Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network

Grapotte, Mathys; Saraswat, Manu; Bessière, Chloé; Menichelli, Christophe; Ramilowski, Jordan A.; Severin, Jessica; Hayashizaki, Yoshihide; Itoh, Masayoshi; Tagami, Michihira; Murata, Mitsuyoshi; Kojima-Ishiyama, Miki; Noma, Shohei; Noguchi, Shuhei; Kasukawa, Takeya; Hasegawa, Akira; Suzuki, Harukazu; Nishiyori-Sueki, Hiromi; Frith, Martin C.; Abugessaisa, Imad; Aitken, Stuart; Aken, Bronwen L.; Alam, Intikhab; Alam, Tanvir; Alasiri, Rami; Alhendi, Ahmad M. N.; Alinejad-Rokny, Hamid; Alvarez, Mariano J.; Andersson, Robin; Arakawa, Takahiro; Araki, Marito; Arbel, Taly; Bornholdt, Jette; Boyd, Mette; Chen, Yun; Coskun, Mehmet; Dalby, Maria; Ienasescu, Hans; Jørgensen, Mette; Kaczkowski, Bogumil; Kere, Juha; Li, Kang; Lilje, Berit; Nepal, Chirag; Nguyen, Quan Hoang; Nielsen, Lars K.; Rennie, Sarah; Sandelin, Albin; Valen, Eivind; Vitezic, Morana; Vitting-Seerup, Kristoffer; FANTOM Consortium

Published in: Nature Communications

DOI: 10.1038/s41467-021-23143-7

Publication date: 2021

Document version
Publisher's PDF, also known as Version of record

Document license: CC BY

Citation for published version (APA):
Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network

Mathys Grapotte\textsuperscript{1,2,3,211}, Manu Saraswat\textsuperscript{1,2,211}, Chloé Bessière\textsuperscript{1,2,211}, Christophe Menichelli\textsuperscript{1,4}, Jordan A. Ramilowski\textsuperscript{5}, Jessica Severin\textsuperscript{5}, Yoshihide Hayashizaki\textsuperscript{6}, Masayoshi Itoh\textsuperscript{6}, Michihira Tagami\textsuperscript{5}, Mitsuyoshi Murata\textsuperscript{5}, Miki Kojima-Ishiyama\textsuperscript{5}, Shohei Nomia\textsuperscript{5}, Shuhei Noguchi\textsuperscript{5}, Takeya Kasukawa\textsuperscript{5}, Akira Hasegawa\textsuperscript{5}, Harukazu Suzuki\textsuperscript{5}, Hiromi Nishiyori-Sueki\textsuperscript{5}, Martin C. Frith\textsuperscript{7,8,9}, FANTOM consortium*, Clément Chatelain\textsuperscript{3}, Piero Carninci\textsuperscript{5}, Michiel J. L. de Hoon\textsuperscript{5}, Wyeth W. Wasserman\textsuperscript{10}, Laurent Bréhélin\textsuperscript{1,4,5} & Charles-Henri Lecellier\textsuperscript{1,2,4,5} & Charles-Henri Lecellier

Using the Cap Analysis of Gene Expression (CAGE) technology, the FANTOM5 consortium provided one of the most comprehensive maps of transcription start sites (TSSs) in several species. Strikingly, \textasciitilde72\% of them could not be assigned to a specific gene and initiate at unconventional regions, outside promoters or enhancers. Here, we probe these unassigned TSSs and show that, in all species studied, a significant fraction of CAGE peaks initiate at microsatellites, also called short tandem repeats (STRs). To confirm this transcription, we develop Cap Trap RNA-seq, a technology which combines cap trapping and long read MinION sequencing. We train sequence-based deep learning models able to predict CAGE signal at STRs with high accuracy. These models unveil the importance of STR surrounding sequences not only to distinguish STR classes, but also to predict the level of transcription initiation. Importantly, genetic variants linked to human diseases are preferentially found at STRs with high transcription initiation level, supporting the biological and clinical relevance of transcription initiation at STRs. Together, our results extend the repertoire of non-coding transcription associated with DNA tandem repeats and complexify STR polymorphism.
NA polymerase II (RNAPII) transcribes many loci outside annotated protein-coding gene promoters\(^1\) to generate a diversity of RNAs, including for instance enhancer RNAs\(^3\) and long non-coding RNAs (lncRNAs)\(^4\). In fact, >70% of all nucleotides are thought to be transcribed at some point\(^5\). Using the Cap Analysis of Gene Expression (CAGE) technology\(^7\), the FANTOM5 consortium provided one of the most comprehensive maps of TSSs in several species\(^8\). Integrating multiple collections of transcript models with FANTOM CAGE datasets, Hon et al. built a new annotation of the human genome (FANTOM CAGE-Associated Transcriptome, FANTOM CAT), with an atlas of 27,919 human lncRNAs, among them 19,175 potentially functional RNAs\(^8\). Despite this annotation, many CAGE peaks remain unassigned to a specific gene and/or initiate at unconventional regions, outside promoters or enhancers, providing an unprecedented mean to further characterize noncoding transcription within the genome “dark matter”\(^9\) and to decode part of the transcriptional “noise”.

Noncoding transcription is indeed far from being fully understood\(^10\) and some authors suggest that many of these transcripts, often faintly expressed, can simply be “noise” or “junk”\(^11\). On the other hand, many non annotated RNAPII transcribed regions correspond to open chromatin and cis-regulatory modules bound by transcription factors (TFs)\(^13\). Besides, genome-wide association studies showed that trait-associated loci, including those linked to human diseases, can be found outside canonical gene regions\(^14\). Together, these findings suggest that the noncoding regions of the human genome harbor a plethora of potentially transcribed functional elements, which can drastically impact genome regulations and functions\(^9\).

The human genome is scattered with repetitive sequences, and a large portion of noncoding RNAs derives from repetitive elements\(^17\) in particular DNA tandem repeats, such as satellite DNAs\(^19\) and minisatellites\(^20\). Minisatelites, also called short tandem repeats (STRs), constitute the third class of DNA tandem repeats. They correspond to repeated DNA motifs of 2–6 bp and constitute one of the most polymorphic and abundant repetitive elements\(^21\). Classes of STRs are defined based on the repeated DNA motif (e.g., (AC)\(_n\) will correspond to all STRs with repeats of the dinucleotide AC). STR polymorphism, which corresponds to variation in the number of repeated DNA motif (i.e., STR length), is presumably due to their susceptibility to slippage events during DNA replication. STRs have been shown to widely impact gene expression and to contribute to expression variation\(^22\). Some constitute genuine expression Quantitative Trait Loci (eQTLs)\(^23\), called eSTRs\(^23\). At the molecular level, STRs can for instance affect expression by inducing inhibitory DNA structures\(^26\) and/or by modulating TF binding\(^27\).

Provided the abundance of STRs on the one hand and the widespread transcription of the genome, including at repeated elements, on the other hand, we hypothesize that transcription initiation also occurs at STRs. To test this hypothesis, we probe CAGE data collected by the FANTOM5 consortium\(^2\) using the STRs catalog built by Willems et al.\(^29\). We specifically show that a significant portion of CAGE peaks (~8.6%) initiate at STRs. This transcription is confirmed by Cap Trap RNA-seq (CTR-seq), a technology that combines cap trapping and long-read MinION sequencing. Transcription of STR-containing RNAs has previously been reported in several species\(^30\)–\(^33\). We report here that thousands of STRs can also initiate transcription in human and mouse, therefore not being only a mere passenger in other RNAs but containing genuine TSSs. We further learn sequence-based Convolutional Neural Networks (CNNs) able to predict these transcription initiation levels with high accuracy (correlation between observed and predicted CAGE signal >0.65 for 14 STR classes with >5000 elements). These models unveil the importance of STR flanking sequences in distinguishing STR classes, one from the other, and also in predicting transcription initiation. We finally show that genetic variants linked to human diseases, are located, not only within, but also around STRs associated with high transcription initiation levels.

**Results**

**CAGE peaks are detected at STRs.** We first intersected the coordinates of 1,048,124 CAGE peak summits\(^2\) with that of 1,620,030 STRs called by HipSTR\(^29\). We found that 89,948 CAGE peaks (~8.6%) initiate at 84,555 STRs (Fig. 1a and Supplementary Fig. 1). As a comparison, only 2.3% of an equal number of randomly selected intervals with equivalent size intersected with CAGE peaks (Fisher’s exact test P value < 2.2e-16). Among CAGE peaks intersecting with STRs, 10,727 correspond to TSSs of FANTOM CAT transcripts\(^4\) and 8823 to enhancer boundaries\(^3\) (Supplementary Data 1). Note that the FANTOM CAT annotation was shown to be more accurate in 5’ end transcript definitions compared to other catalogs (GENCODE\(^34\), Human BodyMap\(^35\), and miTranscriptome\(^36\)), because transcript models combine various independent sources (GENCODE release 19, Human BodyMap 2.0, miTranscriptome, ENCODE and an RNAseq assembly from 70 FANTOM5 samples) and FANTOM CAT TSSs were validated with Roadmap Epigenome DHS and RAMPAGE datasets\(^4\). This transcription does not correspond to random noise because the fraction of STRs harboring a CAGE peak within each class differs depending on the STR class, without any link with their abundance (Fig. 1a, c). Some STR classes with low abundance are indeed more often associated with a CAGE peak than more abundant STRs (Fig. 1a, c, compare for instance (TTTTTT)_n or (AAAAAG)_n vs. (AT)_n or (ATTT)_n). Likewise, the number of STRs associated with CAGE peaks cannot merely be explained by their length, as several STR classes have similar length distribution but very different fractions of CAGE-associated loci (compare for instance (AT)_n and (GT)_n in Fig. 1c and Supplementary Fig. 2).

We computed the tag count sum along each STR ± 5 bp, and averaged the signal across 988 FANTOM5 libraries. We noticed the existence of very low (tag count = 1) CAGE counts along STRs, which artificially increase the signal (see examples in Fig. 1a). Spearman correlation coefficient between sum CAGE tag count along STR and STR length ~0.26. To remove any dependence between STR length and CAGE signal, the mean tag count was normalized by the length of the window used to compute the signal (i.e., STR length + 10 bp). Looking directly at this CAGE signal (not CAGE peaks) along the genome, we observed that some STR classes are more transcribed than others (Fig. 1d, compare (CGG)_n or (CGG)_n vs. (AAGG)_n or (AAAAAT)_n). No drastic difference in terms of CAGE signal was noticed between intra- and intergenic STRs (Supplementary Fig. 3).

Looking at each STR class separately, we confirmed that our CAGE signal computation is not sensitive to the STR length (Supplementary Fig. 4). Supplementary Fig. 4 also shows that STRs with different lengths can be associated with the same CAGE signal while, conversely, two STRs with different CAGE signals can have the same length. Thus, considering transcription, STR polymorphism appears to not only rely on their length (number of repeated elements). Transcription initiation, therefore, appears to complexify STR polymorphism.

**CAGE tags correspond to genuine transcriptional products.** CAGE read detection at STRs faces two problems. First, CAGE tags can capture not only TSSs but also the 5’ ends of post-transcriptionally processed RNAs\(^37\). To clarify this point, we used a strategy described by de Rie et al.\(^38\), which compares CAGE tags
obtained by Illumina (ENCODE) vs. Heliscope (FANTOM) technologies. Briefly, the 7-methylguanosine cap at the 5' end of CAGE tags produced by RNAPII can be recognized as a guanine nucleotide during reverse transcription. This artificially introduces mismatched Gs at Illumina tag 5' end, not detected with Heliscope sequencing, because it skips the first nucleotide. We then evaluated the existence of this G bias in CAGE tags corresponding to peaks detected at STRs, peaks assigned to genes (for positive control), and peaks intersecting the 3' end of precursor microRNAs (pre-miRNAs for a negative control) (Fig. 2). While most CAGE tag 5' ends perfectly match the sequences of pre-miRNA 3'end in all cell types tested, as previously reported, a G
Fig. 1 CAGE peaks are detected at STRs. a Three examples of STRs associated with a CAGE peak. The Zenbu browser was used. top track, hg19 genome sequence; middle track, CAGE tag count as mean across 988 libraries (BAM files with Q3 filter were used); bottom track, CAGE peaks as called in ref. b Number of STRs per STR class. For sake of clarity, only STR classes with >2000 loci are shown. c Fraction of STRs associated with a CAGE peak in all STR classes considered in b. d CAGE signal at STR classes with >2000 loci. CAGE signal was computed as the mean raw tag count of each STR (tag count in STR ± 5 bp) across all 988 FANTOM5 libraries. This tag count was further normalized by the length of the window used to compute the signal (i.e., STR length + 10 bp). The orange bar corresponds to the median value. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The upper whisker extends from the hinge to the largest value no further than 1.5 × IQR from the hinge. Data beyond the end of the whiskers are plotted individually.

Fig. 2 CAGE tags initiating at STRs are truly 5′-capped. G bias in ENCODE CAGE tags (bam files from nuclear fraction, polyA+) was assessed at FANTOM5 CAGE peaks assigned to genes (positive control) and CAGE peaks initiating at STRs. G bias at pre-microRNA 3′ ends was also assessed as a negative control. Five libraries were analyzed corresponding to A549 (replicates 3 and 4), GM12878, HeLa-S3, and K562 cells. The number of intersecting tags in each case is indicated in the bracket.

Bias was clearly observed when considering assigned CAGEs and CAGEs detected at STRs, confirming that the vast majority of STR-associated CAGE tags are truly capped. We also confirmed that STRs located within RNAPII-binding sites exhibit a stronger CAGE signal than STRs not associated with RNAPII-binding events (Supplementary Fig. 5).

Second, because of their repetitive nature, mapping CAGE reads to STRs is problematic and may yield ambiguous results. To circumvent this issue, we developed CTR-seq, which combines cap trapping and long-read MinION sequencing. With this technology, the median read length is >500 bp, thereby greatly limiting the chance of erroneous mapping. Two libraries were generated in A549 cells, including or not polyA tailing. This polyA tailing step before reverse transcription allows the generated in A549 cells, including or not polyA tailing. This polyA tailing step before reverse transcription allows the limitation the chance of erroneous mapping. Two libraries were generated in A549 cells, including or not polyA tailing. This polyA tailing step before reverse transcription allows the generation of polyA-minus noncoding RNAs. Long reads initiating at STRs were readily detected in both libraries (Fig. 3). As expected given the depth of MinION sequencing in only one cell line, the number of STRs associated with long reads is lower than that obtained with CAGE sequencing collected in 988 libraries (n = 5472 and 7812, respectively, with and without polyA tailing with 2291 STRs associated with long reads in both libraries). Among these 2291 STRs, 904 (39%) are also associated with a CAGE peak. Thus, compared to the reproducibility of MinION sequencing in both libraries (only 2291 STRs in common out of 5472 (42%) or 7812 (29%), CAGE and CTR-seq sequencing results are overall in agreement. In fact, STR classes associated with CAGE peaks correspond to those associated with CTR-seq reads (Fig. 3 compared to Fig. 1c). The Spearman correlation ρ between the fractions of STRs associated with CAGE and MinION reads with and without polyA tailing equals 0.88 and 0.89 respectively. Besides, 301 out of 904 STRs associated with both CAGE peak and CTR-seq long read correspond to TSSs of FANTOM CAT transcripts and 54 to enhancer boundaries. Overall, CTR-seq confirms CAGE data and the existence of transcription initiating at STRs. The similarity of the results obtained with and without the polyA tailing step also indicates that RNAs initiating at STRs are mostly polyadenylated.

Transcription initiation at STRs exhibits specific features. We further looked at the subcellular localization of STR-initiating transcripts and used CAGE sequencing data generated after cell fractionation (see “Methods” section). While the majority of CAGE tags, including those assigned to genes, are detected in both the nucleus and cytoplasm, CAGE tags initiating at STRs are mostly detected in the nuclear compartment (Fig. 4a). Functionally distinct RNA species were previously categorized by their transcriptional directionality. We then sought to compute the
A sequence-based deep learning model reveals that features governing transcription initiation depend on the STR classes. We further probed transcription initiation at STRs using a machine-learning approach. We used a deep Convolutional Neural Network (CNN), which is able to successfully predict CAGE signal in large regions of the human genome. This type of machine-learning approach takes as input the DNA sequence directly, without the need to manually define predictive features before analysis. The first question that arose was then to determine the sequence to use as input.

We first sought to build a model common to all STR classes to predict the CAGE signal as computed in Fig. 1d. Note that, because we used mean signal across CAGE libraries, our model is cell-type agnostic. This choice was motivated by the observation that the CAGE signal at STRs in each library is very sparse, thereby strongly reducing the prediction accuracy of our model. As input, we used sequences spanning 50 bp around the 3' end of each STR. Model architecture and constructions of the different sets used for learning are detailed in the “Methods” section and in Supplementary Fig. 7. Source code is available at https://gite.lirmm.fr/ibc/deepSTR. The accuracy of our model was computed as Spearman correlation between the predicted and the observed CAGE signals on held-out test data (see “Methods”). The performance of this global model was overall high (P = 0.72), indicating that transcription initiation at STRs can indeed be predicted by sequence-level features. However, looking at the accuracy for each STR class, we noticed drastic differences with accuracies ranging from 0.6 to 0.81 depending on the STR class (Fig. 5a, blue dots). The global model is notably accurate for the most represented STR class (i.e., (T)n with 766,747 elements), but performs worse in other STR classes. Differences in accuracies are not simply linked to the number of elements available for learning in each STR class. They rather suggest that, as proposed above (Fig. 4b), transcription initiation may be governed by features specific to each STR class.
STR flanking sequences can classify STR classes, independently of the DNA repeated motif. It was previously shown that 50-bp-long sequences flanking (AC)_n have evolved unusually to create specific nucleotide patterns. To determine if such specific patterns hold true for other STRs, we sought to classify STRs based only on their 50-bp surrounding sequences. We trained a CNN model to classify pairs of STR classes (Supplementary Fig. 7). To avoid any problem due to the imprecise definition of STR boundaries, we masked the seven bases located downstream the STR 3' ends (see "Methods"). In that case, model performance is evaluated by the Area Under the ROC (Receiver Operating Characteristics) curve (AUC, Fig. 5b). The AUCs obtained in these pairwise classifications were very high (AUC > 0.7, Fig. 5b), with the notable exceptions of (GTTT)_n vs. (GTTTTT)_n (see below). Thus, STRs can be accurately distinguished, one from each other, using only 50-bp flanking sequences, and not the DNA repeated motif, even in the case of complementary STRs, such as (AC)_n and (GT)_n (Fig. 5b).

Deep learning models unveil the key role of STR flanking sequences. To further probe the sequence-level features for transcription initiation at STRs, we decided to build a model for each STR class with >5000 elements (n = 47). Here, CNN is again used in a regression task to predict the CAGE signal. Sequences...
Fig. 5 Probing STR sequences with CNN models. a Comparison of the accuracies of global vs. class-specific models to predict transcription initiation levels at STRs. A model was learned on all STR sequences, irrespective of their class, and tested on each indicated STR class (accuracies obtained in each case, as Spearman $\rho$, is shown as blue points). Distinct models were also learned for each indicated class, without considering others (accuracies are shown in red). In total, 14 STR classes are shown as representative examples. Example sequence used as input is shown in E. b CNN-based pairwise classification of STRs using only STR flanking sequences (see “Methods” section). The pairs are defined by the line and the column of the matrix (e.g., the bottom left tile represents a classification task between T flanking sequences and GT flanking sequences). The values displayed on the tiles correspond to AUCs measured on the test set with the model trained specifically for the task. Clustering was performed to group pairs of STRs according to AUCs. c CNN performances to predict transcription initiation levels at heterologous STRs evaluated as the Spearman correlation between predicted and observed CAGE signal. The heatmap represents the performance of one model learned on one STR class (rows) and tested either on the same or another class (columns). Clustering is also used to show which models are similar (high correlation) and which ones differ (low correlation). d CNN models were learned on flanking sequences. The models use as an input only the 50-bp-long sequences flanking the STR, with the DNA repeated motif being masked by 9Ns (vectors of zeros in the one-hot encoded matrix). e Example of sequence used as input for each analysis depicted in A, B, C, and D. The pink box highlights the STR. All STRs are replaced by 9Ns in B and D, no matter their lengths. Additional seven bases downstream STR 3’ end are masked in B because this window can contain bases corresponding to the DNA repeat motif, a feature that can easily be learned for STR classification. See details in the “Methods” section.

spanning 50 bp around the 3’ end of each STR were used as input. Longer sequences were tested without improving the accuracy of the model (Supplementary Fig. 8). These class-specific models achieved overall better performances than the global model tested on each STR class separately (Fig. 5a and Supplementary Fig. 9). The only exceptions were classes composed of repetitions of T ((GTTTTT)$_m$, (GTTT)$_m$, and (CTTTT)$_m$). In these cases, global and (T)$_m$-specific models achieved better performance than (GTTTTT)$_m$, (GTTT)$_m$, or (CTTTT)$_m$-specific models. These results have two explanations: (i) compared to (T)$_m$ these classes have less occurrences (18,707 for (GTTTTT)$_m$, 55,898 for (GTTT)$_m$, and 15,433 for (CTTTT)$_m$), making it hard to learn models for these classes and (ii) the classification AUCs to distinguish (GTTTTT)$_m$, (GTTT)$_m$ or (CTTTT)$_m$ from (T)$_m$ was among the lowest observed (Fig. 5b), suggesting the existence of common sequence features that can be used by global and (T)$_m$-specific models. Overall, we estimated that STR class-specific models were accurate for 14 STR classes ($\rho > 0.65$).

We anticipated that class-specific models should not be equivalent and could not be interchangeable. We formally tested this hypothesis by measuring the accuracy of a model learned on one STR class and tested on another one (Fig. 5c). We caution again the fact that the performance of an STR-specific model also depends on the number of sequences available for learning. As observed earlier, the best accuracy is obtained with (T)$_m$ which are overrepresented in our catalog. Overall, the performance of one model tested on another STR class drastically decreases (Fig. 5c), revealing the existence of STR class-specific features predictive of transcription initiation. We also noticed that several models achieved non-negligible performances on other STR classes (Spearman $\rho > 0.5$, Fig. 5c), implying that some features governing transcription initiation at STRs are conserved between these STR classes. Thus, CNN models identified both common and specific features able to predict transcription initiation at STRs.

Our results unveil the importance of STR flanking sequences. We then evaluated the contribution of the sole surrounding sequences in transcription initiation prediction and built a model considering only these sequences (50 bp upstream and downstream STR, masking the STR itself, Fig. 5e). These models were less accurate than the formers but accuracies were still high for several classes (Fig. 5d), confirming that surrounding sequences contain features for transcription initiation prediction. The observed decrease in accuracies (Fig. 5d) implies that the STR itself contains features, which are combined with others present in flanking regions to predict transcription initiation. Remember that the CAGE signal predicted by our CNN models is normalized by the length of the STR (see above), which makes them unable to assess the contribution of STR length in transcription initiation.

Several sequence-level features predicting transcription initiation at STRs are conserved between human and mouse. To test whether transcription at STRs is biologically relevant, we relied on two criteria: conservation and association with diseases. First, we studied conservation in mouse.

The number of loci within each STR class differs in mouse and human HiPSH catalogs (Figs. 1b and 6a and Supplementary Fig. 10). We applied the strategy used in human to compute the CAGE signal (as mean raw tag count in STR ± 5 bp divided by STR length + 10 bp) in mouse using 397 CAGE libraries (Fig. 6b). As observed in human, several STR classes were associated with CAGE signal. This signal appears lower than in human (compare Figs. 1d and 6b). This might be due to the fact that mouse CAGE data are small-scaled in terms of the number of reads mapped and diversity in CAGE libraries, compared to human CAGE data$^2$, making the mouse CAGE signal at STRs probably less accurate than the human one.

We nonetheless tested the correlation of the human and mouse CAGE signals at orthologous STRs. Orthologous STRs were identified converting the mouse STR coordinates into human coordinates with the UCSC liftover tool (see “Methods”). We intersected the coordinates of human STRs with that of orthologous mouse STRs and computed the Pearson correlation between the CAGE signal observed in human and that observed in mouse on the same strand ($n = 18,072$). In that case, Pearson’s $r$ reaches $-0.87$ (Spearman $\rho = 0.51$), suggesting that transcription at STRs is indeed conserved between mouse and human. As expected, no correlation was observed ($r < 0.01$) when randomly shuffling one of the two vectors or when correlating the signals of 18,072 randomly chosen mouse and human STRs.

We then built a CNN model to predict the CAGE signal at mouse STR classes corresponding to the 14 classes shown in Fig. 5a (Fig. 6c, green dots). The performances of the models ranged from 0.4 to 0.8, demonstrating that, as observed for human STRs, transcription at several mouse STR classes can be predicted by sequence-level features. A notable exception is (CTTTT)$_n$ with Spearman $\rho < 0.2$ (see below). The mouse models were overall less accurate than human models (Fig. 6c, compare red and green dots), likely due to differences in the quality of the CAGE signal (i.e., predicted variable), as mentioned above.

We then tested whether the sequence features able to predict STR transcription initiation were conserved between mouse and human. We specifically tested the performances of models learned in one species and tested on another one (Fig. 6c, blue
that several features predictive of transcription initiation at STRs are conserved between human and mouse and that the level of conservation also varies depending on STR classes.

**ClinVar pathogenic variants are found at STRs with high transcription initiation level.** Second, we evaluated the potential implication of transcription initiation at STRs in human diseases and used the ClinVar database, which lists medically important variants. We found that STRs harboring ClinVar variants, located in a window encompassing STR ± 50 bp (n = 34,578), are associated with high CAGE signal compared to STRs without variants (n = 3,068,280, Fig. 7a), indicative of potential biological and clinical relevance for transcription initiation at STRs. Looking at the clinical significance of the variants, as defined in the ClinVar database, we indeed noticed that STRs associated with pathogenic variants exhibit stronger transcription initiation than STRs associated with other variants (Fig. 7b and Supplementary Fig. 12). STRs could be associated with more or less variants linked to a given disease than expected by chance (adjusted P value < 5e-3, Supplementary Data 2) but no clear association with a specific clinical trait was noticed.

We initially sought to identify representations of sequence motifs captured by CNN first layer filters using a strategy inspired by Maslova et al. and identified several influential first layers correlating with JASPAR PMW scores (see “Methods” section and Supplementary Tables provided here at https://gite.lirmm.fr/ibc/deepSTR//first_layer_interpretation). However, it is important to remember that our models were optimized to predict CAGE signal, not to learn interpretable representations from input DNA sequences. Koo and Eddy have indeed demonstrated that tackling these two questions—prediction and interpretation—requires distinct CNN architectures, in particular adapting max-pooling and convolutional filter size. At present, our models likely learn partial motifs and do not limit the ability to learn full interpretable motifs in deeper layers. We then used a perturbation-based approach and randomly created in silico mutations to identify key positions of the models (see “Methods” section). Random variations were directly introduced into STR sequences, and predictions were made on these mutated sequences using the CNN model-specific of the STR class considered. The impact of the variation was then assessed as the difference between the predictions obtained with mutated and reference sequences. Same analyses were performed with ClinVar variants (Fig. 7c and Supplementary Fig. 13). Key positions were defined as positions, which, when mutated, have a strong impact on the prediction changes (i.e., high variance), being either positive or negative. As shown in Fig. 7c, for both random and ClinVar variants, the most important positions appeared located around STR 3′ end (−15 bp/+30 bp) and their distribution is skewed toward the sense orientation of the transcripts. Strikingly, a significant proportion of ClinVar variants are located in the immediate vicinity of the STR 3′ end (Fig. 7d). Hence, the most important positions identified by our models correspond to positions with high occurrences of ClinVar variants (Fig. 7c, d). However, neither the distribution nor the impact of variants appears linked to their pathogenicity because similar results are observed for both benign and pathogenic variants (Supplementary Fig. 14). Note that ClinVar variants are also concentrated around assigned CAGE peak summits and all identified CAGE peak summits (Supplementary Fig. 15). Overall, we conclude that the pathogenicity of ClinVar variants appears to be linked to the transcription initiation level at the targeted STR rather than to the position of the variation or its impact on prediction.

Finally, as machine-learning approaches only unveil correlation between predictive and predicted features, not direct causation, we...
sought to determine whether the features learned by our models correspond to sequence-level instructions for transcription initiation. We looked for gene TSSs located at STRs and harboring variants acting as eQTLs for the corresponding genes, in a scenario similar to that described by Bertuzzi et al. in the case of a minisatellite and the NPRL3 gene\textsuperscript{20}. Gene expression is considered here as a proxy for the measure of transcription initiation at STRs. In that scenario, if our models capture instructions for expression, the difference of the predictions made by our models for the reference and the alternative alleles should have the same sign as the eQTL slope (i.e., gene expression increase (slope > 0) or decrease (slope < 0)) more often than expected by chance. First, to identify STRs potentially acting as TSSs, we selected STRs located in gene promoters (considering 1 kb around FANTOM CAT gene start). We only considered models with accuracy >0.7 (Fig.5c). Second, based on our results depicted in Fig.7c, we selected GTEx eQTLs located in a −15-bp/+30-bp window around STR 3′ end and linked to the expression of the genes associated with STRs in the first step. These selections yielded 86 cases of STR sequence variations linked to gene expression by eQTL. Of note, we first thought to use FANTOM CAT transcript TSSs directly, instead of gene TSSs, but only one case was identified with prediction error (measured as the absolute value of the difference between the predicted and the observed CAGE signals) < 0.2. The alternative alleles corresponding to the selected eQTLs were inserted into their cognate STR sequences and a prediction was made for this modified sequence. The sign of the difference between the two predictions (alternative - reference) was compared to the sign of the eQTL slope. We counted the number of times these signs were identical or different (Supplementary Fig. 16). The prediction errors of the models for these 86 STRs were also computed in the case of the reference genome (Supplementary Fig. 16). As shown in Supplementary Fig. 17, when predictions are accurate on the reference genome (error ≤ 0.2), the models are able to predict the impact of variants on expression i.e., in most cases, the sign of the difference between the predictions made with the alternative and predictive alleles is similar to that of the eQTL slope. Importantly, this is no longer observed when the models poorly perform (error > 0.2). Binomial tests were used to statistically assess the relevance of these findings. Thus, when accurate, our models are able to predict the effects of eQTLs, supporting a causal relationship between the predictive and the predicted variables rather than a mere correlation.

**Discussion**

We report here the discovery of widespread transcription initiation at STRs in human and mouse. These results extend previous findings\textsuperscript{30–33} and reveal that, in addition to being the passenger of
host RNAs initiating at their own TSSs, STRs can also initiate the transcription of distinct and autonomous RNAs. The next main issue is to determine the role(s) of these transcripts. RNA species can be functionally categorized according to transcriptional directionality. In the case of STRs, transcriptional directionality appears to depend on the STR class (Fig. 4b). It is thus likely that RNAs initiating at STRs fulfill distinct functions and many hypotheses could be proposed at this stage. For instance, 10,727 CAGE peaks mapped at STRs correspond to STRs21,58,59. FANTOM CAT transcripts mostly initiate at GC-rich STRs, while enhancer RNAs more often correspond to A/T-rich STRs (Supplementary Data 1). A possible function is provided by $(T)_n$, which are over-represented in eukaryotic genomes51 and have been shown to act as promoter elements by depleting representative nucleosomes52. As a consequence, $(T)_n$ can increase transcription of reporter genes in similar levels to TF-binding sites. The findings that $(A)_n$ and $(T)_n$ represent distinct directional signals for nucleosome removal are very well compatible with differences observed in flanking sequences (Fig. 5b) and transcriptional directionality (Fig. 4b), both able to create asymmetry at $(A)_n$ and $(T)_n$. Besides, we show that most CAGE tags initiating at STRs remain nuclear (Fig. 4a). This observation suggests that, similar to other repeat-initiating RNA, RNAs initiating at STRs could also play roles at the nuclear/chromatin levels, for instance in DNA topology. Note that we also calculated the enrichment of STR classes in FANTOM CAT biotypes (Supplementary Data 3). The strongest enrichments correspond to $(A)_n$, $(AT)_n$, and $(AAAT)_n$ at enhancers, which are known to be GC-poor sequences compared to promoters for instance50. It also remains to clarify whether STR-associated RNAs or the act of transcription per se is functionally important. Dedicated experiments are now required to formally identify the biological functions linked to the transcription of each STR class. These experiments are all the more warranted as STR transcription is associated with clinically relevant genomic variations (Fig. 7).

One key finding of our study is the discovery that STR flanking sequences are not inert but rather contain important features that play critical roles in their biology, as previously suspected. These results call for the development of novel methods able to take these sequences into account in order to revisit STR mapping/genotyping and integrate SNVs located in STR vicinity. These methods should have broad applications in various fields of research and medicine, from forensic medicine to population genetics for instance. STR length variations have notably been shown to influence the results of CAGE models as well as their interpretation.

More broadly, the same rationale could be applied to other methods aimed at predicting CAGE signal along the genome, distinguishing biological entities (genes, enhancers, ...), genomic segments16,62, and/or isochores based on their sequence features. Building a general model increases the risk of designing a model suited for the most represented elements, not for the others. Notably, promoters and enhancers can be distinguished by different CpG content, the presence of polyA signal and of 5’ splice sites50,50 as well as different transcription factor combinations. It is therefore likely that the same filters will not apply similarly to predict transcription in both cases and that one may want to develop a specific model for each of these entities to increase the accuracy of the predictions.

The prediction of transcription initiation based solely on sequence features has long been studied, especially using CAGE data. The high accuracy achieved by CNN models for this task, as illustrated in this study or in refs. 43,44,47, as well as the development of methods aimed at interpreting this type of statistical models48,49,67,68, will certainly accelerate the achievement of this goal, which becomes more than ever “a realistic short-term objective rather than a distant aspiration”.

**Methods**

**Data and bioinformatic analyses.** The bedtools window69 was used to look for CAGE peaks (coordinates available at http://fantom.gsc.riken.jp/5/datafiles/phase1.1/extra/CAGE_peaks/hg19.cage_peak_coord_permissive.bed.gz) at STRs ± 5bp (catalog available at https://github.com/HipSTR-Tool/HipSTR-references/raw/master/human/hg19.hipstr_reference.bed.gz) as follows:

```
```

As a comparison, random intervals were generated using bedtools shuffle69.

```
```

Similar analyses were performed using mouse STR catalog (available at https://github.com/HipSTR-Tool/HipSTR-references/blob/master/mouse/mm10/hipstr_reference.bed.gz) liftovered to mm9 using UCSC liftOver tool70.
lifter mm10.hipstr_reference.bed mm10ToMm9.over.chain.gz mm9.hipstr_reference.bed unlifted.bed

To compute the CAGE signal, we used raw tag count along the genome with a 1-bp binning and Q3 quality mapping filter. At each position of the genome, the mean tag count across 988 libraries for human and 387 for mouse was computed. The values obtained at each position of a window encompassing the STR ± 5 bp were then summarized and normalized (i.e., divided by the STR length + 10 bp) to limit the impact of the CAGE noise signal observed along STRs. CAGE signals at human and mouse STRs are available at https://gite.lirmm.fr/ibc/deepSTR, as, limit the impact of the CAGE noise signal observed along STRs. CAGE signals at centered around STR 3

The bedtools intersect was used to distinguish intra- and intergenic STRs, intersecting their coordinates with that of the FANTOM gene annotation (available at https://fantom.gsc.riken.jp/5/supp/Hon_et_al_2016/data/assembly/v3/robust_FANTOM_CAT33_robust.bed.gz) and enhancers [https://fantom.gsc.riken.jp/5/datasets/latest/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz]. ENCODE RNAPII ChIP-Seq reads were available at https://fantom.gsc.riken.jp/5/datasets/latest/extra/Enhancers/hg19.hipstr_reference.cage.500bp.around3end.fa and mm9.hipstr_reference.MM10_MOUSE_H3K9me3.bed.gz], K562.

Our models were optimized to predict CAGE signal and cannot, as such, be applied to other types of data. However, the methodology used here is generic and could be translated into other pipelines. We plan to release a software pipeline using our methodology and example annotations in the future.

The CAGE signal was computed as explained above. A score equals to 1 or −1 indicates that transcription is strictly oriented towards the (+) or (−) strand, respectively. A score close to 0 indicates that the transcription is balanced and that it occurs equally on the (+) and (−) strands.

1 U1 PWM was built using MEME73 and sequences encompassing −3/10 bp around FANTOM CAT 5′ donor splice sites (exon 3′ end). We then used this PWM and FIMO74 to scan 2 kb regions centered around 3′ ends (T3) STRs (considering the top 50,000 sequences with the highest CAGE signal) and FANTOM CAT TSSs. For polyA sites, we used the UCSC track corresponding to the predictions made by Cheng et al.75, as a bed file and used it in bedtools intersect to look at polyA site distribution in regions encompassing 1 kb around (T3) 3′ ends (top 50,000 with the highest CAGE signal) and FANTOM CAT TSSs.

Convolutional neural network. CNN architecture is described in Supplementary Fig. 7. To build a CNN, we needed aligned sequences of equal length. However, as shown in Supplementary Fig. S1, CAGE peaks are scattered along STRs. We thus decided to align the sequences on STR 3′ ends, as defined by the CAGE data. HipSTR indeed provides a catalog built on the (−) strand but CAGE data are stranded data (see Fig. 1a). CAGE thus allows to orientate each STR of the HipSTR catalog as exemplified here:

**HipSTR catalog (see hg19.hipstr_reference.bed):**
chr1 10001 10468 6 78 Human_STR_1 AACCCT
Same STR with CAGE data (see hg19.hipstr_reference.cage.bed made available at https://gite.lirmm.fr/ibc/deepSTR/chr1 10001 10468 Human_STR_1 AACCCT: 0.410901 + chr1 10001 10468 Human_STR_1_AACCCT: 0.354298

It is then possible to determine the 3′ end of each STR according to the strand considered (here 10468 on the (−) strand and 10002 on the (−) strand). This procedure almost doubles the number of elements in each class.

Sequences spanning 50 bp around the 3′ end of each STR were used as input unless otherwise stated (see Fig. 5e). Longer sequences were tested without improving the accuracy of the model (Supplementary Fig. 8). Note that only 89,189 STRs (out of 1,620,303, −5.5%) are longer than 50 bp and, in only these few cases, the sequence located upstream STR 3′ end only corresponds to the STR itself. The padding of the model was determined by brute force for algorithms using a grid search approach. This approach makes a complete search over all hyperparameters (number of layers, number of neurons, activation functions, different learning rates, shape of convolutional kernels, number of convolutional filters, ...). The grid search algorithm trains and tests all possible models with all combinations of parameters and returns the most accurate model. The model was then trained in PyTorch. The source code of the model, alongside scripts and Jupyter notebooks are available at https://gite.lirmm.fr/ibc/deepSTR.

In order to minimize overfitting, dropout is added to the fully connected layers (probability of dropout = 0.30). The training pipeline is described in Supplementary Fig. 7: we separate training, testing, and validation datasets prior to model training, and these sets are stored on disk. This allows us to carry out analyses on held-out data that has never been seen by the models. We stop the training once the loss function calculated on the validation set drops for five consecutive epochs (early stopping). Relatively good performances on mouse data demonstrated across models trained on DNA polymersase data. Our models were optimized to predict CAGE signal and cannot, as such, be applied to other types of data. However, the methodology used here is generic and could be
applied to other types of data as long as one can associate a numeric signal to a specific genomic region.

To make sure that our models do not overfit due for instance to homologous sequences present in both train and test sets, we used BLASTs76 to look for homology between (T)\(^n\) sequences of the test and train sets. The model learned on (T)\(^n\) STRs was used because it is the most accurate and therefore the more likely to overfit. We found 102,209 sequences from the test set with >60% query cover and >80% identity with at least one sequence of the train set. We separated these sequences (test set 1, homologous sequences) from the rest of the test set (test set 2, 121,808 nonhomologous sequences). We then computed Spearman correlations between the predicted and the observed CAGE signals using these two test sets: 0.75 between test set 1 and 0.78 with test set 2. In both cases, correlations decreased, as compared to correlation computed with the whole test set (0.84). This decrease is due to differences in CAGE signal distribution between the whole test set, test set 1 and 2 (Supplementary Fig. 18) likely linked to mapping issues. However, model performance measured on test set 2 was greater than that obtained with test set 1. This is in contrast to what is expected in the case of model overfitting due to sequence homology. We then conclude that the homology observed between train and test sets is not sufficient to make the model overfit.

For comparison to the baseline model, we computed the correlation between the observed CAGE signal and randomized CAGE signal (equivalent to a predictor that returns a random value drawn from observed values). Randomization was repeated 10 times and Spearman correlation was invariably close to 0 (absolute value (\(\rho\)) < 5e-4).

The models are provided at https://gite.lirmm.fr/bc/deepSTR. They can be used to predict transcription initiation level at STRs using a fasta file. Likewise, impact of genetic variations can be assessed by comparing the predictions obtained for instance with reference and mutated sequences (see Fig. 7 and Supplementary Fig. 17).

Classification. The CNN model can also be set up for a classification task (Fig. 5b and Supplementary Fig. 7). In that case, the only difference with the regression model is the last neuron in the fully connected layer. The classifier CNN uses the same training method. The data are also prepared by separate scripts before training is done and stored on disk. All analyses resulting from the classification are performed on the test sets to avoid optimistic bias in accuracy estimation. Note that 7 bp downstream STR 3’ end were masked and replaced by Ns (Fig 5e) because we noticed that this window can contain bases corresponding to the DNA repeat motif, a feature that can easily be learned by a CNN. The sequences used as input, for classification using flanking sequences only (Fig 5d), are centered around STR 3’ end and consist of 50-bp long upstream sequence + 9 Ns, which mask the STR itself + 7 Ns + 43-bp long downstream sequence (total length = 109 bp, Fig. 5e).

Model swaps between human STR classes. After models are trained on all STR classes, their weights are stored in a .pt file (following the PyTorch convention). Predictions were then computed on all test sets with all models.

Model interpretation. First, for each of the 14 models presented in Fig. 5, we measured the influence of each first layer filters by removing them iteratively and comparing the accuracy of the model (Spearman correlation between observed and predicted CAGE signal) with the 49 remaining filters. We also computed an influence threshold by learning each CNN model ten times and computing a 95% confidence interval (CI). The threshold was calculated as log2(\(I\)C1 length/2). This allows us to focus our analysis on key filters, with performance impact greater than what would have been obtained by chance, simply re-training the model. Influential first layer filters are then ranked according to their influence. Second, on the one hand, we used FIMO74 to scan 101-bp long sequences centered around STR 3’ end (considering all STR sequences if n < 10,000 or 10,000 randomly chosen sequences otherwise) with JASPAR PWM77. For each PWM, we identified a set of STR sequences harboring PWM hits. For each sequence, we kept the PWM maximal score found. On the other hand, we scanned the 10,000 STR sequences with influential first layer filters as defined in step 1 (using matrix multiplication as in convolution) and kept the maximal value obtained for each sequence. We then computed the correlation between JASPAR PWM scores and first layer filter scores. We reasoned that if a filter represents a partial PWM, their score should be correlated. The results of these analyses are provided as Supplementary Tables located on our git repository [https://gite.lirmm.fr/bc/deepSTR/first_layer_interpretation].

Predicting the impact of ClinVar variants. ClinVar vcf file was downloaded January 8th 2019 from this url [ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/] and then converted into bed file. The bed file was found for STRs associated with ClinVar variants (Fig. 7a) using bedtools window\(^{-}\) as follows:

```
bedtools window -w 50 -a clinvar Mutation.bed -b str_coordinates.bed
```

Variants were directly introduced into STR sequences (\(\pm\) 50 bp) using Biopython\(^{78}\) library and the seq/fromontuble() function. To keep sequences aligned, we only considered single nucleotide variants (SNVs). CNN models were then used to predict the CAGE signal of the initial and mutated sequences. The change was computed by the difference between the prediction obtained with the mutated sequence and that obtained with the reference sequence. To insert random variations (Fig. 7c, d), we created mutation position map, which follows a uniform distribution (each position has an equal probability of receiving a mutation). Then, we took sequences in the database and mutated them one by one at a position taken from the mutation map. All possible mutations at the chosen position have an equal probability of occurrence (Fig. 7d).

Reporting summary. Further information on research development is available in the Nature Research Reporting Summary linked to this article.

Data availability


Code availability

Data, alongside source code of the models, a readme.txt file and other instructions for installing and running the analyses are available on our git repository [https://gite.lirmm.fr/bc/deepSTR]. This repository can be downloaded using the following command line:

```
```

Received: 15 July 2020; Accepted: 13 April 2021;
Published online: 02 June 2021

References

Acknowledgements

We thank Cédric Notredame, Anthony Mathelier, Oriol Fornes Crespo, Philip Richard, 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 also thank Takahiro Arakawa for the preparation and provision of cell culture samples. 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 Piechaczky and Edouard Bertrand for their continued support. 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 NUMEY (ANR-10-LAXB-0020). M.G. was supported by a Conventions Industrielles de Formation par la Recherche (CIFRE) PhD fellowship from SANOFI R&D. FANTOM5 was made possible by the following grants: Research Grant for RIKEN Omics Science Center from MEXT to Y.H.; Grant of PhD fellowship from SANOFI R&D. FANTOM5 was made possible by the following grants: Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation by Innovative Technology (Cell Innovation Program) from MEXT to the RIKEN Center for Integrative Medical Sciences; Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation Program from MEXT to Y.H. This work was further supported by a Research Grant from MEXT to the RIKEN Center for Integrative Medical Sciences.

Author contributions


Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23143-7.

Correspondence and requests for materials should be addressed to L.Béhé. or C.-H.L.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

FANTOM consortium
