CFC-seq: identification of full-length capped RNAs unveil enhancer-derived transcription

Authors:
Chi Wai Yip[1,#], Callum Parr[1], Hazuki Takahashi[1], Kayoko Yasuzawa[1], Matthew Valentine[1], Hiromi Nishiyori-Sueki[1], Camilla Ugolini[2], Valeria Ranzani[3], Mitsuyoshi Murata[1], Masaki Kato[1], Wenjing Kang[4], Wing Hin Yip[4], Youtaro Shibayama[1], Andre Darah Sim[5], Ying Chen[5], Xufeng Shu[1,6], Jonathan Moody[1], Ramzan Umarov[1], Jen-Chien Chang[1], Luca Pandolfini[7], Tsugumi Kawashima[1], Michihira Tagami[1], Tomoe Nobusada[1], Tsukasa Kouno[1], Carlos Alfonso Gonzale[8], Roberto Albanese[9], Francesco Dossena[9], Nejc Haberman[10], Kokoro Ozaki[1], Takeya Kasukawa[1], Boris Lenhard[11], Martin Frith[12], Beatrice Bodega[3], Francesco Nicassio[2], Lorenzo Calviello[9], Magda Bienko[4,9], Ivano Legnini[9], Valérie Hilgers[8], Stefano Gustincich[7], Jonathan Göke[5], Charles-Henri Lecellier[13], Jay W. Shin[1,5], Chung-Chau Hon[1,#], Piero Carninci[1,9,#]

Affiliations:

1. RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045 Japan.
2. Center for Genomic Science of IIT@SEMM at the Istituto Italiano di Tecnologia, Via Adamello 16, 20139 Milan, Italy.
3. INGM, Istituto Nazionale di Genetica Molecolare 'Romeo ed Enrica Invernizzi', Milan, Italy.
4. Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, 17177, Sweden; Science for Life Laboratory, Stockholm, 17177, Sweden.
5. Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), 60 Biopolis Street, Genome, Singapore 138672, Republic of Singapore.
6. Department of Computational Biology and Medical Sciences, The University of Tokyo, Kashiwa, Chiba, 277-8562, Japan.
7. Center for Human Technologies, Non-coding RNAs and RNA-based therapeutics, Istituto Italiano di Tecnologia (IIT), Via Enrico Melen 83, 16152 Genova, Italy.
8. Max Planck Institute of Immunobiology and Epigenetics, Freiburg 79104, Germany
9. Human Technopole, via Rita Levi Montalcini 1, Milan, Italy.
10. Division of Neuroscience, Department of Brain Sciences, Imperial College London, London W12 0NN, UK
11. MRC London Institute of Medical Sciences; Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK
12. Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan; Computational Bio Big-Data Open Innovation Laboratory, AIST, Tokyo, Japan; Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Japan.
13. Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France; LIRMM, University of Montpellier, CNRS, Montpellier, France

Correspondence:
#Chi Wai Yip: chiwai.yip@riken.jp
#Chung-Chau Hon: chungchau.hon@riken.jp
#Piero Carninci: carninci@riken.jp

Abstract

Long-read sequencing has emerged as a powerful tool for uncovering novel transcripts and genes. However, existing protocols often lack confidence in identifying the transcription start site (TSS) and fail to capture non-poly(A) RNA, thereby limiting the discovery of novel genes, particularly long non-coding RNAs (lncRNAs). In this study, we introduce Cap-trap full-length cDNA sequencing (CFC-seq), a comprehensive protocol that combines Cap-trapping and poly(A)-tailing with Oxford Nanopore sequencing. This protocol enables precise identification of TSSs and full-length transcripts. Applying CFC-seq to two *in vitro* differentiation time courses resulted in approximately 236 million mappable reads. The transcript Start-site Aware Long-read Assembler (SALA) was developed for *de novo* assembling the transcript models, leading to the identification of 39,425 confident novel genes. Using this dataset, enhancer-derived ncRNAs were re-defined with longer length and more splicing activity, which were correlated with enhancer structure. Compared to enhancers with CpG islands, TATA box enhancers were shown to be more cell type specific with fewer chromatin interaction but produced longer and more stable polyadenylated RNA. A significant proportion of these TATA box-derived eRNAs originated from LTR transposable elements. Overall, this study systematically annotated ~24,000 novel eRNA genes and correlated their transcription properties with enhancer structure.

Keywords:
Long-read sequencing, Transcription start site, enhancer, lncRNA

Highlights

- From 236 million long-reads, CFC-seq identified 39,425 novel genes with genuine TSS support. These include ~24,000 eRNA genes.
- SALA, a long-read assembler, was developed to facilitate genuine TSS incorporation.
- Compared to TATA box enhancers, CGI enhancers are more ubiquitous, enriched with repressive histone mark, with more chromatin connection and are enriched in 2D and super enhancer.
- eRNAs derived from TATA box are longer, more stable, frequently spliced with high splicing efficiency, frequently polyadenylated, and are enriched with LTR retrotransposons.
- The 3'end of non-poly(A) eRNA reveal the cleavage position depleted of secondary structure.

Introduction

In the human genome, pervasive transcription results in primary transcripts covering 75% of the genome, yet only 2% contributes to protein coding (Djebali et al. 2012). Long non-coding RNAs (lncRNAs) are the major product of such pervasive transcription. In contrast to ~20,000 protein coding genes, at least 95,000 lncRNA genes have been identified (Li et al. 2023). Notably, the transcript models of these lncRNAs are mainly constructed by assembling short-read RNA-seq data (Iyer et al. 2015; You et al. 2017; Hon et al. 2017; Jiang et al. 2019). The application of long-read sequencing has proven invaluable in accurately delineating the transcripts of lncRNA (Lagarde et al. 2017; Carbonell Sala et al. 2021). Various long-read protocols have been established to discover novel transcripts (Garalde et al. 2018; Sharon et al. 2013; Pardo-Palacios et al. 2024; Glinos et al. 2022; Uapinyoying et al. 2020). These studies primarily focused on identifying novel isoforms from the annotated gene loci instead of novel genes. Indeed, evidence of independent transcription revealed by CAP-trap protocol provides extra confidence in novel gene identification. Unlike protein coding genes, lncRNAs are usually lowly-expressed and often transcribed from the body of another gene. Detection of an accurate transcription start site (TSS) from the long-read sequencing is important to distinguish them from the genes in the same loci and confirm their independent transcription. Additionally, substantial lncRNA candidates are often excluded from standard long-read protocols which adopt poly-dT priming, due to their lack of a poly(A) tail. Therefore, a tailored protocol of long-read sequencing for lncRNAs is needed to comprehensively annotate them in the human genome.

From the approximately one million *cis* regulatory elements (CREs) predicted from the human genome (The ENCODE Project Consortium et al. 2020), more than 95% of them were enhancer-like. This suggested that enhancer-derived lncRNAs (eRNAs) represent the majority of lncRNAs. While the activity of enhancers is well-investigated (Field and Adelman 2020; Claringbould and Zaugg 2021), non-negligible functional roles of eRNAs have also been identified (Arner et al. 2015; Alvarez-Dominguez et al. 2017; Aguilo et al. 2016; Nair et al. 2019; Gorbovytska et al. 2022). In order to facilitate the functional studies on the RNA derived from enhancers, defining their gene and transcript models is necessary. Genome wide identification of enhancers or CREs is multifaceted (Panigrahi and O'Malley 2021), approaches relying on chromatin state such as H3K27Ac and/or H3K4me1 (The ENCODE Project Consortium et al. 2020), chromatin accessibility showed by DNase or ATAC-seq (Zhang et al. 2021) and the act of transcription revealed by 5' end sequencing (Arner et al. 2015; Hirabayashi et al. 2019; Kristjánsdóttir et al. 2020; Moody et al. 2022) are applied independently or integratively. When identifying enhancers by their acts of transcription, it is essential to distinguish the genuine TSS of the transcripts from technical noise. Thus, a long-read sequencing strategy that identifies enhancers with accurate 5'ends and characterises the eRNAs simultaneously is crucial in identifying novel eRNAs.

Since widespread transcription was observed from enhancers (Kim et al. 2010), transcriptional properties of eRNA have been explored by multiple studies. In general, canonical eRNAs refer to short, non-polyadenylated, unspliced, m7G-capped transcripts that are bi-directionally transcribed (2D) from enhancers (Andersson et al. 2014; Kim et al. 2010; Schwalb et al. 2016). Unidirectional (1D) eRNAs (often referred as enhancer-associated lncRNAs) were also identified and they are often longer, spliced and polyadenylated (De Santa et al. 2010; Vučićević et al. 2015; Gil and Ulitsky 2018). While transcribed 2D enhancers were shown to have less CpG islands (CGI) than promoters (Andersson et al. 2014), chromatin state-defined enhancers were observed with high GC content (Xiong et al. 2018) and functionally cooperate with proximal CGI (Pachano et al. 2021). For non-polyadenylated eRNAs, the termination of the transcripts depends on the integrator complex (Lai et al. 2015; Lykke-Andersen et al. 2021; Welsh and Gardini 2023). Early termination mediated by WDR82, PNUTS/PP1 and SET1 was also reported to control the termination of eRNAs (Austenaa et al. 2015). The half-life of eRNAs is as short as a few minutes (Schwalb et al. 2016) and are subjected to degradation by the nuclear exosome (Pefanis et al. 2015). Therefore, more eRNAs are detected via sequencing nascent transcripts (Schwalb et al. 2016) or chromatin-bound RNA (Hirabayashi et al. 2019).

Here, we developed the CFC-seq protocol for sampling full-length cDNA by adopting Cap-trapping and poly(A)-tailing (PAT) on the Oxford Nanopore technologies (ONT) platform (**Fig. 1a**). We observed the Cap-trap features, including 5' un-encoded G from the reads while these features allow us to identify genuine transcription start sites (TSSs) and transcribed *cis*-regulatory elements (tCREs) from the long-read sequencing. *In vitro* poly(A)-tailing of total RNA increased the sampling of eRNAs and lncRNAs without a natural poly(A) tail. This protocol identifies TSSs and transcript models derived from the same molecules simultaneously. Thereafter, we investigated the transcriptional properties of non-coding RNAs derived from their specific promoters and enhancers. These properties correlated with the enhancer structure including the presence of CGI and TATA box. While the relatively ubiquitous CGI enhancers generate shorter and more exosome-sensitive RNAs, TATA box derived eRNAs are longer, more stable, frequently spliced and polyadenylated. More than 27% of these TATA box-derived eRNAs are originated from long terminal repeats (LTR) retrotransposons. TATA box was shown to be depleted from the transcribed super enhancer (SE) and 2D enhancer, leading to less spliced eRNAs observed from SE and 2D enhancers. Overall, CFC-seq enabled annotation of eRNA transcript models, while the transcription of these models was correlated with the heterogeneity of the enhancer structure.

Results

CFC-seq shows the best TSS coverage and samples non-poly(A) RNA

We developed the CFC-seq protocol by combining Cap-trap technology and PAT with the ONT long-read cDNA sequencing (**Fig. 1a**). After the pre-processing of the CFC-seq data, 236 million mappable reads were obtained from the 5 cell-types (**Fig. 1b & Ext1a**). We first explored the 5' end completeness of the CAP trapping-based CFC-seq. Compared with other long-read protocols collected by ENCODE on the same iPS line (WTC11), CFC-seq showed the highest percentage of reads starting from the SCREEN cCRE regions, chromatin-open regions revealed by ATC-seq and FANTOM5 CAGE clusters (**Fig. 1c**). These results demonstrated the effectiveness of the Cap-trap technology in sampling RNAs with complete 5' ends, which reflect the TSSs. On the other hand, PAT was incorporated into the protocol to capture non-coding RNAs (ncRNAs) devoid of a natural poly(A) tail, by enzymatically adding a stretch of adenines to the 3' end of the extracted RNA. To address the coverage of non-poly(A) RNA, we performed CFC-seq with or without PAT in THP-1. As PAT allowed the protocol to reveal non-poly(A) RNA molecules, a considerable fraction of the library comprised known non-poly(A) genes including histone genes and small nucleolar RNA, suggesting the effectiveness of the PAT (**Ext_Fig. 1b**). Internal priming was known to occur in poly-dT priming protocol by binding to the adenine-rich sequence on the RNA body (Svoboda et al. 2022). We estimated the potential of internal priming of all the sampled 3' ends according to the percentage of downstream adenines (**Fig. 1d**). After excluding known non-poly(A) genes and reads potentially captured by internal priming, ~98% of mRNA reads without PAT carry a poly(A) signal (PAS) motif (**Fig. 1e**), suggesting poly-dT priming in CFC-seq capture full-length molecules effectively. From the protocol with PAT on THP-1, 96%, 80% and 55% of reads derived from mRNA, promoter-derived ncRNAs and enhancer-derived ncRNAs respectively contain the PAS motif. Therefore, CFC-seq captures non-poly(A) RNAs and many of them are non-coding.

Around 4% of the mRNA reads from the PAT protocol were found lacking a PAS motif (**Fig. 1e**). We reasoned that they are the nascent RNAs and identified the location of these captured 3' ends from the coding gene fraction (**Ext_Fig. 1c**). Initially, we explored the mRNA content from the libraries without performing PAT which represents the technical limitation of a poly-dT protocol. From the 2% mRNA reads without a PAS motif, ~28% of these 3' ends are consistent with the 3' ends of GENCODE transcripts, suggesting that these RNAs may acquire the poly(A) tail through less conserved PAS motifs or non-canonical pathways. The other 72% potentially resulted from internal priming at a subthreshold level of downstream adenine nucleotides (**Ext_Fig. 1d upper**). To assess the side effects of including PAT, we subtracted the 3' ends that were captured by the non-PAT THP-1 libraries from the PAT THP-1 libraries. These PAT-specific 3' ends constitute about 2% of the libraries according to the coding gene fraction. They are independent from internal priming (**Ext_Fig. 1d lower**) and were dispersed across the transcript body revealing a relatively higher frequency at the donor sites (**Ext_Fig. 1e**), resembling nascent RNA along the RNA processing pathway. Next, we determined if including PAT increased the ncRNA content. Comparing the lncRNAs sampled from the non-PAT and PAT THP-1 libraries at the read-level, the fraction of lncRNAs increased from 8% to 13% (**Ext_Fig. 1f**). Using the raw output from SALA, 25,612 ncRNA gene models are uniquely identified from the PAT library, versus only 6,963 from the non-PAT library (**Ext_Fig. 1g**). Overall, PAT incorporation increased the ncRNA content and successfully sampled non-poly(A) RNAs which are mainly ncRNAs, with a side effect of introducing ~2% of reads from the transcriptionally incomplete nascent RNAs.

PAS motif and random forest comprehensively annotate the 3' end

To quantitatively assess canonical PolII termination events from the transcript models, we introduced a sequence-based random forest classifier trained on highly curated poly(A)-positive 3' ends

as a ground truth. These 3' ends were derived from a FLAM-seq-based dataset (Legnini et al. 2019; Alfonso-Gonzalez et al. 2023), where transcript ends were defined by the presence of non-templated poly(A) as no oligo-dT priming was involved in the library preparation. Trained on sequence features from this dataset, the classifier was applied to the 3' ends of all the CFC-seq reads and to the GENCODE transcript models, resulting a score for likelihood of each 3' end to derive from a canonical PolII termination event. This poly(A) score revealed high performance with AUC ~ 93% from the THP-1 non-PAT and PAT-specific 3' ends (**Ext_Fig. 1h**). Thresholds for defining poly(A)-positive and -negative were extracted from the ROC curve with 95% confidence. These read-base analyses including the presence of PAS, poly(A) score and identifier were summarised in **Table S1** and were transferred to the 3' ends of the final transcript models.

CFC-seq identifies genuine TSS to indicate primary transcription

A key feature for defining genuine TSSs is the presence of an unencoded G immediately upstream of the 5' end of the mapped read (Cumbie et al. 2015), a characteristic seen in approximately 80% of qualified CFC-seq reads (**Ext_Fig. 2a & 3h**). Combining the CFC-seq libraries of the two differentiation time courses, 159,116 genuine TSS clusters were identified by *SCAFE* (Moody et al. 2022). These TSS clusters were extended and merged into 73,328 transcribed CREs (tCREs) with a median length of 501 bp (**Fig. 2a**). Integrating their genomic locations with the SCREEN cCREs (The ENCODE Project Consortium et al. 2020) revealed 30,695 promoter-like and 32,416 enhancer-like tCREs. Heatmap showing their expression level highlighted the cell-type specific nature of the enhancer-like tRCEs. (**Fig. 2b**). Proximal TSS clusters within 75 nt were merged into extended 5' end clusters (ex5_clusters) to define 5' end transcript classes in SALA and identify strand-specific summits (**Fig. 2a**). Next, the tCREs observed from the Neuron series were intersected with single-cell ATAC-seq peaks derived from the same samples, showing that majority of tCREs overlap with accessible chromatin regions (**Fig. 2c**). From the tCREs without cCRE support (unclassed), ~29% of them are proximal to the starting sites of GENCODE transcript models (**Ext_Fig. 2b**). Apart from these annotated tCREs, around 36% of ATAC-supported unclassed tCREs are specific to NSC, which was not an annotated cell-type in the SCREEN. To further explore the content of the unclassed tCREs, these tCREs were intersected with the *ChromHMM* chromatin states of the Neuron series (Yip et al. 2024). About 52% of non-ATAC unclassed tCRE intersected with the *ChromHMM* promoter or enhancer regions (**Ext_Fig. 2c**). The remaining 48% show a higher incidence of tCRE summits originating from exons (n = 104) and 3' UTRs (n = 246) of annotated genes, regions previously reported as capped 3' UTR-derived RNAs (Haberman et al. 2023). The tCREs with and without GENCODE, cCRE or ATAC support are summarised in **Table S2**. Notably, transcript models originating from unclassified tCREs were included in the final transcriptome, but labeled as non-independent transcripts when lacking ATAC support for open chromatin. Overall, our approach simultaneously identified TSS clusters and full-length long-read molecules, facilitating the precise definition of independent transcription and linking transcript models to well-defined tCREs and promoter types.

The tCRE's summit represents the major TSS of a tCRE. They were extracted and extended in a strand-specific manner to intersect with various promoter features. The genomic location of these tCREs resemble features of promoter, including high GC content and conserved sequence while the summits were consistent with the feature of transcription start sites revealed by the position of TATA box and initiator (**Ext_Fig. 2d**). Enhancer-like tCREs were shown to have all these properties at lower but comparable levels as the promoter-like tCREs, suggesting the identified enhancer-like tCREs are capable to initiate transcription. It is noted that enhancer-like tCREs show enrichment of GC content and CpG island (CGI) downstream of the summit (**Ext_Fig. 2d**). As CGI is more accessible by various regulatory factors, this distribution implies that CGIs may play a crucial role in enhancing the activity of

target promoters rather than driving the transcription of the eRNAs. Coincidently, CGIs proximal to poised enhancers were found to amplify enhancer activity towards CGI promoters in mouse embryonic stem cells (Pachano et al. 2021). To this end, we subset the enhancers which overlap with significant H3K27me3 repressive histone mark identified from our iPSC CUT&Tag data. This group of enhancers represent "transcribed poised enhancers". It is noted that >50% of these poised enhancers produce eRNAs in the iPSC while others are produced from NSC, Neuron or THP-1 (**Fig. 2d**). Poised or bivalent (overlap with both H3K27ac and H3K27me3) enhancers were found significantly enriched with the presence of upstream CGI (uCGI) and downstream CGI (dCGI) (**Fig. 2e**). On the contrary, TATA box enhancers were depleted from H3K27me3 but enriched with active state and the state devoid of both H3K27ac and H3K27me3 (**Fig. 2e**).

Bidirectional enhancers are more conserved, due to CpG islands

Bidirectional transcription was reported as a feature of enhancers (Andersson et al. 2014). Using the 5' end signal from the qualified reads (**Ext_Fig. 3a**, Methods), we have identified 12,984 2D enhancer-like tCREs and 9,635 2D promoter-like tCREs (**Fig. 2d**). Super enhancer (SE) represents clusters of active CREs in genomic proximity that control cell identity (Hnisz et al. 2013). Using *ROSE* (Lovén et al. 2013), we identified 1932, 1278 and 1790 SE regions from iPSC, NSC and Neuron respectively (**Ext_Fig. 3b**), with median SE sizes from 34 to 59 kb (**Ext_Fig. 3c**). Enhancer-like tCREs were then classified as SE or typical enhancer (TE) in each of the three cell-types (**Fig. 2d**). We then explored if directionality and SE clustering showed differences in CRE conservation. Using the sequence 500 bp upstream the tCRE summit, 2D tCREs are significantly more conserved than 1D tCREs (**Fig. 2e & Ext_Fig. 3d**). Super enhancer was reported as more conserved over other enhancers (Khan et al. 2018) in a context of chromatin state-defined enhancers. In our study of "transcribed" enhancers, TE is comparably conserved as SE (**Fig. 2e & Ext_Fig. 3e**). We reasoned that the presence of uCGI/dCGI contributed to the conservation observed in the 2D tCRE and classified the tCRE for the enrichment of CGI and for containing a TATA box (Methods). For both enhancer and promoter, 2D tCREs are enriched with CGI while 1D tCREs are enriched with TATA box (**Ext_Fig. 3f**). By comparing the tCREs with or without CGI and TATA box, both 1D and 2D enhancers carrying CGI are significantly more conserved (**Ext_Fig. 3g**). These results indicated that presence of the CGI in the enhancers contributed to the sequence conservation. It is noted that some distal tCREs are strand-independent and the strand with greater summit count was chosen for the analyses. To obtain the signal from both strands and synchronise the analyses with the transcript models, above analyses were also performed on the ex5_clusters (n = 60351) that linked to the finalised transcript models. These results are consistent with the results derived from the tCREs (**Supplementary Fig. S1 & Table S2**).

CFC-seq identifies repetitive element-enriched tCREs

Employing the same CAP-trapping molecular protocol, identification of tCRE is highly comparable between CFC-seq and CAGE. These include the percentage of reads with un-encoded G (**Ext_Fig. 3h**) and the position of single-nucleotide CTSS (**Ext_Fig. 3i**). Specific to the CFC-seq, a portion of reads (~30%) were revealed with a relatively long 5' region unaligned to the genome. These reads were considered as non-qualified for TSS clusters identification with a cutoff of unaligned length > 3 nt. The transcript models generated by them independently will be filtered in the latter steps if the 5' end locates outside the defines ex5_cluster. Further, pool identification of tCREs with both CFC-seq and CAGE (**Ext_Fig. 2a**) revealed the similar tCRE content and quantification (**Ext_Fig. 3j-k**).

Since long-read sequencing is suited for more accurately distinguishing genomic location of the repetitive elements (Maeng et al. 2023), we compared the identification rate of TSS clusters among CFC-seq, paired-end and single-end CAGE. CFC-seq shows a better performance in identifying TSSs

covered by repeat sequence, compared to CAGE at different MAPQ thresholds (**Fig. 2h**). Further analysis revealed that these repetitive element-enriched TSSs specifically identified by CFC-seq are evolutionarily younger (**Fig. 2i**), concurring with a recent publication (Lee et al. 2023). With the longer read length from the CFC-seq, TSSs covered by repetitive elements are unambiguously and reliably annotated. From the 73,328 tCREs, 64.7% overlapped with any of the highlighted 7 families of repetitive elements (**Ext_Fig. 3l**). Among these tCREs, DNA transposons, LINEs, SINEs and LTRs are enriched in enhancer-like tCREs (**Ext_Fig. 3m**). From the enhancer-like tCREs, LINEs and LTRs are enriched in 1D-enhancers. Conversely, short tandem repeats (STRs) showed enrichment in promoter-like CREs, with CGG/CCG being the most notable motif.

SALA identified 39,425 novel genes from CFC-seq

We subjected approximately 236 million mapped alignments to *TranscriptClean* (Wyman and Mortazavi 2019) to correct splice junctions. These corrected alignments were then processed with SALA to collapse the reads into transcript models (**Fig. 3a**; Methods). The *SALA* raw output excluded the transcript models affected by internal priming and un-detected transcripts from the reference (GENCODE v39) and included a temporary assignment of gene annotation. To establish a confident assembly, we kept only the transcript models with their starting sites linked to the confident ex5_clusters supported by *SCAFE* and applied a tailored filter (**Ext_Fig. 4x**, Methods) Notably, we did not directly require reproducibility from the transcript models but rely on the reproducibility of the TSS clusters (SCAFE regression probability ≥ 0.5). The finalised transcript and gene models were shown in **Table S3**. Next, we classified the transcript models for their promoter type originated from the SCREEN cCRE-classed tCRE, according to their linked TSS clusters. Promoter types at the gene level were assigned according to the transcript ratio (Methods). From this finalised assembly with 194,138 transcript models, 47,984 known full-length transcripts (ENST) were detected with the identification of 146,154 novel transcripts (**Fig. 3b**). The transcript models were tested for their coding potential by *CPAT* (Wang et al. 2013). Over 80% of the group “transcripts from novel genes” were determined as non-coding, compared to around 25% for the ENST (**Fig. 3c**). The novel transcript models of this final dataset are all supported by *SCAFE* (**Ext_Fig. 4a**) while the 3’ ends were annotated by cluster confidence (≥ 5 counts or covering GENCODE 3’ end) and poly(A) possibility (presence of PAS or poly(A) prediction) (**Ext_Fig. 4b**).

To benchmark *SALA*'s performance, the alignments were also processed with *TALON* (Wyman et al. 2019) and *IsoQuant* (Prjibelski et al. 2023). Transcript models affected by putative internal priming were filtered with the same condition from each assembly. In addition to the *SALA* final transcript models, we prepared the raw (**Table S4**) and read-filtered (**Table S5**) transcript models for this comparison (**Ext_Fig. 4c**). These tools are comparable in terms of detection of known transcript models (ENST) (**Ext_Fig. 4d**), except *IsoQuant* excluded models partially matched to ENST. *SALA* was developed to identify completely and partially matched ENST that are distinguishable. The transcript models identified using these different tools with different conditions were classified universally by *SQANTI3* (Pardo-Palacios et al. 2024). *SALA* read-filtered models showed similar percentage of transcript type classification with *TALON* read-filtered models, where both adopted similar assembling algorithm including the use the *TranscriptClean* to pre-adjust the alignments (**Ext_Fig. 4e**). The *SALA* final transcriptomes showed a greater proportion of “intergenic” and “antisense” transcript classes, revealing our strategy in identifying novel confident ncRNAs. Transcripts from novel genes were observed from the raw output of *TALON* and the raw transcriptome of *SALA*, while many of them were maintained by *SALA* final transcriptome for genuine TSS support. (**Ext_Fig. 4f**). While *SALA* showed the highest discovery rate of transcripts from novel genes, a comparable rate of 5’ end support by ATAC open-chromatin regions and SCREEN cCRE epigenetic regions was observed across all the tools (**Ext_Fig. 4g**). When compared with the CFC-seq derived *SCAFE* TSS clusters, only a small fraction of *TALON*

transcript models was supported. This revealed a different strategy in deciding the model 5' ends where *TALON* prioritised reference 5' ends but *SALA* opted to use data-driven 5' ends obtained from the CAP-trap protocol. In summary, *SALA* outperformed the other tools in obtaining transcript models of novel genes, which are predominantly ncRNAs. This information is useful for the in-depth characterization of regulatory regions of the genome.

About 93% of novel genes were classified as ncRNAs

From the group "transcripts from novel genes", transcript models with CPAT score < 0.364 were classed as non-coding (Wang et al. 2013). Novel genes containing > 50% complete reads assigned to non-coding transcripts were classed as non-coding genes. Non-coding genes with median read lengths of all non-coding transcripts ≥ 200 bp were classed as lncRNAs. This resulted in 35,848 novel lncRNA genes and 1,062 short_ncRNA genes (**Ext_Fig. 4h**). Only non-coding transcripts from the non-coding genes were considered as ncRNAs in the downstream analyses. Additionally, 2,948 novel lncRNA transcripts were identified from the "New isoform of GENCODE model". In summary, we have identified 115,275 novel ncRNA transcripts and detected 6559 known lncRNA transcripts (**Ext_Fig. 4i**). These 121,834 ncRNA transcripts were then further subclassed according to the genomic location of GENCODE coding-genes and pseudogenes (**Fig. 3d**, Methods) for the downstream analyses. From the 39,425 novel genes, 35,848, 1,062 and 2,515 genes were defined as lncRNA, short_ncRNA and potentially coding RNA respectively. The ncRNA genes were subclassed according to the transcripts (Methods).

CFC-seq identified 22,506 novel genes out of multiple databases

In order to linked the novel genes identified from CFC-seq with other key databases, we compared the novel transcript models with Refseq, the latest version (v45) of GENCODE, and two lncRNA resources: FANTOM CAT (Hon et al. 2017) and LncBook (Li et al. 2023). At the transcript level, about 6% of the novel transcript models were annotated by at least one of these databases, with 4.4% contributed from the FANTOM CAT (**Ext_Fig. 5a**). To identify transcript models from these 4 databases belonging to the CFC-seq novel genes, transcript models sharing any features (5' end clusters, 3' end clusters and any splicing junction) or having 10% exon overlap with the CFC-seq models were grouped as the same genes. At the gene level, 57% of our novel genes are unannotated from any of the 4 databases (**Ext_Fig. 5b**). Although FANTOM CAT database contains less ncRNAs than LncBook (**Ext_Fig. 5c**), it showed the highest rate of overlap with CFC-seq. This is due to the requirement of TSS in identifying ncRNAs for both FANTOM CAT and CFC-seq. These novel genes observed from CFC-seq are relatively more cell-type specific and lowly expressed (**Ext_Fig. 5d**). The linked transcript ID from these 4 databases to the novel transcript and genes in this study are listed in **Table S6**.

Next, we quantified the short-read RNA-seq of the Neuron series by *kallisto* (Bray et al. 2016), using the final transcript models plus all the GENCODE models (v39) as reference. Over 80% of the novel transcript models and over 97% novel gene models were detected from the short-read RNA-seq data (**Ext_Fig. 5e**), suggesting most of the novel models can be observed from an independent sequencing technique. The undetected novel models revealed with similar proportion of ncRNA subclass as the detected ones (**Ext_Fig. 5f**) while they are expressed lower than the detected ones.

Enhancer transcribes ncRNA with a greater length than previously reported

The ex5_clusters (genuine TSS cluster with extension) were divided into 4 groups according to the promoter types, which were inherited from the tCRE as promoter-like, enhancer-like, CTCF-alone with ATAC support and unclassed with ATAC support. Additionally, ex5_clusters producing GENCODE coding genes alone were grouped as a control. These resulted in 9,749 ex5_clusters with mRNA

outputs, 17,045 promoter-like ex5_clusters with ncRNA outputs (p_ncRNA), 28,718 enhancer-like ex5_clusters with ncRNA (e_ncRNA), 194 ATAC-supported CTCF-alone ex5_clusters with ncRNA (CTCF_ncRNA) and 4,645 ATAC-supported unclassed ex5_clusters with ncRNA (other_ncRNA).

We first investigated the transcript features from the RNAs grouped by their ex5_clusters. The median lengths of mRNAs, p_ncRNAs, e_ncRNAs and other_ncRNAs, were 2120 nt, 617 nt, 659 nt and 722 nt respectively (**Fig. 4a & Ext_Fig. 6a**). Here, the observed length of e_ncRNA with 659 nt is longer than previously reported, with a range of 346 - 449 nt (Andersson et al. 2014; Hou and Kraus 2022). Definition of e_ncRNA from this study is the ncRNAs derived from enhancers which are identified from genuine TSS integrated with enhancer-like chromatin signals (SCREEN). Commonly used representation of enhancers or CREs included ATAC-seq derived accessible CREs distal from GENCODE transcripts (Zhang et al. 2021), 5' end sequencing protocol derived tCREs distal from GENCODE transcripts (Moody et al. 2023), distal bi-directionally transcribed RNAs (Andersson et al. 2014) and chromatin state-defined CREs (Khan et al. 2018). We grouped the ex5_clusters according to these criteria and measured the length of ncRNA products. This resulted in the median lengths between 652-704 nt (**Ext_Fig. 4b**), while these subgroups contain trace amounts of other_ncRNA and p_ncRNA (**Ext_Fig. 4c**). With the same groupings, 22% e_ncRNAs were found to be spliced (**Fig. 4b**) compared to 5% in a previous study (Andersson et al. 2014). From the other enhancer definitions, 19-26% of eRNA show splicing activity (**Ext_Fig. 4d**). Overall, CFC-seq revealed eRNA with a longer transcript length and more frequent splicing activity than previously observed.

Downstream CpG islands and TATA box correlate with eRNA length and splicing

As the promoter structure was found to affect RNA splicing (Cramer et al. 1997; Pandya-Jones 2011), which affect RNA length (described in next section), we investigated if the genomic properties of enhancer correlate with RNA length and splicing activity. These properties include bidirectionality, being within the SE clusters, sequence conservation, and presence of TATA box, uCGI and dCGI, of the ex5_clusters. Among them, presence of uCGI and dCGI correlated with shorter eRNA significantly while presence of TATA box and more conserved sequence correlated with longer eRNA (**Fig. 4c**). While bidirectionality and being within SE clusters showed subtle effect on the length of e_ncRNAs, they significantly correlated with less splicing events (**Fig. 4d**). Additionally, uCGI and dCGI correlated with less splicing events while TATA box correlated with more splicing events (**Fig. 4d**). Since the regulatory elements (TATA box, uCGI and dCGI) sometimes coexist in an enhancer (**Ext_Fig. 3f**) and may affect each other, ex5_clusters with only one of these regulatory elements were grouped and compared. Analyses of transcript length and splicing event verified the role of dCGI for shorter length and less splicing events and TATA box for longer length and higher splicing activity (**Fig. 4e-f**). These analyses were also performed on the p_ncRNAs with less obvious enrichment except for the presence of dCGI showing shorter ncRNAs (**Ext_Fig. 6e-h**). To better illustrate the genomic properties of dCGI and TATA box enhancer, we explored their enrichment against the other genomic properties from the ex5_cluster. Both 2D- and super-enhancer showed depletion of TATA box but subtle enrichment of dCGI (**Fig. 4g**), suggesting the lack of TATA box maybe a factor of the lower splicing activity in 2D- and super-enhancer. Further, ubiquitous enhancers (≥ 1 read from all 5 cell types) are enriched with uCGI/dCGI and depleted with TATA box, consistent to our previous study (Carninci et al. 2006). Overall, enhancer structure (dCGI and TATA box) was found to affect RNA splicing and length while these effects extended to enhancer types (bidirectionality and SE clustering) due to their depletion of TATA box.

Presence of poly(A) tail and splicing enriched with longer ncRNA

It is noted that dCGI directly affect sequence conservation downstream of the TSS (**Fig. 4g**) and the sequence composition of the transcript body. To identify the factors affecting the transcription from

the transcript body, the 4 classes of RNA (mRNA, p_ncRNA, e_ncRNA and other_ncRNA) were grouped at the transcript model level. Protein-coding mRNAs (n = 28,230) were defined as any complete and known transcripts annotated as coding genes by GENCODE. Non-coding RNAs (n = 121,834) were non-coding transcript models derived from non-coding genes annotated in this study or by GENCODE (**Ext_Fig. 4i**). The presence of poly(A) tail for each transcript model was determined by having the poly(A) prediction positive or the presence of PAS from 5 to 50 nt upstream of the 3' end. The occurrence of splicing and poly(A) correlated with longer eRNA (**Fig. 4h**). These transcripts were grouped into the ex5_clusters considering the weighted frequency of transcript models that are poly(A) positive. Dividing the ex5_clusters with the regulatory elements, the ncRNA content derived from dCGI enhancer constituted significantly less poly(A) RNA compared to that of TATA box enhancer (**Fig. 4i**). This result is specific to enhancer and was not observed from p_ncRNA.

Chromatin-bound and exosome sensitivity

As eRNA degrades rapidly in general, we adopted CFC-seq on the chromatin-bound fraction of iPSCs (**Ext_Fig. 1a**) in order to increase eRNA recovery and assess which RNAs that are specifically enriched from this fraction. We applied *SALA* on the chromatin-bound iPSC using the finalised transcriptome with all the GENCODE transcript models as the reference, where newly identified transcript models were labelled independently from the main dataset (**Table S7**). This dataset revealed an addition of 10,857 ncRNA genes while 6,476 of them are derived from enhancer (**Ext_Fig. 6i**). Using exosome knockdown RNA-seq data from another line of iPSC (Yip et al. 2022), chromatin-bound eRNAs were found to be more exosome-sensitive (**Ext_Fig. 6j**). Further, the genes were grouped by ex5_clusters where one gene was allowed to have multiple ex5_clusters but only those ex5_clusters with one gene were kept. The exosome sensitivity was tested against different enhancer features from the ex5_clusters. Enhancers with uCGI, dCGI, 2D, being clustered as SE and ubiquitously active generate eRNAs that are significantly more sensitive to exosome degradation, while TATA box enhancer generates eRNAs that are significantly less sensitive (**Fig. 4j**). These results agree with our discoveries on the splicing, transcript length, and the presence of ploy(A) tail between CGI and TATA box enhancers. Overall, we grouped the ex5_clusters according to the enhancer structure and observed that dCGI enhancers transcribed shorter RNA with high sensitivity to exosome degradation while TATA box enhancers are more frequently spliced, polyadenylated, longer and more stable.

Non-coding RNA showed low splicing efficiency

To extend our observation of ncRNA splicing activity, we investigate the confidence, recurrence and efficiency of the splicing junctions (SJs). The SJs observed from our models were annotated using junctions confidently observed from our short-read sequencing (Methods), junctions reported by Intropolis (Nellore et al. 2016) and GENCODE (v39) junctions (**Ext_Fig. 7a**). Considering the SJs from the final transcriptome, majority of them are annotated by GENCODE while the others are largely supported by Intropolis and short-read SJs (**Fig. 5a**), leaving 11.8% SJs non-supported. For these remaining novel SJs, their counts and base-calling quality scores nearby the junctions (± 3nt) were used as an indicator of technical confidence. With the summarised base-calling score >=10 and counts >=3, 8.9% (both canonical and non-canonical) of the novel SJs was considered confident (**Ext_Fig. 7b**). Overall, 89.3% SJs observed from the final transcriptome were either supported by external databases or passing the quality score. The quality of these SJs were incorporated to the transcript models that were grouped as mRNA, p_ncRNA, e_ncRNA and other_ncRNA as described above. ~90% of the spliced ncRNAs have all of their SJs considered as confident (**Ext_Fig. 7c**). In addition, most of SJs observed from the p_ncRNA and e_ncRNA are canonical (**Ext_Fig. 7d**), echoing the findings of a

previous study (Gil and Ulitsky 2018). Overall, we showed that SJs identified by CFC-seq and SALA are largely supported by external datasets while SJs from ncRNAs are comparable to SJs from mRNA.

Although the SJ motif class from ncRNAs are comparable to that of mRNA, not all the junction splices at the same efficiency (Melé et al. 2017). To address this, we calculated the splicing efficiency of all the junctions and summarised the normalised splicing efficiency at the gene level (**Fig. 5b**). Across several different cutoffs of occurrence as the SJ inclusion criteria, splicing efficiency of ncRNAs is significantly lower than mRNAs, where p_ncRNAs showed a significantly lower efficiency compared to the e_ncRNAs (**Fig. 5c**). Since the accuracy of this splicing efficiency highly depends on the expression level, splicing efficiency of those lowly expressed RNAs was not included. In order to improve the coverage of the analysis, *spliceAI* (de Sainte Agathe et al. 2023) was incorporated as a sequence-base estimation of splicing probability for all the SJs observed. The splicing efficiencies for donor site, acceptor site and junction are correlated with the *spliceAI* score (**Ext_Fig. 7e**). The normalised *spliceAI* score at the gene level resulted with the same observation of less effective splicing event from the ncRNAs (**Ext_Fig. 7f**).

The gene-based splicing efficiency was transferred to the ex5_cluster groups where one gene might assign to multiple ex5_clusters but only the clusters with one gene were kept (Methods). After dividing the ex5_clusters with promoter structures, splicing efficiency of RNA derived from dCGI and uCGI enhancers were found significantly lower than that of TATA box enhancers (**Fig. 5d**). This phenomenon was consistent across e_ncRNA and p_ncRNA (**Fig. 5e**). Since enhancers were shown to enrich with TATA box while promoters were enriched with CGI (**Ext_Fig. 3f**), these promoter structures contributed to the observation of the greater splicing efficiency of e_ncRNA than p_ncRNA in **Figure 5c**. Splicing efficiency for each gene model is listed (**Table S8**).

The 3’ end cleavage site of eRNA shows depletion of secondary structure

From the CFC-seq data, SALA has successfully identified 6,683 poly(A) and 2,239 non-poly(A) 3’ ends from e_ncRNAs having at least 3 counts at the 3’ ends. The 3’end processing of eRNA is partially resolved (Lai et al. 2015) (Welsh and Gardini 2023), where the integrator complex was identified to cleave the eRNA 3’ end. However, the cleavage position or the RNA motif at the cleavage site is unknown. From the recursive non-poly(A) 3’ end of ncRNA, we observed a very different sequence composition from the ones with poly(A) (**Fig. 5f**). In addition to the absence of PAS motif, a relatively high content of cysteine was observed upstream the 3’ end. As the RNA secondary structure was reported to hinder 3’ end cleavage in the canonical poly(A) pathway (Klasens 1998), we measure the secondary structure score by RNAfold (Lorenz et al. 2011) surrounding the 3’ ends of poly(A) and non-poly(A) eRNAs. Similar to the 3’end of poly(A) eRNA, we observed a valley of secondary structure score at the 3’ end of non-poly(A) eRNA, suggesting this as a preferred feature of cleavage site (**Fig. 5g**). Further, the sequence of recursive non-ploy(A) surrounding the 3’ end was subjected to motif search. A C-rich motif and a CA dinucleotide repeat were identified directly upstream and downstream of the 3’ end.

Expansion of non-coding sequence for functional implication

Many annotated eQTL and GWAS SNPs are located outside the protein coding genes and their promoters. These SNPs may affect the nearby CREs or ncRNAs in regulating the protein coding genes. The eQTL and GWAS data were obtained from eQTL Catalogue (Kerimov et al. 2020) and CAUSALdb (Wang et al. 2019) respectively. To investigate if the novel transcript models identified in this study harbour some of these SNPs, they were intersected against GENCODE v39 and our novel models. The exons specific to the novel models revealed 21% and 31% increase of coverage of eQTL and GWAS

SNPs respectively compared to the GENCODE models (**Fig. 6a**), while the majority of the increase detected from the transcript from novel genes (**Fig. 6b**). To explore the effect of these SNPs on the SJs of the transcript model, the donor and acceptor sites were tested against these SNPs in *spliceAI*. The novel transcript models contributed to 4.9-17% increase of splicing interference mediated by eQTL and GWAS SNPs. In **Figure 6b**, only those SNPs reduce the splicing potential were counted while the full list with a cutoff of delta score at 0.5 was shown in **Table S9**. Overall, the novel transcript models identified from this study increase the coverage of annotated eQTL and GWAS SNPs by at least 20% while the splicing interference represents one of the functional mechanisms of these SNPs.

CGI enhancers show greater chromatin connectivity

Due to the different properties of ncRNAs and enhancers derived from dCGI and TATA box, we extracted the number of SNPs from different enhancer structure (**Ext_Fig. 8a**). Enhancer with only dCGI was found to have the highest hit rate of eQTL SNPs against background SNPs, with about 2-fold as the TATA box enhancer (**Fig. 6c**). The eQTL results were similar in exons, where the relatively short eRNAs overlap more with the dCGI that regulates enhancer activity. This result may imply a more established enhancer function from the dCGI enhancer. To explore the enhancer function contributed by different enhancer structures, the interaction between these enhancers and protein-coding gene promoters was predicted by Hi-C performed on the Neuron series. Using the interaction from all 3 cell types, the number of coding gene promoters in contact with dCGI and uCGI enhancers were found significantly greater than that of TATA box enhancer (**Fig. 6d**). This result is consistent when the analysis performed at each cell type separately (**Ext_Fig. 8b**). This finding echoed the observation from the eQTL SNPs and suggested a broader enhancer role of uCGI/dCGI enhancer than TATA box enhancer. This result also concurs with a previous study that CGI enhancers showed higher Hi-C contacts than classical enhancers, where enhancers were epigenetically defined (Bell and Vertino 2017). Further, chromatin accessibility and GC content of the Hi-C contact regions revealed that regions with uCGI/dCGI enhancer were more accessible with higher GC content while both were correlated with contact frequency (**Ext_Fig. 8c-d**). As expected, SE was shown with higher chromatin connectivity (**Ext_Fig. 8e**). This is consistent with our results showing that uCGI/dCGI is enriched with SE (**Fig. 4g**). In addition, uCGI/dCGI enhancers were shown to interact with CGI promoters more frequently (**Ext_Fig. 8f**), which is consistent to a previous study (Pachano et al. 2021).

TATA box derived eRNAs enriched with LTRs

Given the significant contribution of transposable elements to the human genome, and the enrichment of enhancer-like tCREs with SINEs, LINEs and LTRs (**Ext_Fig. 3m**), we reasoned that TATA box-derived eRNAs might be influenced by these transposable elements. We intersected repetitive elements with both the ex5_clusters and transcript models (merged exons) and categorized them as dCGI, uCGI, TATA, and Null. Analysis of both ex5_clusters and transcript models revealed that LTRs were specifically enriched in TATA box enhancers and depleted from uCGI/dCGI enhancers (**Fig. 6e**). This enrichment pattern was consistent across the p_ncRNA, CTCF_ncRNA, and other_ncRNA groups (**Ext_Fig. 8g**). Notably, LTR elements were highly enriched in the other_ncRNA category (unclassified by SCREEN promoter-typing) with TATA box enhancers accounting for 57% (n = 342), suggesting that these elements are a prominent transcription source within this group (**Ext_Fig. 8g**). With a threshold of at least 200 nt overlap, approximately 27% (n = 551) of RNAs transcribed from TATA box enhancers were found to originate from LTRs (**Fig. 6f**). Within the LTR families, ERV1 emerged as the predominant contributor to TATA box eRNAs (**Ext_Fig. 8h**). Quantification of ex5_clusters using the RLE-normalised TSS count revealed the higher cell-type specificity of the TATA box enhancers, but TATA box enhancers

were observed across different cell types (**Ext_Fig. 8g**). Specifically, the LTR subgroup of TATA box enhancers showed pronounced enrichment in iPSCs and NSCs (**Fig. 6g**).

Discussion

This study introduced a novel long-read RNA-sequencing protocol accompanied with the SALA assembler. We demonstrated the annotation of both TSS and full-length transcript model from the same RNA molecule effectively identified novel genes with confident TSS. The incorporation of PAT was shown to successfully include non-poly(A) RNA and almost double the number of ncRNA reads.

Transcription from enhancers and the presence of eRNAs in the human genome are well-established. However, their low abundance hinders transcript model construction using short-read RNA-seq. Although CFC-seq aids in annotating eRNAs with splicing junctions and relatively precise transcript end sites, creating reliable transcript models with limited sequencing depth remains challenging. Since the majority of the eRNAs lacks a canonical polyadenylation signal (**Ext_Fig. 4c**), the concept of typical complete mRNA model with both a defined TSS and poly(A) tail, often does not apply to eRNAs. The use of 3' end clusters to observe recursive transcript end sites provided an additional layer of confidence, albeit hindered by the low expression level of eRNAs. From the analyses showing higher content of "nascent" RNA from the chromatin fraction also suggested that most eRNAs are "incomplete" in nature. Importantly, the eRNAs identified in this study were derived from total RNA under standard cell culture conditions, likely representing the natural, full-length format of these transcripts.

Identification of enhancers in a cell type is generally based on chromatin state and transcriptional activity with transcribed enhancers typically being more active (Andersson et al. 2014). Enhancers are known to exhibit high GC content and are frequently located near orphan CGIs though transcribed enhancers tend to be depleted of CGIs and simple repeats. Because TATA box identification depends on the reference location of the TSS, TATA box-associated enhancers were reliably observed only when transcription was detected. The precise identification of eRNA TSSs also allowed us to observe that enhancer CGIs are often positioned downstream of the TSS, distinguishing them from promoters. By integrating transcriptional activity and chromatin state, our study systematically elucidates these enhancer structures and characterizes the RNA produced from them. From the set of transcribed enhancers, we identified both CGI-specific and TATA box-specific enhancers. TATA box enhancers were observed to produce longer, more stable, and more extensively spliced and polyadenylated eRNAs, though they were less connected compared to CGI enhancers (**Fig. 6h**). CGI enhancers appeared more ubiquitous across cell types, whereas TATA box enhancers were more cell-type-specific, suggesting that ubiquitous enhancers are more connected in enhancer networks. Interestingly, CGI enhancers were enriched with H3K27me3 marks, indicating poised or bivalent states, which implies that they may represent a more readily activated enhancer class. In contrast, TATA box enhancers were characterized as either active or lacking both H3K27Ac and H3K27me3 marks, suggesting that their activation may involve a complex mechanism—potentially initiated by pioneer factor binding followed by subsequent transcription factor regulation.

Based on TSSs identified by CFC-seq, eRNAs were observed to be transcribed upstream of CGIs (**Ext_Fig. 2d**), leading to the identification of a distinct group termed dCGI enhancers. This enhancer group produces the most exosome-sensitive and shortest eRNAs among the identified enhancer classes, indicating that dCGIs may limit eRNA activity while simultaneously enhancing regulatory potential. Further research is needed to elucidate the mechanisms underlying this phenomenon. Conversely, TATA box enhancers produce longer and more stable eRNAs, which enables them to explore the nuclear space in search of target sequences, potentially even in *trans* (Harrison and Bose 2022). A notable proportion of these eRNAs were originated from LTR retrotransposons,

suggesting that their structural characteristics may contribute to this stability and nuclear exploratory function. LTR-derived RNAs have been shown to play roles in key biological processes such as pluripotency (Fort et al. 2014; Grow et al. 2015), neurogenesis (Wang et al. 2020) and neuronal diseases (Dawson et al. 2023).

Limitations

We have performed size selection in the CFC-seq protocol to remove primers and adjust the overall read length distribution in each library, where the final libraries are similar to the initial cDNAs. Since the size selection may introduce bias to the size distribution of different RNA types, our conclusion on RNA length is limited by this protocol. The CFC-seq libraries from the THP-1 were mainly used as protocol development for the effect of poly(A)-tailing. We have included them in building transcript models but avoid comparing them with the Neuron series samples.

Although dCGI, uCGI and TATA box enhancers were identified from this study, majority of the transcribed enhancers (Null, 81%) lacks these enhancer structures. Interesting, about 40% of these “Null” enhancer contains transposable elements (LTR, SINE and LINE) in the RNA body. Further investigation with alternative approaches is required to thoroughly understand the nature of these enhancers.

Acknowledgements

As part of the FANTOM6 project, this study is supported by many other studies of the project. We thank all the member in the FANTOM consortium for excellent support, discussion and inspiration. Sequencing of short-read RNA-seq was supported by RIKEN IMS sequencing platform.

Fundings

- FANTOM6 was made possible by a Research Grant for RIKEN Center for Integrative Medical Sciences (IMS) from MEXT, Japan.
- The Hi-C data pre-processing was enabled by resources in project NAISS 2024/22-849 provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX, funded by the Swedish Research Council through grant agreement no. 2022-06725.
- This work was supported by grants from the Swedish Research Council (grant. no. 2020-02657_3), and the European Union (ERC, RADIALIS, GA n. 101088408) to Dr. Magda Bienko. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.
- This work was supported by the grand from the Hjärnfonden (project no. PS2023-0023) to Dr. Wenjing Kang.
- Mr. Wing Hin Yip received a PhD fellowship from the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie Actions Innovative Training Network (MSCA ITN) Cell2Cell grant agreement number 860675.
- The work performed by Dr. Charles-Henri Lecellier was supported by the CNRS through the MITI interdisciplinary program, as well as from the French National Research Agency (ANR-22-CE45-0031-01)

Supplementary Table

Table S1: Poly(A) prediction for the 3' end of all the reads
Table S2: Annotation of tCREs
Table S3: Transcript information from the finalised transcriptome
Table S4: Transcript information from the raw transcriptome
Table S5: Transcript information from the read-filtered transcriptome
Table S6: Transcript ID and gene ID links to external databases
Table S7: Transcript information from the finalised transcriptome of iPSC chromatin-bound RNA
Table S8: Splicing efficiency of transcript and gene models
Table S9: Results from *spliceAI* showing all affected donor sites and acceptor sites
Table S10: Molecular details of CFC-seq
Table S11: oligo sequence used in CFC-seq
Table S12: TSS regions

Table is available in https://doi.org/10.6084/m9.figshare.27332796.v1

Methods

Cell models

Human iPS line i^3N and THP-1 cells were used in this study. iPS line i^3N is a gift from Dr. Michael Ward from NIH and derived from the WTC11 iPSC by introducing a doxycycline-inducible mNGN2 transgene at the AAVS1 site. THP-1 cells were subcloned by limited dilution and one clone (1-E9) was selected for relatively high sensitivity to PMA. The ability to differentiate relatively homogeneously in response to PMA was evidenced by expression of CD14 and CSF-1R quantified by ssCAGE.

Cell culture

The iPS cells were cultured in StemFit medium (Ajinomoto; Cat. No. RCAK02N) under feeder-free conditions at 37 °C in a 5% CO_2 incubator. The cells were passaged when the culture was ~80% confluent. The cells were washed with DPBS and detached by incubating with Accutase (Sigma; Cat.No. A6964) at 37 °C for 10 min. After centrifuge, cells were resuspended in StemFit medium supplemented with 10 nM Rho-associated kinase (ROCK) inhibitor (Y-27632, Wako; Cat.No. 036-24023), and re-plated onto iMatrix-511 (Nippi; Cat.No. 892012)-coated plates. The following day, ROCK inhibitor was withdrawn by refreshing the StemFit medium. The medium was refreshed every other day. To prepare the cells for transcriptomic profiling, iPSC was plated at a density of 0.5 million cells per 10 cm-dish in 10mL of medium and collected by Accutase after four days. The THP-1_1-E9 cells were cultured in RPMI1640-GlutaMAX-I (Gibco) supplemented with 10% heat inactivated fetal bovine serum (Gibco), 100 U/mL penicillin, and 100 µg/mL streptomycin (Gibco). The cells were maintained in a humidified 37℃ incubator with 5% CO_2.

Neural differentiation from iPSC to Neural stem cell (NSC)

NSCs were generated by using the PSC Neural Induction Medium (NIM) (Gibco; Cat.No. A1647801). iPSCs were seeded onto an iMatrix511-coated 6-well plate at a density of 0.25 million per well in 2 mL of StemFit medium with 10 nM ROCK inhibitor. The following day, medium was replaced with 2.5 mL of NIM. NIM was refreshed every other day. On day 6 of neural induction, the NSCs (P0) were harvested by incubating with Accutase at 37 °C for 10 min and scraping. The cells were centrifuged at 300 x g for 5 min. The supernatant was aspirated and the cell pellet was resuspended with PSC Neural Expansion Medium (NEM) containing 10 nM ROCK inhibitor and seeded onto a iMatrix511-coated 10-cm dish at a density of one million cells per dish. The following day, the medium was refreshed to remove the ROCK inhibitor. Hereafter, NEM was replaced every other day. After 5 days, the expanded P1-NSCs were harvested and stocked with STEM-CELLBANKER (Nippon Zenyaku, Cat.No. CB045). To prepare the cells for transcriptomic profiling, the cryopreserved P1-NSCs were quickly thawed and cultured in NEM with 10 nM ROCK inhibitor on an iMatrix511-coated 10 cm-dish at a density of one million. After four days of culture, P2-NSCs were harvested by incubating with Accutase and scraping.

Neural differentiation from NSC to cortical neuron

The differentiation of cortical neurons from NSC using doxycycline-inducible mNGN2 was performed following the protocol developed in Dr. Peter Heutink's Lab (DZNE), with modifications. Briefly, the cryopreserved P1-NSCs were passaged once. P2-NSCs were then replated onto 0.1mg/mL Poly-L-ornithine (Sigma, Cat.No. P3655)-coated 10cm-dish at a density of 5×10^6 cells per dish in the

differentiation medium I (50% DMEM/F12 (Gibco, Cat.No. 10565018) and 50% Neurobasal Medium (Gibco; Cat.No. 21103049) as the base, supplemented with 0.5 × Non-Essential Amino Acids (NEAA, Gibco; Cat.No. 11140050), 0.5 × GlutaMAX (Gibco; Cat. No.35050-061), 0.5 × N-2 Supplement (Gibco; Cat. No. 17502048), 0.5X B-27 Supplement (Gibco; Cat. No. 17504-044), 2.5 µg/mL human Insulin (Sigma; Cat.No. 19278), 2 µM DAPT (Wako; Cat.No. 043-33581), 50 µM 2-Mercaptoethanol (Gibco; Cat. No. 21985023), 2 µg/mL doxycycline hydrochloride (Sigma; Cat.No. D-9891), 5 µg/mL Mouse Laminin, and 10 nM ROCK inhibitor). The following day, the same medium without ROCK inhibitor was used to replace medium. On day 3, the medium was changed to differentiation medium II (50% DMEM/F12 and 50% Neurobasal Medium as the base, supplemented with 0.5 × Non-Essential Amino Acids, 0.5 × GlutaMAX, 0.5 × N-2 Supplement, 0.5 × B-27 Supplement, 2.5 µg/mL human Insulin, 10 µM DAPT, 50 µM 2-Mercaptoethanol, 2 µg/mL doxycycline hydrochloride, 0.5 µg/mL Mouse Laminin, 10 ng/ml brain-derived neurotrophic factor (BDNF, PeproTech; Cat.No. 450-02), 10 ng/ml glial cell-derived neurotrophic factor (GDNF, PeproTech; Cat.No. 450-10), 10 ng/ml neurotrophin-3 (NT-3, PeproTech; Cat.No. 450-03), and 0.5 µg/ml laminin). On Day 6 onwards, the half of the medium was exchanged every 3 days without Doxycycline. On day 10, the differentiated neurons were harvested for the transcriptomic profiling. Briefly, the cells were incubated with dissociation buffer (50% Accutase, 50% DPBS, and ≥50 units papain (Worthington; Cat.No. LK003178)) at 37 °C for 20 min. Then, wash buffer (DMEM/F12 supplemented with GlutaMax, 10 nM ROCK inhibitor, and ≥ 500 Kunitz units DNaseI (Worthington, Cat.No. LK003172)) was added to the cells. The cells were collected, washed 3 times with 0.1% BSA in DPBS and centrifuged at 150 × g for 10 min to remove the dead cell debris.

THP-1 differentiation

2×10^6 cells of THP-1_1-E9 were seeded in 10 cm dishes and treated with 30 ng/mL of PMA (Sigma) for 24 h and 96 h. As a negative control, THP-1_1-E9 cells were treated with 1/1,000 volume DMSO (Wako) in the culturing medium for 24 h and 96 h.

RNA purification

The harvested cells were subjected to RNA extraction and purification using the RNeasy Kit (Qiagen) with on-column DNaseI treatment according to the manufacturer's instructions. The eluted RNA was quantified by the NanoDrop™ spectrophotometry platform (Thermo Fisher Scientific). The purified RNA samples were stored at -80 °C until they were used for the CFC-seq, ssCAGE and RNA-seq libraries construction.

Cell fractionation

Subcellular fractionation of i^3N cells was performed similarly as previously described (Yip et al. 2022) with minor optimization. Briefly, approximately 10 million cells were used per fractionation experiment to obtain cytoplasmic (CYTO), soluble nucleoplasmic (SNUC) and chromatin (CHR) fractions. Cells detached by Accutase were washed with cold DPBS and lysed for 5 min using a cold lysis buffer containing 0.15% Igepal CA-630, 10 mM Tris pH 7.5, 150 mM NaCl. The lysate was overlaid on a sucrose cushion containing 10 mM Tris pH 7.5, 150mM NaCl, 24% sucrose and centrifuged at 1000 x g for 10 min at 4 °C. The supernatant was removed and centrifuged at 14,000 x g for 1 min, after which the supernatant was taken as the CYTO fraction. The nuclear pellet was washed once in PBS containing 0.5 mM EDTA pH 8.0 and resuspended in a buffer containing 20 mM Tris pH 7.5, 50% glycerol, 75 mM NaCl, 0.5 mM EDTA. An equal volume of nuclear lysis buffer containing 10 mM Tris pH 7.5, 300 mM NaCl, 1M Urea, 1% Igepal CA-630, 7.5 mM $MgCl_2$, 0.2 mM EDTA was added and

incubated on ice for 5 min to lyse the nuclei. After centrifugation at 13,000 x g for 2 min at 4 °C, the supernatant was taken as the SNUC fraction and the pellet as the CHR fraction. The chromatin pellet was washed once in PBS containing 0.5 mM EDTA pH 8.0 and resuspended in the same buffer. RNA from each fraction was isolated using Trizol LS (Invitrogen), according to the manufacturer's instructions. Prior to RNA extraction, chromatin which formed a tight precipitate was solubilized in Trizol LS by first passing through a 27-gauge needle, then treating with a dounce homogenizer. Extracted RNA was subjected to DNase I treatment followed by phenol-chloroform extraction. Average RNA yields were 260 µg, 53 µg and 7 µg from the cytoplasmic, nucleoplasmic and chromatin fractions, respectively. The RNA from the chromatin fraction was subjected to CFC-seq cDNA library construction and sequenced by PromethION.

CFC-seq experimental workflow

For the total RNAs from the iPS, NSC and cortical neuron, 10 µg of RNAs were polyadenylated by incubating with E-coli poly(A) Polymerase (PAP) (NEB M0276) at 37°C for 15 min and purified with AMPure RNA Clean XP beads. For the total RNAs from the THP-1 series (DMSO 24h, DMSO 96h, PMA 24h and PMA 96h), 80 - 100 µg of RNAs were used both for PAP-treated and PAP-nontreated RNAs. For the iPS chromatin fraction RNAs, 5 µg of RNAs were polyadenylated. The RNA was reverse transcribed by Prime Script II Reverse Transcriptase (Takara Bio) at 42°C for 60 min with different poly-dT primer sets dependent on samples. For the PAP treated RNAs from iPS, NSC and cortical neuron, reverse transcription was done by poly-dT primer containing UMI: oligodT_16VN_UMI15 (GAGATGTCTCGTGGGCTCGGN$_{15}$CTACGT$_{16}$VN). The PAP-treated and PAP-nontreated RNAs from THP-1 series and PAP-treated chromatin RNAs were reverse transcribed with poly-dT primer containing ONT barcode: FL-ONT_dT16VN_NB01-NB08 (GAGATGTCTCGTGGGCTCGG[NB01-NB08]CTACGT$_{16}$VN). The product was purified with RNAClean XP beads. Cap-trapping from the RNA/cDNA hybrids was performed as previously described (Takahashi et al. 2012). Then, RNA was digested with RNase H (Takara Bio) at 37°C for 30 min and the cDNA was purified with AMPureXP beads. Double stranded 5’ linker of N6 and GN5 (**Table S10**) were mixed at a ratio of 1:4 and ligated to the cDNA with Mighty Mix (Takara Bio) for overnight and the ligated cDNA was purified with AMPure XP beads. Shrimp Alkaline Phosphatase (Takara Bio) was used to remove phosphates at the ligated linker and purified with AMPureXP beads. The 5’ linker ligated cDNA was then subjected to second strand synthesis with KAPA HiFi mix (Roche) using a second strand primer with an UMI of 25 nucleotides (CTACACTCGTCGGCAGCGTCN$_{25}$GTGGTATCAACGCAGAGTAC) at 95°C for 5 min, 55°C for 5 min and 72°C for 30 min. Exonuclease I (Takara Bio) was added to digest the excessed primers at 37°C for 30 min. After AMPureXP purification, the cDNA/DNA hybrids derived from the neuron series were amplified with PrimerSTAR GXL DNA polymerase (Takara Bio) for 10 cycles. The cDNA/DNA hybrids from the iPS chromatin-bound fraction were amplified with LongAmpR Taq DNA Polymerase (NEB, M0323) for 7 cycles. The amplified products were purified by AMPureXP. The amplified products from the Neuron series and the cDNA/DNA hybrids from the THP-1 series were subjected to library construction using the kit SQK-LSK110 (Oxford Nanopore Technologies) with the manufacturer's instructions. The libraries were sequenced with R9.4.1 flow cell (FLO-MIN106) in MinION sequencer, by adding 50 fmol DNA per flow cell and sequenced for 72 hrs. For the chromatin-bound iPSC, purified product was subjected to library construction using kit V14 SQK-LSK114 (Oxford Nanopore Technologies). The libraries were sequenced in the R10.4.1 flow cell in the PromethION sequencer. For each sample replicate, the technical runs were pooled for further downstream analyses. All the oligo sequences are listed in **Table S10**.

Single-strand CAGE (ssCAGE)

After extracting and purifying the RNA from the 3 cell types, 15 µg of RNA was used from each sample for constructing a single-strand CAGE library (Takahashi et al. 2021). Libraries were subjected to 100-base pair-end sequencing using the Illumina NextSeq 1000/2000 P2 Reagents (200 Cycles) v3. Reads were de-multiplexed and mapped in pair-end to the human genome (hg38) using STAR (Dobin et al. 2013). The median number of mapped CAGE paired reads obtained from each sample is about 60 million.

Pre-processing of the long-read ONT data

Basecalling was processed by *Dorado* v0.2.4 basecaller software provided by Oxford Nanopore Technologies to generate fastq files from FAST5 files using the model "dna_r9.4.1_e8_sup@v3.3". Basecalled reads were filtered to 'pass' based on the average read Q-score > 10. For the rest of the downstream analysis, only Q-score pass reads were used. To prepare clean reads from the fastq files, adapter sequences were trimmed from the raw reads by primer chop (https://gitlab.com/mcfrith/primer-chop), requiring the presence of adaptor sequences from both ends. Primer-chop also orientates the strands based on the order of head and tail linkers appear in the read. Next, the poly(A) tails and cases where the poly-A tail is basecalled to random dinucleotide sequences of adenine and guanine were trimmed. We developed *tail trimmer* (v1.3) to scan with 20bp windows and trim the poly-A tail with maximum 5 As left.

The trimmed and orientated fastq reads were then mapped to the human genome, using the reference GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta, along with accompanying GENCODE v39 annotation. For mapping, minimap2 v2.17-r974-dirty (Li 2018) was used. The raw alignments were then corrected using *TranscriptClean* v2.0.3 (Wyman and Mortazavi 2019) while a junction file containing confident junctions extracted from short-read RNA-seq on the Neuron series and the junctions from the GENCODE v39 were applied. An additional tag to the corrected alignments was then added using the *TALON* v5 (Wyman et al. 2019) *talon_label_reads* function which calculates the fraction of adenines in a 16-bp window within the genome immediately downstream of the last 3'-bp alignment of the read. This is required for running TALON but not other assemblers.

```
[1] minimap2 -I 1000G -k 15 -d reference.fa.mmi reference.fa
[2] minimap2 -t [threads] -2 -ax splice -uf --MD --junc-bed gencode.v39.annotation.bed --secondary=no
GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta.mmi pass_trim_clean.fq.gz >
pass_trim_clean.sam
[3] python TranscriptClean.py -t [threads] --sam pass_trim_clean.sam --genome
GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta -j gencode.v39.annotation.SJs.txt --outprefix
sample --deleteTmp --tmpDir sample/tmp
[4] talon_label_reads --f pass_trim_clean_corrected.sam --ar 16 --fracA 0.5 --g
GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta --t 12 --tmpDir sample/tmp --deleteTmp --o
sample
```

Comparing 5' end of different long-read protocols

The fastq files of 5 different long-read protocols sequencing the same iPSC (WTC11) were downloaded from ENCODE LRGASP (CAP_trap_Pacbio: ENCSR309IKK; uncap_deplete_Pacbio: ENCSR507JOF; TOS_Ont: ENCSR539ZXJ; R2C2_Ont: ENCSR925UQZ; dRNA_Ont: ENCSR392BGY). These fastq files were mapped to hg38 by minimap2 as described above. The alignments of these

libraries and our iPSC libraries were intersected with the exon of coding genes annotated by GENCODE v39, while alignments intersecting with ≥ 90% of any exon were included as mRNA in the analysis. The 5' ends of these alignments (excluding chr M) were intersected with 3 external resources inferring promoter region. These resources included ENCODE SCREEN cCRE regions (either promoter or enhancer regions), the ATAC-peaks called from the single nucleus ATAC-seq we performed on the same iPSC, and the promoter regions (300 bp median) defined from FANTOM5 CAGE data. These regions are listed in **supplementary table S11**.

Comparing 3' ends from libraries with or without PAS motif

The 3' ends from all the alignments were extracted. The DNA sequences ± 50 bp from the 3' ends were obtained from the reference genome GRCh38_no_alt_analysis_set_GCA_000001405.15. The percentage of adenine downstream the 3' end was calculated for potential internal priming. A percentage > 50% from 16 nucleotides downstream or > 75% from 8 nucleotides downstream is marked as internal priming. The 3' ends of transcript models annotated by GENCODE v39 were protected. Position weight matrix was generated from PAS motifs annotated by GENCODE v39. We search for PAS motifs genome wide by scanMotifGenomeWide.pl of Homer v4.11 (Heinz et al. 2010) using a motif file generated from the GENCODE PAS motifs. Only the loci with score > 3 were used for the analysis. These PAS motifs include: "AATAAA", "ATTAAA", "AGTAAA", "TATAAA", "AATATA", "ATTATA", "TTTAAA", "AAAAAA", "AATTAA", "AATAAG", "AATACA", "AAGAAA", "ACTAAA", "AATGAA", "CATAAA", "AATAAT", "GATAAA", "AATAGA". The genomic locations of these PAS motifs were overlapped with the single-nucleotide locations 25 nt upstream of the 3' ends by running bedtools closest. Presence of PAS motif 5 nt to 50 nt upstream the 3' ends, according to the last nucleotide of the PAS motif, was annotated as PAS-positive. The RNA classes: mRNA, p_ncRNA and e_ncRNA were defined from SALA. The nucleotide composition surrounding the 3' ends were plotted according to the reference genome. To subset the 3' ends captured by PAT, 3' ends from PAT THP-1 were subtracted by 3' ends from the non-PAT THP1 libraries.

```
[1] scanMotifGenomeWide.pl PAS.motif hg38 -bed -keepAll -p 10 > output.bed
[2] bedtools closest -a all3n_up25.sort.bed.gz -b
GENCODE_polyA_signal.motif_rng.main.FASTA.6.bed.gz -s -D a | cut -f-6,10- | gzip >
all3n_up25.PAS6.bed.gz
```

Poly(A)-tail prediction

In order to call bona fide 3' ends for the transcripts detected by CFC-seq, we trained a random forest classifier on a curated database of 3' ends identified by FLAM-seq and 3p-seq (Legnini et al. 2019; Alfonso-Gonzalez et al. 2023). Briefly, we collected all 3' ends detected by CFC-seq and flagged those overlapping with the curated (hereon "reference") dataset within a window of 10 nt. We then trained a random forest classifier using the ranger package (v. 0.16.0) in R 4.3.1, considering the 3' ends found in the reference as true positives and defining a set of features to train the model on, including the presence, type and position of polyadenylation signals and the nucleotide composition of a 50 nt window upstream and downstream of each 3' end. We then used the trained model to assign a probability score ("poly(A) score") to each 3' end in the CFC-seq dataset. We defined an upper threshold at which 95% of 3' ends found in the reference and containing a polyadenylation signal were classified as true to define "poly(A)" 3' ends (threshold > 0.34), and a lower threshold at which 95% of all 3' ends were classified as true to flag "non-poly(A)" 3' ends (threshold < 7.6e-5). We then performed motif enrichment analysis with the meme R package (v. 1.7.1) on six classes of 3' ends, collapsed when closer to each other than 5 nt: 3' ends for messenger, enhancer and promoter RNAs were filtered for being observed with more than 4,

2 and 2 unique reads respectively, and stratified for having a poly(A) score higher than 0.34 (threshold for “true” 3’ ends). In addition to the motif enrichment analysis, we also counted and ranked the occurrence of 6-mers in a window of 50 nt upstream and downstream of the 3’ ends for the same classes.

Transcription start site identification by SCAFE

The bam files containing all alignments mapped to the hg38 genome were used as the input for *SCAFE* v1.01 (Moody et al. 2022). Briefly, *scafe.workflow.bk.bam_to_ctss* extracted alignments with ≤ 3 nt softclip at the 5’ end and ≥ 30 matched nt, where ~70% reads passed. These qualified alignments (~165 million) were converted into CTSS, which were grouped into TSS clusters. According to the multiple properties of each TSS cluster (percentage of CTSS with unencoded G, corrected expression, read count, summit count and flanking count), a multiple logistic regression model was trained to distinguish TSS clusters that are likely genuine. The identification of these TSS clusters was performed per cell-type where cutoffs of at least 3 reads and at least 1 read with unencoded G were applied. Next, *scafe.tool.cm.aggregate* defines tCREs by extending and merging overlapped TSS clusters defined from different cell-types. This aggregation step only merges the clusters without pooling signals to generate new TSS clusters. Additional commands including *scafe.tool.cm.directionality* and *scafe.tool.cm.annotate* were used to calculate bi-directional reads in each tCRE and pre-annotate the tCREs with nearby GENCODE transcript model. The tCREs were then classed for the promoter types by intersecting with SCREEN cCREs, in an order of promoter-like, enhancer-like and CTCF-alone. If a tCRE was intersected with both promoter-like and enhancer-like SCREEN cCREs, the tCRE was defined as promoter-like. If no cCRE was found to intersect with the tCREs, these tCREs were defined as “unclassed”. These promoter types were directly transferred to the TSS clusters and transcript models. Expression level of tCREs was measured by CTSS of qualified alignments located inside the genuine TSS clusters, where the per-cluster signal was collapsed into per-tCRE level. The signal was normalised across all the samples with individual replicates by using the RLE method in *edgeR*. In the preparation of running SALA, strand-specific TSS clusters were merged if they are within 75 nt distance. If the signal from the merged clusters revealed bi-modal distribution (>20% from each cluster), the merged clusters were split at the midpoint of the two summits. This extension was designed for concluding reads and reference transcript models into the same read class if their 5’ ends are in a proximity. The codes for running SCAFE in this study are listed:

Convert bam into CTSS with and without unencoded G:

```
[1] scafe.tool.bk.bam_to_ctss
--TSS_mode=softclip \
--bamPath=$bamPath \
--unencoded_G_upstrm_nt=3 \
--max_thread=$max_thread \
--genome=hg38.gencode_v39 \
--max_softclip_length=30 \
--outputPrefix=$outputPrefix \
--outDir=$outDir
```

Aggregate individual library into cell types:

```
[2] scafe.tool.cm.aggregate \
--lib_list_path=$lib_list_path(CTSS per library) \
--max_thread=5 \
```

```
--genome=hg38.gencode_v39 \
--outputPrefix=$outputPrefix \
--outDir=$baseDir/out/aggregate

[3] scafe.tool.cm.ctss_to_bigwig \
--genome=hg38.gencode_v39 \
--ctss_bed_path=aggregate.collapse.ctss.bed.gz \
--outputPrefix=$outputPrefix.all \
--outDir=$baseDir/out/ctss_to_bigwig

[4] scafe.tool.cm.ctss_to_bigwig \
--genome=hg38.gencode_v39 \
--ctss_bed_path=aggregate.unencoded_G.collapse.ctss.bed.gz \
--outputPrefix=$outputPrefix.ung \
--outDir=$baseDir/out/ctss_to_bigwig
```

Identify TSS clusters, tCREs and annotate the tCREs:

```
[5] scafe.tool.cm.aggregate \
--lib_list_path=$lib_list_path(CTSS per cell type) \
--max_thread=5 \
--genome=hg38.gencode_v39 \
--outputPrefix=$outputPrefix \
--outDir=$baseDir/out/aggregate

[6] scafe.tool.cm.cluster \
--overwrite=yes \
--cluster_ctss_bed_path=aggregate.collapse.ctss.bed.gz \
--count_ctss_bed_path=aggregate.unencoded_G.collapse.ctss.bed.gz \
--min_summit_count=0 \
--min_cluster_count=1 \
--outputPrefix=$outputPrefix \
--outDir=$baseDir/out/cluster

[7] scafe.tool.cm.ctss_to_bigwig \
--genome=hg38.gencode_v39 \
--ctss_bed_path=aggregate.collapse.ctss.bed.gz \
--outputPrefix=$outputPrefix.all \
--outDir=$baseDir/out/ctss_to_bigwig

[8] scafe.tool.cm.ctss_to_bigwig \
--genome=hg38.gencode_v39 \
--ctss_bed_path=aggregate.unencoded_G.collapse.ctss.bed.gz \
--outputPrefix=$outputPrefix.ung \
--outDir=$baseDir/out/ctss_to_bigwig

[9] scafe.tool.cm.filter \
--overwrite=yes \
--ctss_bed_path=aggregate.collapse.ctss.bed.gz \
```

```
--ung_ctss_bed_path=aggregate.unencoded_G.collapse.ctss.bed.gz \
--tssCluster_bed_path=tssCluster.bed.gz \
--genome=hg38.gencode_v39 \
--outputPrefix=$outputPrefix \
--outDir=$baseDir/out/filter \

[10] scafe.tool.cm.annotate \
--overwrite=yes \
--tssCluster_bed_path=tssCluster.default.filtered.bed.gz \
--tssCluster_info_path=tssCluster.log.tsv \
--min_CRE_count=3 \
--genome=hg38.gencode_v39 \
--outputPrefix=$outputPrefix \
--outDir=$baseDir/out/annotate

[11] scafe.tool.cm.directionality \
--overwrite=yes \
--CRE_bed_path=CRE.coord.bed.gz \
--CRE_info_path=CRE.info.tsv.gz \
--ctss_bed_path=aggregate.collapse.ctss.bed.gz \
--outputPrefix=$outputPrefix \
--outDir=$baseDir/out/directionality/
```

To assess the content of these tCREs *de novo* identified from CFC-seq, ssCAGE from the corresponding samples were analysed. The reads from ssCAGE were independently subjected to the SCAFE pipeline or aggregated with the CFC-seq CTSS by SCAFE (**Ext_Fig. 2a**). The expression level shown in **Ext_Fig. 3k** was derived from the aggregated tCREs derived from both CFC-seq and ssCAGE signals.

GC content, conservation, CpG island, TATA box and initiator

In order to confirm the accuracy of the identified TSS clusters and tCREs, including their capability of initiating transcription and the precise position of the TSSs, different promoter sequence properties were intersected with the ex5_clusters using the summit (TSS) at the centre position. These ex5_clusters are strand-specific, contain xx TSS clusters on average and only have one parental tCRE.

PhastCon annotations (4way, 17way, 30way and 100way), 5-base GC content and CGI annotation from hg38 were downloaded from UCSC. The motifs of the TATA box (TBP: MA0108.3) and the initiator (INR: POL002.1) were obtained from Jaspar. The motifs were searched against the genome using *scanMotifGenomeWide.pl* of *Homer* v4.11 (Heinz et al. 2010). Summits of the ex5_clusters were extended 5 kb up- and down-stream and intersected with the phastCon annotations. The mean values per position were extracted for the positional plot. The mean value of 17way annotation of the 500 bp upstream the summit was used to represent the sequence conservation for each ex5_cluster. For GC content, summits were extended 5 kb up- and down-stream and intersected with the 5-base GC content annotation. Backgrounds with the same number of ex5_clusters were generated from hg38 masking all the exons from GENCODE v39 and all the ex5_clusters. For CGI, summits were extended 5 kb up- and down-stream and intersected with the annotation. The coverage of CGI for each region was extracted. To determine CGI-positive ex5_clusters, 500 nt before the summit shows > 200 nt overlap with CGI annotation with at least 40 CGI units. For dCGI, the same criteria were required for the 500 nt

downstream of the summits. For TATA box and initiator motifs, summits were extended for 50 bp up- and down-stream and intersected with the Homer genome-wide results. If there is more than one occurrence in the 101 nt region, the ones with higher motif scores were taken. To determine TATA box-positive ex5_clusters, a TATA box with motif score at least 3 is found between -35 to -11 nt according to the first base of the TATA motif.

Transcript model construction using SALA

SALA is divided into 3 steps: features collection, transcript model assembling and gene model annotation. In the feature collection, the 5' end clusters, 3' end clusters and splicing junctions were collected from the reads. The TSS clusters and summits identified from SCAFE were applied directly. The 3' end tags were collapsed into clusters by Paraclu with at least 3 tags in a cluster. Summit and signal were also defined for each 3' end cluster. All the splicing junctions from the bam files after TranscriptClean were extracted. The max quality scores of the 6 nt surrounding the junctions were also extracted. The junctions from sample-matching short-read RNA-seq were obtained from the SJ.out.tab output from STAR. Additionally, the 5' end, 3' end and junction positions from GENCODE v39 were subjected to SALA. The 5' end clusters were merged at a 75 nt distance. According to the long-read signals from the merged clusters, merged bimodal clusters (>10%) were splitted at the midpoint. The GENCODE 5' ends were merged with these clusters with 75 nt. The 3' end clusters were merged similarly at a 150 nt distance while GENCODE 3' ends were also merged with 150 nt. The 5' end clusters with SCAFE support or GENCODE support were considered as "confident". The 3' end clusters with paraclu support or GENCODE support were considered as "confident".

For the transcript model assembling, every read subjected to SALA was assigned with a 5' end cluster and a 3' end cluster while every junction observed from the read was assigned with a junction ID. The reads with both confident 5' clusters and confident 3' clusters, and all the reference transcript models (e.g. GENCODE v39) were considered as "complete". These reads having the same features (5' end cluster, 3' end cluster and all junctions) were grouped as a transcript model. The "incomplete" reads were first assigned to the transcript models generated by the complete reads, becoming a partial support of those models. The remaining incomplete reads will be grouped into independent transcript models. After assembling, the 5' and 3' ends of the transcript models were defined as the most common 5' and 3' ends revealed by the corresponding reads. The following commands were used in this study:
Input files preparation:

Transcript assembling:

```
[1] end5_guided_assembler_v0.1.20231102.pl
--qry_bed_bgz=Neuron_THP1.bed.bgz
--ref_bed_bgz=GENCODEv39.transcript.bed.bgz
--chrom_size_path=chrom.sizes.tsv
--out_dir=20231114_Neuron_THP1
--max_thread=3
--out_prefix=Neuron_THP1.S3
--min_transcript_length=15
--doubtful_end_avoid_summit=yes
--min_exon_length=1
--print_trnscrptID=no
--chrom_fasta_path=hg38.gencode_v39/fasta/genome.fa
--min_output_qry_count=1
--trnscpt_set_end_priority=commonest:summit:longest
```

--doubtful_end_merge_dist=150
--novel_model_prefix=ONTT
--conf_end3_merge_flank=150
--conf_end5_merge_flank=75
--conf_end5_bed_bgz=ontCAGE.Neuron_THP1.end5.cluster.bed.bgz
--conf_end3_bed_bgz=ontCAGE.Neuron_THP1.end3.cluster.bed.bgz
--min_summit_dist_split=50
--retain_no_qry_ref_bound_set=no
--doubtful_end_avoid_summit=yes
--min_size_split=100
--min_frac_split=0.2
--signal_end5_bed_bgz=ontCAGE.Neuron_THP1.end5.signal.bed.bgz
--signal_end3_bed_bgz=ontCAGE.Neuron_THP1.end3.signal.bed.bgz
--conf_end3_add_ref=yes
--conf_end5_add_ref=yes
--conf_junction_bed= Gencode_v39.junct.bed,long_read.hi_qual.junct.bed,short_read.hi_qual.junct.bed
--min_qry_score=0

Filtering of confident transcript models

All the Transcript models generated by *SALA* were classed using GENCODE v39 as reference. These classes include known transcripts (assigned with an ENST ID), novel transcripts from known gene regions (Novel isoforms attached to an ENSG gene ID), and novel transcripts derived from novel gene regions.

we removed the detected partial GENCODE transcripts and kept only the transcript models with their starting sites linked to the confident 5' end clusters supported by SCAFE. Specifically for the novel isoforms grouped into GENCODE annotated genes, 5 supporting reads from each of the replicates and 10% transcript ratio from one of the cell types were required. For the transcripts from novel genes (majority are novel lncRNAs), no further filter was applied.

For filtering the raw TALON transcript models, we applied different criteria for different transcript classes. 1, for the known transcripts of GENCODE v39, no filter was applied. 2, for the novel transcripts of known genes, 5 supporting reads from each of the 2 replicates and 10% transcript ratio from at least one of the cell types were required while the starting sites of the transcripts need to lie inside the genuine TSS clusters or match with the starting sites of any of the known transcripts. Transcript ratio was determined by the full-length reads defined by TALON (iPSC = 75,456,042, NSC = 59,230,321, Neuron = 47,230,302). Transcript ratio for every single transcript model from the raw table was calculated as [number of transcript reads / number of gene reads]. 3, for the transcripts from novel genes (all these transcript models were assigned with a novel gene ID including the "Intronic" we rescued from the Genomic group), transcripts were required to have their starting sites lying inside the genuine TSS clusters. We did not directly require reproducibility from group 3 transcript models but rely on the reproducibility of the TSS clusters (SCAFE regression probability ≥ 0.5). The raw and the filtered transcript models are shown in **Table S1** and maintained in the webpage for this study. Coding potential of all the transcripts were assessed by the Coding Potential Assessment Tool (CPAT v3.0.4) (Wang et al. 2013) with default parameters. For novel transcripts, CPAT score < 0.364 were defined as non-coding, in which transcripts ≥ 200 bp were defined as lncRNAs. For the Novel isoform and Genomic transcripts,

lncRNA transcript was defined by 1, CPAT score < 0.364; 2, length ≥ 200 bp; 3, linked gene annotated as lncRNA by GENCODE.

At the gene level, novel genes were classed as non-coding if more than 50% transcripts were defined as non-coding. Non-coding genes with at least 1 non-coding transcript ≥ 200 bp were defined as lncRNAs, otherwise were defined as short_ncRNAs. The gene classes of known genes were inherited from GENCODE.

Sub-classification of lncRNA transcripts and genes

From the CFC-seq with 5 cell-types, we have detected 6,559 known and complete lncRNA transcripts and identified 115,275 novel lncRNA transcripts. We classed these lncRNA transcripts into divergent lncRNAs, sense overlap lncRNAs, sense intronic lncRNAs, antisense lncRNAs and intergenic lncRNAs in ascending order. Divergent lncRNAs: transcripts with their ex5_cluster summits within ±2 kb on the opposite strand of any ex5_clusters of GENCODE protein coding genes or pseudogenes. Sense intronic lncRNAs: transcripts that intersect with GENCODE protein coding genes or pseudogenes, with their TSS clusters summits and the whole transcript lengths lying inside the intron. For those transcripts partially overlap with exons are classed as sense-overlap lncRNAs. Antisense lncRNAs: genes with ≥50% of their transcript region overlapping with the gene region of GENCODE protein coding genes or pseudogenes on the opposite strand. Intergenic lncRNAs: the remaining lncRNA transcripts that could not be assigned to any of the above categories. For the classification of the novel lncRNA gene, the same ascending order for priority was taken. For example, if a gene included both sense intronic lncRNA transcript and antisense lncRNA transcript, the gene was classed as sense intronic lncRNA.

Classification of ex5_cluster and transcript model into mRNA, p_ncRNA, e_ncRNA, CTCF_ncRNA and other_ncRNA

The ex5_clusters were divided into 4 groups according to the promoter types, which were inherited from the tCRE as promoter-like, enhancer-like, CTCF-alone with ATAC support and unclassed with ATAC support. Only the ex5_clusters producing ncRNAs (> 50% read count) were included while the definition of ncRNAs were described according to the CPAT results or GENCODE annotation. Additionally, ex5_clusters producing GENCODE coding genes alone were grouped as a control. These resulted in 9,749 ex5_clusters with mRNA outputs, 17,045 promoter-like ex5_clusters with ncRNA outputs (p_ncRNA), 28,718 enhancer-like ex5_clusters with ncRNA (e_ncRNA), 194 ATAC-supported CTCF-alone ex5_clusters with ncRNA (CTCF_ncRNA) and 4,645 ATAC-supported unclassed ex5_clusters with ncRNA (other_ncRNA). The ex5_clusters linked to both coding (GENCODE) and non-coding (GENCODE or CPAT) transcript models were excluded from the analyses. Since we have only included complete models in this analysis, the transcript length and exon number were supported by our sequencing data. Conversely, some known and long RNAs (e.g. *KCNQ1OT1*, 91kb) that were partially detected by CFC-seq were excluded from the analyses.

TALON and Isoquant

For comparison with SALA, we analysed the CFC-seq libraries (Neuron series and THP1-series) by TALON. Same as SALA, the bam files after TranscriptClean and TALON label containing only primary alignments were used as input for TALON with the same reference transcriptome (GENCODE v39). TALON internally excludes alignments with percent identity lower than 80% or coverage alignment lower than 90% of the read. After the processing by TALON, transcript IDs were extracted from the TALON database to include all those novel transcripts classed as ‘Genomic’ but excluded all transcripts with evidence of internal priming as determined by the fraction of adenines 3’ of the alignment in the reference genome. All the undetected GENCODE transcript models were also excluded to make it

comparable with the SALA analysis. This dataset was termed “TALON_Raw”. In addition, “TALON_Read_filtered” was generated by applying a filter for the transcript models as having 5 supporting reads from both replicates of the sample.

[1] talon_initialize_database --f gencode.v39.annotation.gtf --g hg38 --a gencode.v39 --o F6_interactome
[2] talon --f config.csv --db F6_interactome.db --build hg38 --threads {threads} --o F6_interactome_run1 && cp F6_interactome.db /scratch/
[3] talon_fetch_reads --db F6_interactome.db --build hg38 --o talon_read_annot/F6_interactome_all
[4] talon_filter_transcripts --db F6_interactome.db --annot gencode.v39 --minCount 1 --minDatasets 1 --allowGenomic --o F6_interactome_permissive_genomic.csv
[5] talon_create_GTF --db F6_interactome.db --build hg38 --annot gencode.v39 --whitelist F6_interactome_permissive_genomic.csv --o F6_interactome_permissive_genomic_talon.gtf
[6] talon_filter_transcripts --db F6_interactome.db --annot gencode.v39 --minCount 5 --minDatasets 2 --datasets iPSC_rep1,iPSC_rep2 --allowGenomic --o F6_interactome_iPSC_robust_genomic.csv

Isoquant

IsoQuant (v3.4.1) assembling was performed at 2 modes (Sensitive and Standard) with the following parameters:
IsoQuant Sensitive:
[1] isoquant.py -d nanopore
--bam_list BAM_list.txt
--reference GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta
-p Neuron_Series_THP1
--model_construction_strategy sensitive_ont
--genedb gencode.v39.primary_assembly.annotation.gtf
--threads 40
-o IsoQuant_NS_THP1_sensitive

IsoQuant Default:
[1] isoquant.py -d nanopore
--bam_list BAM_list.txt
--reference GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta
-p Neuron_Series_THP1
--genedb gencode.v39.primary_assembly.annotation.gtf
--threads 40
-o IsoQuant_NS_THP1_standard

Exosome sensitivity

RNA-seq data of another clone of iPS cell (Yip et al. 2022) treated with knockdown of EXOSC3 and control was annotated by the transcriptome generated in this study using pseudo alignment of Kallisto (Bray et al. 2016). Exosome sensitivity score was calculated at gene level as previously described: ($TPM_{Exosome\text{-}suppressed}$ - $TPM_{Control}$) / $TPM_{Exosome\text{-}suppressed}$. Genes that were not detectable from these short-read libraries (sum TPM of 4 libraries > 0.1) or not detectable from the both the iPSC and chromatin-bound iPSC CFC-seq libraries were excluded from the analysis. Exosome sensitivity smaller than 0 was considered as 0.

Bidirectionality of tCRE

Bidirectionality was acquired from *SCAFE* output for each tCRE. Qualified reads were used for counting the number of CTSS on both strands (**Ext_Fig. 3a**). The D was calculated as absolute difference of the forward and reverse counts divided by the sum of the forward and reverse counts. The tCRE was defined as bidirectional if D < 0.8, unidirectional otherwise. If the sum if the forward and reverse counts is less than 5, the directionality was set as undefined.

Super enhancer identification

Super enhancers have been predicted based on the magnitude of the H3K27ac ChIP-seq signal from various studies (Hnisz et al. 2013; Chapuy et al. 2013; Wang et al. 2023). The histone mark H3K27ac was determined from iPSC, NSC and Neuron by CUT&Tag (Yip et al. 2024). Briefly, reads were mapped to hg38 human genome using *Bowtie2* (v2.2.6). After removing PCR duplicates, the bam files of the two replicates for each cell type were subjected to MACS2 (v2.1.0) for peak calling. The significant peaks ($p < 0.01$) were merged and extended into a minimum window size of 150bp. The peaks were used as the reference loci for counting the read number from the bam files using ROSE (Whyte et al. 2013). The enhancer-like CREs overlapped with the super enhancer regions were marked as super enhancers. The number, activity and size of the SE were shown in **Ext_Fig. 3b&c**.

[1] ROSE_main.py -g hg38 -i [significant H3K27Ac peaks] -r [H3K27Ac bam file] -c [IgG bam file] -o [output folder] -s 10000 -t 2500

Quantification of CFC-seq samples

Bambu (https://github.com/GoekeLab/bambu) was modified to account for the increased presence of full-length reads and enhanced detection of transcript start and end sites. Traditionally, Bambu constructs read classes based on shared splice junctions. Now, we divide these read classes into subgroups by incorporating transcript annotations with refined start and end sites produced by SALA. Briefly, for each first exon of spliced read classes, the start positions of all exons in the provided annotations that share the 3' splice site were used as thresholds for grouping reads. A tolerance of 35 bp was applied to each threshold to accommodate small positional variations, and reads were assigned to the largest exon division they matched. Reads longer than the largest annotated first exon were grouped separately. The same process was applied to the last exon, but based on the 5' splice site. These modifications were implemented in the test_split_read_classes branch. To ensure Bambu recognizes these more granular read classes, we set the opt.discovery = list(min.exonDistance = 0) parameter, reducing the allowance for deviation when assigning read classes to reference transcripts before the Expectation-Maximization (EM) quantification. Additionally, due to the presence of overlapping transcripts from different genes, we preprocessed the SALA gtf file by performing *de novo* gene annotation using the assignGeneIds function from Bambu. This function assigns new gene IDs based on overlapping stranded exon regions, enabling the consideration of overlapping transcripts as a single entity during EM quantification. The quantification was run with the following commands in R:

```
[1] annotations = prepareAnnotations(SALA_table5)
[2] mcols(annotations)$GENEID = bambu:::assignGeneIds(annotations, GRangesList())$GENEID
[3] dir.create(paste0(results_dir,"/rcOut"), recursive=TRUE)
[4] se.new <- bambu(reads =  c(Neuron_Series_bam, Monocytes_bam),
        annotations = annotations,
        genome = fa.file,
        discovery = FALSE,
```

```
        opt.discovery = list(min.exonDistance = 0),
        rcOutDir = paste0(results_dir,"/rcOut"),returnDistTable=TRUE)
[5] saveRDS(se.new, paste0(results_dir,"/se.rds"))
[6] writeBambuOutput(se.new, results_dir)
```

eQTL and GWAS SNPs

The eQTL and GWAS data were obtained from eQTL Catalogue (Kerimov et al. 2020) and CAUSALdb (Wang et al. 2019) respectively. For the eQTL data, a cutoff of pip >0.3 was applied to include SNPs that are putatively casual. This dataset contains 102,614 non-redundant SNP locations. For the GWAS data, credible set version 2.0 was downloaded and converted into hg38 by crossing over the rsid with the SNP 151 dataset obtained from UCSC (n = 710,172,836). The GWAS dataset contains 917,107 non-redundant SNP locations after converting to hg38. These SNP locations were intersected with the regions of the ex5_clusters and transcript models (exons). Ex5_clusters linked to at least one GENCODE transcript model were grouped as ENST, otherwise were novel. The transcript models were intersected as they were. The SNP locations mapped to the GENCODE transcript models were removed from the novel group to reveal only the additional SNP coverage from the novel entries. The SNP vcf files and the transcript models (novel and GENCODE v39) were applied to the *spliceAI* to identify the SNPs that affect the SJs. Only the SNPs that negatively affect the acceptor and the donor score by 0.5 were reported in **Figure 6a&b**.

To divide these results according to the enhancer structures (dCGI, CGI, Null and TATA), intersection was grouped at the ex5_cluster level. Exons were merged from the transcripts that are linked to the same ex5_clusters using *bedtools merge*. In order to obtain a normalized enrichment, “common_all_20180418” dataset downloaded from NIH was used as background (n = 37,302,987, unique SNP locations = 36,363,997). The eQTL and GWAS datasets were limited to this scope of SNPs (unique SNP locations = 81,782 & 878,835 respectively) The merged regions per ex5_cluster were intersected with the common_all dataset, the eQTL dataset and the GWAS dataset. Exons that were shared by more than one ex5_cluster were recounted. The results were normalized as the rate of hit compared to the background SNPs or as the number of SNPs per 1kb. For the rate of hit compared to background SNPs, ex5_clusters containing no background SNPs were excluded.

Hi-C

The cells (iPSC, NSC and Neuron) were cultured and harvested as described above. Two replicates of each cell-type were fixed by 1% formaldehyde and Hi-C library construction was performed with the Arima-HiC+ kit and the Arima Library Prep Module (Arima Genomics) according to the manufacturer’s protocols. The libraries were then subjected to Illumina NovaSeq6000 for sequencing with 150bp paired-end mode. In order to yield a higher proportion of aligned reads, we first trim the 6 bp from the left end of each read as suggested by Arima Hi-C data analysis and then trim the 70 bp from the right end of each read using the command line “seqtk trimfq -b 6 -e 70”. The remaining 75 bp reads were used for the nf-core hic pipeline v 2.1.0 (https://nf-co.re/hic/2.1.0/). Since the two replicates of the Hi-C libraries yield consistent results in a pre Hi-C processing based on the subsampled samples. The two replicates from the same cell type were pooled together for the final nf-core hic pipeline with using the following options:

```
--genome 'hg38'
--restriction_site '[^GATC,G^ANTC]'
--ligation_site '[GATCGATC,GANTGATC,GANTANTC,GATCANTC]'
--digestion 'arima'
```

--res_compartments '500000,250000,100000'
--tads_caller 'insulation,hicexplorer'
--bin_size '1000000,500000,100000,50000,25000,10000,5000,2500,1000'

The processed Hi-C data was transformed into valid genomic interaction pairs at 5kb, 10kb and 25kb resolutions. The statistical significance of these interactions was calculated using the Bioconductor package GOTHiC (version 1.40.0) (Mifsud et al. 2017). Intra-chromosomal (*cis*) interactions supported by at least five read pairs and a q-value (FDR) ≤ 0.01 were deemed significant for downstream analysis.

References

1. Aguilo F, Li S, Balasubramaniyan N, Sancho A, Benko S, Zhang F, Vashisht A, Rengasamy M, Andino B, Chen C, et al. 2016. Deposition of 5-Methylcytosine on Enhancer RNAs Enables the Coactivator Function of PGC-1α. *Cell Rep* **14**: 479–492.

2. Alfonso-Gonzalez C, Legnini I, Holec S, Arrigoni L, Ozbulut HC, Mateos F, Koppstein D, Rybak-Wolf A, Bönisch U, Rajewsky N, et al. 2023. Sites of transcription initiation drive mRNA isoform selection. *Cell* **186**: 2438-2455.e22.

3. Alvarez-Dominguez JR, Knoll M, Gromatzky AA, Lodish HF. 2017. The Super-Enhancer-Derived alncRNA-EC7/Bloodlinc Potentiates Red Blood Cell Development in trans. *Cell Rep* **19**: 2503–2514.

4. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

5. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, Lennartsson A, Rönnerblad M, Hrydziuszko O, Vitezic M, et al. 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**: 1010–1014.

6. Austenaa LMI, Barozzi I, Simonatto M, Masella S, Della Chiara G, Ghisletti S, Curina A, de Wit E, Bouwman BAM, de Pretis S, et al. 2015. Transcription of Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early Termination. *Mol Cell* **60**: 460–474.

7. Bell JSK, Vertino PM. 2017. Orphan CpG islands define a novel class of highly active enhancers. *Epigenetics* **12**: 449–464.

8. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

9. Carbonell Sala S, Uszczyńska-Ratajczak B, Lagarde J, Johnson R, Guigó R. 2021. Annotation of Full-Length Long Noncoding RNAs with Capture Long-Read Sequencing (CLS). *Methods Mol Biol Clifton NJ* **2254**: 133–159.

10. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**: 626–635.

11. Chapuy B, McKeown MR, Lin CY, Monti S, Roemer MGM, Qi J, Rahl PB, Sun HH, Yeda KT, Doench JG, et al. 2013. Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer Cell* **24**: 777–790.

12. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* **20**: 1187–1195.

13. Claringbould A, Zaugg JB. 2021. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol Med* **27**: 1060–1073.

14. Cramer P, Pesce CG, Baralle FE, Kornblihtt AR. 1997. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci U S A* **94**: 11456–11460.

15. Cumbie JS, Ivanchenko MG, Megraw M. 2015. NanoCAGE-XL and CapFilter: an approach to genome wide identification of high confidence transcription start sites. *BMC Genomics* **16**: 597.

16. Dawson T, Rentia U, Sanford J, Cruchaga C, Kauwe JSK, Crandall KA. 2023. Locus specific endogenous retroviral expression associated with Alzheimer's disease. *Front Aging Neurosci* **15**: 1186470.

17. de Sainte Agathe J-M, Filser M, Isidor B, Besnard T, Gueguen P, Perrin A, Van Goethem C, Verebi C, Masingue M, Rendu J, et al. 2023. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. *Hum Genomics* **17**: 7.

18. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei C-L, Natoli G. 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**: e1000384.

19. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.

20. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinforma Oxf Engl* **29**: 15–21.

21. Field A, Adelman K. 2020. Evaluating Enhancer Function and Transcription. *Annu Rev Biochem* **89**: 213–234.

22. Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, et al. 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet* **46**: 558–566.

23. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206.

24. Gil N, Ulitsky I. 2018. Production of Spliced Long Noncoding RNAs Specifies Regions with Increased Enhancer Activity. *Cell Syst* **7**: 537-547.e3.

25. Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. 2022. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**: 353–359.

26. Gorbovytska V, Kim S-K, Kuybu F, Götze M, Um D, Kang K, Pittroff A, Brennecke T, Schneider L-M, Leitner A, et al. 2022. Enhancer RNAs stimulate Pol II pause release by harnessing multivalent interactions to NELF. *Nat Commun* **13**: 2429.

27. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**: 221–225.

28. Haberman N, Digby H, Faraway R, Cheung R, Jobbins AM, Parr C, Yasuzawa K, Kasukawa T, Yip CW, Kato M, et al. 2023. Abundant capped RNAs are derived from mRNA cleavage at 3'UTR G-Quadruplexes. http://biorxiv.org/lookup/doi/10.1101/2023.04.27.538568 (Accessed August 7, 2024).

29. Harrison LJ, Bose D. 2022. Enhancer RNAs step forward: new insights into enhancer function. *Dev Camb Engl* **149**: dev200398.

30. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576–589.

31. Hirabayashi S, Bhagat S, Matsuki Y, Takegami Y, Uehata T, Kanemaru A, Itoh M, Shirakawa K, Takaori-Kondo A, Takeuchi O, et al. 2019. NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat Genet* **51**: 1369–1379.

32. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155**: 934–947.

33. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. 2017. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**: 199–204.

34. Hou TY, Kraus WL. 2022. Analysis of estrogen-regulated enhancer RNAs identifies a functional motif required for enhancer assembly and gene expression. *Cell Rep* **39**: 110944.

35. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al. 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**: 199–208.

36. Jiang S, Cheng S-J, Ren L-C, Wang Q, Kang Y-J, Ding Y, Hou M, Yang X-X, Lin Y, Liang N, et al. 2019. An expanded landscape of human long noncoding RNA. *Nucleic Acids Res* **47**: 7842–7856.

37. Kerimov N, Hayhurst JD, Manning JR, Walter P, Kolberg L, Peikova K, Samoviča M, Burdett T, Jupp S, Parkinson H, et al. 2020. *eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs*. Genomics http://biorxiv.org/lookup/doi/10.1101/2020.01.29.924266 (Accessed December 22, 2020).

38. Khan A, Mathelier A, Zhang X. 2018. Super-enhancers are transcriptionally more active and cell type-specific than stretch enhancers. *Epigenetics* **13**: 910–922.

39. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182–187.

40. Klasens B. 1998. Inhibition of polyadenylation by stable RNA secondary structure. *Nucleic Acids Res* **26**: 1870–1876.

41. Kristjánsdóttir K, Dziubek A, Kang HM, Kwak H. 2020. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. *Nat Commun* **11**: 5963.

42. Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* **49**: 1731–1740.

43. Lai F, Gardini A, Zhang A, Shiekhattar R. 2015. Integrator mediates the biogenesis of enhancer RNAs. *Nature* **525**: 399–403.

44. Lee S, Barbour JA, Tam YM, Yang H, Huang Y, Wong JWH. 2023. *Integrating long-read RNA sequencing improves locus-specific quantification of transposable element expression*. Genomics http://biorxiv.org/lookup/doi/10.1101/2023.03.21.533716 (Accessed March 7, 2024).

45. Legnini I, Alles J, Karaiskos N, Ayoub S, Rajewsky N. 2019. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat Methods* **16**: 879–886.

46. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma Oxf Engl* **34**: 3094–3100.

47. Li Z, Liu L, Feng C, Qin Y, Xiao J, Zhang Z, Ma L. 2023. LncBook 2.0: integrating human long non-coding RNAs with multi-omics annotations. *Nucleic Acids Res* **51**: D186–D191.

48. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.

49. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA. 2013. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**: 320–334.

50. Lykke-Andersen S, Žumer K, Molska EŠ, Rouvière JO, Wu G, Demel C, Schwalb B, Schmid M, Cramer P, Jensen TH. 2021. Integrator is a genome-wide attenuator of non-productive transcription. *Mol Cell* **81**: 514-529.e6.

51. Maeng JH, Jang HJ, Du AY, Tzeng S-C, Wang T. 2023. Using long-read CAGE sequencing to profile cryptic-promoter-derived transcripts and their contribution to the immunopeptidome. *Genome Res* **33**: 2143–2155.

52. Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. 2017. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res* **27**: 27–37.

53. Mifsud B, Martincorena I, Darbo E, Sugar R, Schoenfelder S, Fraser P, Luscombe NM. 2017. GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PloS One* **12**: e0174744.

54. Moody J, Kouno T, Chang J-C, Ando Y, Carninci P, Shin JW, Hon C-C. 2022. SCAFE: a software suite for analysis of transcribed cis-regulatory elements in single cells. *Bioinforma Oxf Engl* **38**: 5126–5128.

55. Moody J, Kouno T, Kojima M, Koya I, Leon J, Suzuki A, Hasegawa A, Akiyama T, Akiyama N, Amagai M, et al. 2023. A single-cell atlas of transcribed *cis* -regulatory elements in the human genome. http://biorxiv.org/lookup/doi/10.1101/2023.11.13.566791 (Accessed October 29, 2024).

56. Nair SJ, Yang L, Meluzzi D, Oh S, Yang F, Friedman MJ, Wang S, Suter T, Alshareedah I, Gamliel A, et al. 2019. Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. *Nat Struct Mol Biol* **26**: 193–203.

57. Nellore A, Jaffe AE, Fortin J-P, Alquicira-Hernández J, Collado-Torres L, Wang S, Phillips RA, Karbhari N, Hansen KD, Langmead B, et al. 2016. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol* **17**: 266.

58. Pachano T, Sánchez-Gaya V, Ealo T, Mariner-Faulí M, Bleckwehl T, Asenjo HG, Respuela P, Cruz-Molina S, Muñoz-San Martín M, Haro E, et al. 2021. Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nat Genet* **53**: 1036–1049.

59. Pandya-Jones A. 2011. Pre-mRNA splicing during transcription in the mammalian system. *Wiley Interdiscip Rev RNA* **2**: 700–717.

60. Panigrahi A, O'Malley BW. 2021. Mechanisms of enhancer action: the known and the unknown. *Genome Biol* **22**: 108.

61. Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De María M, Adams MS, Balderrama-Gutierrez G, et al. 2024. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods*.

62. Pefanis E, Wang J, Rothschild G, Lim J, Kazadi D, Sun J, Federation A, Chao J, Elliott O, Liu Z-P, et al. 2015. RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell* **161**: 774–789.

63. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. 2023. Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* **41**: 915–918.

64. Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, Gagneur J, Cramer P. 2016. TT-seq maps the human transient transcriptome. *Science* **352**: 1225–1228.

65. Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**: 1009–1014.

66. Svoboda M, Frost HR, Bosco G. 2022. Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. *NAR Genomics Bioinforma* **4**: lqac035.

67. Takahashi H, Lassmann T, Murata M, Carninci P. 2012. 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7**: 542–561.

68. Takahashi H, Nishiyori-Sueki H, Ramilowski JA, Itoh M, Carninci P. 2021. Low Quantity Single Strand CAGE (LQ-ssCAGE) Maps Regulatory Enhancers and Promoters. *Methods Mol Biol Clifton NJ* **2351**: 67–90.

69. The ENCODE Project Consortium, Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Ai R, Aken B, Akiyama JA, Jammal OA, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: 699–710.

70. Uapinyoying P, Goecks J, Knoblach SM, Panchapakesan K, Bonnemann CG, Partridge TA, Jaiswal JK, Hoffman EP. 2020. A long-read RNA-seq approach to identify novel transcripts of very large genes. *Genome Res* **30**: 885–897.

71. Vučićević D, Corradin O, Ntini E, Scacheri PC, Ørom UA. 2015. Long ncRNA expression associates with tissue-specific enhancers. *Cell Cycle Georget Tex* **14**: 253–260.

72. Wang J, Huang D, Zhou Y, Yao H, Liu H, Zhai S, Wu C, Zheng Z, Zhao K, Wang Z, et al. 2019. CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. *Nucleic Acids Res* gkz1026.

73. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**: e74–e74.

74. Wang T, Medynets M, Johnson KR, Doucet-O'Hare TT, DiSanza B, Li W, Xu Y, Bagnell A, Tyagi R, Sampson K, et al. 2020. Regulation of stem cell function and neuronal differentiation by HERV-K via mTOR pathway. *Proc Natl Acad Sci* **117**: 17842–17853.

75. Wang Y, Song C, Zhao J, Zhang Y, Zhao X, Feng C, Zhang G, Zhu J, Wang F, Qian F, et al. 2023. SEdb 2.0: a comprehensive super-enhancer database of human and mouse. *Nucleic Acids Res* **51**: D280–D290.

76. Welsh SA, Gardini A. 2023. Genomic regulation of transcription and RNA processing by the multitasking Integrator complex. *Nat Rev Mol Cell Biol* **24**: 204–220.

77. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**: 307–319.

78. Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, Matheos D, Zeng W, Williams B, Trout D, et al. 2019. *A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification*. Genomics http://biorxiv.org/lookup/doi/10.1101/672931 (Accessed March 6, 2023).

79. Wyman D, Mortazavi A. 2019. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinforma Oxf Engl* **35**: 340–342.

80. Xiong L, Kang R, Ding R, Kang W, Zhang Y, Liu W, Huang Q, Meng J, Guo Z. 2018. Genome-wide Identification and Characterization of Enhancers Across 10 Human Tissues. *Int J Biol Sci* **14**: 1321–1332.

81. Yip CW, Hon C-C, Yasuzawa K, Sivaraman DM, Ramilowski JA, Shibayama Y, Agrawal S, Prabhu AV, Parr C, Severin J, et al. 2022. Antisense-oligonucleotide-mediated perturbation of long non-coding RNA reveals functional features in stem cells and across cell types. *Cell Rep* **41**: 111893.

82. Yip CW, Yasuzawa K, Kouno T, Chang J-C, Lehmann R, Cassan O, Abugessaisa I, Nobusada T, Takahashi H, Kato M, et al. 2024. Single cell bimodal analyses reveal the mode of activity of transcription factors on enhancers and promoters. *Prep*.

83. You B-H, Yoon S-H, Nam J-W. 2017. High-confidence coding and noncoding transcriptome maps. *Genome Res* **27**: 1050–1062.

84. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, Qiu Y, Li YE, Gaulton KJ, Wang A, et al. 2021. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**: 5985-6001.e19.

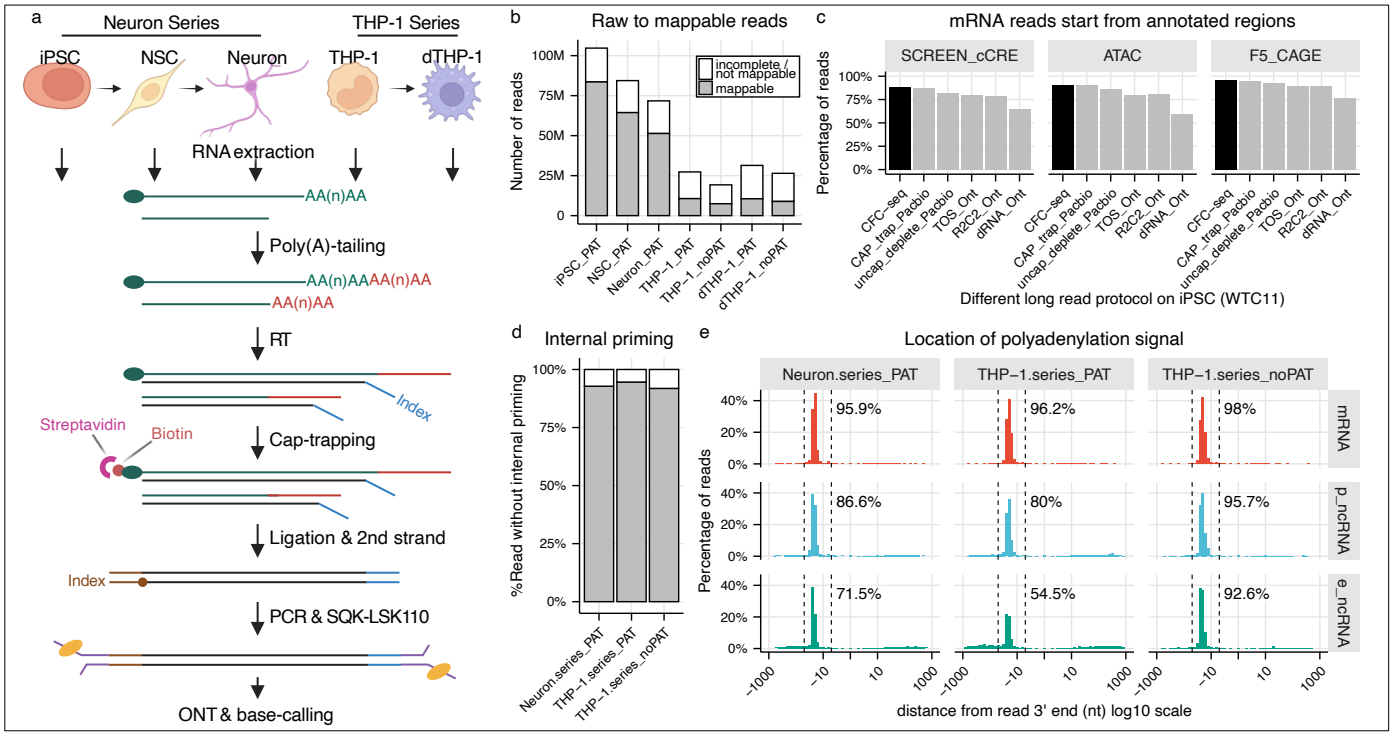

Figure 1 Output of 236 million long-reads from CFC-seq. a, Schematic of the experimental pipeline. **b**, Among the reads passed the base-calling score (≥ 10), the number of incomplete reads (lacking adaptors), non-mappable and mappable reads. **c**, The 5' end sites of the reads derived from mRNA were intersected with different annotated regions (SCREEN cCRE (The ENCODE Project Consortium et al. 2020), ATAC-seq peaks from WTC11 and FANTOM 5 CAGE clusters (ref)). The plot shows the percentage of reads with intersection. Long-read data of iPSC WTC11 was downloaded from ENCODE LRGASP (CAP_trap_Pacbio: ENCSR309IKK; uncap_deplete_Pacbio: ENCSR507JOF; TOS_Ont: ENCSR539ZXJ; R2C2_Ont: ENCSR925UQZ; dRNA_Ont: ENCSR392BGY) (Pardo-Palacios et al. 2024). **d**, Read-base assessment of potential internal priming, percentage in grey represent reads without internal priming. **e**, CFC-seq was performed with and without poly(A)-tailing (PAT) on THP-1 cells. Reads without internal priming and without known non-poly(A) genes (**Ext_Fig. 1b**) were extracted and located for the closest PAS motif at 25 nt upstream of the 3' end of the reads. The presence of a PAS motif located between -50 to -5 bp upstream of the 3' end of the read is considered as PAS-positive. The reads were assigned to their raw transcript models to categorise mRNA (excluded histone genes), promoter-derived ncRNA and enhancer-derived ncRNA, using the gene class.

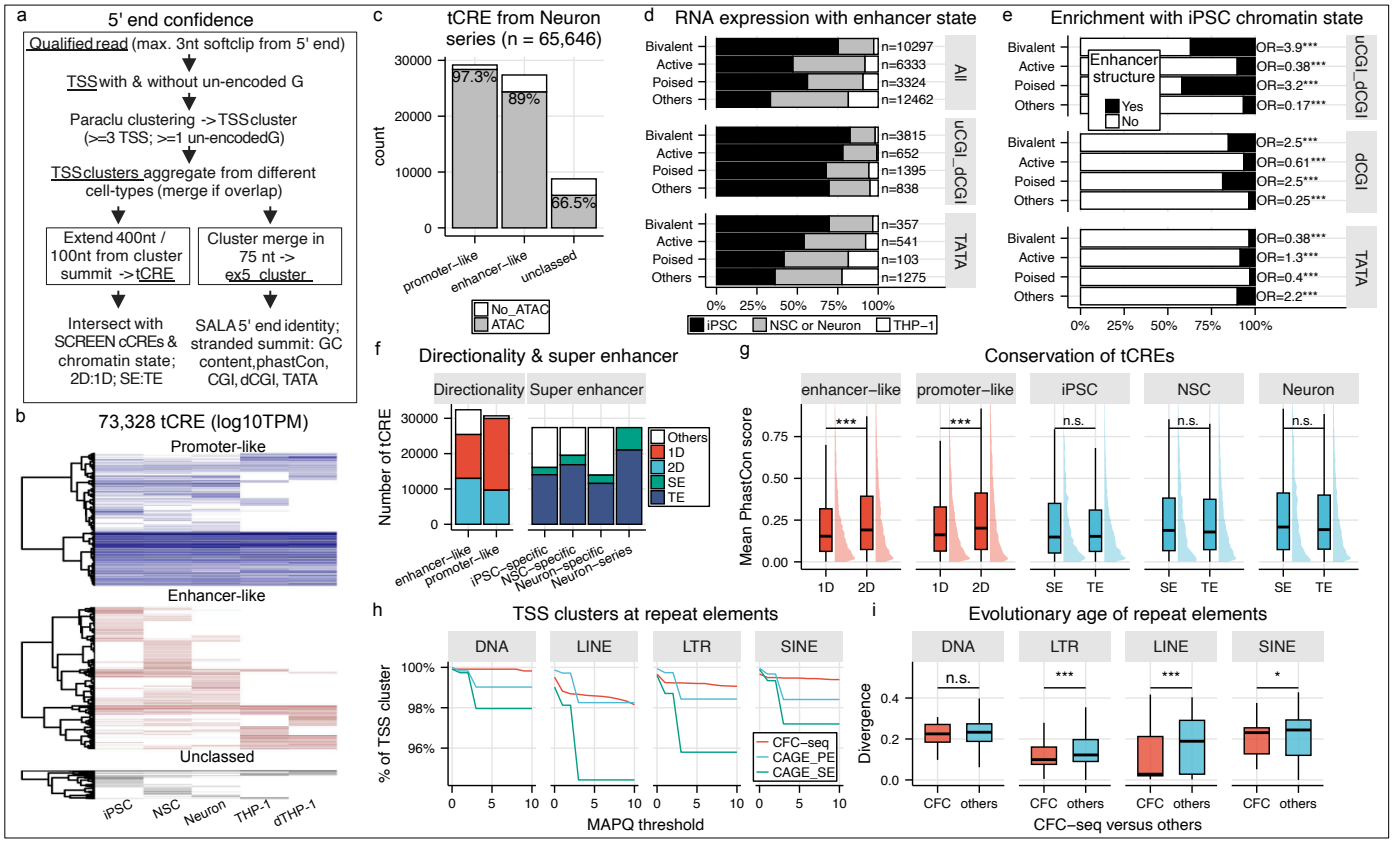

Figure 2 Identification of genuine TSS and tCRE from CFC-seq. a, Flow chart showing the pipeline of SCAFE in identifying genuine TSS, TSS clusters, tCREs and ex5_clusters. **b**, 73,328 tCREs were identified from CFC-seq using 5 cell-types. They were classified as promoter-like, enhancer-like and unclassed according to the epigenetic information summarised by SCREEN cCRE. **c**, tCRE from the Neuron series overlap with open-chromatin regions defined by ATAC-seq from the same sample. **d**, The epigenetic signals H3K27ac and H3K27me3 of iPSC were intersected to the enhancer-like tCREs. The enhancer-like tCREs were split into 4 groups as Bivalent (both H3K27ac and H3K27me3), Active (H3K27ac alone), Poised (H3K27me3 alone) and Others (none of the two). Qualified TSS were counted into the tCREs while >=1 count was considered expression from that cell type. The upper panel shows the percentage of all the tCRE that produce eRNA in each enhancer state while the middle panel shows a subset which is uCGI/dCGI positive. The lower panel shows tCREs with TATA box. **e**, The percentage of tCRE groups that is uCGI/dCGI-positive (upper panel), dCGI-positive-alone (negative for both TATA box and CGI, middle panel) and TATA box-positive-alone (lower panel). Statistics were performed by the Fisher's exact test, where OR represents odds ratio. **f**, Number of tCREs identified as unidirectional (1D), bidirectional (2D), cell-type dependent super enhancer (SE) and typical enhancer (TE). **g**, Sequence conservation analysis comparing 1D and 2D enhancers/promoters, and comparing SE and TE, based on tCRE grouping. **h**, Identification coverage of TSS clusters by CFC-seq, CAGE pair-end (PE) and CAGE single-end (SE) along an increasing MAPQ threshold. The union tCREs of all methods at MAPQ=0 is set as 100%. **i**, The divergence, which reflects the evolutionary age of the repetitive element, was plotted against repetitive elements found in CFC-specific TSS clusters (or others) at MAPQ = 10. Except specified, all the statistics in this figure were determined by Wilcoxon test with two-sided setting, where * represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < 0.001$, n.s. represents not significant.

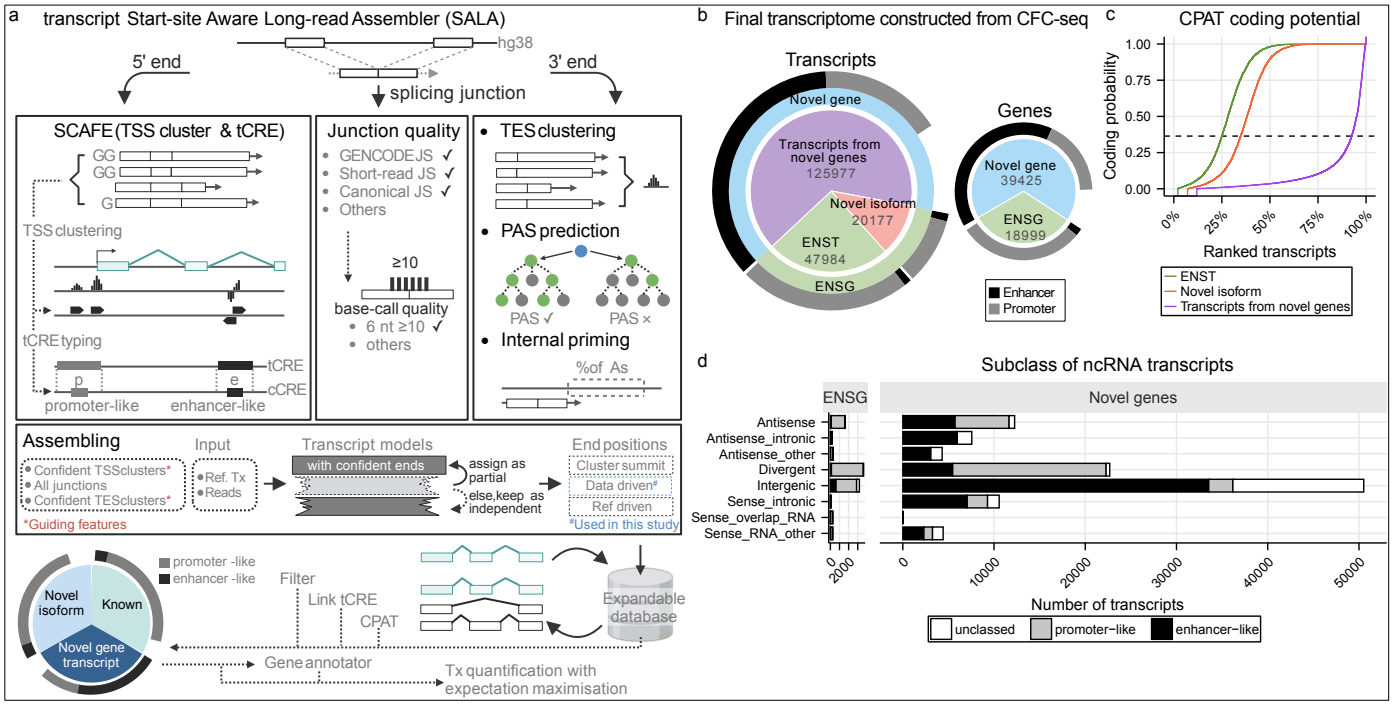

Fig. 3 Transcript models identified by SALA. a, Schematic of the pipeline post genome mapping. Three key features—5' end, splicing junction, and 3' end—of all reads were extracted, aggregated, labelled, and qualified. The 5' ends were clustered into Transcription Start Site (TSS) clusters, with confident TSS clusters identified by *SCAFE*. Splicing junctions were compared with junctions from GENCODE (v39) and short-read sequencing of the same samples. Base-call quality for each junction was extracted and summarised. The 3' ends were clustered into Transcription End Site (TES) clusters, with confident clusters selected based on the read count. Additional labels for the 3' end included the presence of a poly(A) signal (PAS) and internal priming. These features, along with the reference transcriptome and all reads, were applied to assembling. Confident TSS and TES clusters acted as guiding features, while other features were used to label the transcript models. The method for determining the end positions of the transcript models could be chosen according to the data structure. The resulting transcript models were linked to tCREs (transcriptional cis-regulatory elements) with promoter-typing and filtered for downstream analyses. The final transcriptome was quantified using expectation maximisation in *Bambu* (Chen et al. 2023). **b**, Number of transcript models and gene models identified by SALA. Transcript models were shown in three groups: ENST (annotated transcript models from GENCODE v39), novel isoforms (novel transcript models that are grouped under GENCODE gene models) and transcripts from novel genes. The promoter types of transcript models were determined by their derived TSS clusters. **c**, Coding potential of the three transcript groups were determined by *CPAT* (Wang et al. 2013). **d**, All the ncRNAs in this dataset including detected GENCODE lncRNAs and novel ncRNAs were subclassed according to the genomic location of coding genes and pseudogenes.

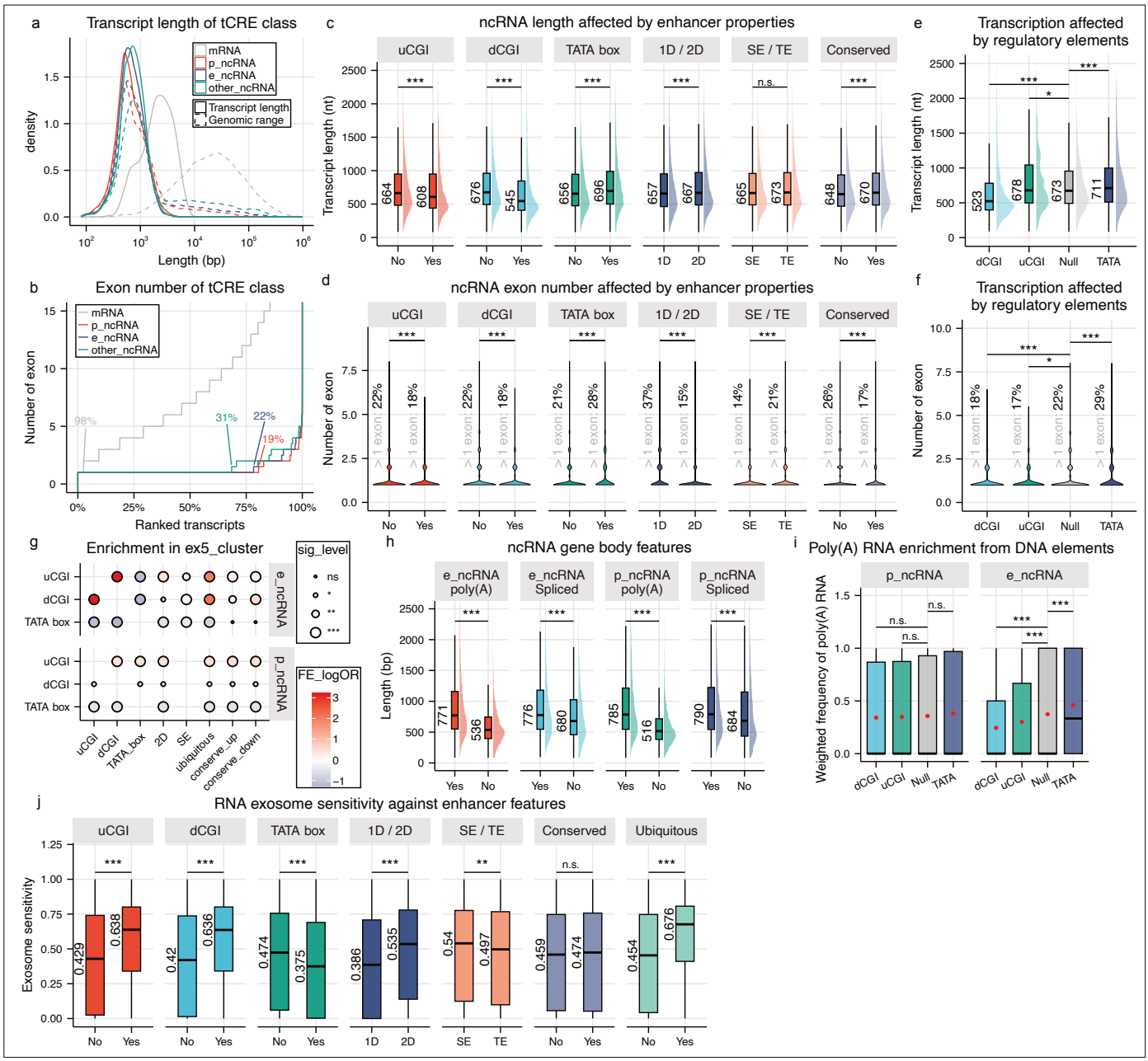

Fig. 4 Factors that affect the transcript model of eRNAs. a, Transcript length and genomic range of different RNA classes grouped by extended 5' clusters (ex5_clusters). **b**, The median number of exons of different RNA classes grouped by ex5_clusters. Percentage inside the plot represents the percentage of groups having more than one exon. **c**, Transcript length of e_ncRNA comparison across a series of enhancer properties: CGI (presence of CGI within 500 nt upstream the summit), downstream CpG island (dCGI, presence of CGI within 500 nt downstream the summit), TATA box (presence of TATA box 5' location within -35 to -11 nt upstream the summit), directionality (unidirectional 1D or bidirectional 2D), super enhancer (SE / TE: enhancers that are clustered within the super enhancer regions defined from iPSC, NSC or Neuron, enhancers from the THP-1 series are excluded), conservation (using the 17way phastCon annotation, the mean conservation score of the 500nt upstream the summit, ex5_cluster of e_ncRNA having a score > median is assigned as "Yes", others as "No"). **d**, Median number of exons comparison across various enhancer features. Percentage inside the plot represents the percent of groups with more than one exon. **e**, The transcript length of 4 different groups of e_ncRNAs while each feature is exclusive from the others: e.g. the CGI group only shows CGI positive without dCGI or TATA box. Null represents the ex5_clusters with none of the above features. **f**, Median number of exons of 4 groups of e_ncRNAs. **g**, Enrichment of CGI, dCGI and TATA box with other promoter features for e_ncRNA and p_ncRNA. Enrichment was tested by Fisher's exact in two-tail. FE_logOR represents the natural log of the odds ratio. **h**, RNAs were grouped at transcript-level while the presence of poly(A) and at least one splicing event were determined for each transcript model. Transcript length was compared for each feature. **i**, The transcript models were grouped to ex5_clusters while the frequency of poly(A) were weighted with read count. The red dot represents the mean value. **j**, Exosome sensitivity was determined at the gene level using RNA-seq data with and without knockdown of EXOSC3 in iPSC and quantified by our novel transcriptome. The e_ncRNA genes were grouped by ex5_clusters and subgrouped by various enhancer features. The exosome sensitivity was compared for each feature. Except being specified, all the statistics in this figure were determined by Wilcoxon test with two-sided setting. Here, * represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < 0.001$, n.s. represents not significant.

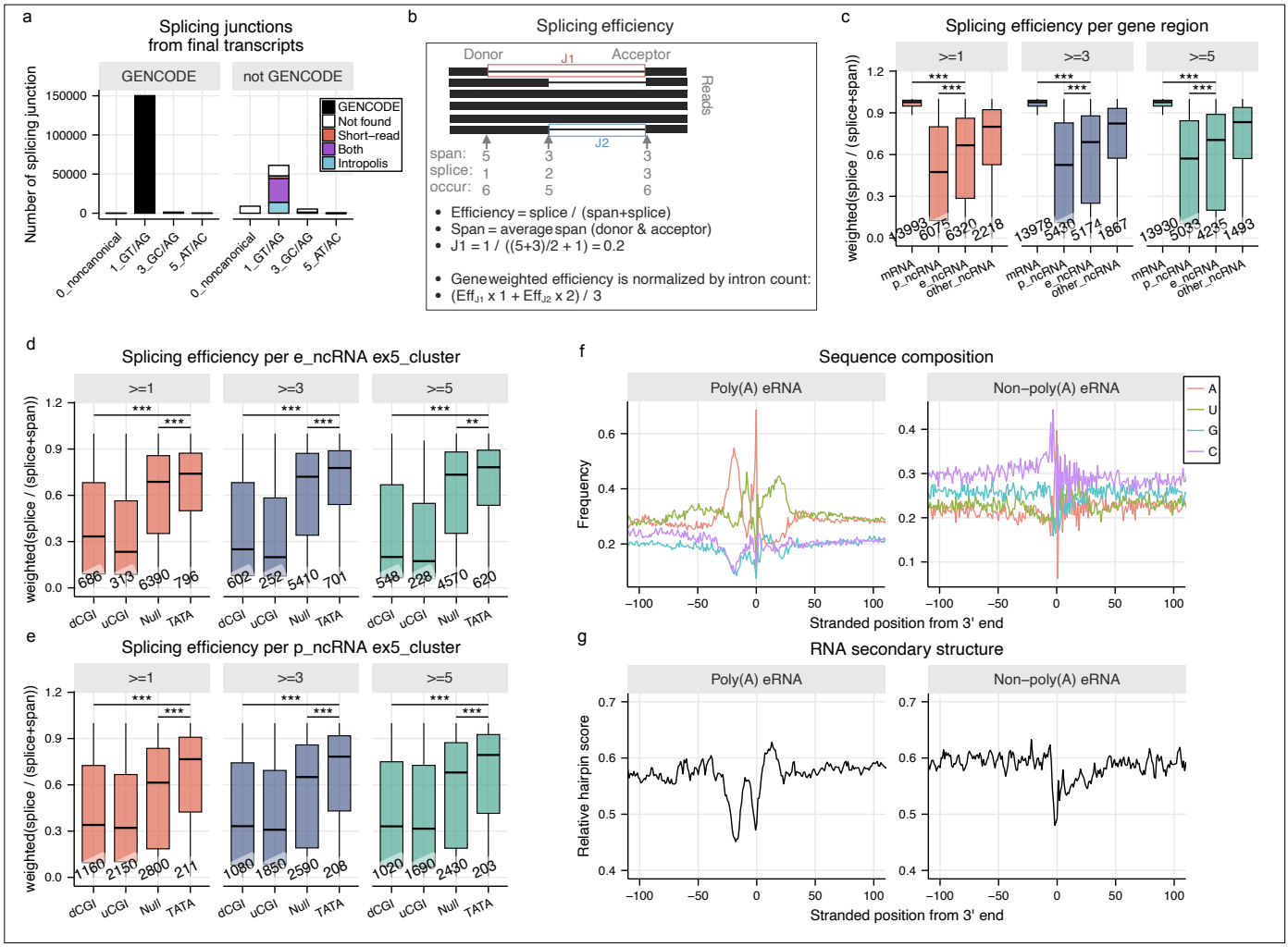

Figure 5 Splicing efficiency and 3' cleavage site of non-poly(A) ncRNAs. a, All the splicing junctions (SJs) observed from the final transcriptome. Majority of them has been annotated by GENCODE (v39), while most of the others SJs are canonical. **b**, Schematic of determining junction splicing efficiency per junction and weighted splicing efficiency per gene region. **c**, Splicing efficiency of the junctions was summarised at gene level normalised by intron count (Methods, **Fig. 5b**). Gene models without SJ were excluded from the analysis. The included SJs were filtered by three levels (>=1, >=3, >=5) of occurrence (count of span + count of splice, for both donor site and acceptor site). **d**, Gene models were grouped by ex5_clusters where only ex5_clusters assigned with one gene model were kept. The splicing efficiency were tested versus enhancer structure (dCGI, CGI, Null and TATA) for enrichment. **e**, Same as **d** for the p_ncRNA group. **f**, Nucleotide composition surrounding the 3' ends of poly(A) eRNAs and non-poly(A) eRNAs. **g**, RNA secondary structure obtained from RNAfold by submitting 300 nt sequence from the 3' ends of poly(A) eRNAs and non-poly(A) eRNAs. The relative hairpin score was summarised for 100 nt surrounding the 3' ends.

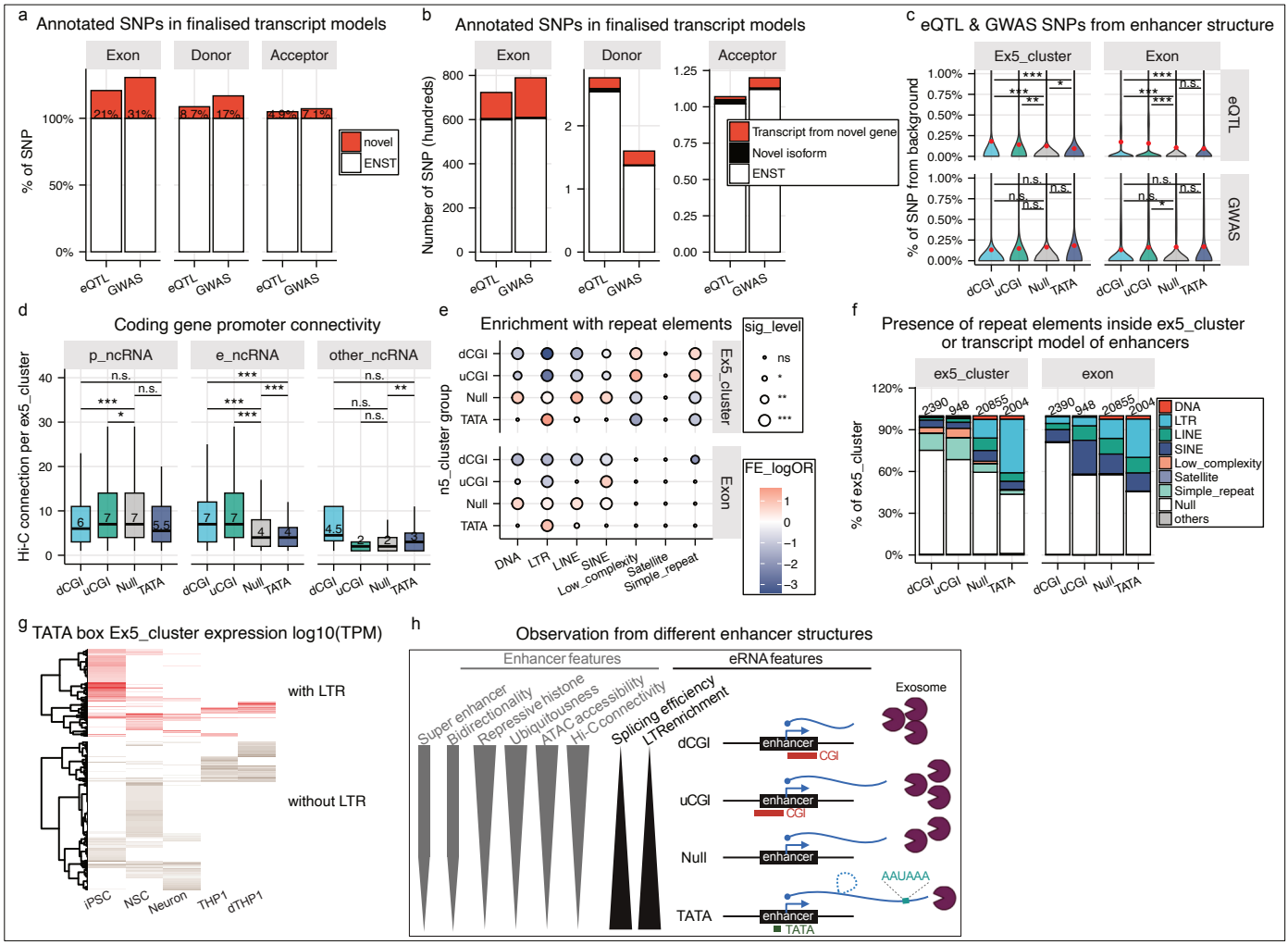

Figure 6 Functional aspects of the expanded annotated sequence a-b, The eQTL and GWAS data were obtained from eQTL Catalogue (Kerimov et al. 2020) and CAUSALdb (Wang et al. 2019) respectively while a cutoff of pip 0.3 was applied to the eQTL data. The SNPs were intersected with the exon, donor site and acceptor site from GENCODE (v39) transcript and the novel models from the finalised transcriptome. The SNPs from each dataset revealed in the plot are non-redudant. For the donor and acceptor sites, a cutoff of 0.5 measured by SpliceAI was applied. Percentage increase from the novel entries were shown in (**a**) while the RNA groups were shown in (**b**). **c**, The presence of eQTL and GWAS SNPs found in (**a**) were subgrouped for the e_ncRNA according to the mutually exclusive regulatory element of the ex5_clusters. The size of the ex5_cluster and the transcript model were normalized to number of SNP per 1kb. **d**, Hi-C interaction at 5 kb resolution with FDR 0.01 and at least 5 reads was used to infer connectivity between ex5_clusters of ncRNA and mRNA. Only the ex5_clusters showing at least one significant connection were included in the plot where the median number of connection per ex5_cluster was shown. Positive interaction from any of the 3 cell types was considered positive. **e**, Fisher's exact tests were performed to test the enrichment of different repeat elements located in different e_ncRNA ex5_cluster groups. For the cluster region, 6 nt coverage was required while for the exon region 200 nt was required. **f**, The percentage and number of ex5_clusters and exon grouped by ex5_clusters intersected with different repeat elements. At least 6 nt coverage is required for ex5_cluster while at least 200 nt coverage was required for exon. **g**, Expression profile of TATA box e_ncRNA ex5_cluster across the 5 cell types. The heatmap was divided into LTR positive or LTR negative. **h**, Overall summary of the features of enhancer and eRNA grouped by enhancer structures. Except specified, all the statistics in this figure were determined by Wilcoxon test with two-sided setting. Here, * represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < 0.001$, n.s. represents not significant.