Sequence-level instructions direct transcription at polyT short tandem repeats.

Chloé Bessière[1,2,10], Manu Saraswat[1,2,10], Mathys Grapotte[1,2], Christophe Menichelli[1,3], Jordan A. Ramilowski[4], Jessica Severin[4], Yoshihide Hayashizaki[5], Masayoshi Itoh[5], Akira Hasegawa[6], Harukazu Suzuki[7], FANTOM consortium, Piero Carninci[8], Michiel J.L. de Hoon[4], Wyeth W. Wasserman[9], Laurent Bréhélin[1,3,11] & Charles-Henri Lecellier[1,2,11]

[1]*Institut de Biologie Computationnelle, Montpellier, France*

[2]*Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France*

[3]*LIRMM, CNRS, Univ. Montpellier, Montpellier, France*

[4]*Laboratory for Applied Computational Genomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan*

[5]*RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Saitama, Japan*

[6]*Large Scale Data Managing Unit, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan*

[7]*Laboratory for Cellular Function Conversion Technology, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan*

[8]*RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan*

[9]*Centre for Molecular Medicine and Therapeutics at the Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver BC, Canada*

[10]*contributed equally to this work*

[11] *correspondence: brehelin@lirmm.fr, charles.lecellier@igmm.cnrs.fr*

Abstract

Using the Cap Analysis of Gene Expression technology, the FANTOM5 consortium provided one of the most comprehensive maps of Transcription Start Sites (TSSs) in several species. Strikingly, $\sim 72\%$ of them could not be assigned to a specific gene and initiate at unconventional regions, outside promoters or enhancers. To determine whether these unconventional TSSs, sometimes referred to as 'transcriptional noise' or 'junk', are relevant nonetheless, we look for novel and conserved regulatory motifs located in their vicinity. We show that, in all species studied, a significant fraction of CAGE peaks initiate at short tandem repeats (STRs) corresponding to homopolymers of thymidines. Biochemical and genetic evidence further demonstrate that several of these CAGEs correspond to TSSs of mostly sense and intronic non-coding RNAs, whose transcription rate can be predicted with $\sim 81\%$ accuracy by a sequence-based deep learning model. Excitingly, our model further predicts that genetic variants linked to human diseases affect this STR-associated transcription. Together, our results extend the repertoire of non-coding transcription and provides a valuable resource for future studies of complex traits.

Introduction

RNA polymerase II (RNAP-II) transcribes many loci outside annotated protein-coding gene (PCG) promoters [1,2] to generate a diversity of RNAs, including for instance enhancer RNAs [3] and long non-coding RNAs [4]. Non-coding transcription is far from being fully understood [5] and some authors suggest that many of these transcripts, often faintly expressed, can simply be 'noise' or 'junk' [6]. On the other hand, many non annotated RNAP-II transcribed regions correspond to open chromatin [1] and *cis*-regulatory modules (CRMs) bound by transcription factors (TFs) [7]. Besides, genome-wide association studies showed that trait-associated loci, including those linked to human diseases, can be found outside canonical gene regions [8–10]. Together, these findings suggest that the non-coding regions of the human genome harbor a plethora of functionally significant elements which can drastically impact genome regulations and functions [10,11]. Notably, short tandem repeats (STRs), which constitute one of the most polymorphic and abundant repetitive elements [12], have been shown to widely impact gene expression and to contribute to expression variation [13,14]. At the molecular level, STRs affect transcription by inducing inhibitory DNA structures [15] and/or by modulating the binding of transcription factors [16,17]. Provided that a growing number of regulatory elements is associated with non-coding transcription [5], we looked at STRs from another perspective and asked whether local transcription can be detected at these repetitive elements.

Using the Cap Analysis of Gene Expression (CAGE) technology [18,19], the FANTOM5 consortium provided one of the most comprehensive maps of TSSs in several species [2]. Integrating multiple collections of transcript models with FANTOM CAGE datasets, Hon *et al.* built an at-

las of 27,919 human lncRNAs, among them 19,175 potentially functional RNAs, and provided a new annotation of the human genome (FANTOM5 CAGE Associated Transcriptome, FANTOM CAT) [4]. Despite this annotation, many CAGE peaks cannot be assigned to a specific gene and/or initiate at unconventional regions, outside promoters or enhancers, providing an unprecedented mean to further characterize non-coding transcription within the genome 'dark matter' [11] and to decode part of the transcriptional noise.

Here, we probed CAGE data collected by the FANTOM5 consortium [2] and specifically looked for repeating DNA motifs around CAGE peak summits. In all species studied, we showed that a fraction of CAGE peaks (between 2.22% in rat and 6.45% in macaque) initiate at repeats of thymidines (Ts) i.e. STRs of Ts. Biochemical and genetic evidence demonstrate that many of these CAGEs do not correspond to technical artifacts [20] but rather exhibit genuine features of TSSs. Using sequence-based Convolutional Neural Network (CNN), we were able to predict this transcription with high accuracy ($\sim$ 81% in human) and to reveal that many genetic variants linked to human diseases potentially affect this STR-associated transcription. Together, our results extend the repertoire of non-coding transcription and advance the capacity to interpret regulatory variants [13].

Results

A significant fraction of CAGE peaks initiates at polyT repeats in various species. Using the motif enrichment tool HOMER [21], we looked at repeating DNA motifs in 21bp-long sequences

centered around the FANTOM5 CAGE peak summits. As shown in Figure 1, the first motif identified in both human and mouse is the canonical initiator element INR [22,23], demonstrating the relevance of our strategy to unveil specific sequence-level and TSS-associated features. A second motif corresponding to a polyT repeat starting at -2 bases (0 being the summit of the CAGE peak) is identified. This motif is present in 61,907 human and 8,274 mouse CAGE peaks, thereafter called polyT CAGEs (Figure 1A and B). In human, the median size of CAGE-associated polyT repeats is 17 bp with a minimum of 9 bp and a maximum size of 64 bp. Similar results are obtained in mouse (median = 21 bp, minimum = 8 bp and maximum = 58 bp). Looking directly at the number of polyT repeats of $\geqslant$ 9 Ts in the human genome, we found that 63,974 T repeats out of 1,337,561 exhibit a CAGE peak located at their 3'end + 2bp. Thus, the vast majority of polyT-associated CAGEs are identified by motif enrichment and not all T repeats are associated with CAGE peaks. In mouse, only 8,825 T repeats out of 834,954 are associated with a CAGE peak but, compared to human CAGE data, mouse data are small-scaled in terms of number of reads mapped and diversity in CAGE libraries [2].

We further looked at CAGE peaks in dog, chicken, macaque and rat (Supplementary Figure S1) and polyT repeat is invariably detected by HOMER (sometimes even before INR motif). The position of the CAGE summit at -2 is presumably artificially introduced by the hybridization step on the flow cell (Supplementary Figure S2) : sequencing of any CAGE tags initiating within T repeats will preferentially start just after the T repeats. The motif enrichment is therefore more indicative of transcription initiation at T repeats than of the existence of a true motif located precisely 2bp after T repeats (only 216 CAGE summits fall into a repeat of $\geqslant$ 9Ts in human). In line

with this, a small fraction of Start-seq [24] and DECAP-seq [25] TSSs can also initiate in sequences with a preference for T and this clearly does not represent a genuine motif (Supplementary Figure S3A and B). Note that the number of mouse CAGE summits [2] located within a 10bp window around Start-seq [24] and DECAP-seq [25] TSSs is low: $\sim$ 34.7% of CAGE TSSs have same coordinates as Start-seq TSSs but only $\sim$ 6.8% of CAGE TSSs share coordinates with DECAP-seq TSSs and $\sim$ 7.2% of Start-seq TSSs share coordinates with DECAP-seq TSSs. However, we noticed that mouse repeats of $\geqslant$ 9Ts associated with CAGE peaks are more associated with Start-seq and DECAP-seq peaks than T repeats not associated with CAGE peaks: $\sim$ 2% of CAGE-associated T repeats are also associated with Start-seq peaks (94 out of 4,600 T repeats considering only forward data) with only $\sim$ 0.3% T repeats not associated with CAGE but associated with Start-seq peaks (1256 out 413475, Fisher's exact test p-value $<$ 2.2e-16). Likewise, $\sim$ 0.36% of CAGE-associated T repeats are also associated with DECAP-seq peaks (32 out 8825) with only $\sim$ 0.03% T repeats not associated with CAGE but associated with DECAP-seq peaks (307 out 826129, Fisher's exact test p-value $<$ 2.2e-16). TSS-seq data collected in *Arabidopsis thaliana* [26] indicate that some plant TSSs can also initiate at repeats of Ts (Supplementary Figure S4, motif #2). As observed in animal FANTOM CAGE data, these TSS-seq peak summits never map a T (Supplementary Figure S4).

Heliscope sequencing used by FANTOM5 can be internally primed at polyT repeats (Supplementary Figure S2), thereby questioning the relevance of CAGE peaks at T repeats [20]. Several features are indeed in agreement with the idea that polyT CAGEs could for instance arise from internal priming within introns of messenger RNAs (Supplementary Figures S5 to S9 and Supplementary Tables S1 and S2). However, the artifact scenario could not explain all polyT CAGEs as

4,098 (out of 63,974, $> 6\%$) are not located within one of the 53,220 genes of the FANTOM CAT annotation, one of the largest gene annotations so far. Besides, we observed some concordance between several technologies (Supplementary Figures S3 and S4). Together these observations raise the possibility that at least a fraction of polyT CAGEs represent genuine TSSs. The corresponding RNAs appear stable [28] according to the CAGE exosome sensitivity score previously computed by FANTOM [4] (median sensitivity score = 0.08, Supplementary Figure S6B), suggesting that they do not correspond to cryptic transcription [5]. To clarify the existence of these TSSs, we further investigated whether some polyT CAGEs exhibit canonical TSS features.

Several polyT CAGE tags are truly capped. We used a strategy described by de Rie *et al.* [29], which compares CAGE sequencing data obtained by Illumina (ENCODE) vs. Heliscope (FANTOM) technologies. Briefly, the 7-methylguanosine cap at the 5' end of CAGE tags produced by RNA polymerase II can be recognized as a guanine nucleotide during reverse transcription. This artificially introduces mismatched Gs at Illumina tag 5' end, which is not detected with Heliscope [29]. Although such G bias is not clearly observed when considering the whole CAGE peak, it is readily detectable when considering the 3' end of polyT repeat (position -2 from FANTOM CAGE summit) using Illumina ENCODE CAGE data produced in Hela-S3 nuclei (Figure 2A). This bias is also observed in other cell types and is comparable to that observed with CAGE tags assigned to gene TSSs (Figure 2B and Supplementary Figure S10). In the case of assigned CAGEs, the G bias is observed in all reads mapping to the CAGE tags. Conversely, most CAGE tag 5' ends perfectly match the sequences of pre-miRNA 3'end, as previously reported [29], or that of 61,907 randomly chosen genomic positions (Figure 2B and Supplementary Figure S10). Mismatched Gs at the 3'

end of all T repeats located within the same genes as polyT CAGEs are also detected (Figure 2B and Supplementary Figure S10), though the abundance of tags is higher at polyT CAGEs (Figure 2B), suggesting the potential existence of false negatives (see below). Overall, these analyses show that polyT CAGEs are truly capped as opposed to post-transcriptional cleavage products of pri-mRNAs [29] or to random genomic positions.

Many polyT CAGEs are associated with transcription-related epigenetic marks. Because CAGE technology can capture not only TSSs but also the 5' ends of post-transcriptionally processed RNAs [30], we determined whether polyT CAGEs are associated with typical epigenetics marks associated with transcription at the chromatin level. Using the ENCODE DNaseI Hypersensitive Site (DHS) Master list, we noticed that $\sim$ 11% of polyT CAGEs (7028 out of 63,974) lie in DHSs, while this is true for only $\sim$ 6% of T repeats without CAGE peaks (76,616 out of 1,273,587, Fisher's exact test p-value $<$ 2.2e-16). We next investigated potential transcription at T repeats at the gene level and library-wise in order to preclude the effect of global chromatin/gene environment. To limit the influence of genomic context and associated epigenetics modifications, we created two sets of 'expressed' and 'non-expressed' polyT CAGEs : polyT CAGEs are considered as 'expressed' if (i) associated with a detectable CAGE signal in the sample considered (TPM $>$ 0) and (ii) located in a gene containing at least one 'non-expressed' polyT CAGEs. Conversely, 'non-expressed' polyT CAGEs are (i) not detected in the sample considered but detected in other samples and (ii) located in a gene containing at least one 'expressed' polyT CAGEs. Genome segmentation provided by combined ChromHMM and Segway [31,32] shows that 'expressed' polyT CAGEs are systematically more enriched in regions corresponding to predicted transcribed regions

than 'non expressed' polyT CAGEs (Figure 3A, Fisher's exact test p-value $<$ 2.2e-16 in HeLa-S3 (CNhs12325) and in GM12878 (CNhs12331), p-value = 2.053e-13 in K562 (CNhs12334)), despite being in the same genes and chromatin environment. Using the same comparison, transcription at T repeats was also confirmed with GRO-seq data (Figure 3B and Supplementary Figures S11 and S12).

We then looked at epigenetics marks individually and used data collected by the Roadmap Epigenome consortium in H1 embryonic stem cells to compare the epigenetics status of 'expressed' vs. 'non-expressed' polyT CAGEs. In three replicates of untreated H1 cells CAGE libraries, H3K36me3 and H3K4me3 peaks are invariably enriched in 'expressed' polyT CAGEs (Figure 3C and Supplementary Figure S11). Similar profiles are obtained with ENCODE ChIP-seq data, although less pronounced in GM12878 (Supplementary Figure S14). H3K4me3 levels at polyT CAGEs are low (Supplementary Figures S13 and S14), as observed with lncRNAs [33]. H3K36me3 is a histone modification mark enriched in the gene body region and associated with transcription elongation [33]. H3K4me3 is a mark classically associated with active or poised transcription start sites [33]. Hence, 'expressed' polyT CAGEs , as opposed to 'non-expressed' ones, are associated with H3K4me3/H3K36me3 domains. Interestingly, this type of 'K4-K36' domains have been previously used to characterize lncRNAs [34]. Figure 3C also shows an enrichment in H3K4me2 (see also Supplementary Figure 13 and S14), which was previously associated with intragenic *cis*-regulatory [35] and TF binding [36] regions. We concluded that, overall, detection of CAGEs at polyT repeats is associated with transcription-related chromatin marks.

Several polyT CAGEs correspond to annotated transcript and enhancer TSSs. We further assessed the presence of polyT-associated transcription among annotated TSSs. Motif enrichment around FANTOM CAT TSSs [4] shows that $\sim$ 1.5% of them initiate downstream T repeats (Figure 4A). Looking directly at TSS coordinates of FANTOM CAT robust transcripts [4], we noticed that 6,734 TSSs corresponding to 10,606 robust transcripts (out of 525,237, $\sim$ 2%) initiate 2bp after a T repeat, with a clear enrichment in lncRNA_intergenic, lncRNA_sense_intronic or sense_overlap_RNA (hypergeometric test p-values $<$ 2.2e-16 for the 3 RNA classes). A list of these TSSs is provided in Supplementary Table S3A. Similar results were obtained for 5,889 stringent transcripts (n = 402,813) [4]. Importantly, transcript models in FANTOM CAT combine various independent sources (GENCODE release 19, Human BodyMap 2.0, miTranscriptome, ENCODE and an RNA-seq assembly from 70 FANTOM5 samples) and FANTOM CAT TSSs were validated with Roadmap Epigenome DHS and RAMPAGE data sets [4]. As expected, these TSSs only moderately contribute to gene expression (Supplementary Figure S15). No specific Open Reading Frame could be detected 2kb downstream polyT CAGEs (Supplementary Figure S16).

PolyT motif is also observed for $\sim$ 4% of FANTOM enhancer TSSs [3] (Figure 4B), though no enrichment in enhancer epigenetic marks are predicted by chromHMM/Segway (Figure 3A). Looking directly at genomic coordinates, 4,976 enhancers (out of 65,423, $\sim$ 7.6%) are defined with at least one polyT-associated CAGE (Supplementary Table S3B) and 173 enhancers are defined by two polyT-associated CAGEs [3]. Hence, enhancer TSSs are more often associated with polyT that transcript TSSs ($\sim$ 7.6% and $\sim$ 2% respectively, Fisher's exact test p-value $<$ 2.2e-16). Similar results were obtained with mouse enhancers (Supplementary Figure S17) with 1,171 enhancers

(out of 44,459, $\sim 2.6\%$) involving at least one polyT-associated CAGE. These results reinforce the idea that a number of polyT CAGEs correspond to genuine TSSs.

PolyT CAGEs interact with gene TSSs. We used ENCODE Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) directed against RNAP-II in K562 cell line [38] to better characterize the elements interacting with polyT CAGEs thanks to DNA looping. As in Figure 3, we used the sets of 'expressed' and 'non-expressed' polyT CAGEs in K562 cells. The 'expressed' polyT CAGEs are enriched in ChIA-PET data compared to 'non-expressed' polyT CAGEs (Figure 5A), confirming the results depicted in Figure 3. Second, using combined chromHMM/Segway, we observed that regions interacting with 'non-expressed' polyT CAGEs mostly correspond to 'Predicted Repressed or Low Activity' region (Figure 5B and Supplementary Figure S18). On the other hand, regions interacting with 'expressed' polyT CAGEs mostly correspond to 'Predicted promoter region including TSS' regions (Figure 5B and Supplementary Figure S18). Third, we used the FANTOM5 CAGE TSS classification [2] and showed that CAGEs interacting with 'expressed' polyT CAGEs mostly correspond to 'strict' and 'weak' TSSs, while 'non-expressed' polyT CAGEs preferentially interact with 'no' TSSs (Figure 5C and Supplementary Figure S19).

Transcription at T repeats is triggered by sequence-level instructions. We built a deep Convolutional Neural Network (CNN) model to predict the transcription at T repeats. We considered CNN as a regression problem and decided to predict the mean raw tag count of each T repeat across all FANTOM libraries (Supplementary Figure S20). Plotting the densities of mean raw tag counts at T repeats associated or not with CAGE peaks shows the existence of potential false neg-

atives i.e. T repeats not associated with CAGE peaks but associated with CAGE tags nonetheless (Supplementary Figure S21), likely explaining results depicted in Figure 2B and Supplementary Figure S10. CNN model architecture and the constructions of the different sets used for learning are shown Supplementary Figure S22. As input, we used sequences spanning $\pm$ 50bp around the 3' end of each T repeat. Longer sequences were tested without improving the accuracy of the model (Supplementary Figure S23). The accuracy of our model, computed as Spearman correlation between the predicted and the observed tag counts, is remarkably high (median Spearman r over 10 trainings = 0.81, Supplementary Figure S24A) and the error low (median absolute error for one model = 1.82, Supplementary Figure S24B). This demonstrates that transcription at T repeats is mainly directed by sequence-level instructions.

Probing sequence-level instructions for transcription at T repeats. We further used our deep CNN model to probe these instructions. Because the length of the T repeats associated with CAGE peaks is greater than that of T repeats not associated with CAGE peaks but located within the same genes (Figure 6A), we first compared predictions with varying lengths of polyT track. We chose a T repeat with a low error and increased the length of its polyT repeat ranging from 9 to 51. The predicted tag count increases with the length of the polyT track (Figure 6B). We next tested whether our model captures additional instructions, lying downstream the T repeat. The accuracy of our model increases considering increasing lengths of sequence located downstream T repeat with a plateau at 40-50 bp (Figure 6C). We further considered two specific T repeats, one, A, with a low predicted tag count (0.9, observed tag count = 1.0) and another, B, with a high predicted tag count (39.37, observed tag count = 39.44) (Supplementary Figure S25). Sequences located downstream

each T repeat were swapped and the prediction was assessed in each case. The prediction of the artificial sequence A drastically increased while the prediction of the artificial B decreased (Supplementary Figure S25), strengthening the existence of sequence-level instructions downstream T repeats. To more precisely characterize these instructions, we considered 50 sequences with low error and high tag count (set 'high') and 50 sequences with low error and low tag count (set 'low'). For all pairs of 'high' and 'low' sequences, we sequentially replaced 5-mers from one sequence with 5-mers from the other. We then predicted the tag count of the 2 new sequences for each 5-mer swapping and assessed transcription change as the difference between tag counts before and after 5-mer swapping. Inserting 5-mers of 'high' sequences in 'low' sequences induces overall positive changes (i.e. increased transcription) (Figure 6D), while inserting 5-mers of 'low' sequences in 'high' sequences overall decreases transcription (Figure 6E). Besides, these 5-mer segmentation reveals that the instructions located near the end of T repeats (i.e. close to the CAGE peak summit) are crucial (i.e. can induce more change in the prediction) compared to the rest of the sequence (Figures 6D and E). However, except their lengths (Figure 6A), no specific motif clearly distinguished T repeats with high tag count (top 5,000) from T repeats with low tag count (top 5,000) (Supplementary Figure S26). Moreover, we randomly modified sequences in order to maximize their predicted tag count and to computationally generate an optimized sequence (see methods). Pairwise comparisons of these optimized sequences (2,000 comparisons) reveal that the only feature common to all optimized sequences is the length of the T repeat and no specific and unique subsequence (i.e. motif) could be identified by this approach (Supplementary Figure S27). Motif search indeed confirms the importance of T repeat length and absence of a single motif able to dis-

tinguish 'high' and 'low' sequences (Supplementary Figure S26). As expected, a simple 1-layer CNN, which is similar to finding enriched k-mers or motifs, poorly performed (median Spearman r = 0.2). Together these results show that a single representation cannot model the intricate structure of the features captured by CNN.

Alternative methods were tested, namely LASSO [39] and Random Forest [40], using, as predictive variables, mono-, di- and tri-nucleotides rates in the region located upstream (-50bp) or downstream (+50bp) T repeat ends, the T repeat lengths and the first nucleotide (A,C,G or T) downstream the T repeats (-1 when considering 0 as the CAGE summit). In that case, the model accuracies are 51% and 61% for LASSO and Random Forest respectively. These models are not as accurate as CNN but they confirm that T repeat length is the main feature being used from the sequence located upstream T repeat end, while other features lie downstream the T repeat (Supplementary Table S4 and Supplementary Figure S28).

Predicting transcription at T repeats in other species. Deep CNN could also predicts transcription at mouse T repeats (Spearman r = 0.77 and median absolute error = 1.82). We then asked whether the human and mouse models (i.e. learned features) were similar. While the human model can predict transcription at mouse T repeats (Spearman r = 0.72), the mouse model poorly performs on human CAGE data (Spearman r = 0.39). However, compared to human CAGE data, mouse data are small-scaled in terms of number of reads mapped and diversity in CAGE libraries [2], making the signal in mouse less accurate than in human. It is therefore possible that a CNN trained on human data will learn more features than a model learned on mouse data.

We also tried to build a model in chicken, human's most distant species considered in FANTOM5 CAGE data. The prediction, though not null, is much less accurate than in human or in mouse (Spearman r = 0.61 and median absolute error = 1.09). Only 58 libraries are available in chicken and the mean signal across these library may not be as robust as in mouse or in human. To confirm that the robustness of the CAGE signal is directly linked to the number of libraries considered, we computed the mean raw tag count in only 58 randomly chosen human libraries and learned a new model using the same architecture as in Supplementary Figure S22. In that case, the signal appears very sparse compared to a signal computed on 988 libraries (Supplemental Figure S29) and the model accuracy falls to 0.61. These results reveal that the CAGE signal at T repeats is noisy when considered library-wise but becomes robust when averaged on numerous libraries.

Clinically relevant variants are predicted to impact transcription at T repeats. Previous studies showed that genomic variations at polyT repeats can impact gene expression. For instance, the variant rs10524523, located in TOMM40 intron 6, which corresponds to a T repeat associated with a CAGE signal in several FANTOM libraries, is linked to the age of onset of cognitive decline, Alzheimer's disease and sporadic inclusion body myositis [41–46]. The TOMM40 mRNA brain expression appears to be linked to the length of the polyT repeat with the longer variant the higher expression [41,42]. Conversely, the KIAA0800/ENSG00000145041 promoter is repressed by an upstream element involving a polyT repeat and two Alu repeats [47]. In that case, the CAGE signal detected at this particular polyT repeat has been annotated as one TSS of KIAA0800/ENSG00000145041 in FANTOM CAT [4]. More widely, STRs of T homopolymers have been reported to impact gene expression in a directional and strand-specific manner [48].

We used our model to predict the effects of genomic variations located within sequences surrounding T repeats (end of T repeat $\pm$ 50bp). First, we observed, in the case of TOMM40 intron 6 variant, that transcription is positively correlated to the length of the T repeat (Figure 7A). Second, we compared the effect of variants listed in dbSNP, the world's largest database for nucleotide variations [49], and variants listed in ClinVar, which pinpoints medically important variants [50]. Almost all T repeats harbor variations listed in dbSNP (962,241 out of 1,169,145), while 2,577 T repeats are associated with variations listed in ClinVar. Strikingly, these T repeats tend to be associated with high tag counts compared to T repeats with dbSNP variants (Figure 7B), indicative of potential clinical relevance of transcription at T repeats. The ClinVar variants were not equally distributed around T repeats (Figure 7C). Overall, ClinVar variants are frequently found around T repeats, with a peak at position 0 and +1 (0 being the end of T repeat), close to the CAGE summit (+2 in that case). This particular distribution was not observed when considering all permissive CAGE peak summits (Supplementary Figure S30). The variants located at the vicinity of T repeats are predicted by our model to induce extensive changes in transcription, being overall positive (i.e. transcription increase) when located on both sides of T repeat while negative (i.e. transcription decrease) when located within T repeat (Figure 7D). These differential effect are not linked to the nature of the variants (i.e. SNP or insertion/deletion) as both types of variants can have positive and negative effects on transcription with insertions/deletions having broader effects (Supplemental Figure S31). In fact, genomic variations within a T repeat will more likely disrupt the repeat and decrease its length, thereby inducing an overall decrease in transcription prediction. Variations linked to certain diseases were enriched in the 101bp-long regions centered around T repeat ends,

although no clear enrichment for specific types of diseases was noticed (Supplementary Table S5).

Discussion

We looked at recurring DNA motifs associated with FANTOM CAGE peaks [2] and showed that a significant portion of them initiate at polyT repeats in several species. We do not exclude that some CAGE peaks initiating at polyT repeats are generated by internal priming during Heliscope sequencing. Yet we provide evidence that a fraction represents genuine TSSs as (i) polyT CAGEs are truly capped and associated with typical epigenetic marks, (ii) several of them correspond to annotated long non-coding transcripts and enhancer RNAs detected by various technologies in different laboratories and (ii) their expression can be predicted by a deep learning model, which uses features located downstream T repeats. We show that GRO-seq coincides with CAGE at T repeats (Figure 3B and Supplementary Figures S11 and S12). Li *et al.* suggested that CAGEs detected at polyA/T repeats may correspond to technical artifacts because they are not detected by a GRO-seq method that enriches for 5'-capped RNAs (GRO-cap) [20]. However, in comparison to CAGE, GRO-cap generates fewer reads mapping to introns [51], making hard to confirm the existence of faintly expressed and intronic CAGEs, such as polyT CAGEs, with GRO-cap. In fact, non-coding transcription appear overall much less efficient than that of mRNAs [5] (Supplementary Figure 7A). It is also worth noticing that several features insinuating that polyT CAGEs are technical artifacts may constitute specific properties of lncRNAs. For instance, the majority of lncRNAs are enriched in the nucleus [52–54].

The human genome is scattered with repetitive sequences, and the vast majority of the genomic DNA is supposed to be transcribed [1]. Implicitly, the human transcriptome should contain a large portion of RNAs derived from repetitive elements (called repeat-derived RNAs, repRNAs in [58]). We provide here an example of such repRNAs. Homopolymers of $>$ 10 Ts appear at frequencies well above chance in the non-coding regions of several organisms, suggesting selective advantages that lead to their overrepresentation [59]. In human and mouse genomes, polyT repeats can correspond to the polyA tails of short interspersed nuclear elements (SINEs) [60]. However, according to RepeatMasker, only a fraction of human T repeats associated with CAGE peaks correspond to SINEs ($\sim$ 37% vs. $\sim$ 50% for T repeats without CAGE peak, Supplementary Table S6). PolyT-CAGEs are also detected in chicken (Supplementary Figure S1) where SINE elements are rare [61] (Repeats of $\geqslant$ 9 Ts represent 0.04% of the chicken genome, akin to human). These observations argue against the existence of a link between polyT CAGEs and SINEs but rather reinforce the intrinsic relevance of these homopolymers [59].

Whether polyT-associated non-coding RNAs or the act of transcription *per se* is functionally important [5] remain to be clarified. This is all the more warranted as T repeat transcription is associated with clinically relevant genomic variations (Figure 7B). This non-coding transcription, mostly sense and intronic, may specifically regulate certain coding - as opposed to non-coding - genes (Supplementary Figure 7B, Supplementary Tables S1 and S2). In line with this idea, we show here that polyT CAGEs preferentially interact with gene TSSs (Figure 5).

PolyT repeats have been shown to act as promoter elements favoring transcription by deplet-

ing repressive nucleosomes [62], specifically by orientating the displacement of nucleosomes [63]. As a consequence, polyT repeats can increase transcription of reporter genes in similar levels to TF binding sites [64].

Besides, eSTRs located within the gene body are more likely to have T on the template strand and show higher gene expression with increasing repeat number [48]. This directional and strand-specific bias [48] can be molecularly explained by transcription at T repeats. These observations, in addition to overrepresentation of long T repeats in eukaryotic genomes [59], suggest functional role(s) for transcription at T repeats but dedicated experiments are required to formally assess its contribution in gene expression.

Methods

Datasets and online resources. Publicly availble data used in this study can be found at the urls provided in Supplementary Table S7. The genome assemblies considered are human hg19, mouse mm9, chicken galGal5, dog canFam3, rat rn6, rhesus macaque rheMac8. Intron coordinates were downloaded from UCSC table browser. Human RepeatMasker coordinates were downloaded from UCSC table browser.

Motif analyses. The HOMER motif analysis tool [21] was used to find 21bp long-motifs in regions spanning 10bp around CAGE peakmax (*findMotifsGenome.pl* with options -len 21, -size given, -noknown and -norevopp). HOMER was further used to identify, CAGEs harboring specific motifs (*annotatePeaks.pl* with option -m) in particular INR motif (motif1 in Figure 1) and polyT repeat (motif2 in Figure 1).

Characterization of T repeats in human and mouse. T repeats of more than 9Ts were identified in the human, mouse and chicken genome using the tool developed by Martin *et al.* [65]. The human, mouse and chicken coordinates are provided as supplementary bed files.

Bioinformatics tools. Intersections between genomic coordinates were computed with BEDTools [66]. Signal were extracted from bigwig files using bwtool (*agg* and *extract* subcommands) [67] and deepTools [68]. Statistical tests were done using R [69] (*fisher.test, wilcox.test, cor*) or Python (*fisher_exact, mannwhitneyu, spearmanr* from scipy.stats (http://www.scipy.org/)). Graphs were made using ggplot2 R package [70] or Python matpotlib library (https://matplotlib.org/api/pyplot_api.html). All scripts are available upon request.

Evaluating mismatched G bias at Illumina 5'end CAGE reads. Comparison between Helliscope vs. Illumina CAGE sequencing was performed as in de Rie *et al.* [29]. Briefly, ENCODE CAGE data were downloaded as bam file (Supplementary Table S7) and converted into bed file using samtools view [71] and unix awk as follow:

```
samtools view file.bam | awk '{FS="\t"}BEGIN{OFS="\t"}{if($2=="0") print $3,
    $4-1,$4,$10,$13,"+"; else if($2=="16") print $3,$4-1,$4,$10,$13,"-"}' >
    file.bed
```

. The bedtools intersect [66] was further used to identify all CAGE reads mapped at a given position (see Figure 2B and C). The unix awk command was used to count the number and type of mismatches as follow:

```
intersectBed -a positions_of_interest.bed -b file.bed -wa -wb -s |
awk '{if(substr($11,1,6)=="MD:Z:0" && $6=="+") print substr($10,1,1)}' | grep
    -c "N"
```

with N = {A, C, G or T}, positions_of_interest.bed being for instance polyT-associated CAGE coordinates and file.bed being the Illumina CAGE tag coordinates.
Absence of mismatch focusing on the plus strand were counted as:

```
intersectBed -a positions_of_interest.bed -b file.bed -wa -wb -s |
 awk '{if(substr($11,1,6)!="MD:Z:0" && $6=="+") print $0}' |wc -l
```

As a control, 61,907 random positions were generated with bedtools *random* (-l 1 and -n 61907). We also used the 3' end of the pre-miRNAs, which were defined, as in de Rie *et al.* [29], as the 3' nucleotide of the mature miRNA on the 3' arm of the pre-miRNA (miRBase V21, see Supplementary Table S7), the expected Drosha cleavage site being immediately downstream of

this nucleotide (pre-miR end + 1 base).

Epigenetic marks at T repeats. Average chromatin CAGE signal around polyT CAGEs (Figure **??**D) were computed using the *agg* subcommand of the bwtool [67]. For RNAP-II ChIA-PET, interacting regions within the same chromosome were extracted using the column name $4 of the bed file. Roadmap and GENCODE ChIP-seq as well as GROseq data were extracted from URLs indicated in Supplementary Table S7. To look at epigenetic marks around T repeats into H1-hESC cell line, the coordinates of 'expressed' and 'non-expressed' T repeats ends+2 bp (see results for details) were intersected with Roadmap bedfiles using bedtools intesect [66].

```
intersectBed -wb -a epigenetic_mark.bed -b expressed_polyT_CAGEs.bed | sort -u
    | wc -l
```

The fractions of intersecting T repeats within each set were compared for each available epigenetic mark and each of the 3 H1-hESC samples using Fisher's exact test. The same procedure was used in other cell lines with ENCODE broadpeaks data for the H3K4me1, H3K4me2, H3K4me3 and H3K36me3.

For GRO-seq data, we used bigwig files from the 'gro-seq.colorado' repository (Supplementary Table S7). In H1-hESC cell line, we compared the fraction of intersecting 'expressed' and 'non-expressed' T repeats similar to the procedure used for GENCODE and Roadmap data. The GRO-seq signal around these two sets was also compared as the mean signal into a 100 bp window centered on the T repeat end using deeptools multiBigwigSummary [68].

```
multiBigwigSummary BED-file -b CellLine_Nascent_RNAseq_files_forward/*.bw
-o out.npz --BED sample_polyT.end_minus.plus50_coordinates_plusStrand.bed --
   outRawCounts sample_polyT.end_minus.plus50_meanSignal_plusStrand.txt

multiBigwigSummary BED-file -b CellLine_Nascent_RNAseq_files_backward/*.bw
```

```
-o out.npz --BED sample_polyT.end_minus.plus50_coordinates_minusStrand.bed --
    outRawCounts sample_polyT.end_minus.plus50_meanSignal_minusStrand.txt
```

We compared the fractions of T repeats with a no-null signal into both 'expressed' and 'non-expressed' sets using Wilcoxon test. We also calculated the sum of these mean signals across all the samples for each T repeat with similar results.

Convolutional Neural Network. A Zenbu [72] xml custom script was used to calculate the mean tag count of each base genome-wide in 988 FANTOM libraries. We then summed these means along each T repeat + 5bp. This values were used as predicted variables in a CNN model built on sequences centered around each T repeat end (n = 1,169,236). The same procedure was repeated with mouse (397 libraries) and chicken data (58 libraries) using the same architecture. The coordinates of the T repeats with the mean tag count values (score column) used to train our models are provided as supplementary bed files. The fasta file corresponding to T repeats sequences with their tag count was parsed using Biopython [73] library. To build the convolutional neural network, we used Keras [74] library in Python with Tensorflow [75] backend. The subsequent analyses of altering length of T repeats and swapping downstream sequence was performed by treating sequences as strings. The complete code can be found at `https://gite.lirmm.fr/ibc/t-repeats`.

Feature extraction. We looked at motif by random optimization of the CNN prediction using the following algorithm : first, two sequences are randomly chosen in the test set. For each sequence, the effect of changing one nucleotide at a random position into A,T, C or G is assessed. The nucleotide with the maximum predicted tag count is kept and the procedure continues testing

another position. This procedure is repeated until no further increase of predicted tag count is observed over 3,500 randomly selected positions. The two optimized sequences are then compared position-wise in order to return 1 for each position with the same nucleotide in both sequences and 0 otherwise. This process was repeated for 2,000 pairs of random sequences. Same procedure was applied minimizing the tag count for each sequence. The complete code can be found at `https://gite.lirmm.fr/ibc/t-repeats`.

Predicting impact of ClinVar variants ClinVar and dbSNP vcf files were downloaded and then converted to bed files.They were intersected with coordinates of all T repeats using bedtools intersect [66] as follows:

```
bedtools intersect -a clinvar_mutation.bed -b t_repeats.bed -wa -wb >
    t_repeats_clinvar.bed

bedtools intersect -a db_snp_mutation.bed -b t_repeats.bed -wa -wb >
    t_repeats_dbSNP.bed
```

The bed files generated above were then converted into Pandas [76] dataframe in python and variants were introduced in T repeats sequences.The CNN model developed previously was then used to predict the mean tag count of the mutated sequences. The change was computed using the following formula:

$$\frac{(a-b)}{b} \times 100$$

where a is the predicted tag count after variation and b the predicted tag count before variation. The complete code can be found at `https://gite.lirmm.fr/ibc/t-repeats`.

Comparison with other models. To compare our deep learning regression model with other parametric and non parametric approaches, we used linear regression with l1-norm penalty via LASSO [39] and Random Forest [40] to predict the mean tag count signal. For both approaches, the predictive variables are the nucleotide, dinucleotide and trinucleotide content computed into the 50 bp downstream the polyT end and 50 bp upstream the polyT end excluding the T repeat. We also integrated the length of the T repeat and the nature of the nucleotide following the T repeat end (A, C or G at position -1 referring to the CAGE summit at 0). LASSO inference was performed using the function cv.glmnet from the R package glmnet and Random Forest was performed using the randomForest R package. To compute the distribution of the minimal depth we used the *min_depth_distribution* function from the R package randomForestExplainer.

1. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

2. Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

3. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

4. Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199–204 (2017).

5. Ard, R., Allshire, R. C. & Marquardt, S. Emerging Properties and Functional Consequences of Noncoding Transcription. *Genetics* **207**, 357–367 (2017).

6. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Front Genet* **6**, 2 (2015).

7. Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* (2017).

8. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).

9. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

10. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6131–6138 (2014).

11. Clark, M. B., Choudhary, A., Smith, M. A., Taft, R. J. & Mattick, J. S. The dark matter rises: the expanding world of regulatory RNAs. *Essays Biochem.* **54**, 1–16 (2013).

12. Willems, T., Gymrek, M., Highnam, G., Mittelman, D. & Erlich, Y. The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).

13. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).

14. Press, M. O., McCoy, R. C., Hall, A. N., Akey, J. M. & Queitsch, C. Massive variation of short tandem repeats with functional consequences across strains of Arabidopsis thaliana. *Genome Res.* **28**, 1169–1178 (2018).

15. Rothenburg, S., Koch-Nolte, F., Rich, A. & Haag, F. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 8985–8990 (2001).

16. Contente, A., Dittmer, A., Koch, M. C., Roth, J. & Dobbelstein, M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat. Genet.* **30**, 315–320 (2002).

17. Martin, P., Makepeace, K., Hill, S. A., Hood, D. W. & Moxon, E. R. Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 3800–3804 (2005).

18. Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**, 1150–1159 (2011).

19. Murata, M. *et al.* Detecting expressed genes using CAGE. *Methods Mol. Biol.* **1164**, 67–85 (2014).

20. Li, C., Lenhard, B. & Luscombe, N. M. Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome. *Genome Res.* **28**, 676–688 (2018).

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

22. Smale, S. T. & Baltimore, D. The "initiator" as a transcription control element. *Cell* **57**, 103–113 (1989).

23. Butler, J. E. & Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592 (2002).

24. Scruggs, B. S. *et al.* Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol. Cell* **58**, 1101–1112 (2015).

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

26. Nielsen, M. *et al.* Transcription-driven Chromatin repression of Intragenic transcription start sites. *PLoS Genet.* **15**, e1007969 (2019).

27. Wei, W., Pelechano, V., Jarvelin, A. I. & Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet.* **27**, 267–276 (2011).

28. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat Commun* **5**, 5336 (2014).

29. de Rie, D. *et al.* An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.* **35**, 872–878 (2017).

30. Fejes-Toth, K. *et al.* Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).

31. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–216 (2012). URL `http://www.nature.com/nmeth/journal/v9/n3/full/nmeth.1906.html`.

32. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**, 473–476 (2012). URL `http://www.nature.com/nmeth/journal/v9/n5/full/nmeth.1937.html`.

33. Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* **46**, 1–19 (2012).

34. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).

35. Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res.* **20**, 1493–1502 (2010).

36. Wang, Y., Li, X. & Hu, H. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics* **103**, 222–228 (2014).

37. Andersson, R., Sandelin, A. & Danko, C. G. A unified architecture of transcriptional regulatory elements. *Trends Genet.* **31**, 426–433 (2015).

38. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).

39. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288 (1996).

40. Liaw, A. & Wiener, M. Classification and regression by randomforest. *R News* **2**, 18–22 (2002). URL `https://CRAN.R-project.org/doc/Rnews/`.

41. Payton, A. *et al.* A TOMM40 poly-T variant modulates gene expression and is associated with vocabulary ability and decline in nonpathologic aging. *Neurobiol. Aging* **39**, 1–7 (2016).

42. Linnertz, C. *et al.* The cis-regulatory effect of an Alzheimer's disease-associated poly-T locus on expression of TOMM40 and apolipoprotein E genes. *Alzheimers Dement* **10**, 541–551 (2014).

43. Maruszak, A. *et al.* TOMM40 rs10524523 polymorphism's role in late-onset Alzheimer's disease and in longevity. *J. Alzheimers Dis.* **28**, 309–322 (2012).

44. Greenbaum, L. *et al.* The TOMM40 poly-T rs10524523 variant is associated with cognitive performance among non-demented elderly with type 2 diabetes. *Eur Neuropsychopharmacol* **24**, 1492–1499 (2014).

45. Bernardi, L. *et al.* Role of TOMM40 rs10524523 polymorphism in onset of alzheimer's disease caused by the PSEN1 M146L mutation. *J. Alzheimers Dis.* **37**, 285–289 (2013).

46. Mastaglia, F. L. *et al.* Polymorphism in the TOMM40 gene modifies the risk of developing sporadic inclusion body myositis and the age of onset of symptoms. *Neuromuscul. Disord.* **23**, 969–974 (2013).

47. Zhao, L. J., Zhang, S. & Chinnadurai, G. Sox9 transactivation and testicular expression of a novel human gene, KIAA0800. *J. Cell. Biochem.* **86**, 277–289 (2002).

48. Feupe Fotsing, S. *et al.* Multi-tissue analysis reveals short tandem repeats as ubiquitous regulators of gene expression and complex traits. *bioRxiv* (2018). URL `https://www.biorxiv.org/content/early/2018/12/13/495226`. `https://www.biorxiv.org/content/early/2018/12/13/495226.full.pdf`.

49. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

50. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–868 (2016).

51. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).

52. Bergmann, J. H. *et al.* Regulation of the ESC transcriptome by nuclear long noncoding RNAs. *Genome Res.* **25**, 1336–1346 (2015).

53. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).

54. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

55. Guo, Q. *et al.* Comprehensive analysis of lncRNA-mRNA co-expression patterns identifies immune-associated lncRNA biomarkers in ovarian cancer malignant progression. *Sci Rep* **5**, 17683 (2015).

56. Wu, L. *et al.* A new avenue for obtaining insight into the functional characteristics of long noncoding RNAs associated with estrogen receptor signaling. *Sci Rep* **6**, 31716 (2016).

57. Zhao, Z. *et al.* Co-LncRNA: investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. *Database (Oxford)* **2015** (2015).

58. Matylla-Kulinska, K., Tafer, H., Weiss, A. & Schroeder, R. Functional repeat-derived RNAs often originate from retrotransposon-propagated ncRNAs. *Wiley Interdiscip Rev RNA* **5**, 591–600 (2014).

59. Dechering, K. J., Cuelenaere, K., Konings, R. N. & Leunissen, J. A. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res.* **26**, 4056–4062 (1998).

60. Goodier, J. L. Restricting retrotransposons: a review. *Mob DNA* **7**, 16 (2016).

61. Gao, B. *et al.* Low diversity, activity, and density of transposable elements in five avian genomes. *Funct. Integr. Genomics* **17**, 427–439 (2017).

62. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19**, 65–71 (2009).

63. Krietenstein, N. *et al.* Genomic Nucleosome Organization Reconstituted with Pure Proteins. *Cell* **167**, 709–721 (2016).

64. Weingarten-Gabbay, S. *et al.* Systematic interrogation of human promoters. *Genome Res.* **29**, 171–183 (2019).

65. Martin, D., MAILLOL, V. & Rivals, E. Fast and accurate genome-scale identification of DNA-binding sites. In *BIBM: Bioinformatics and Biomedicine*, 201–205 (Madrid, Spain, 2018). URL `https://hal-lirmm.ccsd.cnrs.fr/lirmm-01967466`. This is the author version of the article published in the conference proceedings. It includes supplementary information. A software called MOTIF is available on the ATGC bioinformatics platform.

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

67. Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618–1619 (2014). URL `http://bioinformatics.oxfordjournals.org/content/30/11/1618`.

68. Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–165 (2016).

69. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017). URL `https://www.R-project.org/`.

70. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016). URL `http://ggplot2.org`.

71. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

72. Severin, J. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat. Biotechnol.* **32**, 217–219 (2014).

73. Dalke, A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009). URL `https://doi.org/10.1093/bioinformatics/btp163`. `http://oup.prod.sis.lan/bioinformatics/article-pdf/25/11/1422/944180/btp163.pdf`.

74. Chollet, F. *et al.* Keras. `https://keras.io` (2015).

75. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL `http://tensorflow.org/`. Software available from tensorflow.org.

76. McKinney, W. Data structures for statistical computing in python. In van der Walt, S. & Millman, J. (eds.) *Proceedings of the 9th Python in Science Conference*, 51 – 56 (2010).

Acknowledgements We thank Cédric Notredame, Anthony Mathelier, Oriol Fornes Crespo, Philip Richmond, Jean-Christophe Andrau, Diego Garrido Martin, Dimitri D. Pervouchine, Roderic Guigo, Charles Plessy and Chung Hon for their help in analyzing the data and for insightful suggestions. We are indebted to the researchers around the globe who generated experimental data and made them freely available. C-H.L. is grateful to Marc Piechaczyk and Edouard Bertrand for continued support.

Author contributions C.B., M.S., M.G. W.W.W., M.d.H, L.B and C-H.L. analyzed and interpreted data. M.S. and M.G. developed CNN models and extracted features. C.M. identified T repeats in human, mouse, chicken genome. J.R., Y.H., A.H., H.S., S.N., I.M. generated CAGE data used in this study. M.d.H., J.S. and C-H.L . generated Zenbu tracks. M.d.H and C-H.L. studied G bias at ENCODE read 5' ends. P.C., W.W.W, L.B. and C-H.L acquired fundings. C-H.L. wrote the manuscript. All authors have read and approved the manuscript.

Funding The work was supported by funding from CNRS (International Associated Laboratory "miREGEN"), INSERM-ITMO Cancer project "LIONS" BIO2015-04, *Plan d'Investissement d'Avenir* #ANR-11-BINF-0002 *Institut de Biologie Computationnelle* (young investigator grant to C-H.L.) and GEM Flagship project funded from Labex NUMEV (ANR-10-LABX-0020).

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to L.B. and C-H.L.. (email: brehelin@lirmm.fr and charles.lecellier@igmm.cnrs.fr).

Figure 1 Motif discovery around CAGE peakmax. HOMER [21] was used to find 21bp-long motifs in 21bp-long sequences centered around the peak max of FANTOM5 CAGEs. **A.** 1,048,124 human permissive CAGEs were used. The top 5 is shown. 61,907 CAGEs are associated with polyT repeat starting at -2 (0 being the peakmax). **B.** Among 158,966 mouse CAGEs, 8,274 are associated with polyT repeat. Note that another repeat of 2-5 Gs is also identified in FANTOM CAGE data with strong similarities in both human and mouse (A and B). However, this repeat is less strict and is not conserved in dog, chicken, macaque and rat (Supplementary Figure S1) and is absent in Start-seq and DECAP-seq data (Supplementary Figure S3). Though this motif may represent a biologically relevant signal in human and mouse, these features make it less relevant for our study. We then focused our analyses on polyT CAGEs .

Figure 2 polyT CAGEs are 5'-capped. A. G bias in ENCODE CAGE reads was assessed at the indicated positions (x-axis) around polyT CAGEs (0 corresponds to end of poly repeat + 2bp, n = 63,974). The number of intersecting reads on the positive strand is indicated. **B.** Same analyses as in A but considering position -2 only and other CAGEs/positions indicated in x-axis: CAGE peaks assigned to gene, pre-microRNA 3' ends, 61,907 random positions, T repeat without CAGE. Note that the number of polyT CAGEs on + strand is 33,209 and that of T repeats without CAGE 636,993 indicating that the number of reads at T repeats with CAGE is relatively higher than that at T repeats without CAGE.

Figure 3 Several polyT CAGEs are associated with promoter-related epigenetics marks A. Distribution of ChromHMM/Segway genome segments containing 'expressed' or 'non-expressed' polyT CAGEs (see text for definitions) in HeLa-S3, GM12878 and K562. 61,907 random position were also considered for comparison. Note the absence of 'Predicted promoter flanking region' class in K562. **B.** Fraction of polyT CAGEs 'expressed' (blue) or not (red) in H1 embryonic stem cells (CNhs13964 FANTOM library) and intersecting with nascent RNAseq signal from 26 gro-seq colorado H1-ESC samples (SRR*) (considering a window of 101bp centered around T repeat end). Boxplots show the results of the intersection for 26 samples (see Supplementary Table S7 for individual analyses). Fisher's exact test indicates that the fraction of polyT CAGEs with no-null GROseq signal obtained for 'expressed' polyT CAGEs is higher than that obtained for 'non expressed' polyT CAGEs (p-value $<$ 2.2e-16). **C.** Intersection of 'expressed'/'non expressed' polyT CAGEs coordinates with that of Roadmap epigenetics data collected in H1 embryonic stem cells. The fraction obtained for 'expressed' and 'non expressed' polyT CAGEs were compared using Fisher's exact test. The color correspond to p-value $<$ 5e-3 (green) or $\geqslant$ 5e-3 (grey).

Figure 4 Motif discovery around FANTOM CAT transcript and enhancer TSSs. HOMER [21] was used to find 21bp-long motifs in 21bp-long sequences centered around TSSs of FANTOM CAT robust transcripts (A) or enhancers (B). **A.** 525,237 transcript starts as defined in [4] were used. The top 5 is shown. **B.** 65,423 enhancers defined in [3] were used corresponding to 130,846 TSSs. Only 5 motifs were identified by HOMER.

Figure 5 polyT CAGEs interact with gene TSSs A. The number of polyT CAGEs associated with a CAGE signal ('expressed') or not ('non-expressed') in K562 cells located (red) or not (green) in ChIA-PET interacting regions was calculated intersecting their genomic coordinates. Results are shown in the case of CNhs12334 CAGE and RNAP-II ENCODE K562 ChIA-PET replicate 2 data. Other replicates are shown Supplementary Figure S16. A Fisher's exact test was computed to assess the statistical significance of the results (p-value = 1.166e-05). Similar results were observed with other CAGE and /or ChIA-PET replicates. **B.** The coordinates of the regions interacting with 'expressed' or 'non-expressed' polyT CAGEs as defined by ChIA-PET were intersected with that of the genome segments provided by combined chromHMM/Segway in K562. Fisher's exact tests were performed to assess potential enrichments in the indicated segments (**, $<$ 0.001). Similar results were observed with other CAGE and /or ChIA-PET replicates. E, 'Predicted enhancer' ; R, 'Predicted Repressed or Low Activity region' ; T, 'Predicted transcribed region' ; TSS, 'Predicted promoter region including TSS'. **C.** The coordinates of the regions interacting with 'expressed' or 'non-expressed' polyT CAGEs were FANTOM5 CAGE coordinates and the number of CAGEs in each class was calculated. Fisher's exact tests were performed to assess potential enrichments in the indicated classes (**, $<$ 0.001).

Figure 6 Probing sequence-level instructions of transcription at T repeats A. Length of polyT repeat associated (blue) or not (red) with CAGE peaks. For sake of comparison, only T repeats without CAGE peak and located within the same genes as T repeats with

CAGE peaks were considered. **B.** T repeat whose transcription is predicted with low error was chosen and its length was gradually increased from 9 to 51. The prediction was assessed at each step. **C.** Deep CNN models were trained using 50 bp upstream T repeat ends and varying lengths of sequences located downstream, ranging from 0 to 50 bp (x-axis). For each trained model, the Spearman correlation was computed on the test set (30% of entire set of T repeats i.e. 350,770 sequences) between the prediction vector and that of observed tag counts (i.e. accuracy of the model, x axis). **D and E.** 50 sequences with low error and high tag count (set 'high') and 50 sequences with low error and low tag count (set 'low') were chosen. For all pairs of 'high' and 'low' sequences, we sequentially replaced 5-mers from one sequence with 5-mer from the other and predicted the tag count of the 2 new sequences for each 5-mer. Transcription change was assessed as the difference between tag counts before and after 5-mer swapping. D shows the results of insertions of 5-mers of 'high' sequences in 'low' sequences, while E shows the results obtained inserting 5-mers of 'low' sequences in 'high' sequences.

Figure 7 Evaluating the effect of genetic variants on transcription at T repeats A. Increasing the length of the T repeat located in TOMM40 intron 6 as in Figure 6B increases the predicted tagcount. **B.** Tag count distribution of all T repeats (black), T repeats with dbSNP variants (red) and T repeats associated with ClinVar variants (green). textbfC. Frequency of ClinVar variants around T repeat ends (position 0). The y-axis shows the fraction of all ClinVar variants located at the position indicated on the x-axis. Low variant frequency within T repeats is presumably due to mapping issues. **D.** Effect of prediction

changes induced by ClinVar variants (see methods for details). The changes are shown as percentages of the original tag count in a position-wise manner (see 'methods' section).

A

rank	motif	p-value	% in foreground	% in background
1		1e-125974	42.37	12.49
2		1e-99010	5.91	0.06
3		1e-27467	5.39	0.76
4		1e-13220	8.24	3.19
5		1e-10229	10.53	5.21

-10 -5 0 5 10

position relative to CAGE peakmax

B

rank	motif	p-value	% in foreground	% in background
1		1e-19871	53.17	19.13
2		1e-9244	5.20	0.15
3		1e-6244	7.45	0.95
4		1e-405	0.26	0.01
5		1e-70	0.03	0.00

-10 -5 0 5 10

position relative to CAGE peakmax

Figure 1

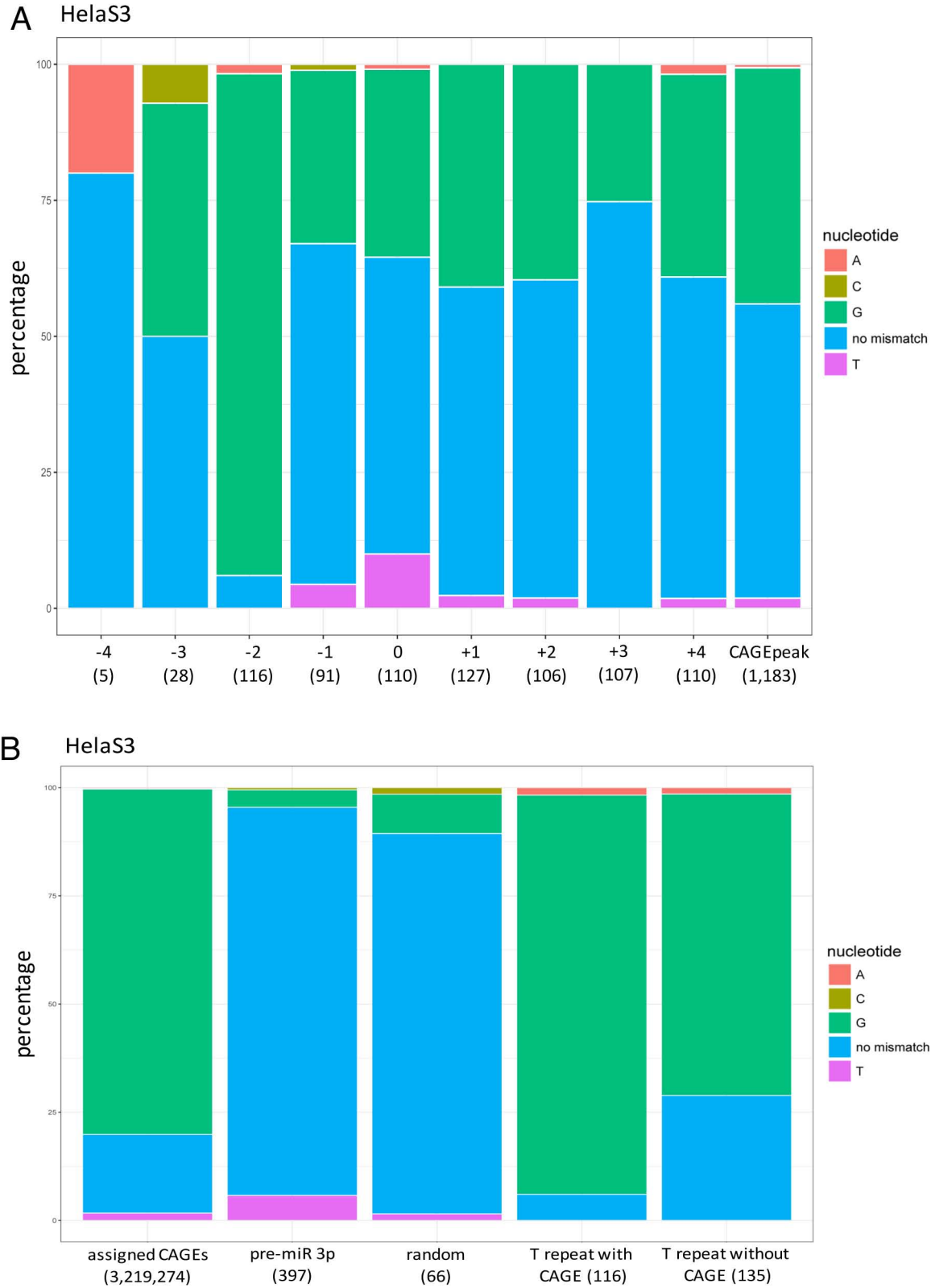

A
HelaS3
percentage
100
75
50
25
0
nucleotide
A
C
G
no mismatch
T
-4 (5)
-3 (28)
-2 (116)
-1 (91)
0 (110)
+1 (127)
+2 (106)
+3 (107)
+4 (110)
CAGEpeak (1,183)
B
HelaS3
percentage
100
75
50
25
0
nucleotide
A
C
G
no mismatch
T
assigned CAGEs (3,219,274)
pre-miR 3p (397)
random (66)
T repeat with CAGE (116)
T repeat without CAGE (135)

Figure 2

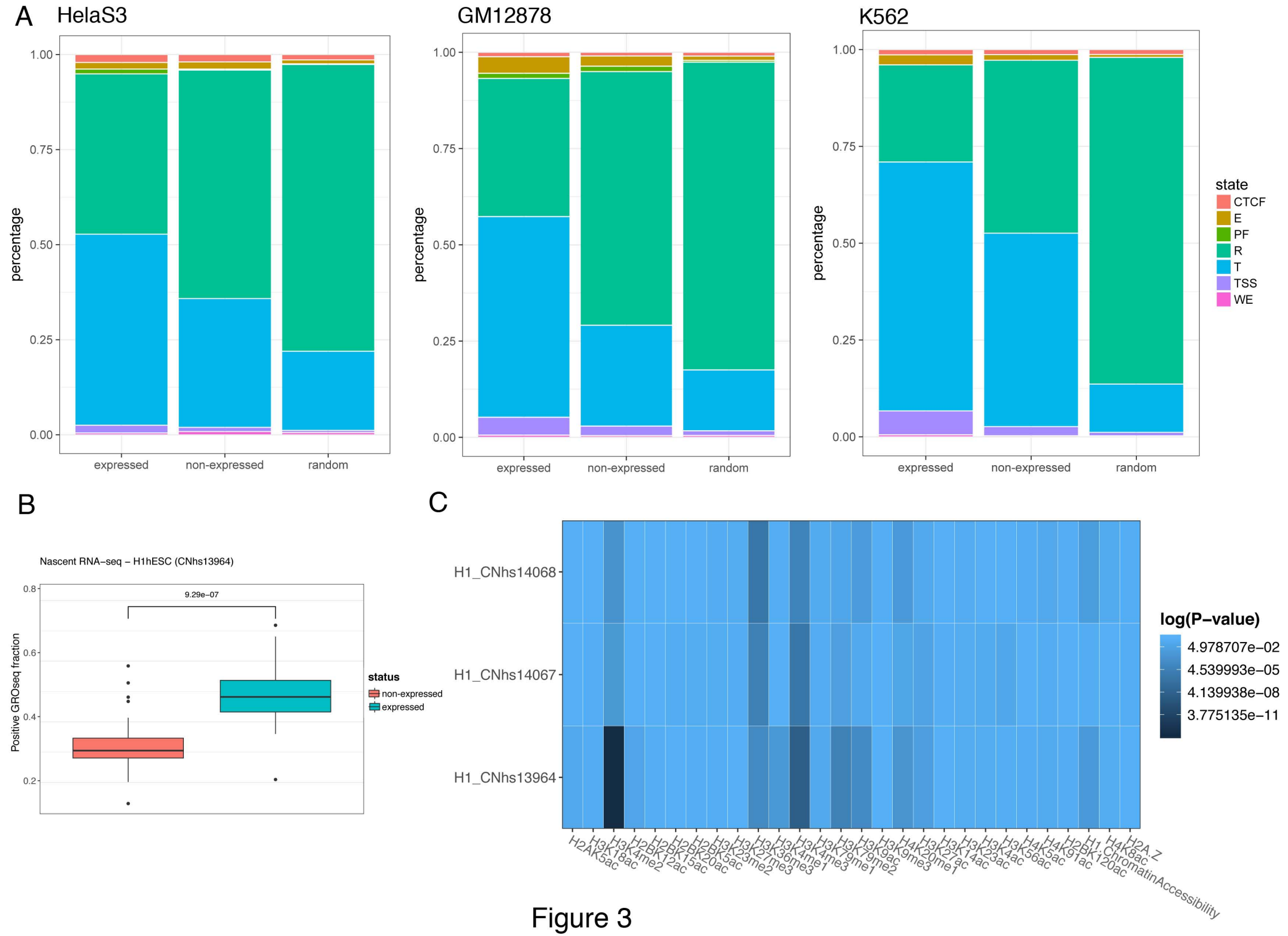
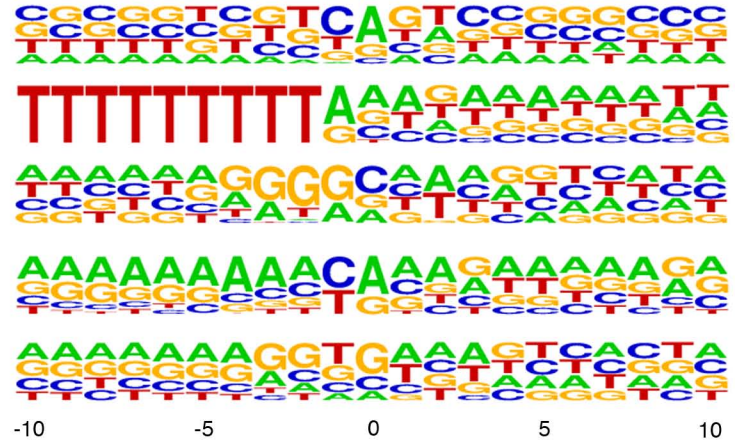

A
HelaS3
GM12878
K562
percentage
1.00
0.75
0.50
0.25
0.00
expressed
non-expressed
random
state
CTCF
E
PF
R
T
TSS
WE
B
Nascent RNA−seq − H1hESC (CNhs13964)
9.29e−07
Positive GROseq fraction
0.8
0.6
0.4
0.2
status
non-expressed
expressed
C
H1_CNhs14068
H1_CNhs14067
H1_CNhs13964
H2AK5ac
H3K18ac
H3K4me2
H2BK12ac
H2BK15ac
H2BK20ac
H2BK5ac
H3K23me2
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K79me1
H3K79me2
H3K9ac
H3K9me3
H4K20me1
H3K27ac
H3K14ac
H3K23ac
H3K4ac
H3K56ac
H4K5ac
H4K91ac
H2BK120ac
H1.ChromatinAccessibility
H4K8ac
H2A.Z
log(P−value)
4.978707e−02
4.539993e−05
4.139938e−08
3.775135e−11

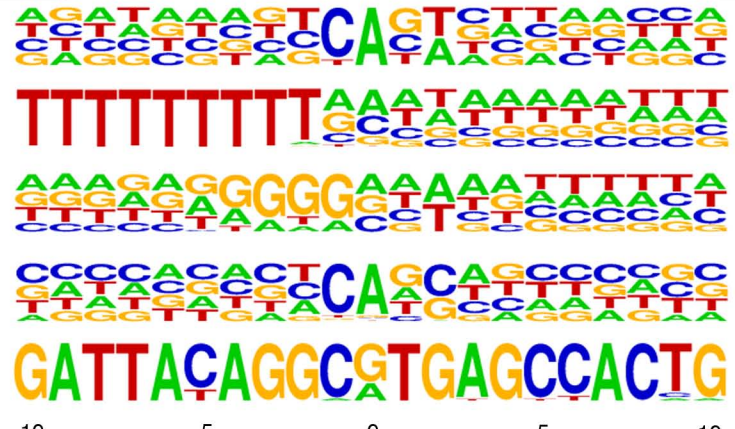
Figure 3

A FANTOM CAT TSSs



rank	motif	p-value	% in foreground	% in background
1		1e-60788	69.05	33.26
2		1e-12440	1.54	0.02
3		1e-10854	5.56	1.08
4		1e-10052	8.90	2.79
5		1e-5756	7.78	3.15

B enhancer TSSs



rank	motif	p-value	% in foreground	% in background
1		1e-9144	22.77	5.71
2		1e-6594	3.96	0.08
3		1e-1358	2.95	0.60
4		1e-399	4.46	2.43
5		1e-19	0.02	0.00

Figure 4

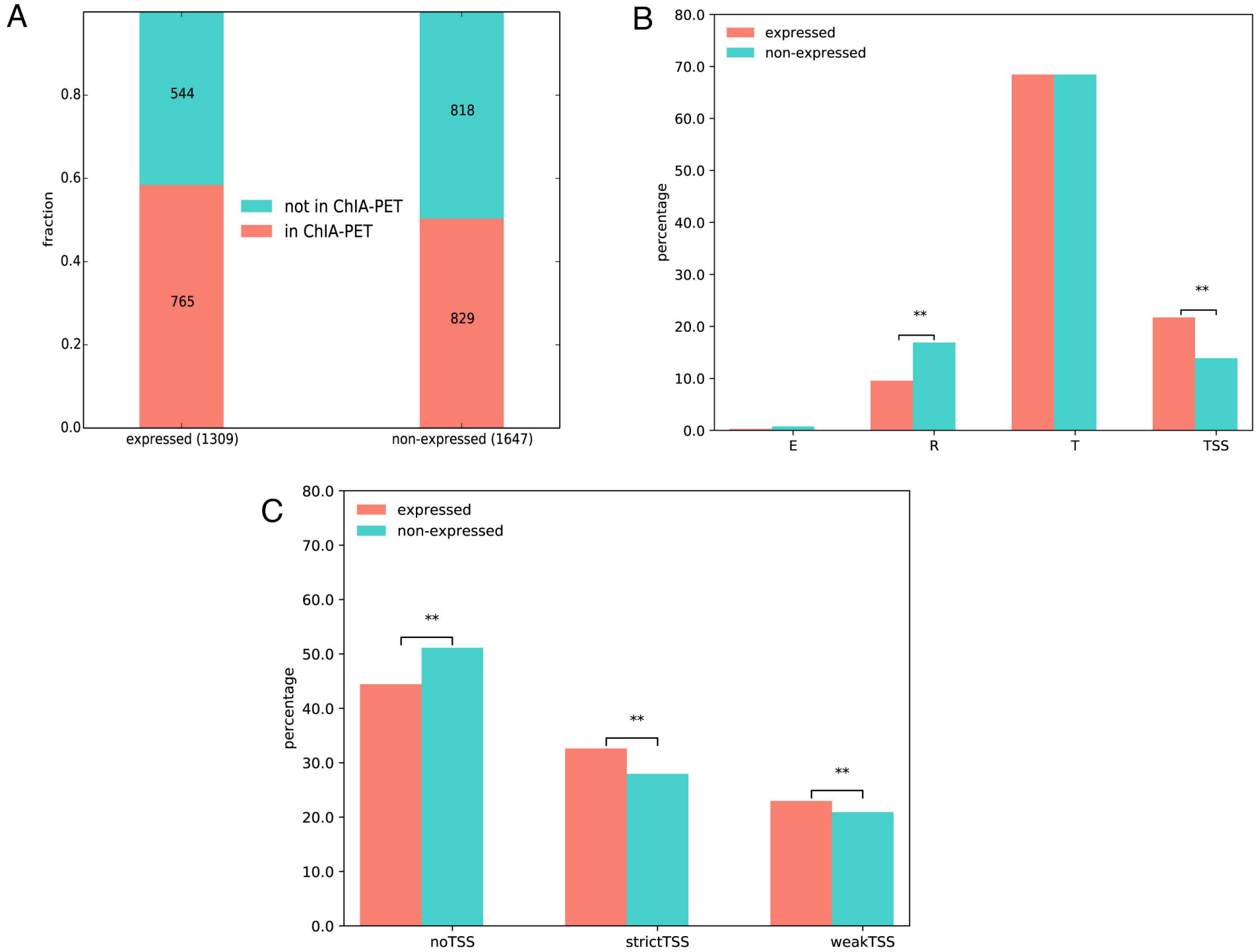

A
fraction
0.0
0.2
0.4
0.6
0.8
544
765
818
829
not in ChIA-PET
in ChIA-PET
expressed (1309)
non-expressed (1647)
B
percentage
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
expressed
non-expressed
**
**
E
R
T
TSS
C
percentage
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
expressed
non-expressed
**
**
**
noTSS
strictTSS
weakTSS

Figure 5

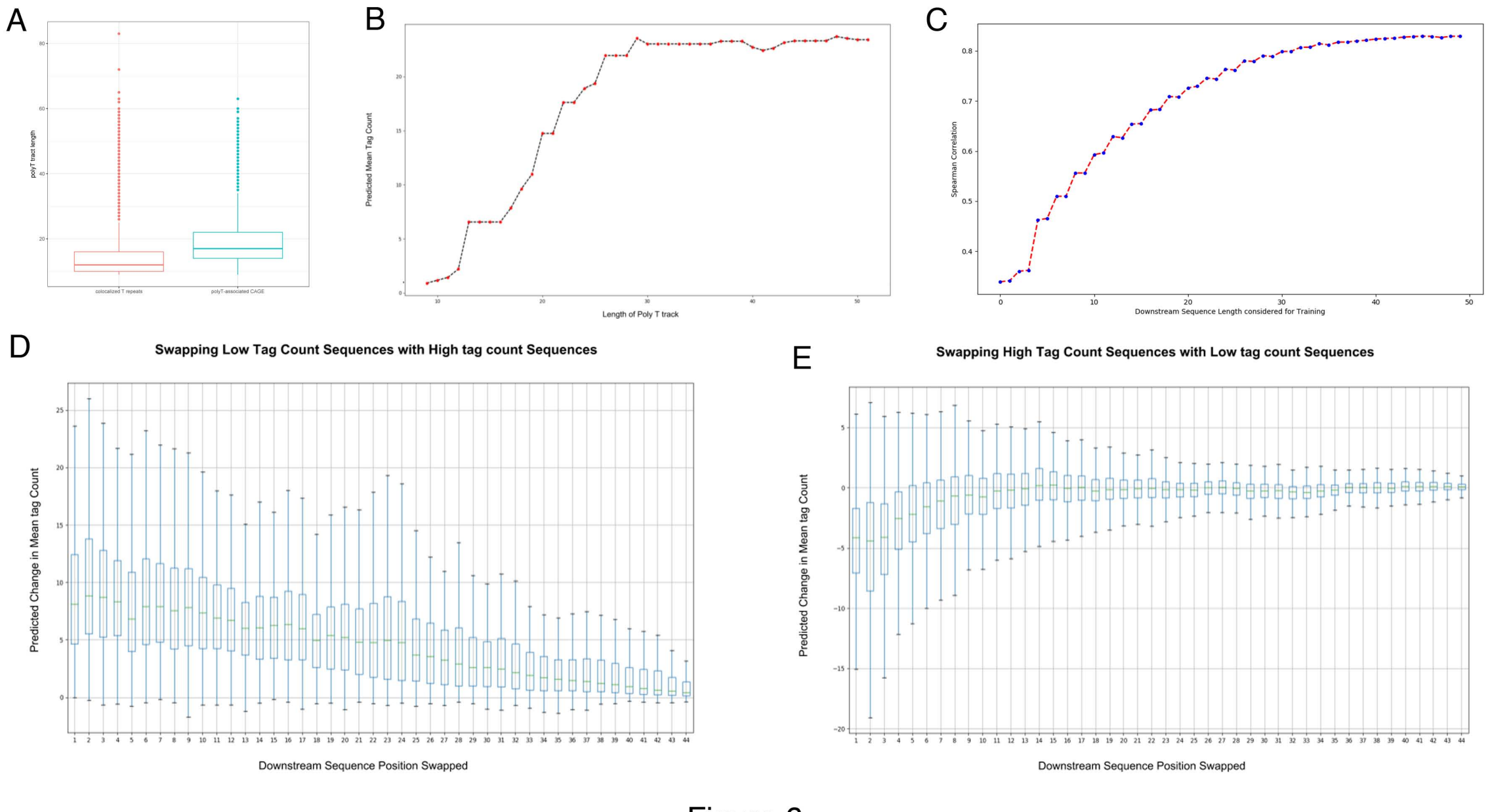

A
polyT tract length
colocalized T repeats
polyT-associated CAGE
B
Predicted Mean Tag Count
Length of Poly T track
C
Spearman Correlation
Downstream Sequence Length considered for Training
D
Swapping Low Tag Count Sequences with High tag count Sequences
Predicted Change in Mean tag Count
Downstream Sequence Position Swapped
E
Swapping High Tag Count Sequences with Low tag count Sequences
Predicted Change in Mean tag Count
Downstream Sequence Position Swapped

Figure 6

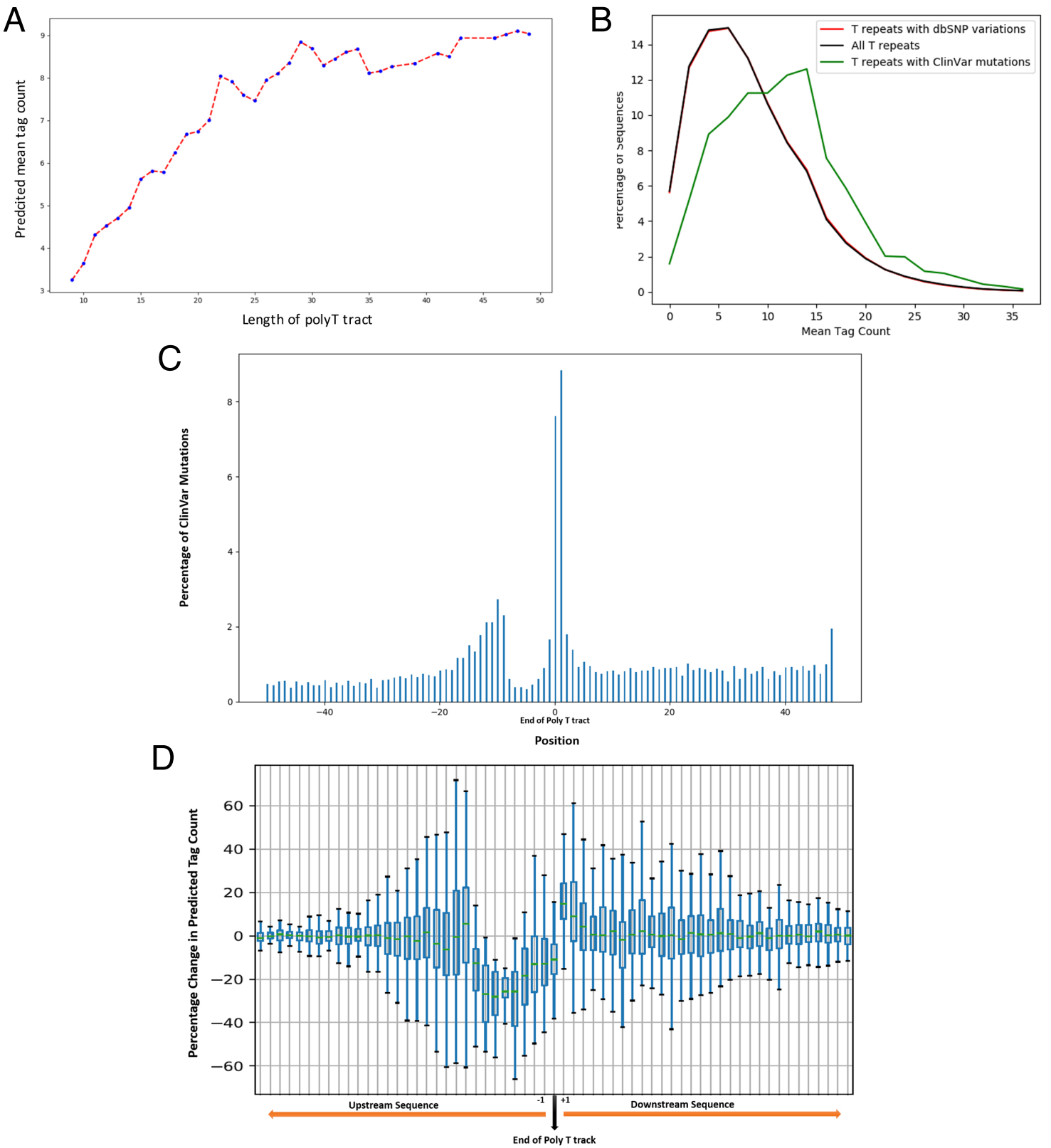

A
Predicted mean tag count
Length of polyT tract
B
T repeats with dbSNP variations
All T repeats
T repeats with ClinVar mutations
Percentage of Sequences
Mean Tag Count
C
Percentage of ClinVar Mutations
End of Poly T tract
Position
D
Percentage Change in Predicted Tag Count
Upstream Sequence
-1
+1
Downstream Sequence
End of Poly T track

Figure 7

chicken

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-2611	-6.014e+03	4.28%	0.02%
2		1e-2455	-5.655e+03	57.76%	29.03%
3		1e-780	-1.797e+03	11.12%	3.45%

dog

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-2967	-6.833e+03	4.97%	0.00%
2		1e-2013	-4.637e+03	66.57%	35.31%
3		1e-646	-1.489e+03	12.50%	3.86%
4		1e-545	-1.256e+03	6.36%	1.23%
5		1e-37	-8.709e+01	0.30%	0.04%

macaque

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-4405	-1.014e+04	6.45%	0.00%
2		1e-3479	-8.012e+03	47.63%	14.39%
3		1e-1632	-3.759e+03	26.63%	8.31%
4		1e-709	-1.633e+03	6.18%	0.98%

rat

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-4009	-9.231e+03	64.62%	26.08%
2		1e-1869	-4.305e+03	21.69%	5.46%
3		1e-1442	-3.321e+03	2.22%	0.01%
4		1e-50	-1.167e+02	0.39%	0.06%

Figure S1: **Motif discovery around CAGE summit.** The procedure described in Figure 1 was repeated with CAGE peak summits collected in chicken, dog, macaque and rat.

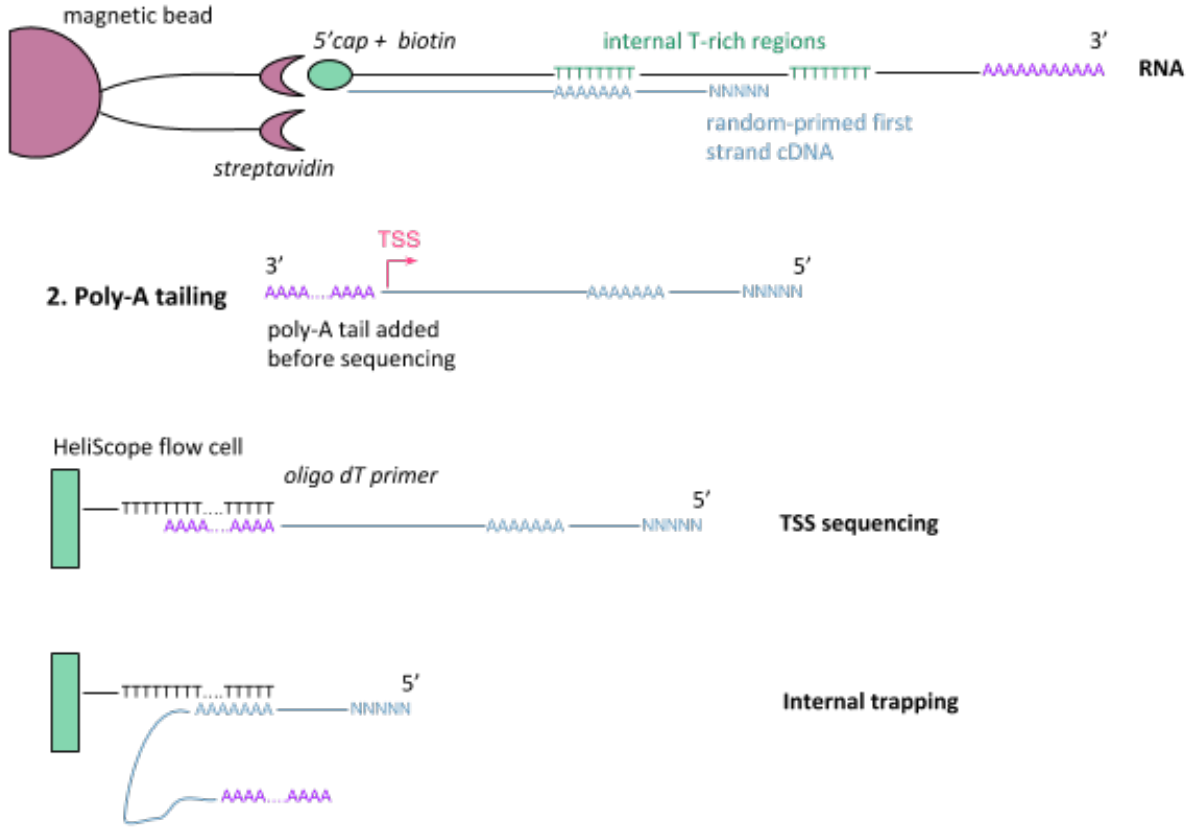


Figure S2: **Internal priming at internal polyT tracts.** In the HeliScopeCAGE protocol, capped RNA will be randomly primed and the RNA/DNA duplex will be captured on beads by the CAP Trapper reaction. The first-strand cDNAs are then released and a poly-A tail is added to prepare them for sequencing. The HeliScope platform primes the sequencing reactions with poly-dT oligonucleotides grafted to the surface of its flow cells. In the HeliScope CAGE, it is intended that the sequencing reactions will start after the poly-A tail added to the first-strand cDNA, thus producing reads that align at the TSS. If the first-strand cDNA contains internal A-rich regions, the priming can also happen internally and yield CAGE peaks at internal poly-T tracts.

A. STARR-seq (GSE63151)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-83123	-1.914e+05	26.67%	7.22%
2		1e-33	-7.804e+01	0.02%	0.01%

B. DECAP-seq (GSM2422532 and GSM2422533)

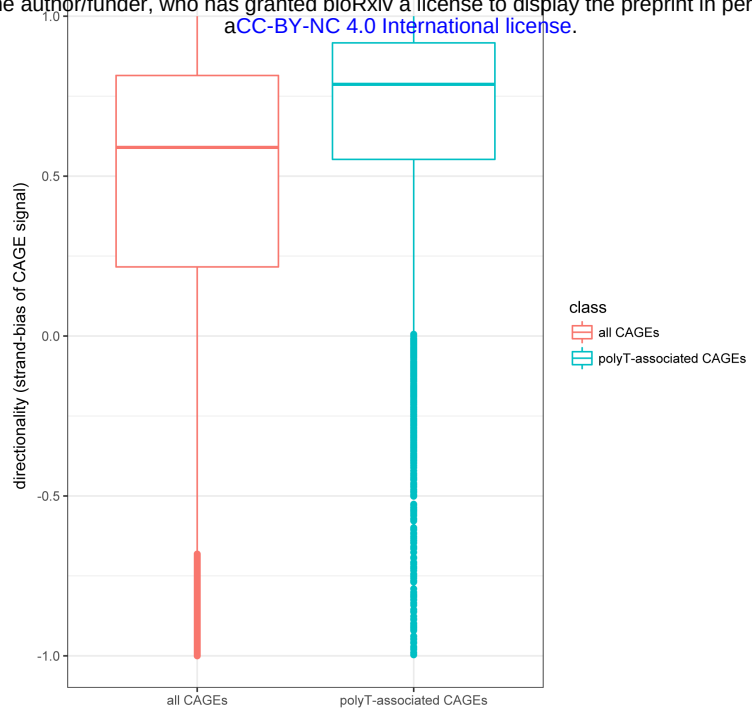
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-117	-2.710e+02	5.32%	3.35%
2		1e-28	-6.492e+01	2.07%	1.46%
3		1e-16	-3.690e+01	0.04%	0.01%
4 *		1e-3	-7.261e+00	0.01%	0.00%

Figure S3: **Motif discovery around Start-seq and DECAP-seq peak summits. A.** The procedure described in Figure 1 was repeated on TSSs detected by Start-seq [1] (forward data, n = 1,086,787) normalizing by total GC-content with -gc option. **B.** Same procedure applied on DECAP-seq TSSs [2] (n = 106,742). GSM2422532 and GSM2422533, which are not stranded data, were merged. Note that the INR motif is not detected with this procedure and that motif 4 (poly tract) is labelled as possible false positive.

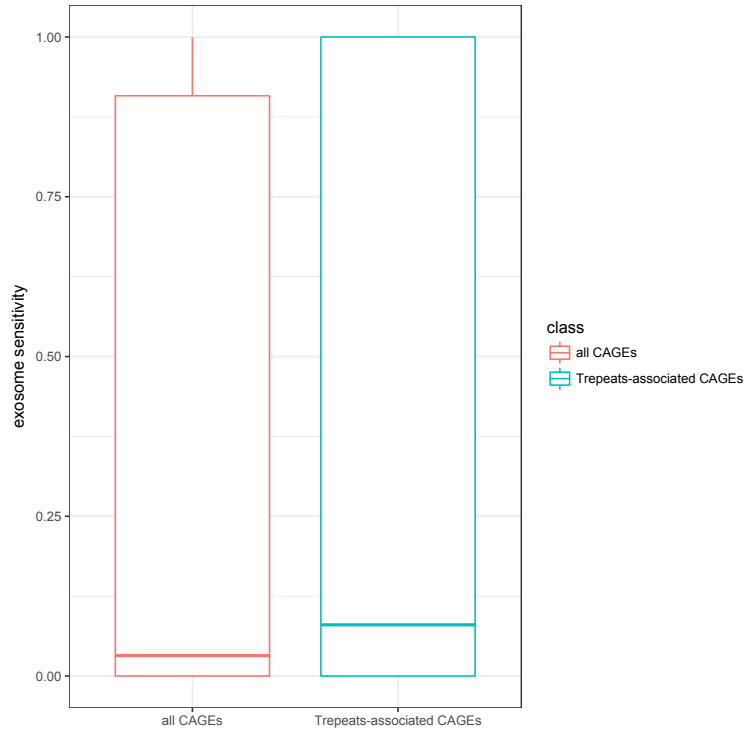
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-9984	-2.299e+04	33.63%	8.84%
2		1e-3125	-7.196e+03	10.37%	2.42%
3		1e-2414	-5.559e+03	12.34%	4.08%
4		1e-47	-1.104e+02	0.74%	0.41%
5 *		1e-9	-2.092e+01	0.02%	0.00%

Figure S4: **Motif discovery around TSS-seq peak summits collected in Arabidopsis thaliana.** The procedure described in Figure 1 was repeated on TSS-seq peaks [3] normalizing by total GC-content with -gc option.

Figure S5: **Distribution of CAGE peaks at homopolymers of A/T A.** 81,080 STRs from STR reference catalog [4] are associated with a CAGE peak (considering a window of 5bp around the STR) : 36,062 correspond to homopolymers of Ts and 33,802 to homopolymers of As. Most CAGE peaks associated with STRs of T are on the '+' strand (35,653 out of 36,185), while CAGE peaks associated with homopolymers of As are on the '-' strand (33,313 out of 33,922) **B.** CAGE peaks are not equally distributed around polyA/T STRs with 35,974 and 686 located respectively downstream and upstream of polyT STRs (CAGE tags within STRs are here counted in both cases explaining why 35,974+686 > 36062) and 679 and 33,723 located respectively downstream and upstream of polyA STRs. Hence, CAGE peaks are almost exclusively located downstream polyT STRs. Since no amplification is involved in CAGE sequencing [5], imbalance for polyT vs. polyA supports the idea of internal priming (note that FANTOM CAT genes contain 628,762 T repeats and 499,419 A repeats).



A



B

Figure S6: **polyT CAGEs harbor specific features. A.** Directionality, extracted from [6], of all CAGEs (red) and that of 10,926 polyT CAGEs (blue) are shown as boxplots. Directionality close to 1 indicates that sense transcription is more abundant than antisense transcription within a region of -800bp/+200bp centered around the CAGE summit while directionality of -1 indicates that antisense transcription is more abundant. These results argue against the presence of divergent (i.e. upstream antisense) RNAP-II transcription associated with polyT CAGEs , as widely observed for canonical TSSs [7] **B.** Exosome sensivity score were extracted from [6]. Exosome sensitivity of a CAGE cluster is measured as the relative fraction of CAGE signal observed after exosome knockdown in HeLa-S3 cells as previously described in [8]. The exosome sensitivity of all CAGEs (red) and that of 10,926 polyT CAGEs are shown as boxplots.

gene class	CAGE-associated T repeats (63,480)	all T repeats (699,048)
coding mRNA	51,774 (81.55%)	486,730 (69.62%)
lncRNA_antisense	604 (0.95%)	13,407 (1.92%)
lncRNA_divergent	1,921 (3.02%)	40,295 (5.76%)
lncRNA_intergenic	5,704 (8.98%)	117,056 (16.74%)
lncRNA_sense_intronic	1,744 (2.74%)	11,067 (1.58%)
pseudogene	808 (1.27%)	16991(2.43%)
sense_overlap_RNA	386 (0.61%)	1,940 (0.28%)
short_ncRNA	18 (0.03%)	241 (0.03%)
small_RNA	169 (0.27%)	1,838 (0.26%)
structural_RNA	4 (0.006%)	121 (0.02%)
uncertain_coding	347 (0.55%)	9,362 (1.34%)

Table S1: **CAGE-associated T-repeats are preferentially located in coding genes.** The coordinates of CAGE-associated T repeats (n = 63,974) and that of all repeats of more than 9 Ts (n = 1,337,561) were intersected with the annotation provided by the FANTOM5 CAGE Associated Transcriptome (CAT). The total number of intersections obtained with CAGE-associated T repeats and T repeatss is 63,480 and 699,048 respectively. The location of all T repeats are shown for comparison. For each gene class, the number of intersections and the corresponding percentage is shown. The median number of CAGE-associated T repeats per gene is 2. In contrast, $\sim$ 47% of all T repeats are located in FANTOM CAT genes with a median number of T repeats per gene of 7. 45,058 out of 63,974 CAGE-associated T repeats are located in introns (> 70%), while only 34% of all T repeats are intronic (453,428 out of 1,337,561, Fisher's exact test p-value < 2.2e-16). Likewise, in mouse, 6,500 out of 8,825 CAGE-associated T repeats ($\sim$ 74%) but 204,328 out of 834,954 T repeats ($\sim$ 24%) are located in introns.

gene_class	CAGE-associated T repeats (49,312)	T repeats (477,492)
3prime_overlapping_ncrna	7 (0.014%)	65 (0.014%)
TR_C_gene	1 (0.002%)	9 (0.002%)
TR_V_gene	1 (0.002%)	6 (0.001%%)
antisense	1,270 (2.57%)	26,135 (5.48%)
lincRNA	2,059 (4.17%)	43,120 (9.03%)
miRNA	2 (0.004%)	27 (0.005%)
polymorphic_pseudogene	21 (0.04%)	508 (0.11%)
processed_transcript	592 (1.20%)	6,986 (1.46%)
protein_coding	44,682 (90.61%)	386,036 (80.85%)
pseudogene	379 (0.77%)	11,801 (2.47%)
sense_intronic	168 (0.34%)	1,225 (0.26%)
sense_overlapping	130 (0.26%)	1,503 (0.31%)
snRNA	0 (0%)	7 (0.001%)
snoRNA	0 (0%)	6 (0.001%)

Table S2: **CAGE-associated T-repeats are preferentially located in coding genes.** The coordinates of polyT CAGEs (n = 63,947) and that of all repeats of more than 9 Ts (n = 1,337,561) were intersected with the annotation provided by GENCODEv19. The total number of intersections polyT CAGEs and polyT tracts is 49,312 and 477,492 respectively. The location of all T repeats are shown for comparison. For each gene class, the number of intersections and the corresponding percentage is shown.

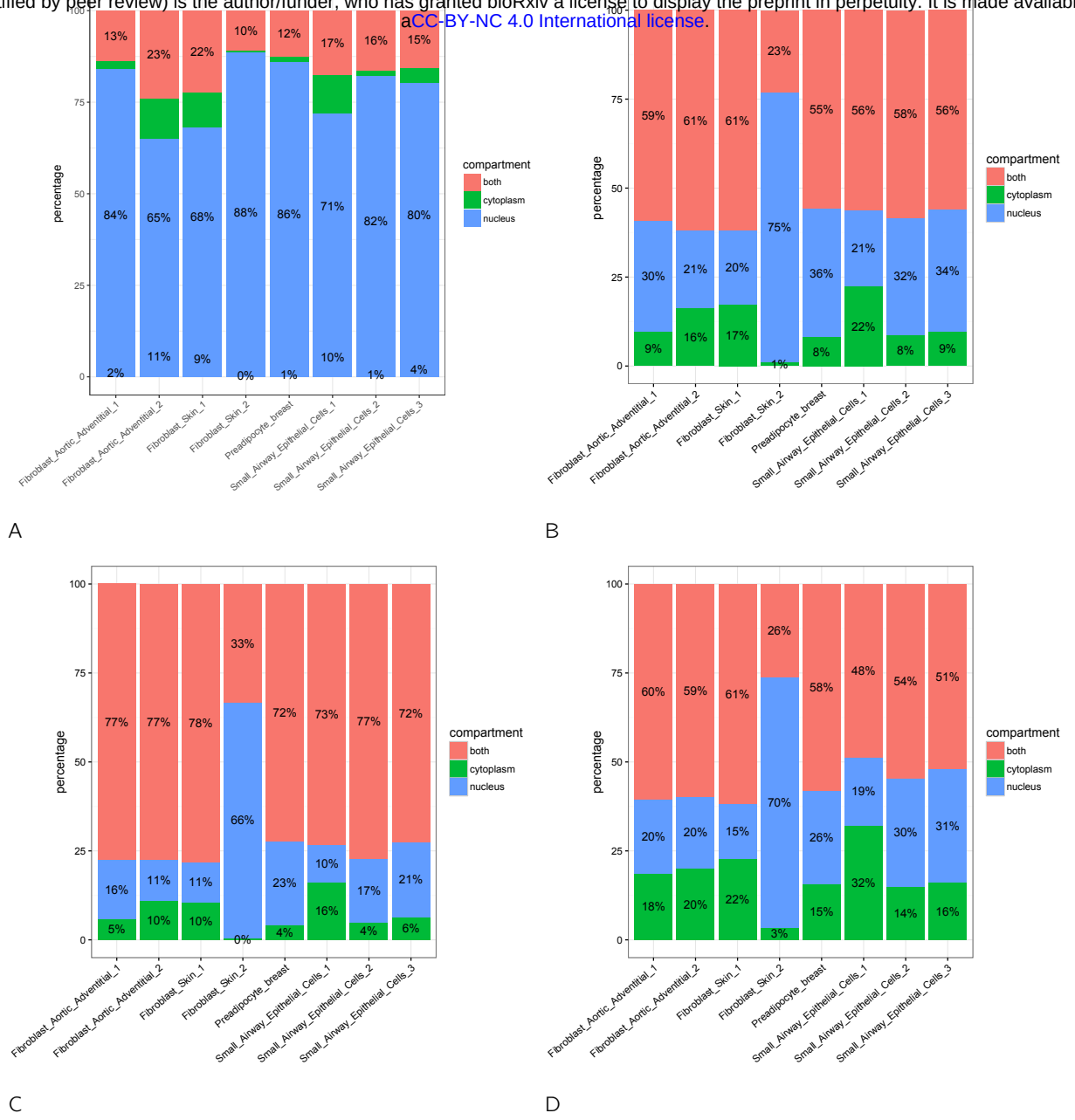


Figure S8: **polyT CAGEs are preferentially detected in the nuclear compartment.** For each indicated library, CAGE expression (RLE normalized) was measured in nuclear and cytoplasmic fractions. Each CAGE was then assigned to nucleus, cytoplasm or both compartments. The number of CAGEs in each class is shown for each sample as a fraction of all detected CAGEs. The sample *Fibroblast_Skin_2* likely represents an outlier. **A.** 63,974 CAGEs associated with T repeats **B.** 1,048,124 CAGEs detected **C.** 130,286 CAGEs assigned to PCGs **D.** 68,731 CAGEs harboring the third motif shown in Figure 1.

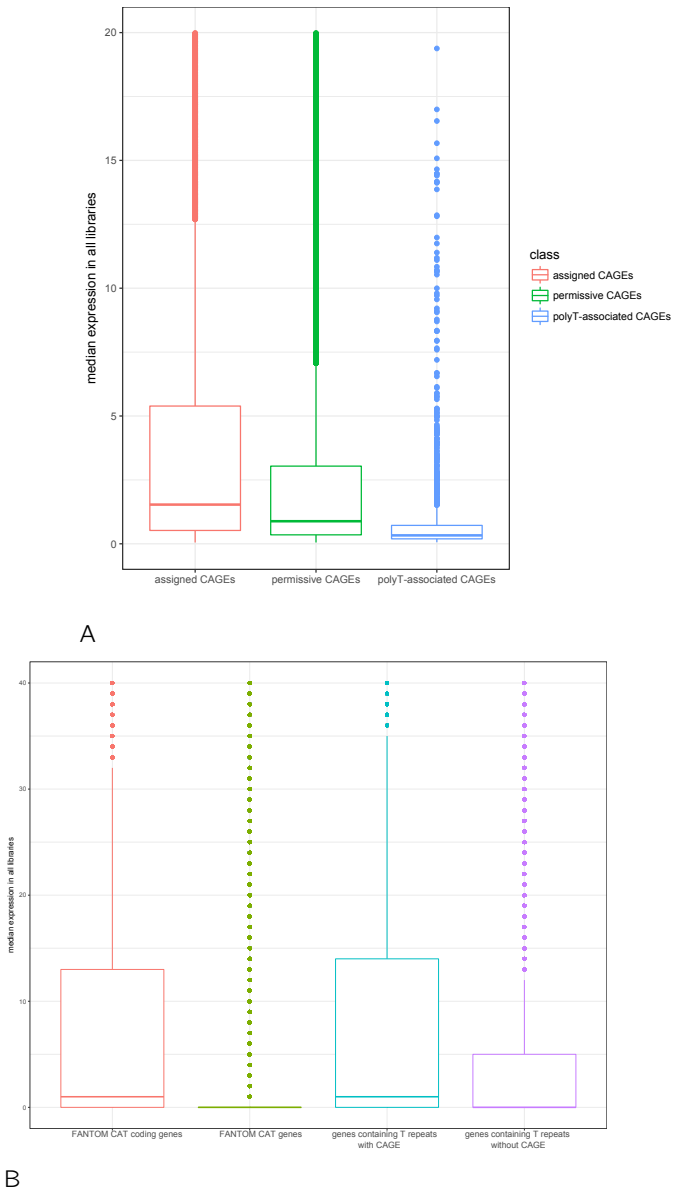


Figure S7: **polyT CAGEs are faintly expressed and detected in highly expressed genes. A.** Expression medians were computed for each permissive CAGE, CAGE assigned to genes and polyT CAGEs across 1,829 libraries. The Phase 1 and 2 combined Tag Per Million (TPM) normalized data were used (Supplementary Table S7). The y-axis was limited to 20. Overall polyT CAGEs appear faintly expressed (median of median CAGE TPM expression in all samples = 0.3347) compared to all CAGEs detected (median = 1.117, Wilcoxon test p-value < 2.2e-16) or to CAGEs assigned to gene TSS (median = 2.396, Wilcoxon test < 2.2e-16) **B.** polyT CAGEs are preferentially detected in highly expressed genes. Distribution of median expression was computed across 1,829 FANTOM5 libraries of genes containing T repeats with (blue) or without (purple) CAGE (Wilcoxon test p-value < 2.2e-16). The median expressions of all FANTOM CAT genes (green) and that of PCGs (red) are shown for sake of comparison. The y-axis was limited to 40.

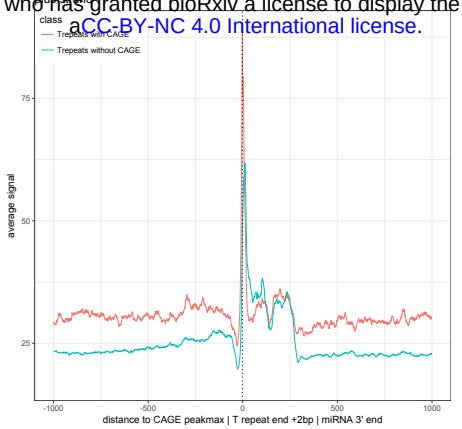
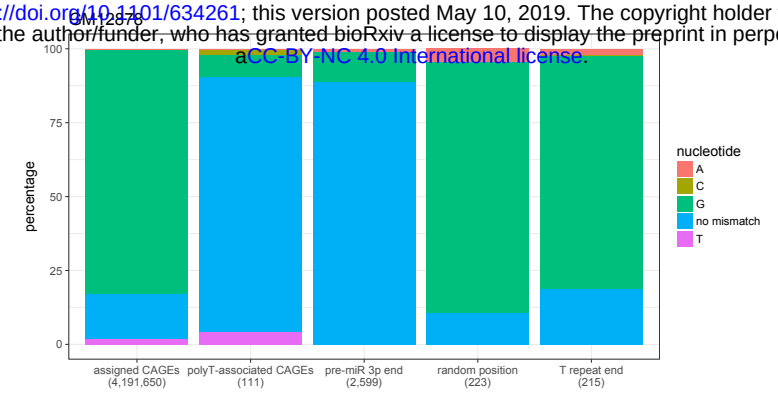
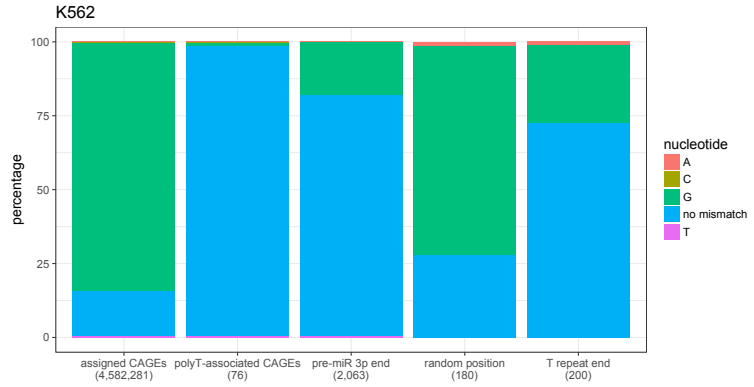


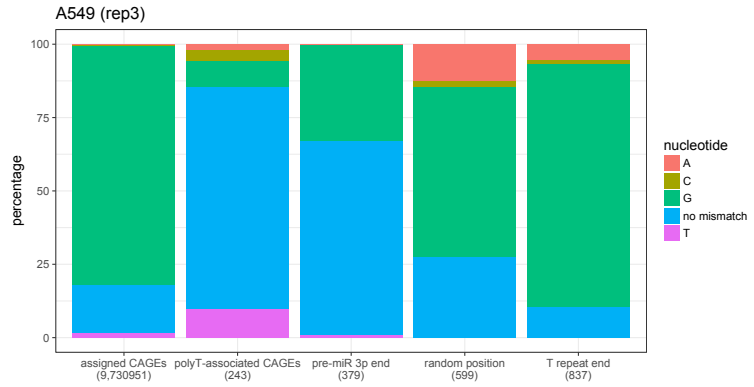
Figure S9: **polyT CAGEs are associated with chromatin.** Distribution of chromatin-associated RNAs extracted from K562 cells around T repeats with (green) or without (blue). Total signal on + strand is shown.



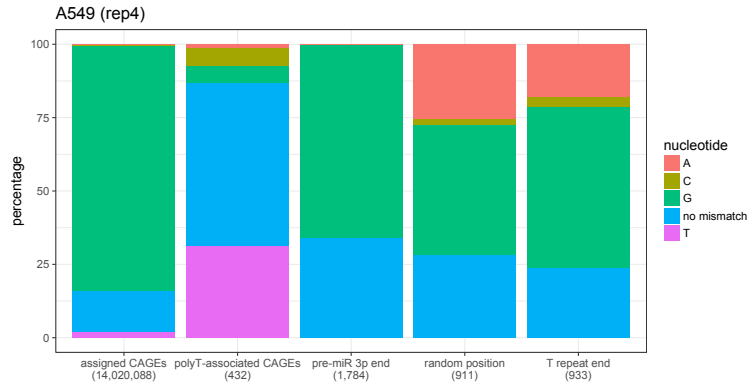
A



B

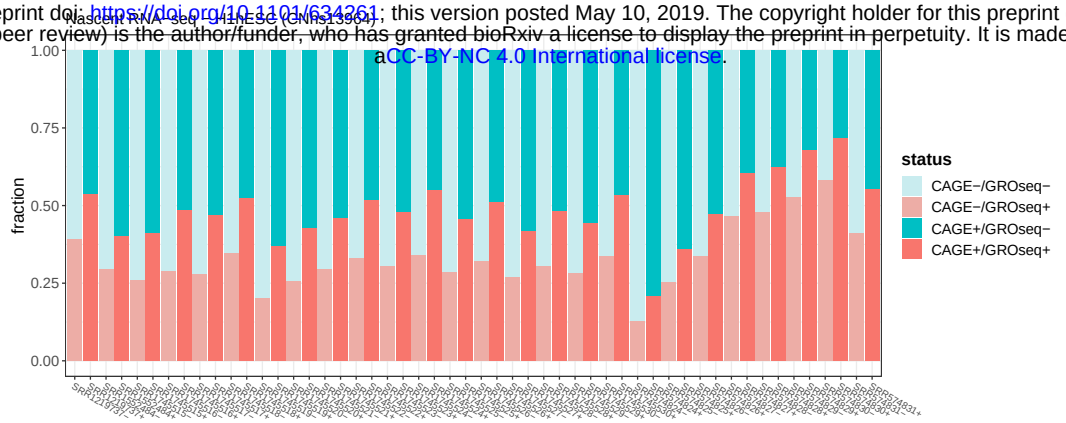


C

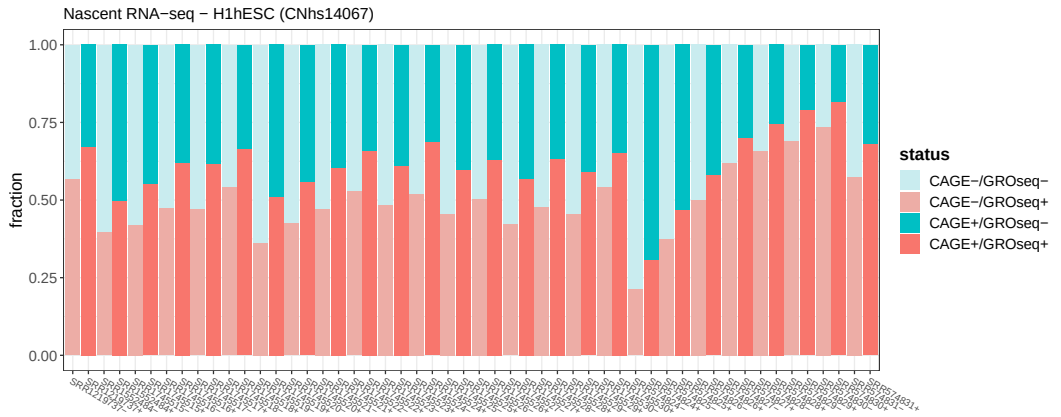


D

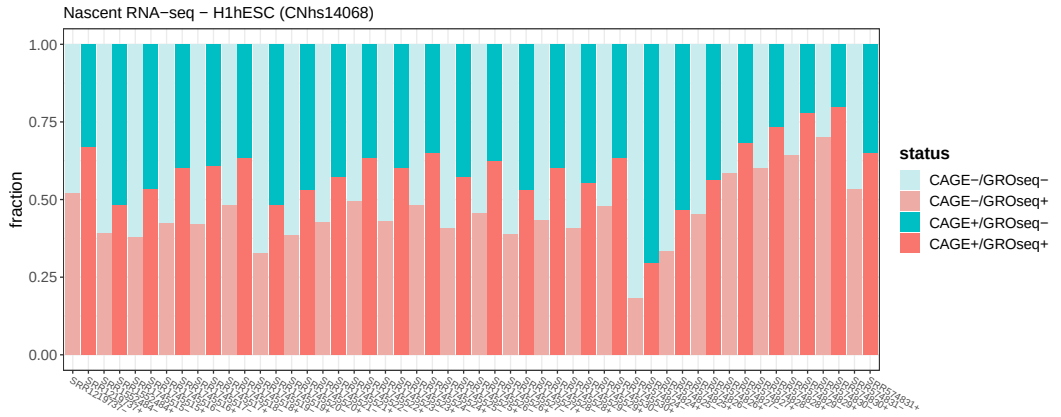
Figure S10: **G bias around polyT CAGEs .** G bias in ENCODE CAGE reads was assessed at -2 of CAGE peakmax as in Figure 3 in GM12878 (A), K562 (B) and A549 (C and D).



A

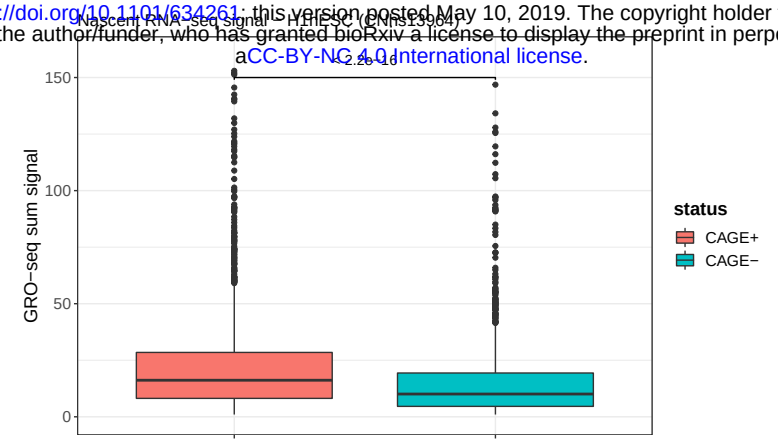


B

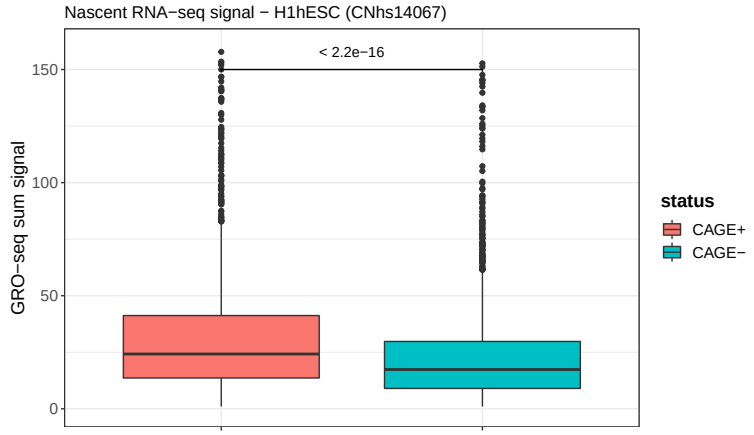


C

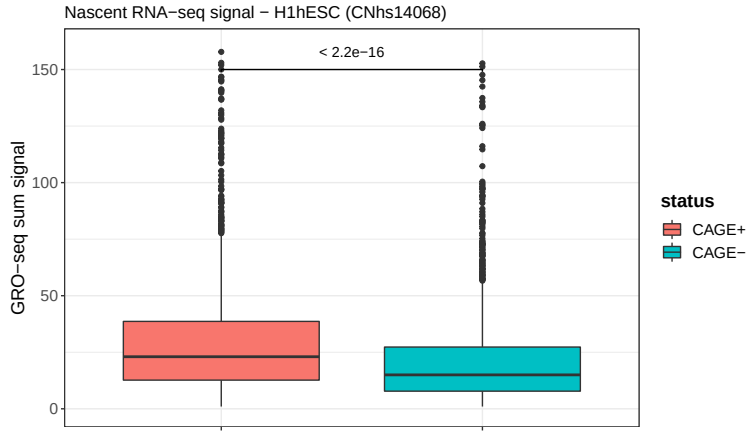
Figure S11: **Intersection with GRO-seq peaks.** Intersection of 'expressed' (CAGE+)/'non expressed' (CAGE-) polyT CAGEs (end ± 50bp) with nascent RNAseq peaks from 3 'gro-seq colorado' samples (SRR*) obtained in H1 embryonic stem cells (CNhs13964, CNhs14067 and CNhs14068 FANTOM libraries). Fisher's exact tests indicate that the fraction of polyT CAGEs with no-null GROseq signal obtained for 'expressed' polyT CAGEs is invariably higher than that obtained for 'non expressed' polyT CAGEs (p-value < 2.2e-16).



A



B



C

Figure S12: **Intersection with GRO-seq signal.** Intersection of 'expressed' (CAGE+)/'non expressed' (CAGE-) polyT CAGEs (end ± 50bp) with nascent RNAseq signal from 3 'gro-seq colorado' samples (SRR*) obtained in H1 embryonic stem cells (CNhs13964, CNhs14067 and CNhs14068 FANTOM libraries). The sum of the mean signal for each of the 'expressed' (CAGE+) and 'non expressed' (CAGE-) polyT CAGEs is represented on the y-axis. For instance, in CNhs13964, signal was computed for 2,013 'expressed' polyT CAGEs and 2,001 'non expressed'. Wilcoxon test was used to asses significant difference (p-value is indicated).

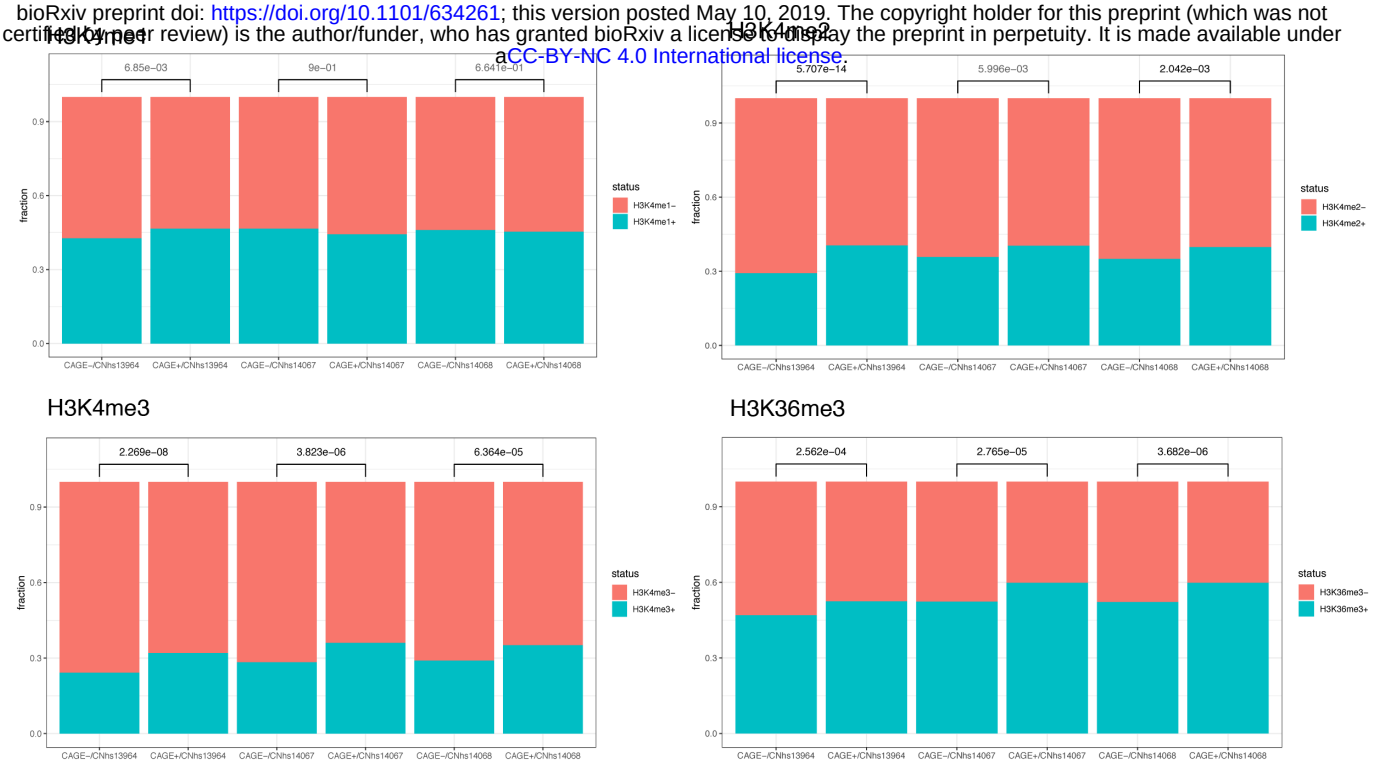


Figure S13: **Intersection with Roadmap epigenetics ChIP-seq data.** Fraction of 'expressed' / 'non expressed' polyT CAGEs in H1-hESC cells intersecting with H3K4me1 (top), H3K4me2 (middle), H3K4me3 (middle) and H3K36me3 (bottom) Roadmap ChIP-seq data.

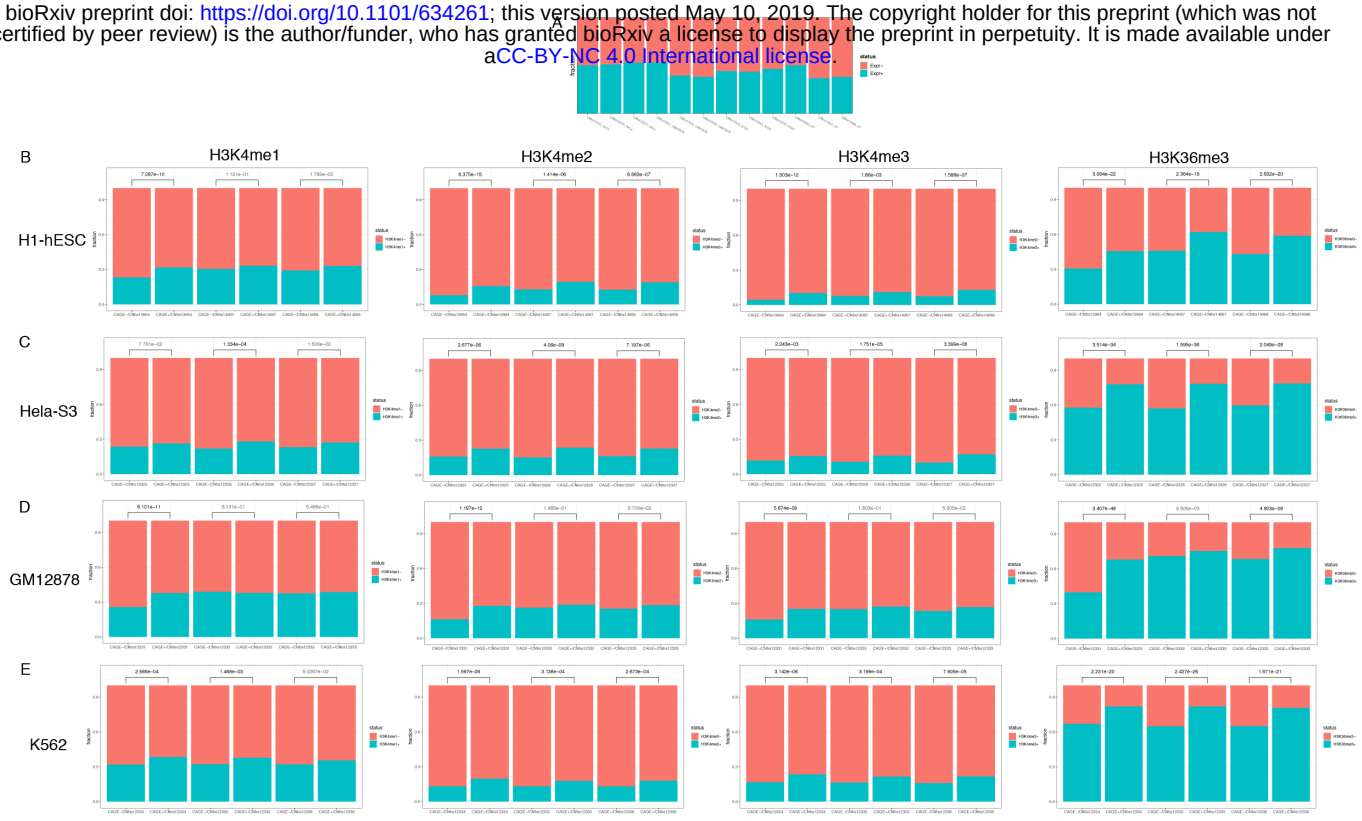
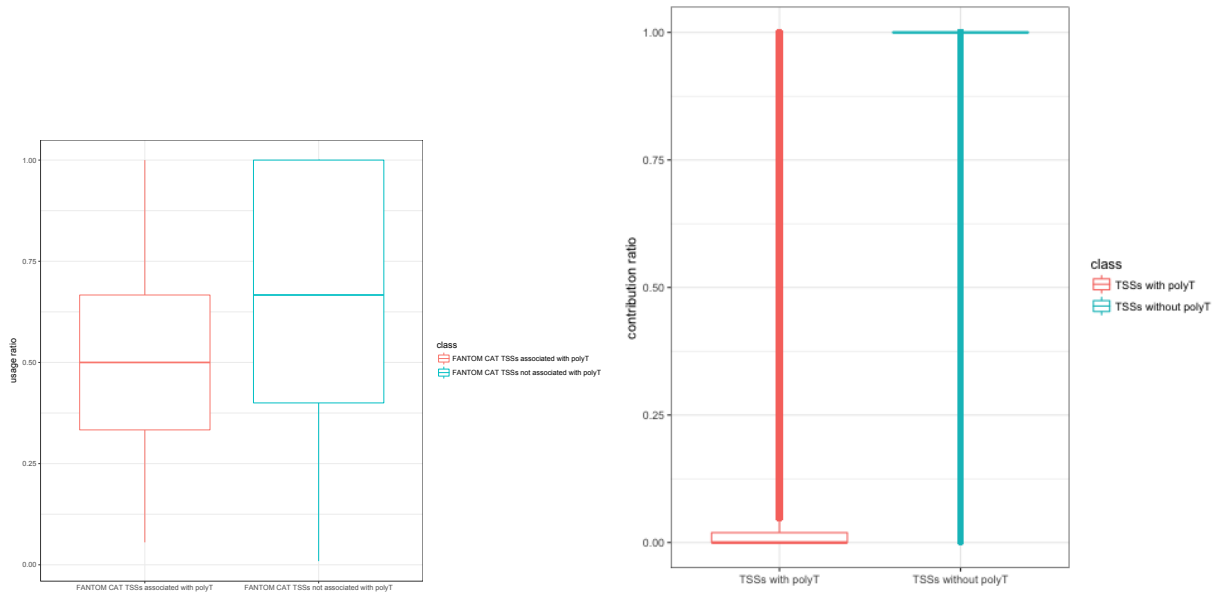


Figure S14: **Intersection with ENCODE epigenetics ChIP-seq data. A.** Fraction of 'expressed'/'non expressed' polyT CAGEs within each library considered. **B-E.** Fraction of 'expressed' / 'non expressed' polyT CAGEs intersecting with H3K4me1, H3K4me3 and H3K36me3 ENCODE ChIP-seq data in H1-hESC (B), Hela-S3 (C), GM12878 (D) and K562 (E). The 3 replicates of each cell line were considered in each case.



A B

Figure S15: **Usage and contribution of polyT-associated FANTOM CAT TSSs.** We first assessed the position of polyT-associated TSSs among each promoter: 937 TSSs out of 10,606 correspond to the most upstream TSS, arguing against internal priming and artifact hypothesis in these cases. **A.** We evaluate the usage of FANTOM CAT TSSs associated (red) or not (blue) with poly tract for each gene with more than 1 TSS. For each assigned TSS, we counted the number of samples wherein the considered TSS is expressed. We then sorted all TSSs for each gene according to this usage value and divided the rank of each polyT-associated TSS by the length of the list of all TSS assigned to each gene (usage ratio indicated in y-axis). TSSs associated with polyT repeats are slightly less used than TSSs not associated with polyT repeat (Wilcoxon test p-value < 2.2e-16). However, the median usage ratio for polyT-associated TSSs is 0.5, indicating that these TSSs are far from being the least used TSSs. **B.** We further evaluated the contribution of polyT-associated TSSs in gene expression. For each gene in each library, we divided the sum tag count (CPM RLE normalized) of all TSS associated (red) or not (blue) with polyT tract by the sum tag counts of all TSSs assigned to the gene considered (contribution ratio indicated in y-axis). As expected provided faint expression of polyT CAGEs , polyT-associated FANTOM CAT TSSs contribute only poorly to gene expression with a median contribution ratio of 0 (the median contribution of TSSs not associated with polyT repeat logically equals 1).

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-8380	-1.930e+04	24.52%	4.98%
2		1e-3444	-7.930e+03	2.51%	0.03%
3		1e-1305	-3.007e+03	4.43%	0.96%
4		1e-56	-1.303e+02	0.08%	0.01%
5		1e-32	-7.539e+01	0.07%	0.01%

Figure S17: **Motif discovery around FANTOM mouse enhancer TSSs.** HOMER [9] was used to find 21bp-long motifs in 21bp-long sequences centered around TSSs of 44,459 mouse enhancers defined in [10] and corresponding to 88,918 TSSs. Only the top 5 motifs are shown.

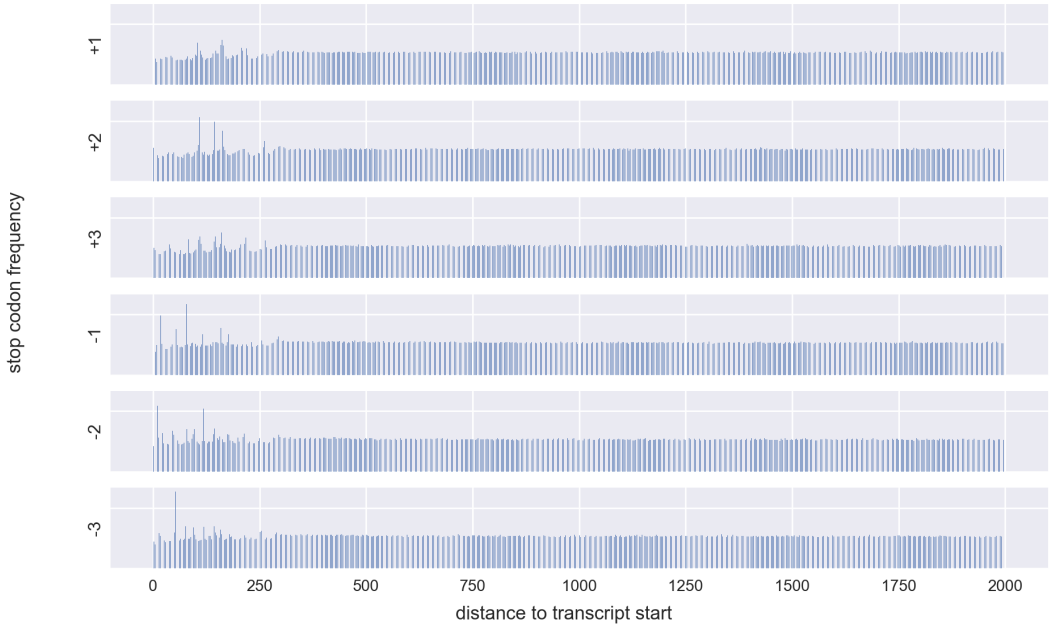
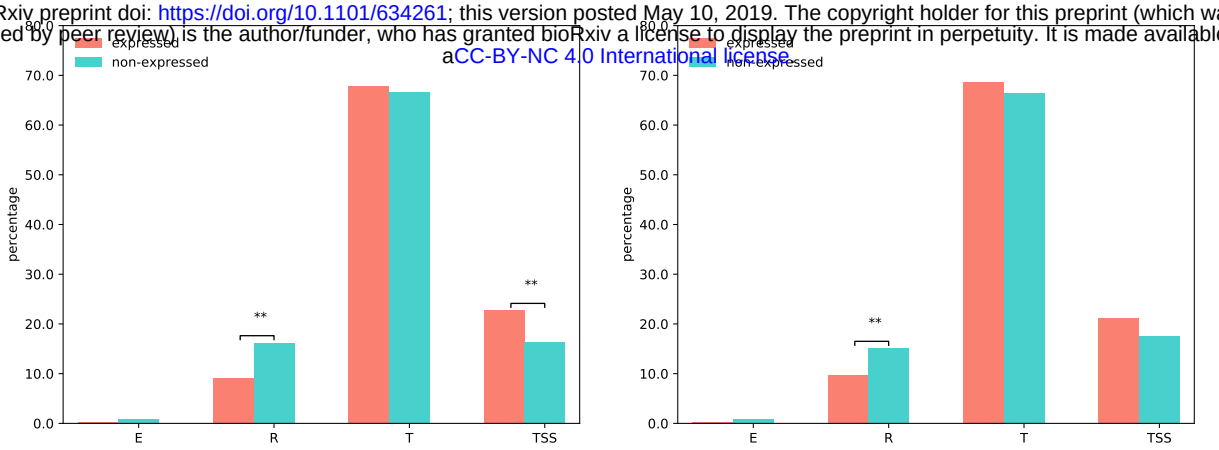
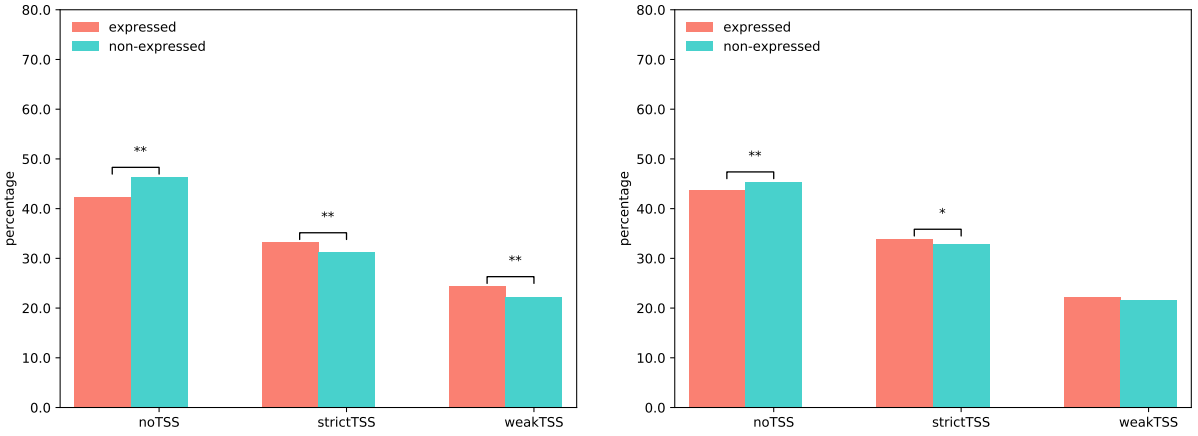


Figure S16: **No specific Open Reading Frame could be detected 2kb downstream polyT CAGEs.** Stop codon distribution was assessed along 2,000 nt downstream polyT-associated CAGE summits (n = 63,974). The x-axis represents the distance to transcript start and the y-axis represents the frequency of the three stop codons TAA, TAG and TGA. All possible 6 frames were considered (see y-axis labels).



A B

Figure S18: **Regions interacting with polyT CAGEs** . The coordinates of the regions interacting with 'expressed' or 'non-expressed' polyT CAGEs in CNhs12335 (A) and CNhs12336 (B) were intersected with that of the genome segments provided by combined chromHMM/Segway in K562. Fisher's exact tests were performed to assess potential enrichments in the indicated segments (**, < 0.001). E, 'Predicted enhancer' ; R, 'Predicted Repressed or Low Activity region' ; T, 'Predicted transcribed region' ; TSS, 'Predicted promoter region including TSS'.



A B

Figure S19: **TSSs interacting with polyT CAGEs** . The coordinates of the regions interacting with 'expressed' or 'non-expressed' polyT CAGEs in CNhs12335 (A) and CNhs12336 (B) were FANTOM5 CAGE coordinates and the number of CAGEs in each class was calculated. Fisher's exact tests were performed to assess potential enrichments in the indicated classes (**, < 0.001 ; *, < 0.05)

Figure S20: **Example of signal used to train CNN.** To compute the raw mean tag count of each T repeat, we first calculated the mean tag count of each base genome-wide in 988 FANTOM libraries (upper track). We then summed these values along the T repeat + 5bp (bottom track).

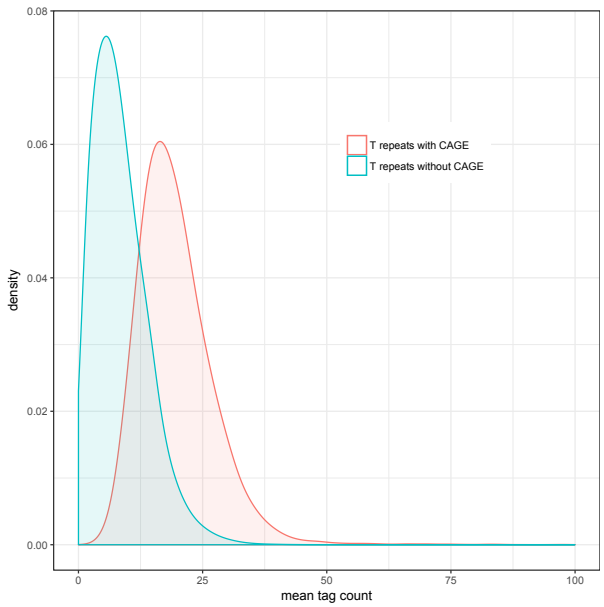
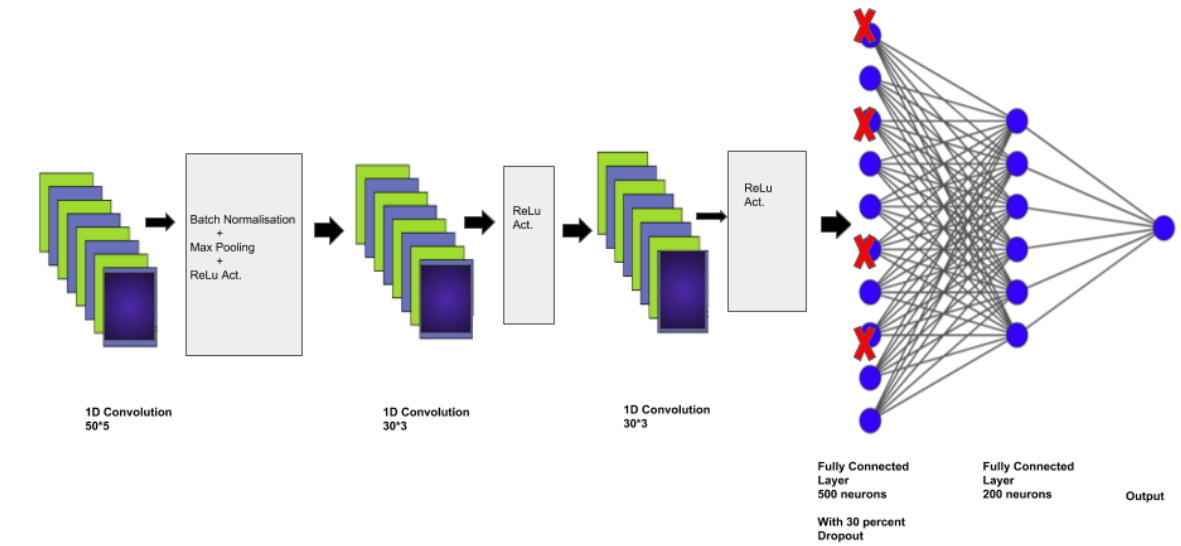
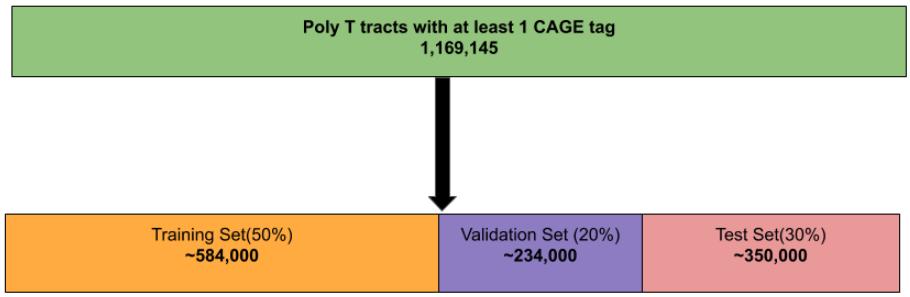


Figure S21: **Mean tag counts along T repeats.** To compute the raw mean tag count of each T repeat, we first calculated the mean tag count of each base genome-wide in 988 FANTOM libraries. We then summed these values along the T repeat + 5bp. The distribution of these tag counts are plotted for T repeats initially defined as associated (red) or not (blue) with CAGE peaks as defined in [11]. The overlapping area likely represents the existence of potential false negatives i.e. T repeats not associated with CAGE peaks but associated with CAGE tags nonetheless. The y-axis was limited to 100 for sake of clarity.



A



B

Figure S22: **Deep learning to predict transcription at T repeats**. **A.** CNN model architecture. **B.** Constructions of the different sets for learning. Only T repeats with at least 1 tag and no undetermined sequence ('N') were considered (n = 1,169,145) because T repeats without tag may represent false negatives.

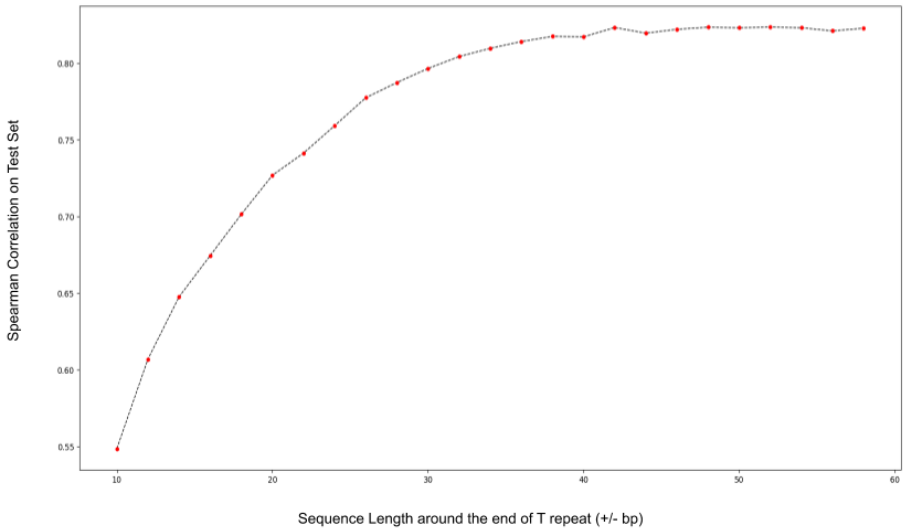
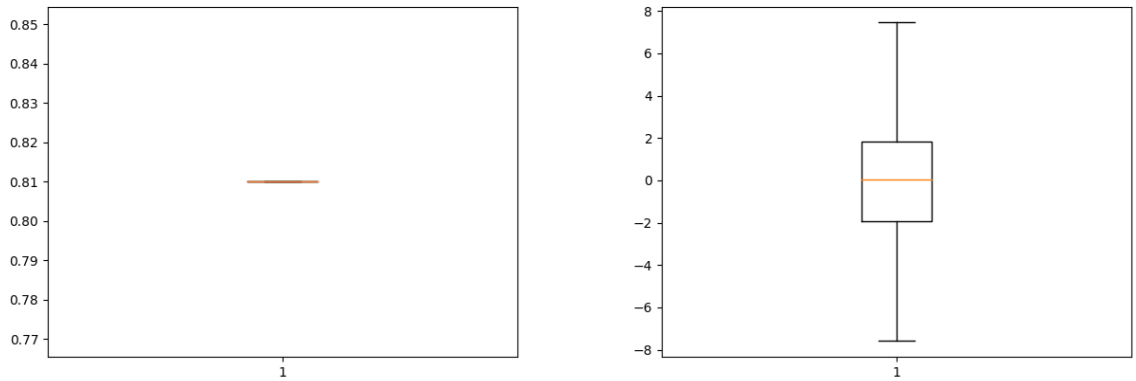


Figure S23: **Finding the optimal sequence length for deep learning.** The architecture depicted Supplementary Figure S19 was used to learn features in increasing lengths of sequences centered around T repeat ends (x-axis). The accuracy of each model is shown (y-axis).



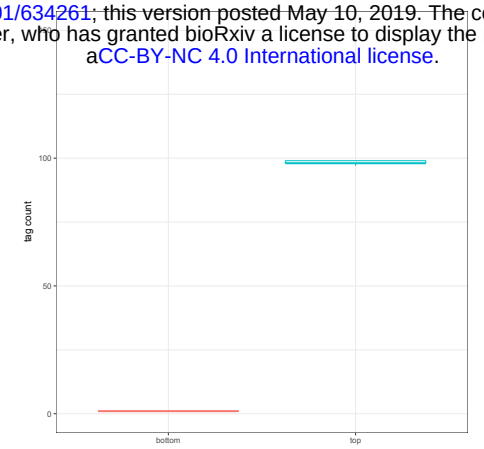
A B

Figure S24: **Predicting transcription at T repeats with CNN. A.** Accuracy of the deep CNN model was assessed as the Spearman correlation between observed and predicted tag counts of 1,169,145 considered T repeats. Learning was repeated 10 times to ensure model robustness. Results are shown as boxplot (median = 0.81). **B.** Error of one model was assessed as the difference between observed and predicted tag counts. Results are shown for 1,169,145 T repeats as boxplots (median = 1.82).

Figure S25: **Instructions lie downstream T repeats.** Two specific T repeats, one A with a low predicted tag count (0.9, observed tag count = 1.0) and another B with a high predicted tag count (39.37, observed tag count = 39.44). Sequences located downstream each T repeat were swapped and the prediction was assessed in each case. The prediction of the artificial sequence C drastically increased while the prediction of the artificial D decreased.

selected variables (top10)
T length
GAG downstream
AGC downstream
CTA downstream
GpA downstream
TpA downstream
T downstream
CCT downstream
TAG downstream
TGG downstream

Table S4: **Predicting transcription at T repeats using LASSO.** The predictive variables are the nucleotide, dinucleotide and trinucleotide content computed into the 50 bp downstream the polyT end and 50 bp upstream the polyT end excluding the T repeat. We also integrated the length of the T repeat and the nature of the nucleotide following the T repeat end (A, C or G at position -1 referring to the CAGE summit at 0). The selected top 10 variables are shown.



A

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-288	-6.644e+02	72.47%	47.14%
2		1e-229	-5.293e+02	35.69%	16.67%
3		1e-92	-2.123e+02	24.12%	13.38%
4		1e-85	-1.965e+02	14.10%	6.35%
5		1e-80	-1.859e+02	1.92%	0.13%
6		1e-77	-1.777e+02	32.73%	21.34%
7		1e-74	-1.719e+02	4.02%	0.80%
8		1e-68	-1.570e+02	5.84%	1.75%
9		1e-67	-1.564e+02	9.60%	3.95%
10		1e-64	-1.493e+02	20.78%	12.22%
11		1e-59	-1.365e+02	5.38%	1.67%
12		1e-59	-1.361e+02	10.58%	4.91%

B

Figure S26: **Motif enrichment comparing T repeats with high and low tag count. A.** Tag count distribution of the top 5,000 sequences with high (blue) and low (red) tag count. **B.** Motif enrichment was computed with HOMER [9] considering 101bp-long sequences centered around T repeat ends with high and low tag count (top 5,000). The top12 motifs out of 33 are shown.

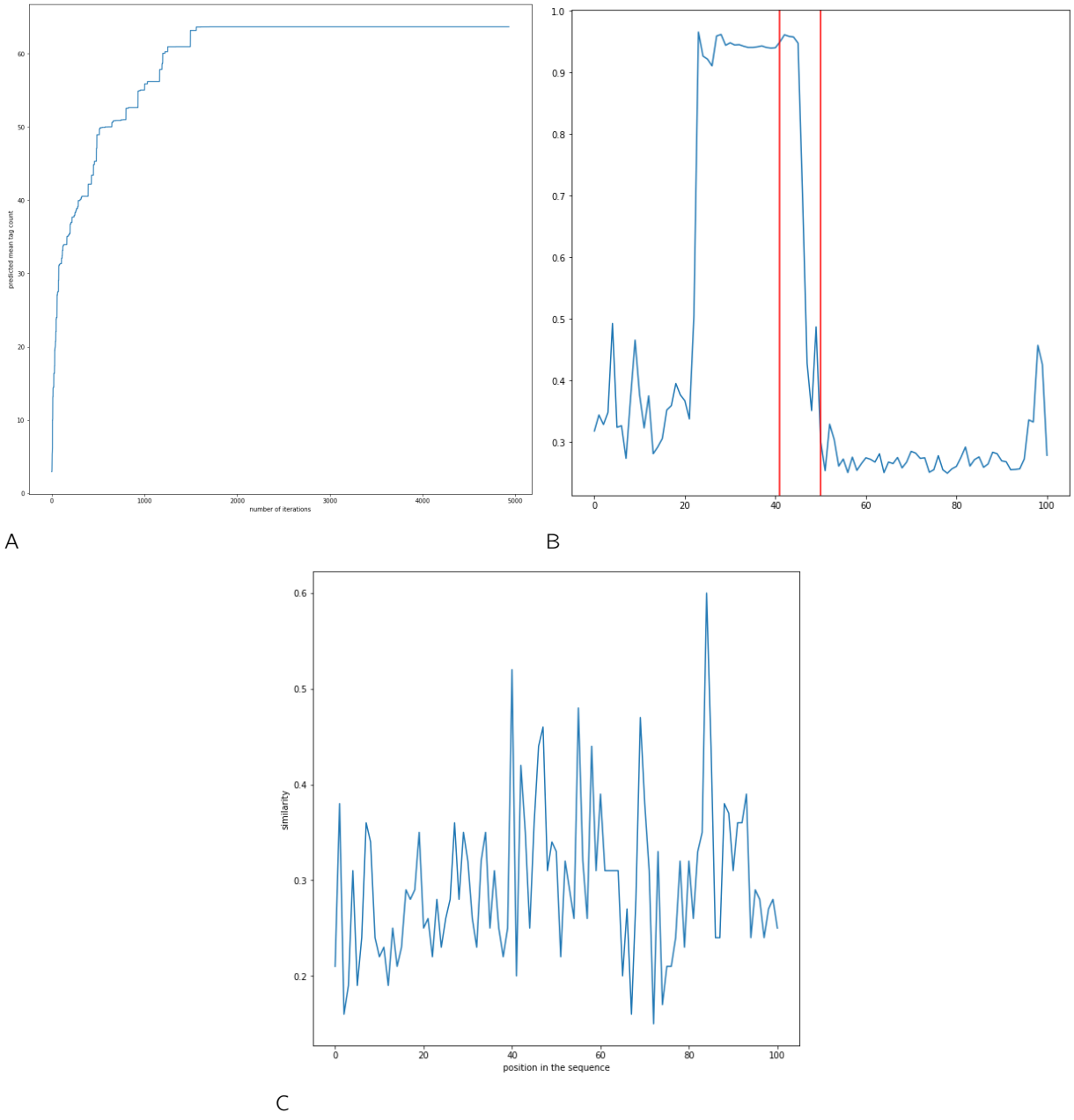


Figure S27: **Feature extraction. A.** For each sequence, the effect of changing one nucleotide at a random position into A,T, C or G was assessed. The nucleotide with the maximum predicted tag count is kept and the procedure continues testing another position. This process is repeated over 3,500 randomly selected positions (x-axis) until no further increase of predicted tag count (y-axis) is observed . **B.** Two sequences were randomly chosen in the test set. The two sequences optimized as in A are then compared position-wise in order to return 1 for each position with the same nucleotide in both sequences and 0 otherwise. This process was repeated for 2,000 pairs of random sequences. y-axis: fraction of identical pairs ; x-axis: position (The end of T repeat is located at position 50 ; repeat of 9 Ts is indicated by red vertical lines) **B.** Same procedure as in B minimizing the tag count for each sequence. The process was repeated for 100 pairs of random sequences.

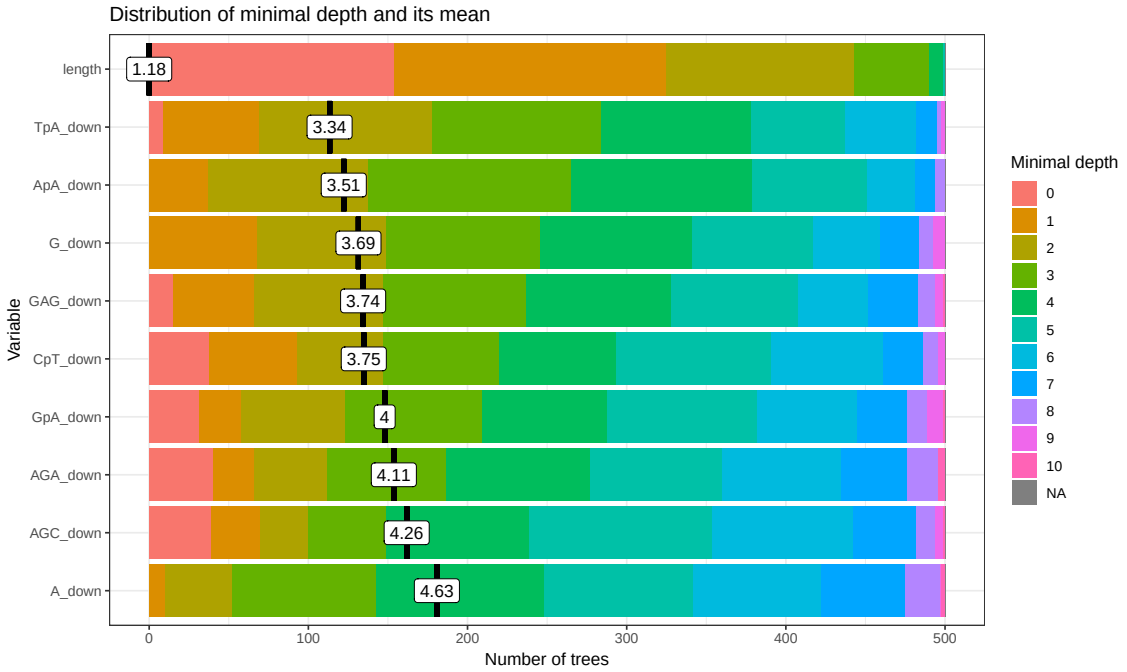


Figure S28: **Predicting transcription at T repeats using Random Forest.** The predictive variables are the nucleotide, dinucleotide and trinucleotide content computed into the 50 bp downstream the polyT end and 50 bp upstream the polyT end excluding the T repeat. We also integrated the length of the T repeat and the nature of the nucleotide following the T repeat end (A, C or G at position -1 referring to the CAGE summit at 0). Variable importance was assessed computing the minimal node depth, which reflects the predictiveness of a variable by a depth calculation relative to the root node of a tree (smaller values correspond to more predictive variables). The distribution of minimal depth is represented for top 10 variables according to mean minimal depth. The mean minimal depth was calculated using only non-missing values.

Figure S29: **Predicting transcription at human T repeats.** The mean tag count of each base genome-wide is computed from 988 (up) or 58 (bottom) FANTOM libraries. Arrows (middle panel) represent T repeats.

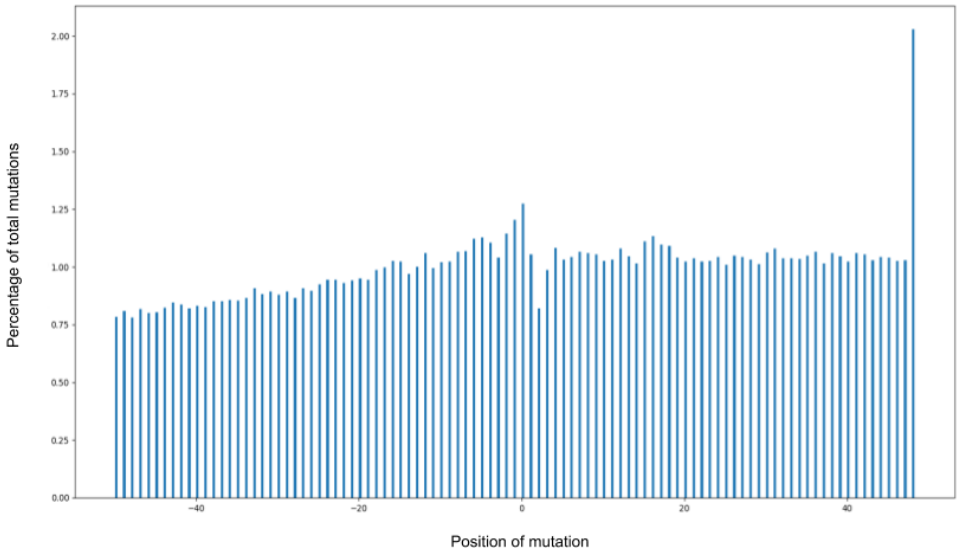


Figure S30: **Distribution of ClinVar variants around permissive CAGE summit.** Position 0 indicates the CAGE summit.

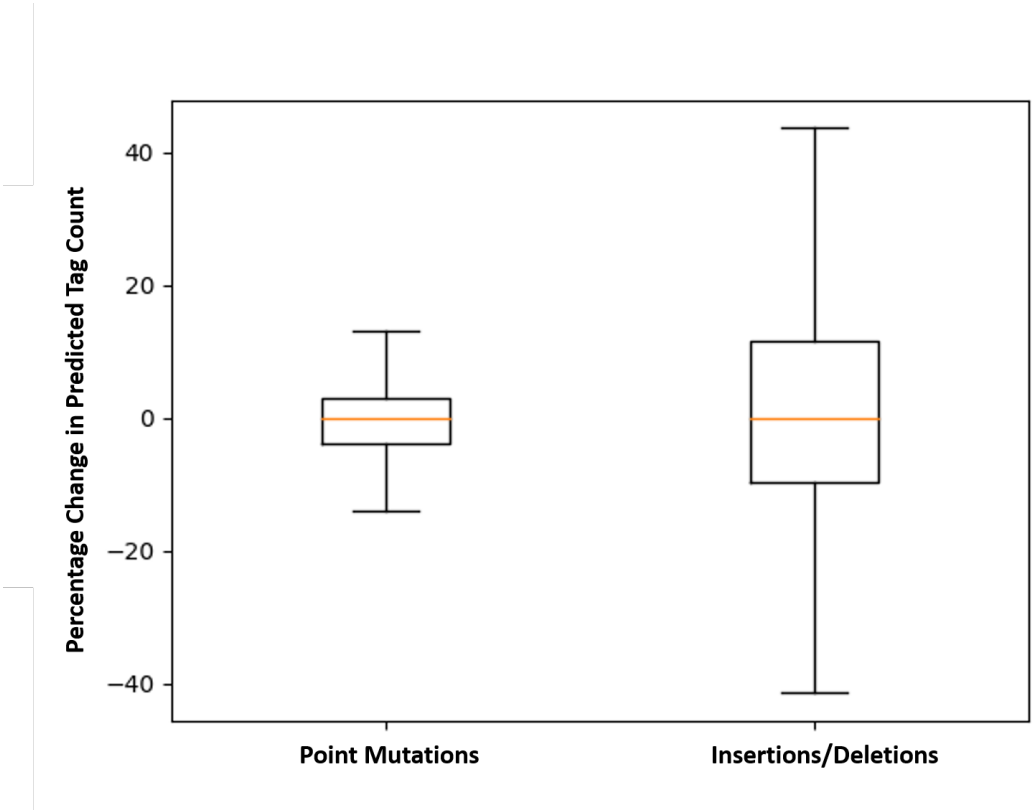


Figure S31: **ClinVar variants induce changes in transcription prediction at T repeats.** Changes were measured as the difference between models scores before and after variant insertion.

repeat_class	T repeats with CAGE (31,445)	T repeats without CAGE (844,086)
DNA	651	13202
LINE	5611	151870
LTR	756	15182
Low$_c$*omplexity*	851	11705
Other	3	2090
RC	6	43
RNA	1	10
Low_complexity	851	11705
SINE	23418	642986
Satellite	4	874
Simple_repeat	99	4692
Unknown	6	112
rRNA	2	32
scRNA	18	201
snRNA	8	176
srpRNA	9	259
tRNA	2	5

Table S6: **polyT CAGEs in repeat elements.** The coordinates of repetitive elements, according to RepeatMasker, were intersected with that of T repeats associated (n = 63,974) or not (n = 1,273,587) with CAGE peaks. The total number of intersections is indicated in brackets.

dataset	url and file name
human CAGE coordinates	http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/CAGE_peaks/hg19.cage_peak_coord_permissive.bed.gz
human CAGE annotation	http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/CAGE_peaks/hg19.cage_peak_tpm_ann.osc.txt.gz
human CAGE expression (TPM)	http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_tpm_ann.osc.txt.gz
human CAGE expression (RLE)	http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_rel_expr.txt.gz
mouse CAGE coordinates	http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/CAGE_peaks/mm9.cage_peak_coord_permissive.bed.gz
mouse CAGE annotation	http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/CAGE_peaks/mm9.cage_peak_tpm_ann.osc.txt.gz
dog CAGE coordinates	http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/canFam3.cage_peak_coord.bed.gz
chicken CAGE coordinates	http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/galGal5.cage_peak_coord.bed.gz
macaque CAGE coordinates	http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/rheMac8.cage_peak_coord.bed.gz
rat CAGE coordinates	http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/rn6.cage_peak_coord.bed.gz
mouse START-seq	https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE62151&format=file&file=GSE62151\%5FAdelman\%5FB16\%5FMEF\%5FstartRNA\%2Dseq\%5F5pr\%5Frep1and2\%5Fforward\%2Ebedgraph\%2Egz
mouse DECAP-seq	https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSM2422532&format=file&file=GSM2422532\%5FDECAPseq\%5FESC\%5FWT\%5Frep1\%2EbedGraph\%2Egz
mouse DECAP-seq	https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSM2422533&format=file&file=GSM2422533\%5FDECAPseq\%5FESC\%5FWT\%5Frep2\%2EbedGraph\%2Egz
A. thaliana TSS-seq	https://journals-plos-org.insb.bib.cnrs.fr/plosgenetics/article/file?type=supplementary&id=info:doi/10.1371/journal.pgen.1007969.s016
FANTOM_CAT	http://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/assembly/lv3_robust/FANTOM_CAT.lv3_robust.info_table.gene.tsv.gz
FANTOM_CAT gene expression (CPM)	https://fantom5-collaboration.gsc.riken.jp/webdav/FANTOM_CAT/expression/gene_based/count/individual/FANTOM.robust_CAT_gene.count.gene.without_anno.linear.tsv.gz
FANTOM_CAT robust transcripts	http://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/assembly/lv3_robust/FANTOM_CAT.lv3_robust.gtf.gz
GENCODEv19	ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.chr_patch_hapl_scaff.annotation.gtf.gz
GENCODEm1	ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_mouse/release_M1/gencode.vM1.annotation.gtf.gz
TSS classification	http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/TSS_human.bed.gz
TATAbox/CpG F5 annotation	http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks_annotation/DPIcluster_hg19_20120116.permissive_set.TATA_CpG_annotated.osc.gz
exosome sensitivity score	http://fantom.gsc.riken.jp/cat/?fd=other_data\%2Fexsome_sensitiviy/FANTOM_CAT.exsome_sensitiviy.CAGE_cluster.tsv.gz
CAGE directionality	http://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/other_data/directionality/FANTOM_CAT.CAGE_directionality.CAGE_cluster
STR reference catalog	http://strcat.teamerlich.org
ENCODE CAGE data	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRikenCage/
Master list of DHSs	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgDnaseMasterSites/wgEncodeAwgDnaseMasterSites.bed.gz
Roadmap epigenetics data	ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/by_sample/H1_cell_line/
ENCODE histone mark ChIP-seq data	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/
GRO-seq	https://gro-seq.colorado.edu
RNAP-II ChIA-PET K562 rep1	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeGisChiaPet/wgEncodeGisChiaPetK562Pol2InteractionsRep1.bed.gz
RNAP-II ChIA-PET K562 rep2	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeGisChiaPet/wgEncodeGisChiaPetK562Pol2InteractionsRep2.bed.gz
chromHMM/Segway K562 segmentation	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgSegmentation/wgEncodeAwgSegmentationCombinedHelas3.bed.gz
miRBase V21	ftp://mirbase.org/pub/mirbase/21/genomes/hsa.gff3
dbSNP	
ClinVar	

Table S7: **Datasets and online resources.**

References

[1] Scruggs, B. S. *et al.* Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol. Cell* **58**, 1101–1112 (2015).

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

[3] Nielsen, M. *et al.* Transcription-driven Chromatin repression of Intragenic transcription start sites. *PLoS Genet.* **15**, e1007969 (2019).

[4] Willems, T., Gymrek, M., Highnam, G., Mittelman, D. & Erlich, Y. The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).

[5] Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**, 1150–1159 (2011).

[6] Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199–204 (2017).

[7] Wei, W., Pelechano, V., Jarvelin, A. I. & Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet.* **27**, 267–276 (2011).

[8] Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat Commun* **5**, 5336 (2014).

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

[10] Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

[11] Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).