments of smaller molecular weights were generated. As a control the DNase-digested RNA template did not yield any PCR amplification. **C:** RNA was isolated from pGL3-control LoB5 construct transfected and untransfected cells. To study atypical promoter activity of the LoB5-SV40 promoter strand-specific cDNA were synthesised from DNaseI digested RNA by using promoter P1 (f1Ori Forward) and P6 (SV40 Promoter Reverse). As expected, no specific PCR product (1.97 Kb) was detected by using RNA as a template from untransfected CT and KD cells. However, low levels of the specific PCR product were observed by using DNaseI digested RNA from LoB5 pGL3 control construct transfected CT and KD cells. The specific amplification was comparatively higher in KD. Such a weak amplification may be due to weak reverse transcriptase activity of the Taq DNA Polymerase. Strand Specific transcriptions were compared at f1Ori and SV40 Promoter. Very low transcriptional levels at f1Ori were observed as compared to SV40 promoter by using cDNA from P1 (f1Ori Forward). KD has shown higher transcript levels compared to CT at SV40 promoter and f1Ori. Similarly, no specific PCR products were observed at f1Ori and weak transcriptional activity were observed as SV40 promoter by using cDNA from P6 (SV40 Promoter Reverse).

REFERENCES

Agarwal, P., Enroth, S., Teichmann, M., Jernberg Wiklund, H., Smit, A., Westermark, B., and Singh, U. (2016). Growth signals employ CGGBP1 to suppress transcription of Alu-SINEs. Cell Cycle *15*, 1558–1571.

Arzate-Mejía, R.G., Recillas-Targa, F., and Corces, V.G. (2018). Developing in 3D: the role of CTCF in cell differentiation. Development *145*.

Barkess, G., and West, A.G. (2012). Chromatin insulator elements: establishing barriers to set heterochromatin boundaries. Epigenomics *4*, 67–80.

Becker, J.S., Nicetto, D., and Zaret, K.S. (2016). H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. Trends Genet. *32*, 29.

Benoist, C., and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promotor region. Nature *290*, 304–310.

Bulut-Karslioglu, A., De La Rosa-Velázquez, I.A., Ramirez, F., Barenboim, M., Onishi-Seebacher, M., Arand, J., Galán, C., Winter, G.E., Engist, B., Gerle, B., et al. (2014). Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. Mol. Cell *55*, 277–290.

Byrne, B.J., Davis, M.S., Yamaguchi, J., Bergsma, D.J., and Subramanian, K.N. (1983). Definition of the simian virus 40 early promoter region and demonstration of a host range bias in the enhancement effect of the simian virus 40 72-base-pair repeat. Proc. Natl. Acad. Sci. U. S. A. *80*, 721–725.

Cardiello, J.F., Kugel, J.F., and Goodrich, J.A. (2014). A new twist on cell growth control. Cell Cycle *13*, 3474–3475.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.-Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. *19*, 24–32.

Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110–1121.

Farrell, C.M., West, A.G., and Felsenfeld, G. (2002). Conserved CTCF Insulator Elements Flank the Mouse and Human β-Globin Loci. Molecular and Cellular Biology *22*, 3820–3831.

Galupa, R., and Crocker, J. (2020). Enhancer–Promoter Communication: Thinking Outside the TAD. Trends in Genetics *36*, 459–461.

Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. Genes Dev. *30*, 881–891.

Gidoni, D., Kadonaga, J.T., Barrera-Saldaña, H., Takahashi, K., Chambon, P., and Tjian, R. (1985). Bidirectional SV40 transcription mediated by tandem Sp1 binding interactions. Science *230*, 511–517.

Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E., et al. (2016). Structural organization of the inactive X chromosome in the mouse. Nature *535*, 575–579.

Gruss, C., Wetzel, E., Baack, M., Mock, U., and Knippers, R. (1988). High-affinity SV40 T-antigen binding sites in the human genome. Virology *167*, 349–360.

Hahn, M., Dambacher, S., Dulev, S., Kuznetsova, A.Y., Eck, S., Wörz, S., Sadic, D., Schulte, M., Mallm, J.-P., Maiser, A., et al. (2013). Suv4-20h2 mediates chromatin compaction and is important for cohesin recruitment to heterochromatin. Genes Dev. *27*, 859–872.

Han, L., Lee, D.-H., and Szabó, P.E. (2008). CTCF Is the Master Organizer of Domain-Wide Allele-Specific Chromatin at the H19/Igf2 Imprinted Region. Molecular and Cellular Biology *28*, 1124–1135.

Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., Buchholz, F., et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell *163*, 712–723.

Hertz, G.Z., and Mertz, J.E. (1988). The enhancer elements and GGGCGG boxes of SV40 provide similar functions in bidirectionally promoting transcription. Virology *163*, 579–590.

Holwerda, S.J.B., and de Laat, W. (2013). CTCF: the protein, the binding partners, the

binding sites and their chromatin loops. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20120369.

Hou, C., Dale, R., and Dean, A. (2010). Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proceedings of the National Academy of Sciences *107*, 3651–3656.

Ichiyanagi, K. (2014). Regulating Pol III transcription to change Pol II transcriptome. Cell Cycle *13*, 3625–3626.

Iglesias, N., and Moazed, D. (2017). Heterochromatin: Silencing repetitive DNA.

Kadesch, T., and Berg, P. (1986). Effects of the position of the simian virus 40 enhancer on expression of multiple transcription units in a single plasmid. Mol. Cell. Biol. *6*, 2593–2601.

Kelly, J.J., and Wildeman, A.G. (1991). Role of the SV40 enhancer in the early to late shift in viral transcription. Nucleic Acids Research *19*, 6799–6804.

Kim, Y.J., Cecchini, K.R., and Kim, T.H. (2011). Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. Proc. Natl. Acad. Sci. U. S. A. *108*, 7391–7396.

Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E., and Lobanenkov, V.V. (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. Mol. Cell. Biol. *13*, 7612–7624.

van Kruijsbergen, I., Hontelez, S., Elurbe, D.M., van Heeringen, S.J., Huynen, M.A., and Veenstra, G.J.C. (2017). Heterochromatic histone modifications at transposons in Xenopus tropicalis embryos. Dev. Biol. *426*, 460.

Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W., and Ohlsson, R. (2006). CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc. Natl. Acad. Sci. U. S. A. *103*, 10684–10689.

Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V., and Goodwin, G.H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene *5*, 1743–1753.

Lu, Y., Shan, G., Xue, J., Chen, C., and Zhang, C. (2016). Defining the multivalent functions of CTCF from chromatin state and three-dimensional chromatin interactions. Nucleic Acids Res. *44*, 6200–6212.

Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E.O., and Reinberg, D. (2016). CTCF-mediated topological boundaries during development foster appropriate gene regulation. Genes Dev. *30*, 2657–2662.

Nichols, M.H., and Corces, V.G. (2015). A CTCF Code for 3D Genome Architecture. Cell *162*, 703–705.

Ninova, M., Fejes Tóth, K., and Aravin, A.A. (2019). The control of gene expression and cell identity by H3K9 trimethylation. Development *146*.

Nishana, M., Ha, C., Rodriguez-Hernaez, J., Ranjbaran, A., Chio, E., Nora, E.P., Badri, S.B., Kloetgen, A., Bruneau, B.G., Tsirigos, A., et al. (2020). Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation. Genome Biol. *21*, 1–34.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nature Reviews Genetics *15*, 234–246.

Patel, D., Patel, M., Westermark, B., and Singh, U. (2018). Dynamic bimodal changes in CpG and non-CpG methylation genome-wide upon CGGBP1 loss-of-function. BMC Research Notes *11*.

Patel, D., Patel, M., Datta, S., and Singh, U. (2019). CGGBP1 regulates CTCF occupancy at repeats. Epigenetics & Chromatin *12*.

Patel, M., Patel, D., Datta, S., and Singh, U. (2020). CGGBP1-regulated cytosine methylation at CTCF-binding motifs resists stochasticity. BMC Genet. *21*, 84.

Phillips, J.E., and Corces, V.G. (2009). CTCF: Master Weaver of the Genome. Cell *137*, 1194–1211.

Ren, G., and Zhao, K. (2019). CTCF and cellular heterogeneity. Cell & Bioscience *9*.

Sassone-Corsi, P., Dougherty, J.P., Wasylyk, B., and Chambon, P. (1984). Stimulation of in Vitro Transcription from Heterologous Promoters by the SV40 Enhancer. Transfer ‖ Expression of Eukaryotic Genes 7–21.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell *148*, 335–348.

Shaw, P.E., Bohmann, D., and Sergeant, A. (1985). The SV40 enhancer influences viral late transcription in vitro and in vivo but not on replicating templates. The EMBO Journal *4*, 3247–3252.

Singh, U., and Westermark, B. (2015). CGGBP1—an indispensable protein with ubiquitous cytoprotective functions. Upsala Journal of Medical Sciences *120*, 219–232.

Singh, U., Roswall, P., Uhrbom, L., and Westermark, B. (2011). CGGBP1 regulates cell cycle in cancer cells. BMC Mol. Biol. *12*, 28.

Ulaner, G.A., Yang, Y., Hu, J.-F., Li, T., Vu, T.H., and Hoffman, A.R. (2003). CTCF Binding at the Insulin-Like Growth Factor-II (IGF2)/H19 Imprinting Control Region Is Insufficient to Regulate IGF2/H19 Expression in Human Tissues. Endocrinology *144*, 4420–4426.

Valadez-Graham, V. (2004). CTCF-dependent enhancer blockers at the upstream region of the chicken -globin gene domain. Nucleic Acids Research *32*, 1354–1362.

Van Bortle, K., Ramos, E., Takenaka, N., Yang, J., Wahi, J.E., and Corces, V.G. (2012). Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. Genome Research *22*, 2176–2187.

Wasylyk, B., Wasylyk, C., Augereau, P., and Chambon, P. (1983). The SV40 72 bp repeat preferentially potentiates transcription starting from proximal natural or substitute promoter elements. Cell *32*, 503–514.

Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell *171*, 1573.

Yusufzai, T.M., Tagami, H., Nakatani, Y., and Felsenfeld, G. (2004). CTCF Tethers an Insulator to Subnuclear Sites, Suggesting Shared Insulator Mechanisms across Species. Molecular Cell *13*, 291–298.

Cell culture
All the experiments on cells were performed in the HEK293T cells. These cells were cultured in serum (10% FBS) supplemented DMEM (AL007A). CGGBP1 depletion in these cells was achieved by lentiviral transduction of the lentiviral shmiR constructs (four different sites in the ORF) targeting CGGBP1 (KD) or Control-shmiR (CT) obtained from Origene as described earlier (PMID: 31547883) and not shown. The process of lentiviral production involved co-transfection of the third generation packaging plasmids (from Addgene): pRSV-Rev (12253), pMDLg/pRRE (12251) and pMD2.G (12259) and lentiviral constructs (from Origene) in equimolar ratios. For efficient transfection, Fugene (Promega) was used (3 µl/µg of DNA). Higher transduction yield was obtained with the help of Polybrene (Sigma), used at 1:10000 dilution of 10 mg/ml stock. Further, these cells were selected using 0.4 mg/ml of puromycin for stable transduction.
For the downstream experiments in HEK293T cells containing normal levels of CGGBP1 (CT) or depleted levels of CGGBP1 with the three different episomal constructs LoB5, LoB3 and GoB4 respectively along with the pRV-CMV plasmid (transfection control).

Cloning of CGGBP1-dependent CTCF-binding sites into pGL3-Control vector
LoB5, LoB3 and GoB4 regions along with approximately 250 base flanking sequences were cloned between *Kpn*I and *Xho*I restriction enzyme sites in pGL3-control vector (PMID: 31547883). The LoB and GoB peaks and immediate flanking 250 base regions were amplified from genomic DNA of HEK293T cells by using primers described in Table 1. Clones were confirmed by Sanger sequencing (not shown). The PCR amplified genomic regions were double digested with *Kpn*I and *Xho*I. Similarly, pGL3-Control vector was double digested with *Kpn*I and *Xho*I and cloned in pGL3-control. The primer sequences are mentioned below (5´3´)

Primer	Sequence 5´-3´
LoB5 Cloning Forward	AAAGGTACCACGAAGTTGAGGGTGACCAG
LoB5 Cloning Reverse	ATAACTCGAGTCAGACCAGGGGTTTGTCTC
GoB4 Cloning Forward	AAAGGTACCCCTAACCGGAAAACCACTCA
GoB4 Cloning Reverse	ATAACTCGAGTGCATTGCCAGTTTATCCAA
LoB3 Cloning Forward	ATAACTCGAGGGGAGCATCTTGGTCTGTGT
LoB3 Cloning Reverse	AAAGGTACCGAGACCTGAGGAGCAAGTGG

Chromatin immunoprecipitation (ChIP)
Non-targeting or CGGBP1-targeting shRNA expressing lentivirus-transduced HEK293T cells (as described above) were grown in DMEM supplemented with 10% FBS. pGL3-LoB3, pGL3-LoB5 or pGL3-GoB4 constructs were transfected [3 µg per 10 cm plates] into CT and KD cells. A pRL-CMV plasmid (E2231, Promega) [2 µg per 10 cm plates] was co-transfected with these constructs as a transfection control. HEK293T cells were harvested after 48 h of transfection and subjected to RNA-isolation and ChIP-qPCR.
ChIP for the respective antibodies was performed as previously described (PMID: 31547883). Cells were harvested after crosslinking for 10 min at 37° C with 1% formaldehyde and quenching with Glycine (125 mM). Cells were washed with PBS twice and lysed in an SDS lysis buffer containing a cocktail of 1x protease phosphatase inhibitors

(PI78441, Invitrogen). Based on our previous experience, the chromatin thus obtained was sonicated for 21 cycles of 30 seconds on/30 seconds off to obtain DNA fragments in the range of 0.5 kb - 1 kb length. The clear fraction of sonicated chromatin was obtained after centrifugation (16000 rcf, 5 minutes, 4° C). 30 µl of fragmented chromatin was kept aside as input and remaining (150 µl) was further diluted in ChIP-dilution buffer (0.01% SDS, 1.1% Triton X- 100, 1.2 mM EDTA, 16.7 mM Tris-HCl, pH 8.1, 167 mM NaCl) containing 1x protease phosphatase inhibitors cocktail. subjected for immunoprecipitation. The chromatin was pre-cleared by incubating with protein G sepharose beads for 4 hrs at 4° C followed by overnight incubation with targeted primary antibody with mild tumbling at 4° C. Subsequently, the protein G sepharose beads were added and incubated for 1 hour followed by gentle centrifugation. Beads pellet was washed with different buffers in the following order: low-salt IP wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl and 150 mM NaCl), high-salt IP wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl and 500 mM NaCl), LiCl IP wash buffer (0.25 M LiCl, 1% IGEPAL, 1% Na deoxycholate, 1 mM EDTA and 10 mM Tris–HCl) and two washes of TE buffer (10 mM Tris–HCl and 1 mM EDTA). The immunoprecipitated DNA was eluted using an elution buffer (1% SDS and 0.1 M NaHCO3) and was reverse-crosslinked (addition of 20 µl of 5 M NaCl and incubation at 65° C for 6 hrs). Further, the DNA was subjected to proteinase-k digestion for 1 hour (addition of 10 µl of 0.5 M EDTA pH 8.0 and and 20 µl of 1 M Tris-HCl pH 6.8 followed by 2 µl of 10 mg/ml Proteinase K (P2308, Sigma). The DNA was purified following the column-based purification protocol (A1460, Promega). The set of different primary antibodies used in the ChIP assays are mentioned below:

ChIP-target	No of confluent 10 cm plates	Antibody used
CTCF	2	5 µg (2.5 µg of SC-28198 and 2.5 µg of SC-271514)
RNA Polymerase II	2	3 µg (05-623 Merck)
H3K9me3	1	3 µg of SC-130356
H3K27me3	1	2 µg of ABCAM ab6002

RNA extraction, cDNA synthesis and qRT-PCRs

RNA was isolated by using the TRIzol reagent as per manufacturer's instructions. RNA was isolated from pGL3-control construct transfected cells (as described above). Cells were washed with DEPC treated ice-chilled PBS twice and lysed by using TRIzol. Cells were scraped and cell lysate were collected in 1.5 ml eppendorf. Chloroform (250 µl) was added to lysed cells and vigorously mixed and centrifuged at 10,000 rpm for 5 min. The aqueous layer was carefully transferred to the new tube and 550 µl isopropanol was added to aqueous phase. The immunoprecipitated RNA was washed with 1 ml 75% ethanol in DEPC treated H_2O and RNA was dissolved in 40 µl of DEPC treated water. The isolated RNA was digested with DNaseI (M0303S, 4 Units) at room temperature for 15 minutes. DNaseI digested RNA was isolated by using the TRIzol method.

Luciferase assays

Lentiviral transduced HEK293T cells with normal and depleted levels of CGGBP1 were seeded in 96 well plates. LoB (pGL3-LoB3 and pGL3-LoB5), GoB (pGL3-GoB4) or pGL3-control empty vectors were transfected [100 ng per well] in CT and KD cells. An pRL-CMV plasmid (E2231, Promega) [25 ng per well] was co-transfected as transfection control. HEK293T cells were harvested after 72 h of transfection. The dual luciferase assay was performed as per manufacturer protocol (E2920, Promega). Cells were washed with PBS and 100 µl of 1x Passive lysis buffer. 100 µl of Luciferase Assay Substrate resuspended in Luciferase Assay Buffer II was added to each well. Firefly luciferase activity was measured at 550 to 570 nm wavelength and followed by 100 µl of 1X Stop & Glo Reagent (part of the kit E1910, Promega) was added to each well. Renilla Luciferase activity was measured at 470 to 490 nm wavelength.

cDNA synthesis

For cDNA synthesis using random primer the following method was used:
cDNA synthesis was performed by using SuperScript VILO cDNA Synthesis Kit (invitrogen 11754050) as per manufacturer's protocol. RNA (2 µg) was incubated with 1X VILO Reaction Mix and 1X SuperScript Enzyme Mix. The reaction mix was incubated at room temperature for 15 minutes and followed by incubation at 42^0 C for 60 minutes.
For cDNA synthesis using oligo-dT or primers P1 or P6 the following method was used:
DNaseI digested RNA from LoB and GoB construct transfected cells CT and KD cells was used for cDNA synthesis. RNA ((2 µg) was incubated with primer at 65^0 C for 5 minutes and snap-chilled immediately. M-MLV Reverse-transcriptase (Promega M1701, 200 Unit) in 1x reverse transcriptase buffer was added to primer annealed RNA and incubated at 40^0 C for 60 minutes.

Strand-specific PCR

cDNA synthesis by using oligo-dT primer or strand-specific primer P1 (f1Ori Forward) and P6 (SV40 Promoter Reverse) were used as template for PCR using primer P1 (f1Ori Forward) and P6 (SV40 Promoter Reverse). The specific amplifications were compared by Agarose gel electrophoresis.

Sanger sequencing of PCR products

The PCR products obtained from the strand specific PCR were subjected for the Sanger sequencing using sequencing primers P4 and P5 (Fig 4A).

Statistical analysis

All the statistical analyses were performed on Graphpad Prism 8 and open office.

Genome browser views

The genome-browser views of the three different CGGBP1-dependent CTCF-binding sites were obtained using Integrated Genomic Viewer (IGV). Repeat sequences were identified in 10kb flanks of the CGGBP1-dependent CTCF-binding sites by using RepeatMasker. The genomic coordinates of the identified repeat elements were used to generate bigwig signal files by using deepTools bamCoverage tool.

ChIP-qPCR

All quantitative PCR reactions were performed at 57℃ annealing temperature and the specific template amplification was confirmed by agarose gel electrophoresis. Following are the PCR conditions: 95° C-5 minutes, (95° C- 20 seconds, 55° C- 20 seconds, 72° C-30

seconds, 80° C-30 seconds (signal capturing)) x45 followed by melting curve analysis (55 to 95 °C constant signal recording).
Following is the list of primers used for ChIP-qPCRs or RT-PCRs:

PBX1 Upstream Forward	CTTGATTGTCCCAGCCTCAT
PBX1 Upstream Reverse	CGTCAGTTTAGGCCCCATTA
PBX1 Downstream Forward	TGCCTTCCCCCTTTTATCTT
PBX1 Downstream Reverse	AGGGCAATGGAATAGGAACC
NRXN2 Upstream Forward	GATGAAGCGGAGGTAAACCA
NRXN2 Upstream Reverse	TGCAGAGAGACCCCTAGCTC
NRXN2 Downstream Forward	GGCTACGCCCATCTATGAAG
NRXN2 Downstream Reverse	CGAGATTCTGGACCTGGAGA
SPTBN Upstream Forward	GCGCCTTCTCTTTGTTTTTG
SPTBN Upstream Reverse	TGGCTTTCGATTGGAAAAAT
SPTBN1 Downstream Forward	CCGTTACATGAGGCTCAACA
SPTBN1 Downstream Reverse	GCAGGCTATACCACGAGGAG
OPRL Upstream Forward	CGGTAGTAGATGGGGGTGTG
OPRL Upstream Reverse	ACAGACAGGCGGCTCTTAAC
OPRL Downstream Forward	CGGCTATGAGGAGACCTTTG
OPRL Downstream Reverse	AGAGGTGAGGGGTCCTGAGT
COMT1 Upstream Forward	AACACTGGGATAGGGTGTGG
COMT1 Upstream Reverse	AAGGTCCGAATTCCCCAGTT
COMT1 Downstream Forward	CAGCAGGGCTCCAGTAAGAC
COMT1 Downstream Reverse	TTGATCTCCTGACCTCGTGA
P1 (f1Ori Forward)	ACATTTCCCCGAAAAGTGC
P2 (f1Ori Reverse)	AGCCCCCGATTTAGAGCTT
P3 (CTCF LoB5 Forward)	CGTAGTTGGGCAGGTTGTCT
P4 (CTCF LoB5 Reverse)	CAGCTAGGGGGCTACTTCCT
P5 (SV40 Promoter Forward)	GCCCCATCGCTGACTAATTT

P6 (SV40 Promoter Reverse)	AAAGCCTAGGCCTCCAAAAA
P7 (FF Luciferase Forward)	GTGTTGGGCGCGTTATTTAT
P8 (FF Luciferase Reverse)	ACCCCTTTTTGGAAACGAAC
P9 (SV40 Enhancer Forward)	GACTTTCCACACCTGGTTGC
P10 (SV40 Enhancer Reverse)	AAGGAGCTGACTGGGTTGAA
P11 (ori Forward)	AGTTCGGTGTAGGTCGTTCG
P12 (ori Reverse)	GCCTACATACCTCGCTCTGC
P13 (AmpR Forward)	GCGCCACATAGCAGAACTTT
P14 (AmpR Reverse)	TTTTGCCTTCCTGTTTTTGC
LoB5 Upstream Forward	GGAATATCCAGGCACATGCT
LoB5 Upstream Reverse	TGGTTTTTCAGGCTGGTTTT
LoB5 Downstream Forward	CGAAGTGTCAAGGGAACAGC
LoB5 Downstream Reverse	GCAGAAGCCTCAGTGGTCTC
GoB4 Upstream Forward	GCCCTGCTTCATACTTTGGA
GoB4 Upstream Reverse	CACTGGCCCCAGTACTTGTT
GoB4 Downstream Forward	AAGGCTCCTGGGATGAGAAT
GoB4 Downstream Reverse	CCCCCAGACAGAAATGGTTA
LoB3 Upstream Forward	CTCACCGTCCTCCTTCTCTG
LoB3 Upstream Reverse	CACGGAGGTGACTCCAAGTT
LoB3 Downstream Forward	TCCATCCACTGCACAGACTC
LoB3 Downstream Reverse	AGCCTAGGCAGCCACAAATA
GoB4 CTCF Forward	CAGTGACCCCAGAGACCATT
GoB4 CTCF Reverse	ATCGCTCCTCCATTCAAAAA
LoB3 CTCF Forward	CCTGCTGTCCTTCAGAGCAT
LoB3 CTCF Reverse	GCTCCAGAGGAGGAAAAAGG

Table 1

Genes with TSSs separated by CGGBP1-dependent CTCF-binding site	Type of change in CTCF-binding upon CGGBP1 depletion	Upstream distance between the TSS and CTCF-binding site (kb)	Downstream distance between the TSS and CTCF-binding site (kb)
NRXN2	Loss	21.5	13.6
PBX1	Loss	13.5	8.1
OPRL	Loss	1.6	2.8
SPTBN1	Gain	12.7	87.7
COMT1	Gain	5.8	2.9

Table 1: Table represents the distance between CGGBP1 regulated chromatin barrier element and permissive TSSs used for PCR assay.

Figure 1

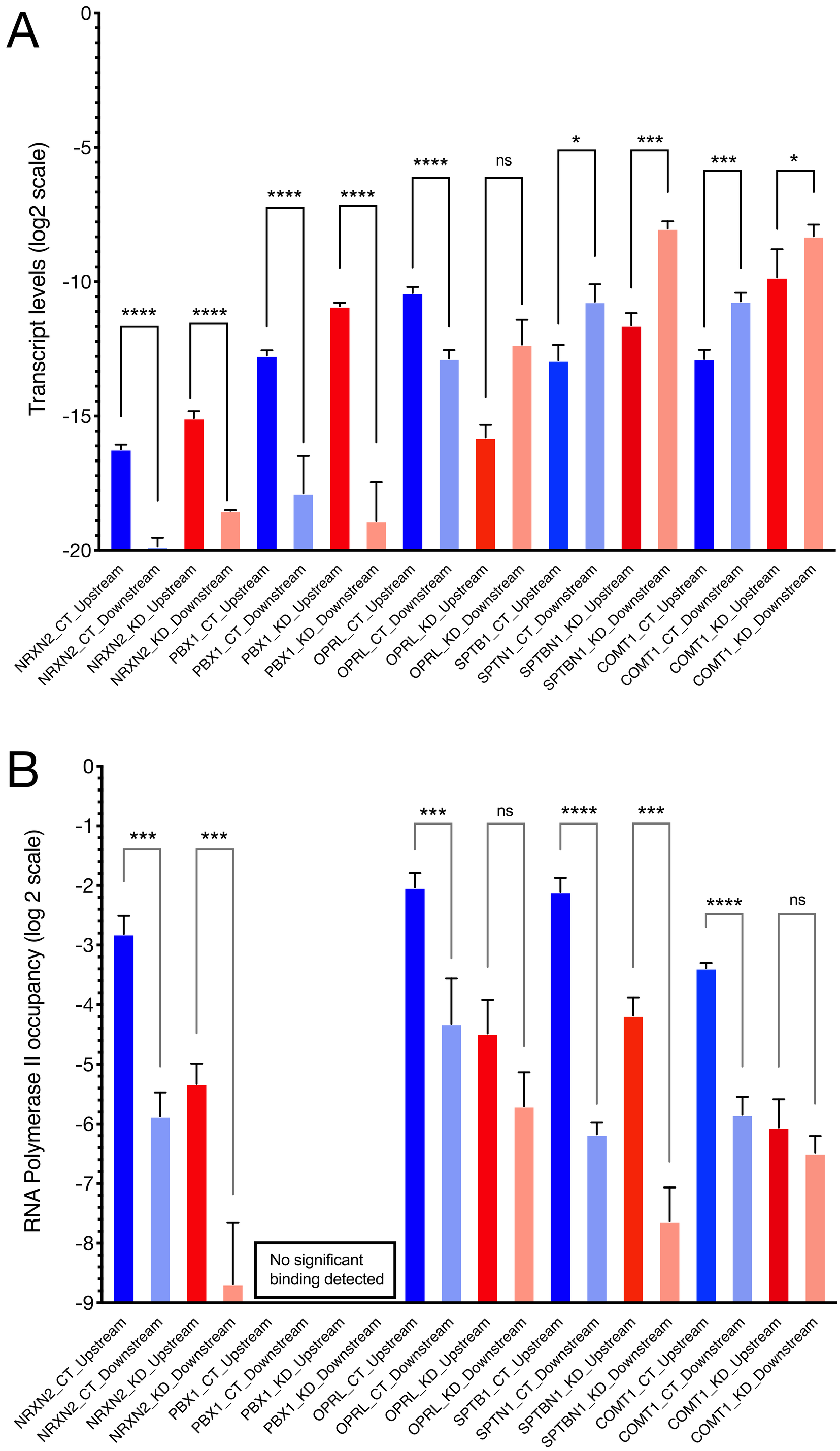
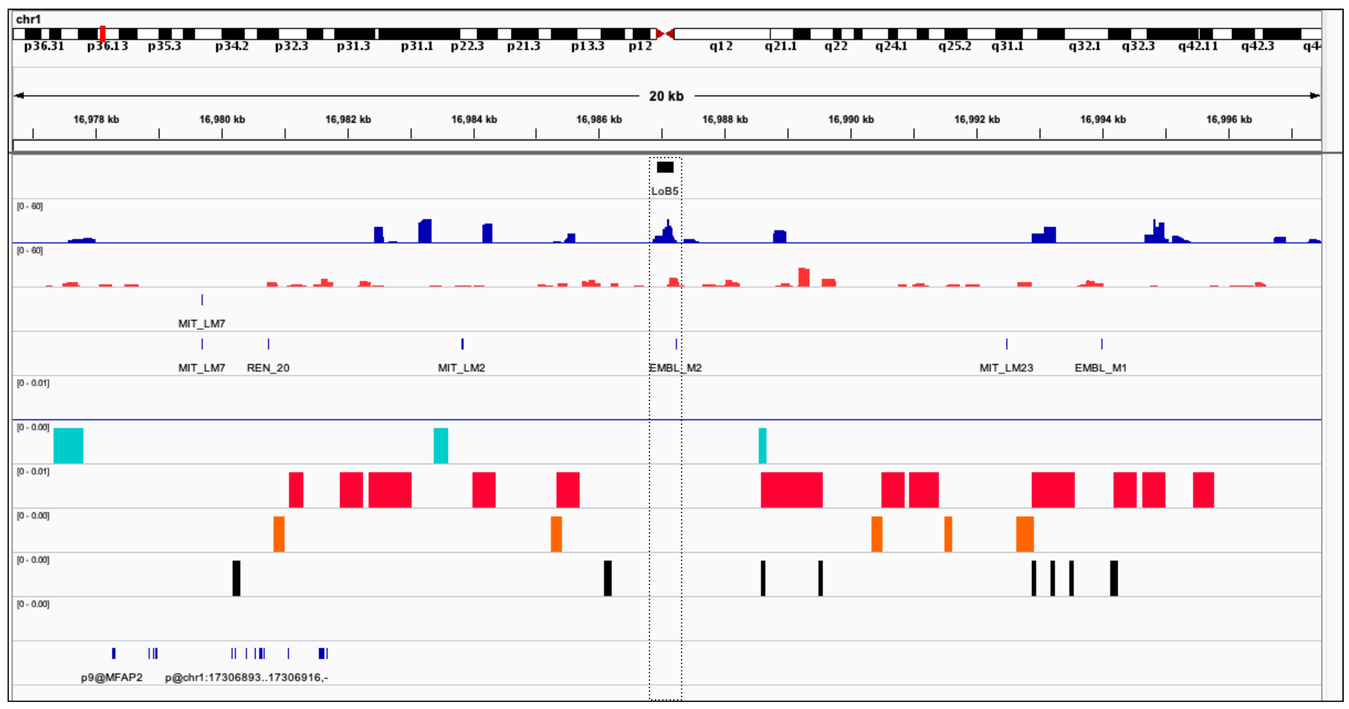
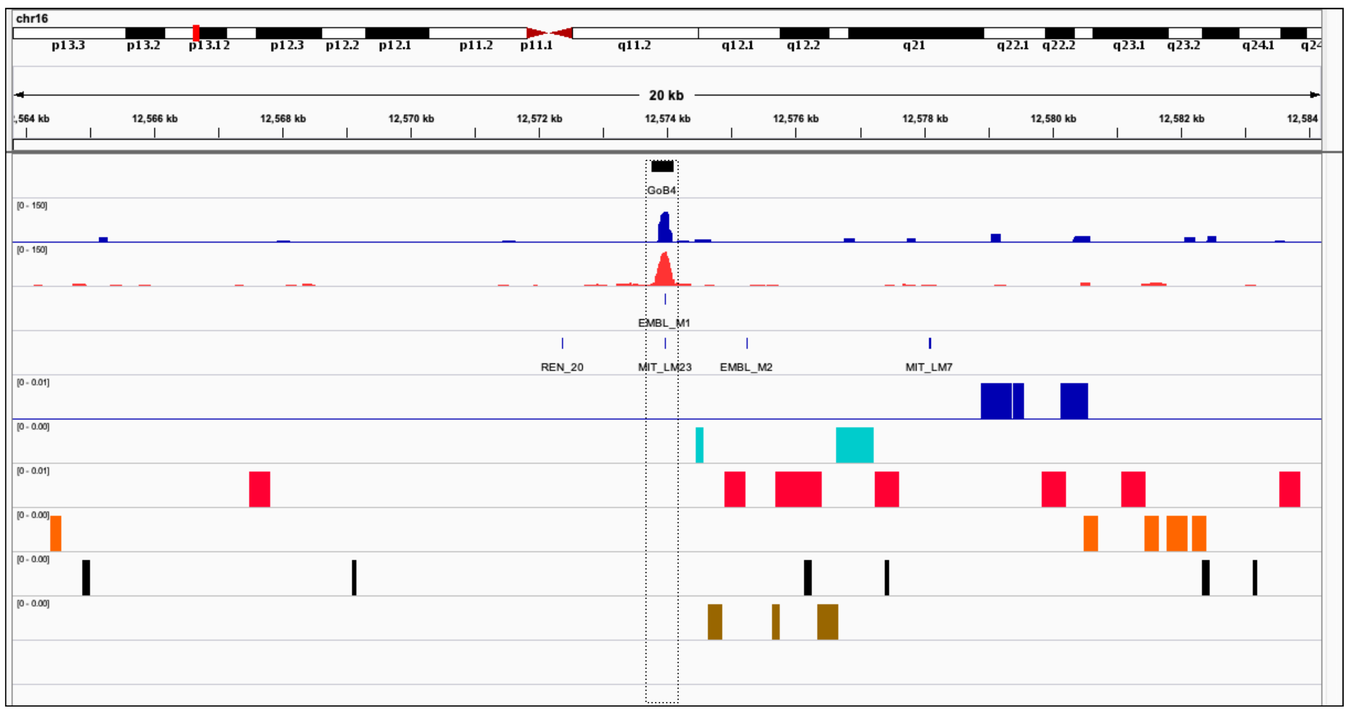

A
Transcript levels (log2 scale)
0
-5
-10
-15
-20
****
****
****
****
****
ns
*
***
***
*
NRXN2_CT_Upstream
NRXN2_CT_Downstream
NRXN2_KD_Upstream
NRXN2_KD_Downstream
PBX1_CT_Upstream
PBX1_CT_Downstream
PBX1_KD_Upstream
PBX1_KD_Downstream
OPRL_CT_Upstream
OPRL_CT_Downstream
OPRL_KD_Upstream
OPRL_KD_Downstream
SPTB1_CT_Upstream
SPTN1_CT_Downstream
SPTBN1_KD_Upstream
SPTBN1_KD_Downstream
COMT1_CT_Upstream
COMT1_CT_Downstream
COMT1_KD_Upstream
COMT1_KD_Downstream
B
RNA Polymerase II occupancy (log 2 scale)
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
***
***
***
ns
****
***
****
ns
No significant binding detected
NRXN2_CT_Upstream
NRXN2_CT_Downstream
NRXN2_KD_Upstream
NRXN2_KD_Downstream
PBX1_CT_Upstream
PBX1_CT_Downstream
PBX1_KD_Upstream
PBX1_KD_Downstream
OPRL_CT_Upstream
OPRL_CT_Downstream
OPRL_KD_Upstream
OPRL_KD_Downstream
SPTB1_CT_Upstream
SPTN1_CT_Downstream
SPTBN1_KD_Upstream
SPTBN1_KD_Downstream
COMT1_CT_Upstream
COMT1_CT_Downstream
COMT1_KD_Upstream
COMT1_KD_Downstream

Figure 2

A


Peak representing LoB5
CTCF ChIP-seq reads (CT)
CTCF ChIP-seq reads (KD)
CTCF binding motifs (JASPAR)
CTCF binding motifs (CTCFDB)
LINE-1 repeats (UCSC)
LINE-2 repeats (UCSC)
Alu-SINEs (UCSC)
MIR SINEs (UCSC)
Simple repeats (UCSC)
DNA transposons (UCSC)
Permissive TSSs (Fantom)
chr1
20 kb
LoB5
MIT_LM7
REN_20
MIT_LM2
EMBL_M2
MIT_LM23
EMBL_M1
p9@MFAP2
p@chr1:17306893..17306916,-

B


Peak representing GoB4
CTCF ChIP-seq reads (CT)
CTCF ChIP-seq reads (KD)
CTCF binding motifs (JASPAR)
CTCF binding motifs (CTCFDB)
LINE-1 repeats (UCSC)
LINE-2 repeats (UCSC)
Alu-SINEs (UCSC)
MIR SINEs (UCSC)
Simple repeats (UCSC)
DNA transposons (UCSC)
Permissive TSSs (Fantom)
chr16
20 kb
GoB4
EMBL_M1
REN_20
MIT_LM23
EMBL_M2
MIT_LM7

C

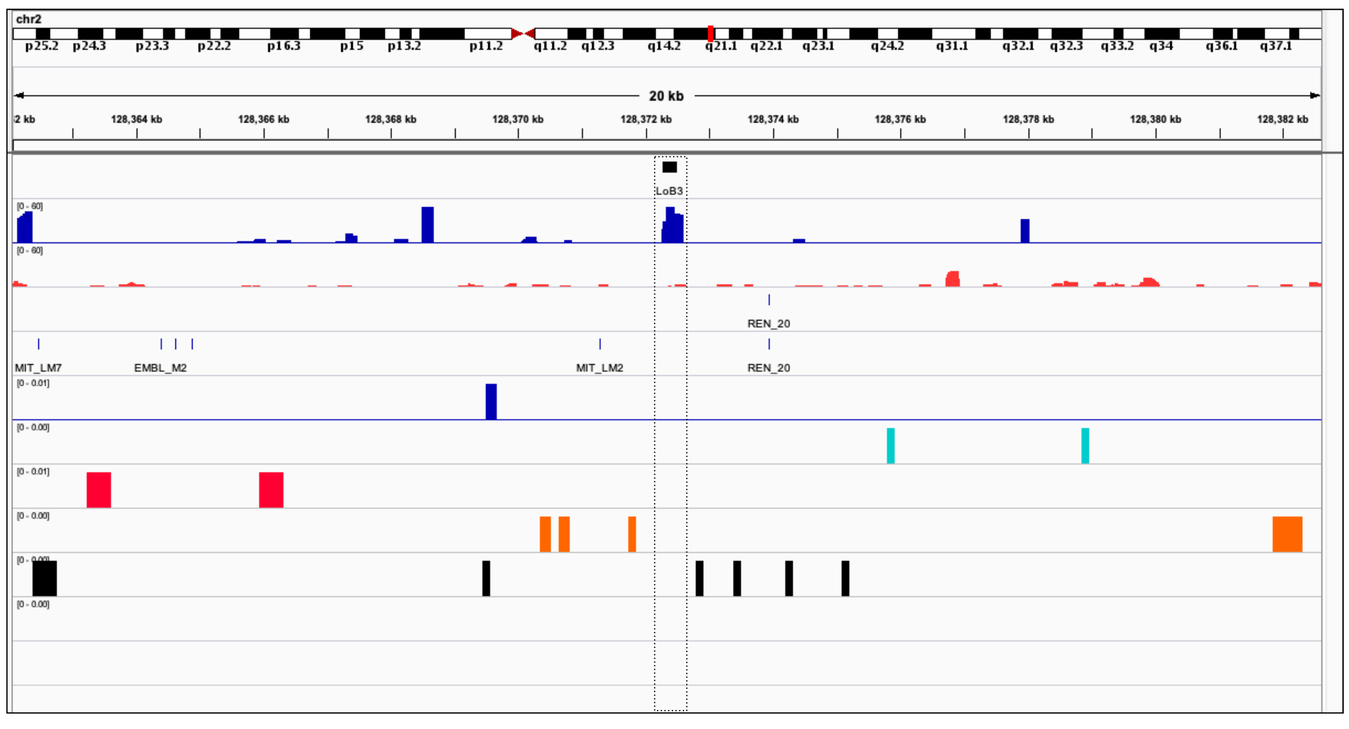

Peak representing LoB3
CTCF ChIP-seq reads (CT)
CTCF ChIP-seq reads (KD)
CTCF binding motifs (JASPAR)
CTCF binding motifs (CTCFDB)
LINE-1 repeats (UCSC)
LINE-2 repeats (UCSC)
Alu-SINEs (UCSC)
MIR SINEs (UCSC)
Simple repeats (UCSC)
DNA transposons (UCSC)
Permissive TSSs (Fantom)
chr2
20 kb
LoB3
REN_20
MIT_LM7
EMBL_M2
MIT_LM2
REN_20

Figure 3

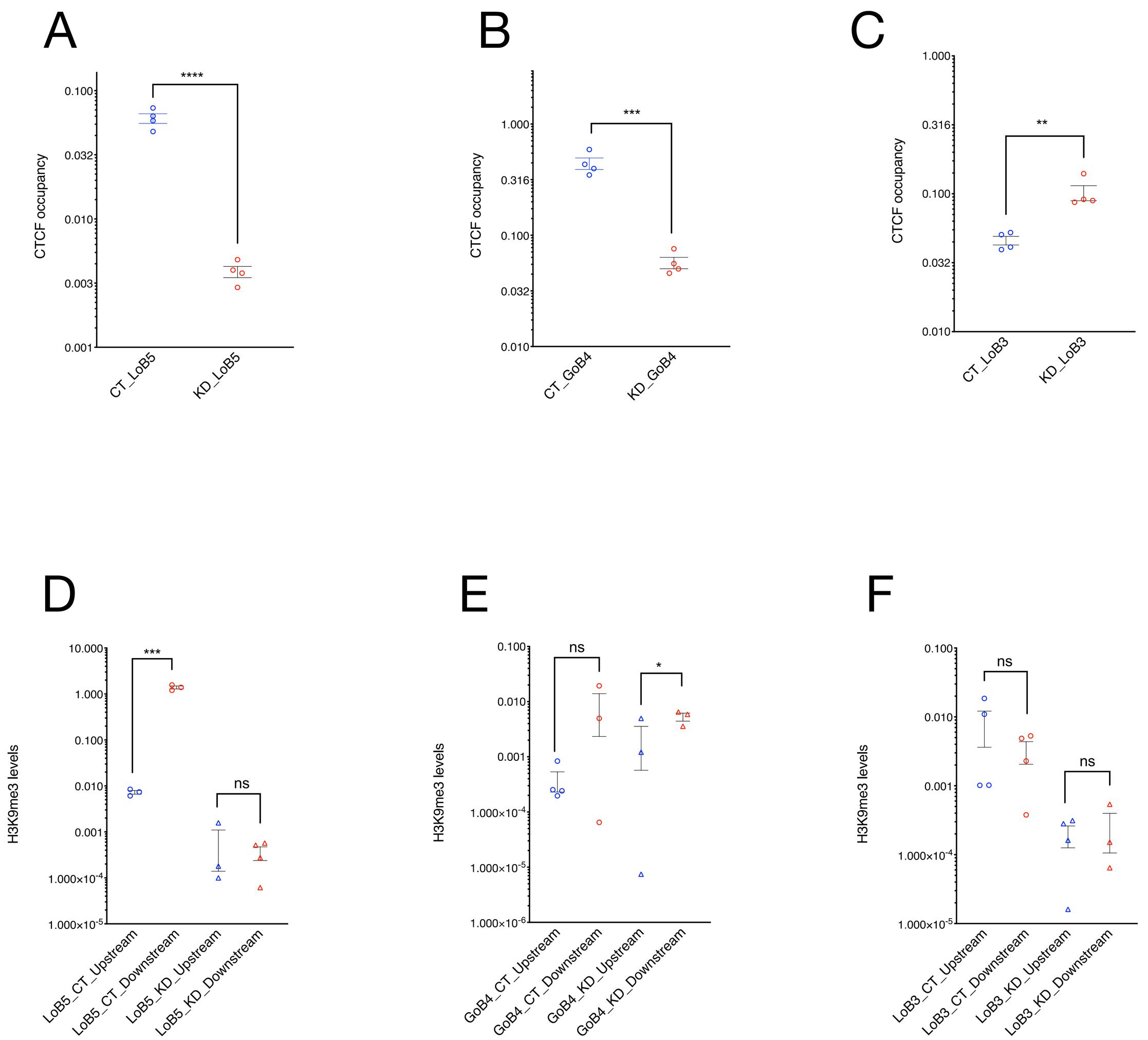
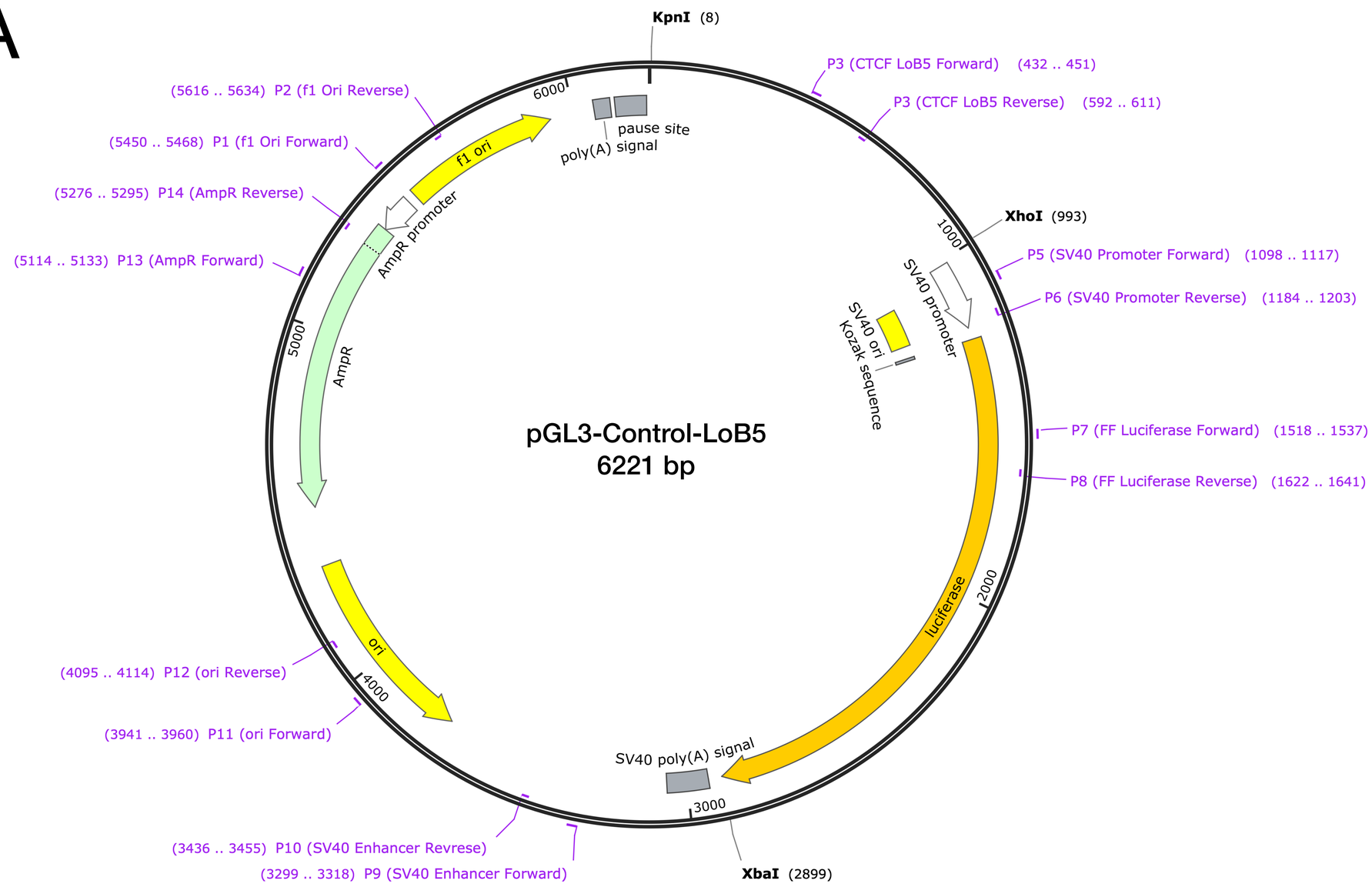
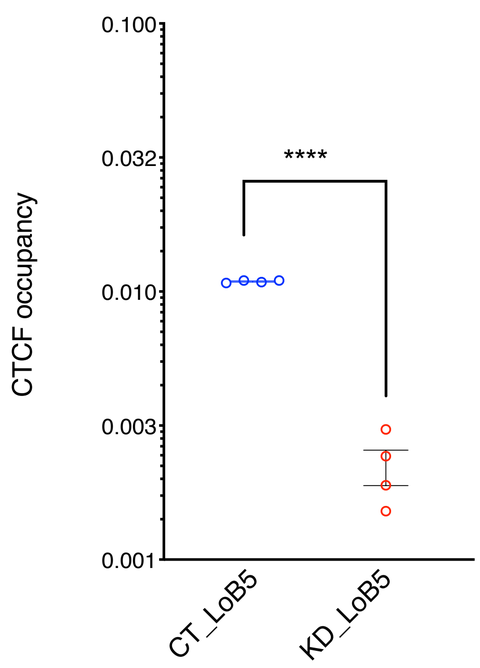

A
CTCF occupancy
0.100
0.032
0.010
0.003
0.001
****
CT_LoB5
KD_LoB5
B
CTCF occupancy
1.000
0.316
0.100
0.032
0.010
***
CT_GoB4
KD_GoB4
C
CTCF occupancy
1.000
0.316
0.100
0.032
0.010
**
CT_LoB3
KD_LoB3
D
H3K9me3 levels
10.000
1.000
0.100
0.010
0.001
1.000×10-4
1.000×10-5
***
ns
LoB5_CT_Upstream
LoB5_CT_Downstream
LoB5_KD_Upstream
LoB5_KD_Downstream
E
H3K9me3 levels
0.100
0.010
0.001
1.000×10-4
1.000×10-5
1.000×10-6
ns
*
GoB4_CT_Upstream
GoB4_CT_Downstream
GoB4_KD_Upstream
GoB4_KD_Downstream
F
H3K9me3 levels
0.100
0.010
0.001
1.000×10-4
1.000×10-5
ns
ns
LoB3_CT_Upstream
LoB3_CT_Downstream
LoB3_KD_Upstream
LoB3_KD_Downstream

Figure 4

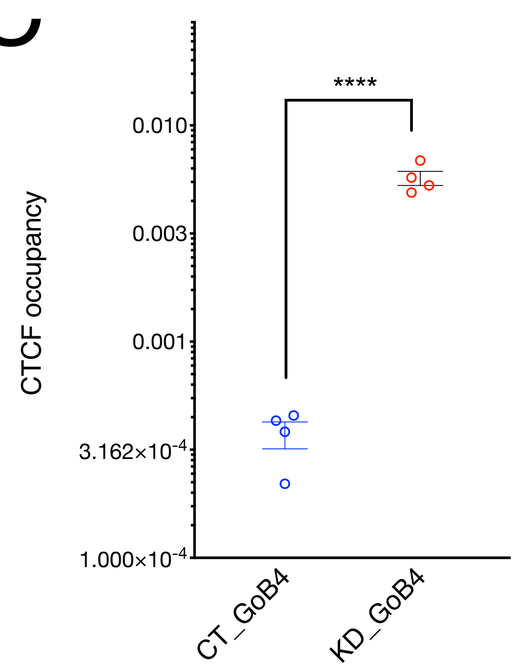
A



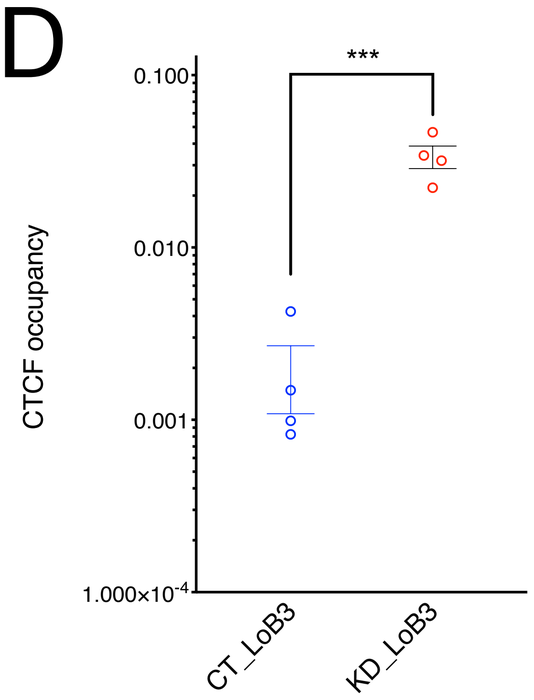
B



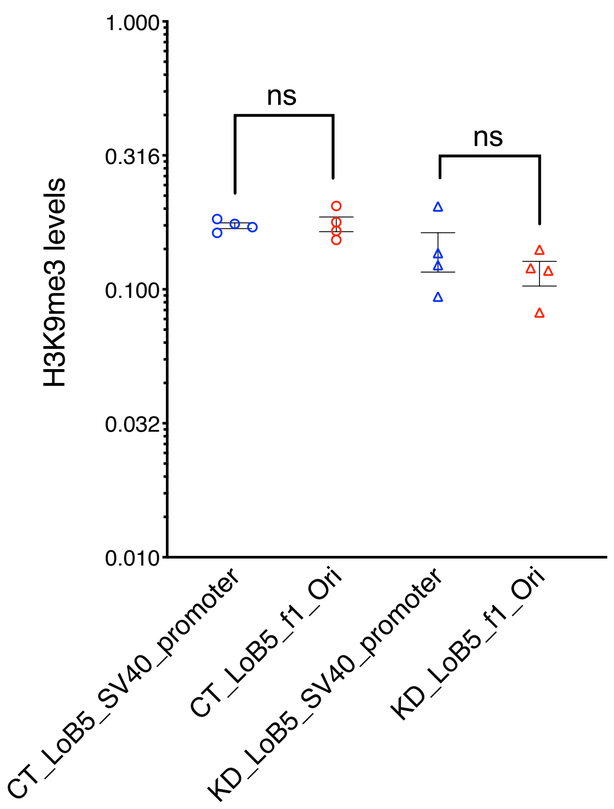
C



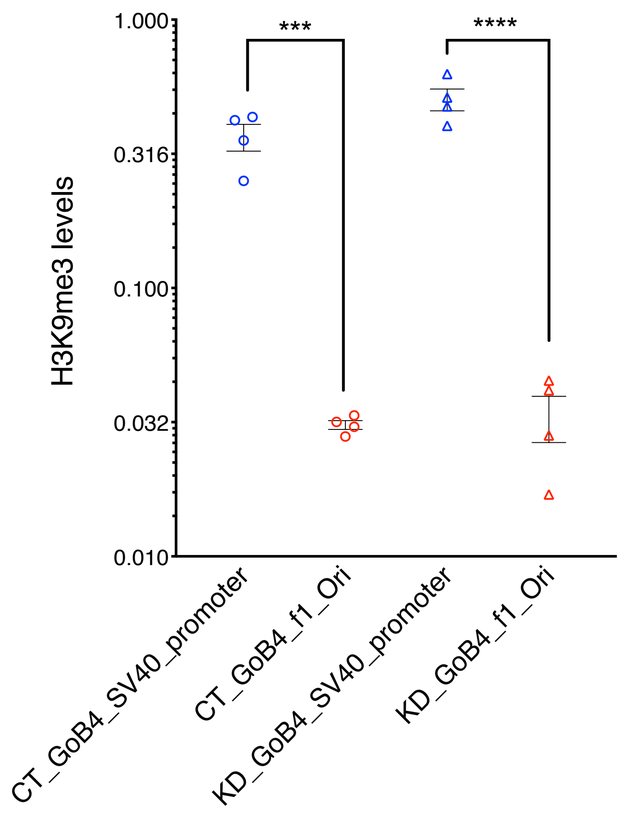
D



E



F



G

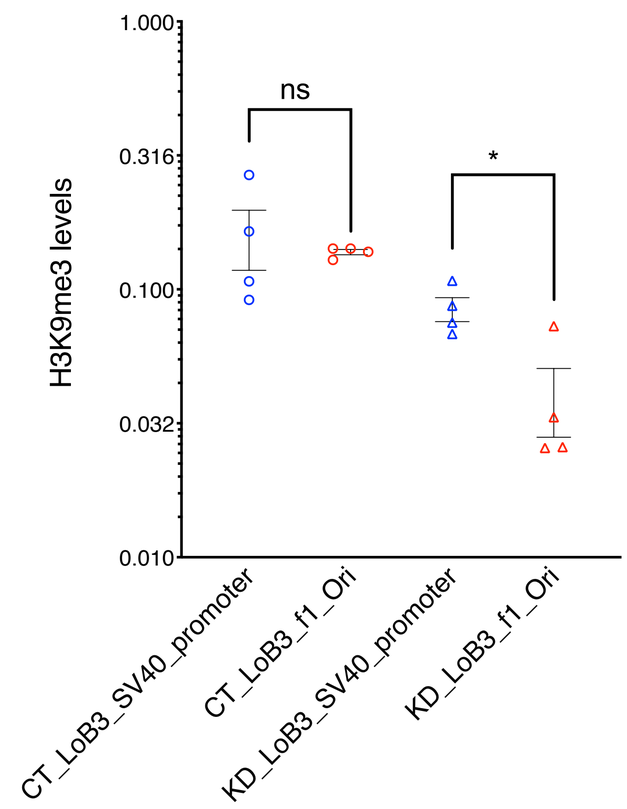


Figure 5

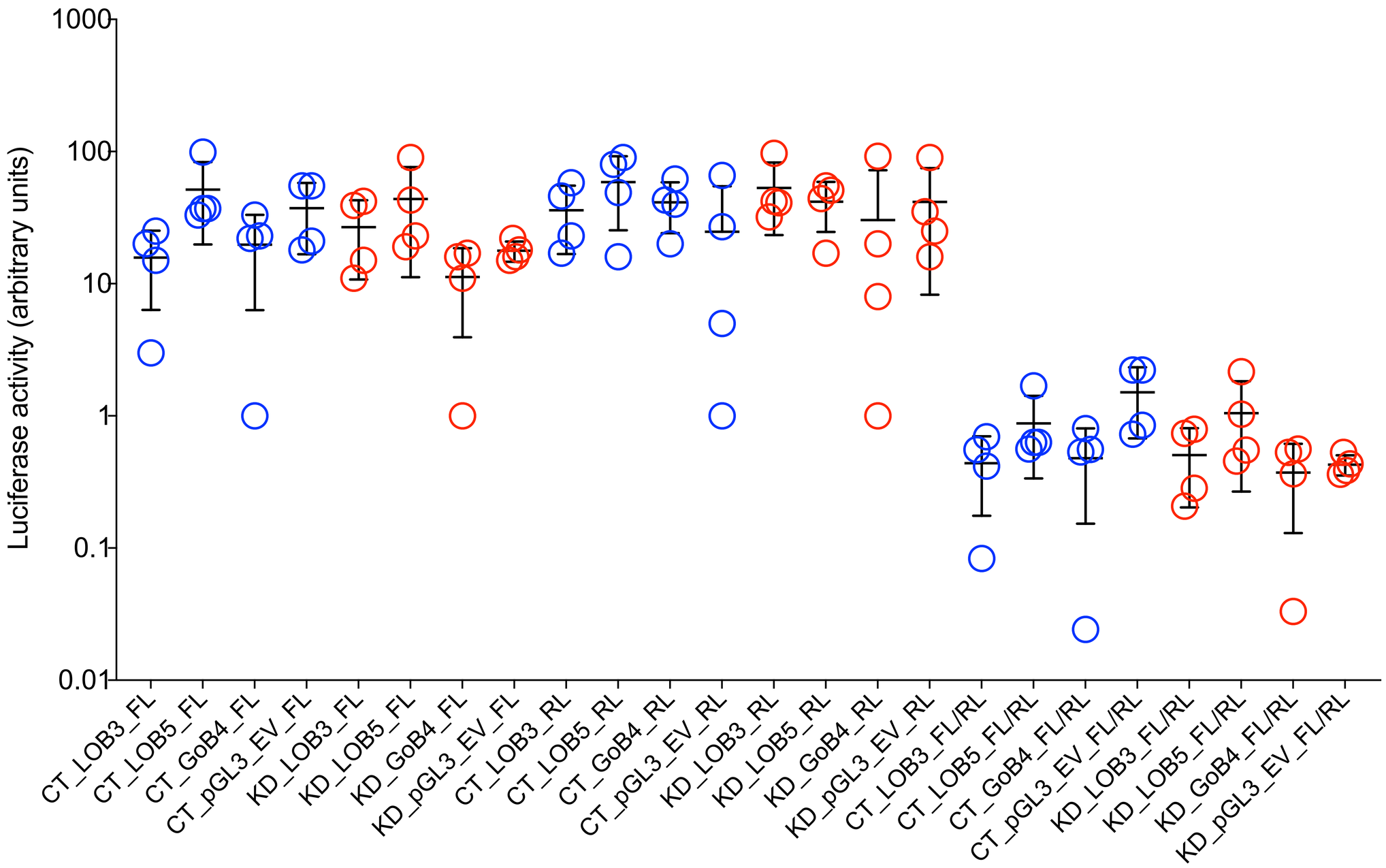
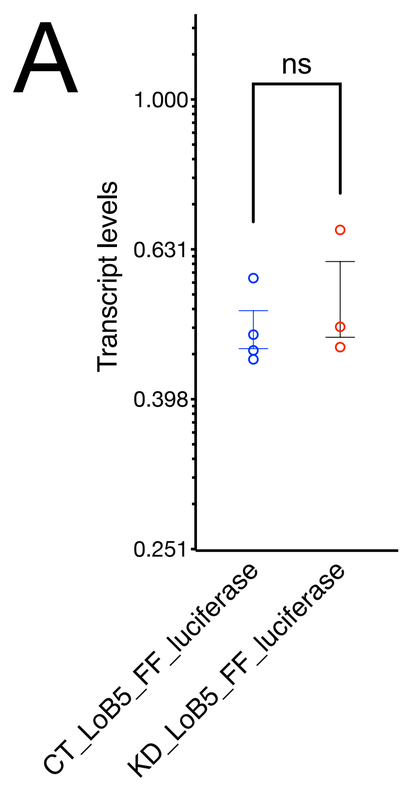
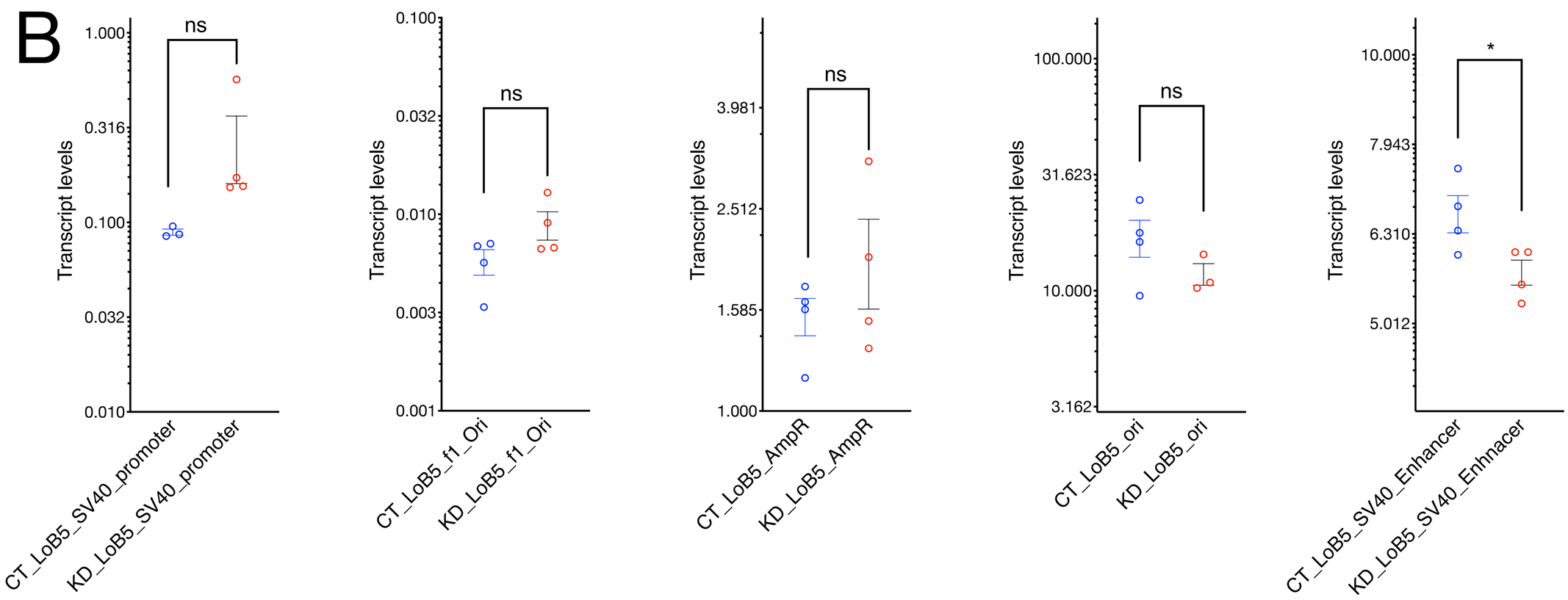

Luciferase activity (arbitrary units)
1000
100
10
1
0.1
0.01
CT_LOB3_FL
CT_LOB5_FL
CT_GoB4_FL
CT_pGL3_EV_FL
KD_LOB3_FL
KD_LOB5_FL
KD_GoB4_FL
KD_pGL3_EV_FL
CT_LOB3_RL
CT_LOB5_RL
CT_GoB4_RL
CT_pGL3_EV_RL
KD_LOB3_RL
KD_LOB5_RL
KD_GoB4_RL
KD_pGL3_EV_RL
CT_LOB3_FL/RL
CT_LOB5_FL/RL
CT_GoB4_FL/RL
CT_pGL3_EV_FL/RL
KD_LOB3_FL/RL
KD_LOB5_FL/RL
KD_GoB4_FL/RL
KD_pGL3_EV_FL/RL

Figure 6


A
ns
1.000
0.631
0.398
0.251
Transcript levels
CT_LoB5_FF_luciferase
KD_LoB5_FF_luciferase


B
ns
1.000
0.316
0.100
0.032
0.010
Transcript levels
CT_LoB5_SV40_promoter
KD_LoB5_SV40_promoter
ns
0.100
0.032
0.010
0.003
0.001
Transcript levels
CT_LoB5_f1_Ori
KD_LoB5_f1_Ori
ns
3.981
2.512
1.585
1.000
Transcript levels
CT_LoB5_AmpR
KD_LoB5_AmpR
ns
100.000
31.623
10.000
3.162
Transcript levels
CT_LoB5_ori
KD_LoB5_ori
*
10.000
7.943
6.310
5.012
Transcript levels
CT_LoB5_SV40_Enhancer
KD_LoB5_SV40_Enhnacer

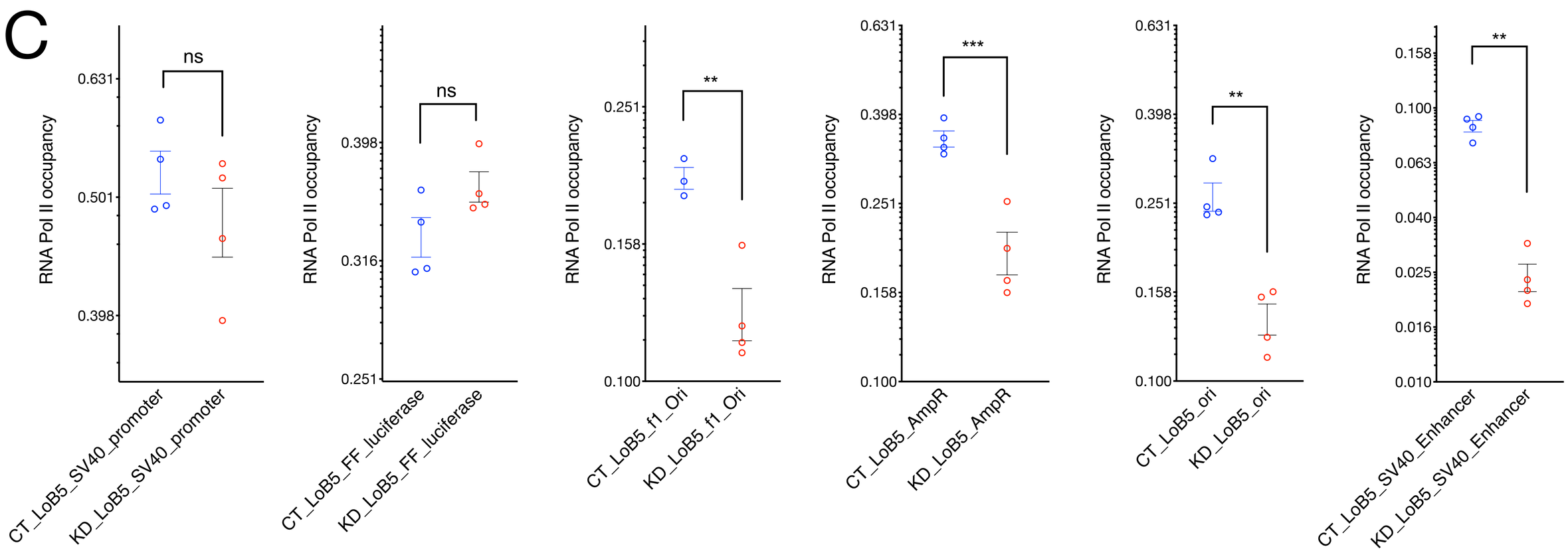
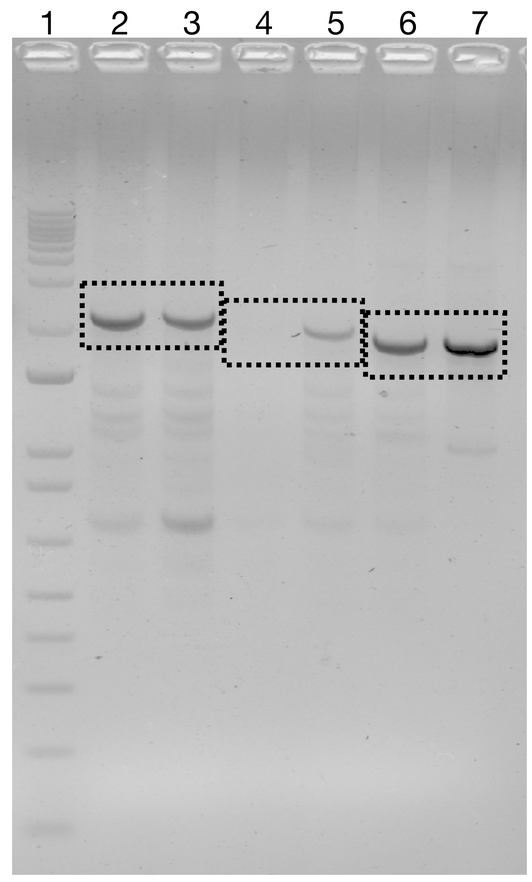

C
ns
0.631
0.501
0.398
RNA Pol II occupancy
CT_LoB5_SV40_promoter
KD_LoB5_SV40_promoter
ns
0.398
0.316
0.251
RNA Pol II occupancy
CT_LoB5_FF_luciferase
KD_LoB5_FF_luciferase
**
0.251
0.158
0.100
RNA Pol II occupancy
CT_LoB5_f1_Ori
KD_LoB5_f1_Ori
***
0.631
0.398
0.251
0.158
0.100
RNA Pol II occupancy
CT_LoB5_AmpR
KD_LoB5_AmpR
**
0.631
0.398
0.251
0.158
0.100
RNA Pol II occupancy
CT_LoB5_ori
KD_LoB5_ori
**
0.158
0.100
0.063
0.040
0.025
0.016
0.010
RNA Pol II occupancy
CT_LoB5_SV40_Enhancer
KD_LoB5_SV40_Enhancer

Figure 7

A

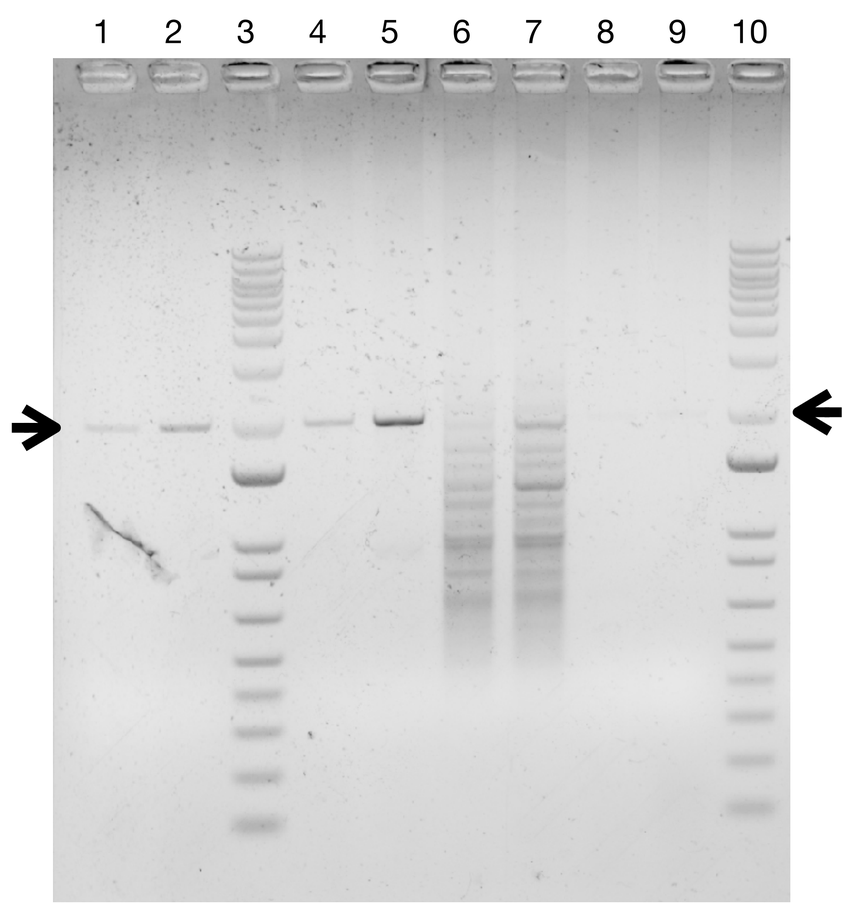


LANE DESCRIPTION

- 1: DNA ladder.
- 2 and 3: cDNA using random primers and PCR using P1 and P6; CT (2) and KD (3); plasmid pGL-Control-LoB3
- 4 and 5: cDNA using random primers and PCR using P1 and P6; CT (4) and KD (5); plasmid pGL-Control-LoB5
- 6 and 7: cDNA using random primers and PCR using P1 and P6; CT (6) and KD (7); plasmid pGL-Control-GoB4

Expected molecular weight of PCR products (2.15 kb for LoB3 lanes 2 and 3, 1.97 kb for LoB5 lanes 4 and 5, and 1.86 kb for GoB4, lanes 6 and 7) are indicated by boxes.

B

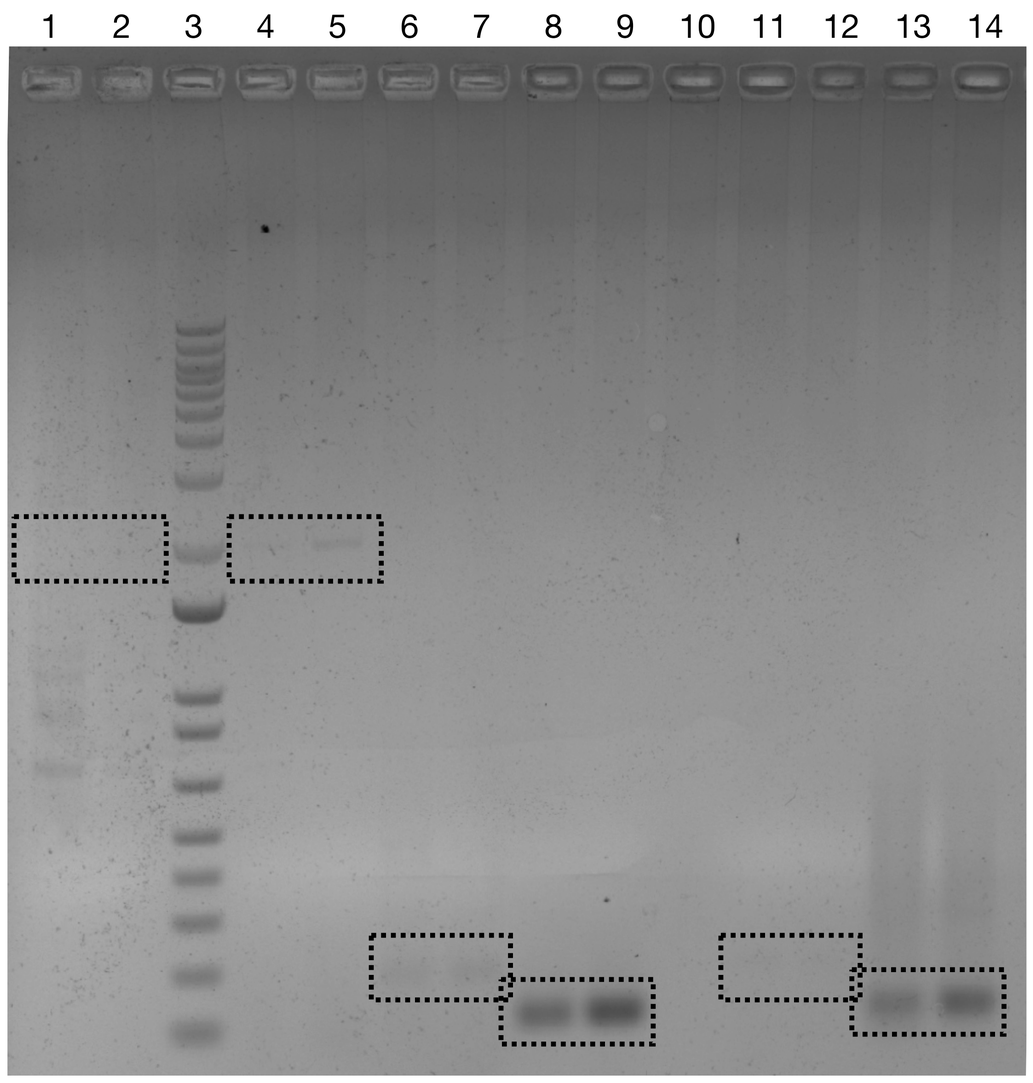


LANE DESCRIPTION

- 1 and 2: cDNA using oligo-dT, PCR using P1 and P6; CT (1) and KD (2).
- 3: DNA ladder.
- 4 and 5: cDNA using P1 and PCR using P1 and P6; CT (4) and KD (5).
- 6 and 7: cDNA using P6 and PCR using P1 and P6; CT (6) and KD (7).
- 8 and 9: RNA (pGL3-Control-LoB5 transfected), PCR using P1 and P6; CT (8) and KD (9).
- 10: DNA ladder.

Expected molecular weight of PCR products (1.97 kb) are indicated by an arrow.

C



LANE DESCRIPTION

- 1 and 2: RNA (no pGL3-Control-LoB5), PCR using P1 and P6; CT (1) and KD (2).
- 3: DNA ladder.
- 4 and 5: RNA (pGL3-Control-LoB5 transfected), PCR using P1 and P6; CT (4) and KD (5).
- 6 and 7: cDNA using P1 and PCR using P1 and P2; CT (6) and KD (7).
- 8 and 9: cDNA using P1 and PCR using P5 and P6; CT (8) and KD (9).
- 10: Empty lane.
- 11 and 12: cDNA using P6 and PCR using P1 and P2; CT (11) and KD (12).
- 13 and 14: cDNA using P6 and PCR using P5 and P6; CT (13) and KD (14).

Expected molecular weight of PCR products are indicated by boxes.