BioRxiv
Research Article

ReCappable Seq: Comprehensive Determination of Transcription Start Sites derived from all RNA polymerases.

Bo Yan[*], George Tzertzinis[*], Ira Schildkraut[&], Laurence Ettwiller[&#]

Abstract

Methodologies for determining eukaryotic Transcription Start Sites (TSS) rely on the selection of the 5' canonical cap structure of Pol-II transcripts and are consequently ignoring entire classes of TSS derived from other RNA polymerases which play critical roles in various cell functions. To overcome this limitation, we developed ReCappable-seq and identified TSS from Pol-II and non-Pol-II transcripts at nucleotide resolution. Applied to the human transcriptome, ReCappable-seq identifies Pol-II TSS with higher specificity than CAGE and reveals a rich landscape of TSS associated notably with Pol-III transcripts which have been previously not possible to study on a genome-wide scale. Novel TSS consistent with non-Pol-II transcripts can be found in the nuclear and mitochondrial genomes. By identifying TSS derived from all RNA-polymerases, ReCappable-seq reveals distinct epigenetic marks among Pol-II and non-Pol-II TSS and provides a unique opportunity to concurrently interrogate the regulatory landscape of coding and non-coding RNA.

New England Biolabs Inc., 240 County Road, Ipswich, MA, 01938, USA
[*] *Both authors contributed equally to this work* [&] *Co-last authors* [#] *Correspondence should be addressed to ettwiller@neb.com*

Introduction

Current methods to characterize transcriptomes such as whole transcriptome shotgun sequencing (RNA-seq) [1] fall short in providing accurate descriptions of transcriptional landmarks such as Transcription Start Sites (TSS), termination sites and isoform composition. The identification of TSS is essential to study gene regulation because it permits the association between RNA transcription and the underlying genomic landmarks such as promoters and histone marks. Alternative TSS have been detected in more than 50% of human genes [2] driving most of the transcript isoforms differences across tissues [3]. Genomic positions of alternative TSS are often found within other isoforms of the same gene, confounding their detection using conventional RNA-seq.

To date Cap Analysis of Gene Expression (CAGE)[4] is considered the best performing and the most extensively used technology for mRNA TSS determination in eukaryotes [5]. CAGE offers a detailed view of the eukaryotic TSS landscape derived from the RNA polymerase II (Pol-II) by specifically capturing the canonical cap structure of Pol-II-derived transcripts. Likewise, other methods for TSS determination based on template switching [6] or enzymatic removal of non-capped ends [7] identify only Pol-II-derived TSS. Inherent to the principle of targeting canonical cap structures, these methods entirely exclude the large number of TSS derived from eukaryotic RNA polymerase I (Pol-I), RNA polymerase III (Pol-III) and mitochondrial RNA polymerase (POLRMT) which produce uncapped non-coding RNA. These uncapped primary transcripts display a 5' triphosphate identical to the 5' ends of prokaryotic primary transcripts. With the growing body of literature highlighting the key role of non-coding RNA in regulating biological processes and diseases [8], it is crucial that a technology be available that comprehensively identifies TSS for all eukaryotic RNA polymerases.

We have previously developed Cappable-seq to identify TSS in prokaryotic species [9]. Cappable-seq is based on the ability of the Vaccinia Capping Enzyme (VCE) to add a biotinylated guanosine to 5' triphosphorylated RNA. Applied to eukaryotes, Cappable-seq would identify transcripts derived from Pol-I, Pol-III and POLRMT (generally defined as non-Pol-II transcripts) but would miss Pol-II transcripts. In this study, we have adapted Cappable-seq to also identify canonical G-capped transcripts derived from Pol-II. To achieve this, we used the property of the yeast scavenger decapping enzyme (yDcpS) to decap capped RNA [10] leaving a di-phosphorylated 5' end that can be recapped by the vaccinia capping enzyme [11], hence the name ReCappable-seq. Thus, ReCappable-seq can identify TSS for RNA transcripts derived from all RNA polymerases. Further stratification in the TSS identification can be obtained by comparing the ReCappable-seq dataset with an additional ReCappable-seq dataset derived from RNA which has been first dephosphorylated using calf intestinal alkaline phosphatase (CIP). The comparison of those two datasets permits the discrimination between the CIP-resistant capped 5' ends and the CIP-sensitive triphosphorylated 5' ends, effectively differentiating Pol-II versus non Pol-II TSS. Furthermore, the sequencing of a control library for which the streptavidin enrichment step has been omitted allows the refinement of positions into high confidence TSS.

We applied ReCappable-seq to the transcriptome of the A549 human cancer cell line and identified 33,468 and 5,269 high confidence Pol-II and non-Pol-II TSS respectively. Pol-II TSS identified by ReCappable-seq are in good agreement with CAGE TSS with 76% of ReCappable-seq TSS located within 5bp of a CAGE TSS. Furthermore, ReCappable-seq shows an increased specificity for cap structures compared to CAGE, generating fewer false positive TSS particularly for highly expressed genes. Among the non-Pol-II TSS, we detected the known Pol-I TSS located upstream of the annotated 45S RNA locus and 2 out of the 3 known mitochondrial TSS. Recappable-seq also identified TSS for more than 80% of the known Pol-III transcripts with unprecedented granularity. Importantly, ReCappable-seq detected thousands of novel non-Pol-II TSS revealing a rich landscape of the non-coding primary transcriptome.

Results

ReCappable-seq

We have developed ReCappable-seq (Figure 1a) to comprehensively identify TSS of all eukaryotic genes transcribed by Pol-I, Pol-II, Pol-III and POLRMT RNA polymerases. We used the Vaccinia Capping Enzyme (VCE) which can add a biotinylated G-cap structure to either a 5' triphosphate or 5' diphosphate RNA and has been used previously to identify TSS and primary transcripts in prokaryotes [9] [12] [13]. In eukaryotes the 5' ends of transcripts derived from Pol-II are capped and therefore cannot be directly biotinylated with VCE. In order to include transcripts containing a canonical cap structure characteristic of Pol-II transcripts, we performed a prior decapping reaction to render capped RNA 5' ends amenable to capping by VCE. For this reaction, we used the property of the yDcpS decapping enzyme to hydrolyze the phosphodiester bond between the gamma and beta phosphates of the G-cap [11]. The reaction leaves a diphosphate-terminated 5' end that can be recapped by VCE with a biotinylated GTP derivative. Importantly, cap0 and cap1 as well as m7Gpppm6A and m7Gpppm6Am were all found to be suitable substrates for yDcpS [11] indicating that transcripts containing canonical G-cap structures can be recapped and sequenced.

We first tested whether an endogenous capped transcript such as ACTB can be recovered using the enrichment strategy of ReCappable-seq. RT-qPCR results reveal nearly quantitative recovery of ACTB transcripts after streptavidin enrichment, while 18S ribosomal RNA is essentially depleted (Figure 1b). Importantly, if the decapping step is omitted, only a small fraction of ACTB transcript is recovered, demonstrating the requirement of the yDcpS decapping step for efficiently recovering capped transcripts (Figure 1b). These RT-qPCR results indicate that this approach greatly enriches capped transcripts as compared to processed RNAs such as ribosomal RNA (rRNA).

Thus, ReCappable-seq permits both the G-capped mRNAs from Pol-II and the 5' triphosphate RNAs from Pol-I, Pol- III and POLRMT (non-Pol-II) to be tagged with a biotin labelled cap. These tagged RNAs can be enriched using streptavidin and sequenced using direct adaptor ligation and short read, high throughput sequencing resulting in genome-wide identification of TSS derived from all RNA polymerases at single nucleotide resolution.

Genome-wide identification of TSS in human lung cells

We applied ReCappable-seq to 5 microgram of total RNA isolated from human A549 cells. In order to evaluate the reproducibility of ReCappable-seq, we performed technical replicates and sequenced the resulting libraries using Illumina sequencing platform yielding approximately 32 million single end reads per library that were mapped to the human genome using STAR (Materials and Methods)[14]. In parallel, we used RNA from the same sample for CAGE analysis in order to obtain data for comparisons (see below Comparing the performance of ReCappable-seq to CAGE). Analysis of the ReCappable-seq technical replicates reveals a high correlation (Pearson corr=0.96, P-value $< 2.2e-16$) between replicates at single nucleotide resolution (Supplementary Figure 1a) demonstrating a high reproducibility of the technique. ReCappable-seq replicates were combined and downsampled to 63 million mappable reads for consistency with subsequent analysis.

The percentage of mapped reads from processed rRNA drops from 74% in the unenriched control libraries to only 3% in the ReCappable-seq libraries (Supplementary Figure 2a). In eukaryotes, rRNAs are formed by the processing of a single pre-rRNA 45S transcript to form the mature 18S, 5.8S and 28S rRNAs [15]. Because the processed rRNA accounts for the vast majority of the RNA in the cell, their depletion in ReCappable-seq libraries is an indicator of the specificity of ReCappable-seq for primary transcripts that have retained their original 5' end structure (G-cap or triphosphate). These results further demonstrate that ReCappable-seq efficiently removes transcripts with processed or degraded 5'ends.

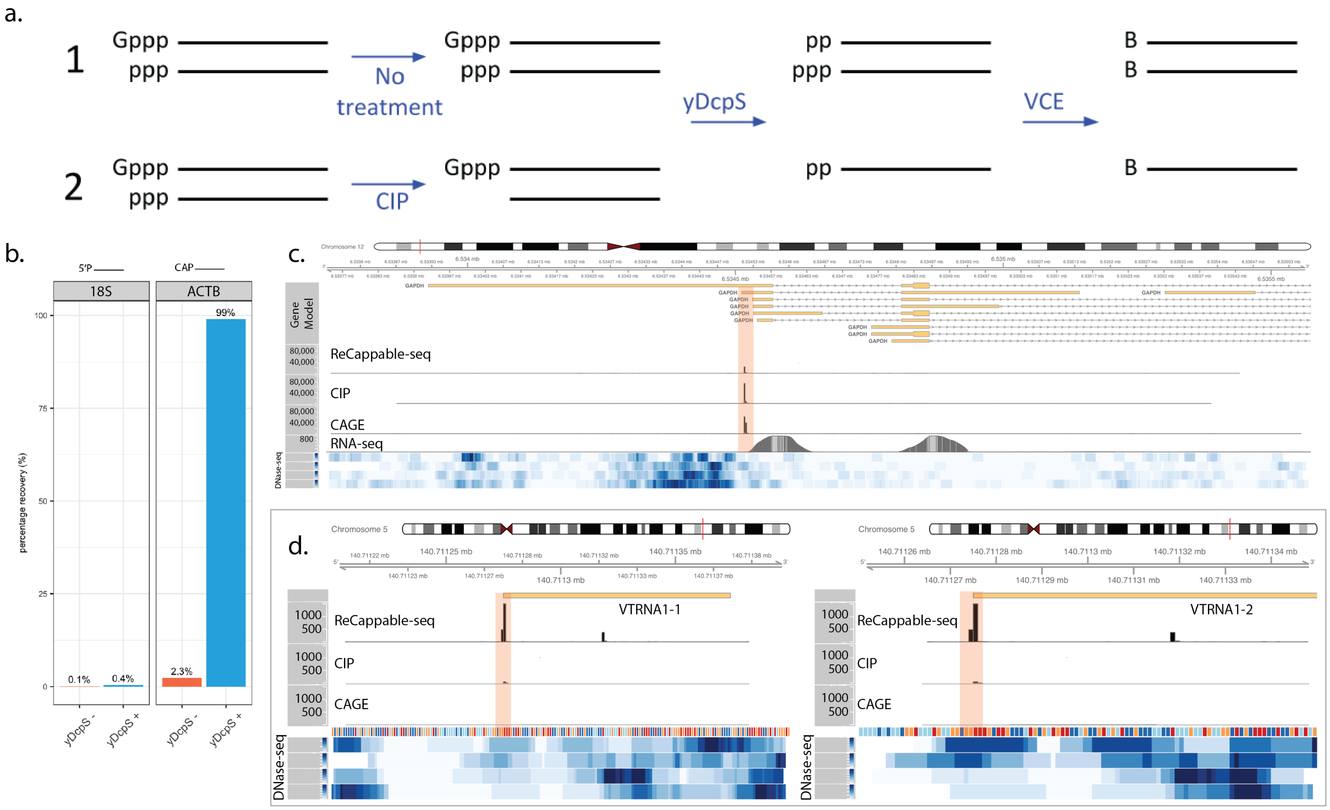
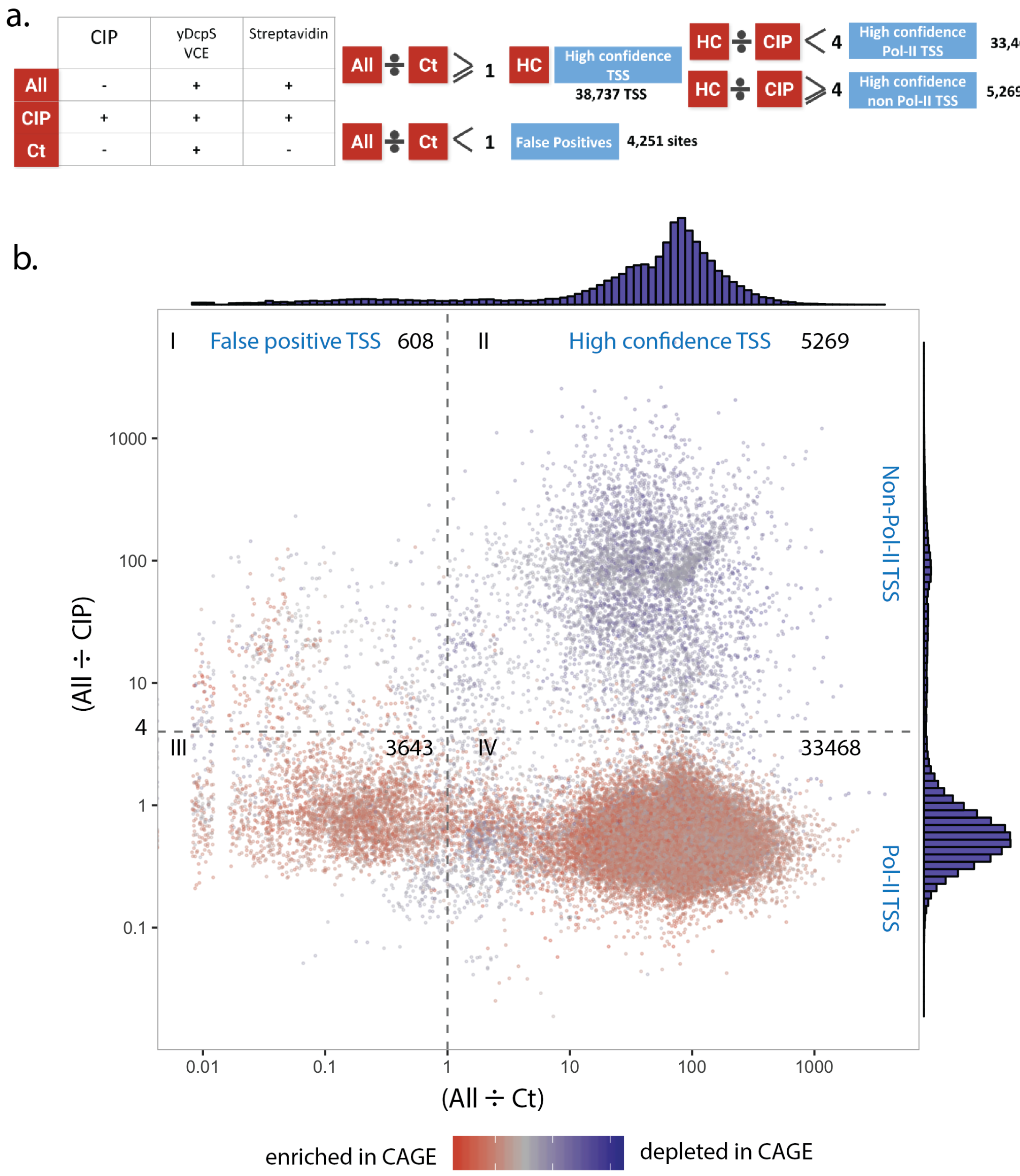


Figure 1. ReCappable-seq. a. Principle of ReCappable-seq. Enrichment of primary RNA transcripts containing 5' ends originating from either a cap structure (Gppp) or a triphosphate (ppp). Samples can either be enriched for both capped (Pol-II) and triphosphorylated (non-Pol-II) transcripts (1) or first dephosphorylated and enriched for only capped (Pol-II) transcripts (2). B denotes a desthiobiotin labelled cap structure. See text for details. b. RT-qPCR assay measuring the recovery after streptavidin enrichment of 18S rRNA (left panel) and the protein coding transcript for ACTB (right panel) with (blue) or without (orange) the decapping step. c. Example of a Pol-II TSS in the GAPDH locus: the same positions (highlighted in red) are found in the CAGE dataset. CIP treatment intensifies the signal, consistent with a Pol-II TSS. d. Example of Pol-III TSS corresponding to the start of two Vault RNAs (Vault RNA 1-1 and Vault RNA 1-2) located on Chr.5. The positions (highlighted in red) are missing in the CAGE dataset. CIP treatment reduces the signal, consistent with non-Pol-II TSS. In c. and d. the panels correspond to ReCappable-seq, CIP-Recappable-seq, CAGE and RNA-seq (A549 rRNA-depleted RNA-seq) read coverage. The four bottom panels correspond to read density from public ENCODE DNase-seq done on A549 cells.

As expected, mapped reads are found near the 5' end of annotated protein coding transcripts (Supplementary Figure 2c) as well as non-coding transcripts known to be transcribed by Pol-III (see Figure 1c and 1d for examples) suggesting high specificity of this methodology for all TSS. To further investigate the specificity of ReCappable-seq for primary 5' ends, we tested ReCappable-seq on highly degraded RNA samples obtained by magnesium ion-mediated fragmentation. RNA degradation has been proven to be challenging for the determination of TSS because the vast majority of the 5'ends are generated from fragmentation and do not correspond to TSS. Only a small minority of fragments have a cap or triphosphate characteristic of 5' ends of primary transcripts. Thus, results from these samples are informative in accurately estimating the specificity of ReCappable-seq for capped and triphosphate ends.

The profile of mapped reads demonstrates a very good correlation between pre-fragmented replicates (Supplementary Figure 1b) and pre-fragmented and intact starting material (Pearson corr=0.76, P-value < 2.2e-16, Supplementary Figure 2b and d). Furthermore reads from pre-fragmented material are predominantly mapping to the start of annotated genes consistent with the positioning of TSS (Supplementary Figure 2c). Importantly, reads derived from processed rRNA drop from 65% in the unenriched control libraries to 1.3% in the ReCappable-seq libraries from pre-fragmented RNA (Supplementary Figure 2a). Thus, ReCappable-seq is not affected by the large excess of uncapped 5' ends as the result of fragmentation, and genuine primary 5' ends were predominant in the ReCappable-seq libraries. Together, these results show that ReCappable-seq performs equally well on intact and fragmented RNA, adding further support to its high specificity for identifying 5'ends of primary RNA transcripts.



	CIP	yDcpS VCE	Streptavidin
All	-	+	+
CIP	+	+	+
Ct	-	+	-

Figure 2. a. Summary of sequencing experiments: Experiments were performed resulting in 3 dataset types: ReCappable-seq (ALL), CIP treated ReCappable-seq (CIP), and unenriched control for which the streptavidin step has been omitted (Ct). ALL corresponds to ReCappable-seq and defines 42,988 TSS. Comparison of ALL with Ct enables the definition of 38,737 high confidence (HC) TSS. Comparison of HC with CIP enables the definition of high confidence Pol-II TSS (33,468) and the high confidence non-Pol-II TSS (5,269). b. 42,988 candidate TSS positions distributed according to the ratio between ALL and Ct (x axis) and the ratio between ALL and CIP (y axis). Dotted lines define 4 quadrants as follows : ALL ÷ CIP less than 4 (pol-II TSS), ALL ÷ CIP above or equal to 4 (non-Pol-II TSS), ALL ÷ Ct above or equal to 1 (high confidence TSS) and ALL ÷ Ct less than 1 (false positive TSS). Color of the TSS denotes the ratio between ALL and CAGE with red (enriched in CAGE relative to ReCappable-seq), blue (depleted in CAGE relative to ReCappable-seq).

Classification of TSS into Pol II and non-Pol II TSS

Candidate TSS were identified and quantified at single nucleotide resolution using the TSS analysis pipeline (Materials and Methods). In short, candidate TSS are defined as positions in the genome where the number of 5’ ends of the reads mapping to those single nucleotide positions is above a defined threshold (TSS tags per million (TPM) greater or equal to 1) with TPM reflecting the strength of the TSS (see Materials and Methods for more details). From 63 million primary mappable reads, a total of 42,988 candidate TSS (TPM >=1) were identified genomewide.

To evaluate the specificity of ReCappable-seq in capturing cap and triphosphate structures, we constructed unenriched control libraries (Ct) for which the streptavidin enrichment step was omitted (Figure 2a). We calculated an enrichment score for each candidate TSS by dividing the TPM in the ReCappable-seq libraries with the TPM in the Ct library for each position. Candidate TSS depleted (ratio less than 1) in the ReCappable-seq library are considered false positives (Figure 2b). The enriched candidate TSS (ratio equal to or greater than 1, Figure 2b quadrant II and IV) are considered high confidence TSS. We found 38,737 (90.1%) of the candidate TSS were high confidence indicating that the majority of the TSS identified by ReCappable-seq alone are true positives. Unless otherwise stated, below we refer to the high confidence TSS as TSS.

To enable the discrimination of TSS derived from canonical capped transcripts from those derived from triphosphate transcripts, total RNA was treated with CIP prior to performing ReCappable-seq (Figure 1a, Supplementary Figure 1c). CIP converts tri/di and monophosphorylated ends to 5'OH [16], abolishing or greatly diminishing the number of reads mapping at non-Pol-II TSS positions. Comparison of the ReCappable-seq dataset with the CIP dataset reveals two populations composed of CIP resistant and CIP sensitive TSS, consistent with Pol-II and non-Pol-II transcripts respectively (Figure 2b).

Pol-II TSS are found at 33,468 positions for 7,186 annotated genes. Consistent with the function of Pol-II polymerase, the majority (89%) of ReCappable-seq Pol-II TSS are found either upstream or within protein coding genes (Figure 3a). There are 5,269 non-Pol-II TSS (Supplementary data 1) accounting for about 50% of the total reads in spite of representing only one seventh of all the TSS positions. This is consistent with the fact that typical non-Pol-II transcripts are very abundant in the cell. Non-Pol-II TSS are detected within or upstream of only 326 genes corresponding to rRNA (mostly 5S RNA, 5.6% of TSS), lincRNA (2.5%), protein coding genes (2.5%) and miscellaneous RNA (2.6%) (Figure 3a). In contrast to Pol-II TSS, the vast majority (82.9%) of the non-Pol-II TSS cannot be assigned to any GENCODE annotation (Figure 3a). We therefore investigated the origin of the unannotated non-pol-II TSS by using specific annotation datasets. We first compared our non-Pol-II TSS with the comprehensive list of 387 previously published Pol-III genes (including tRNA prediction)[17] and found that ReCappable-seq detects non-Pol-II TSS for 327 of these 387 genes. Nonetheless, ReCappable-seq identifies 2,569 additional non-Pol-II TSS that cannot be assigned to the list of published Pol-III genes [17].

Around one third of the non-Pol-II TSS (34%) are derived from reads mapping to multiple positions in the genome, consistent with the role of these polymerases in transcribing repeat elements [18]. Pol-III RNA polymerase for example, has been previously shown to transcribe repeat genes such as 5S, 7SK and 7SL [19]. We therefore proceeded with a genome-wide investigation of TSS relative to repeat genes and found non-Pol-II TSS at the starts of 5S and 7SL genes as well as upstream of tRNA, HY, U6 and MIR repeat genes (Supplementary Figure 5a). This result is in agreement with previous literature that has assigned transcription of these repeat genes to Pol-III [18]. Interestingly, a closer look at the 7SL gene identifies 2 highly expressed TSS at +1 and -1 bp relative to the annotated gene (Figure 4b). The -1 TSS is mostly eliminated by CIP treatment consistent with the triphosphorylated nature of the 5'end. Conversely, and to our surprise, the +1 TSS is not affected by CIP treatment and is also found by CAGE, consistent with a possible canonical cap structure at the 5' end of 7SL, but sharply contrasting with the body of literature that describe 7SL genes as being transcribed by Pol-III [19]. We experimentally confirmed the presence of both TSS using 5' RACE followed by Sanger sequencing (Figure 4c, Materials and Methods).

Likewise, TSS identified at the start of 7SK genes appear resistant to CIP treatment and as such were classified as pol-II TSS (Supplementary Figure 5a). Studies have shown that the 5' end of the 7SK transcript contains a unique structure composed of a gamma-phosphate monomethylation on the triphosphate [20] conferring resistance to alkaline phosphatase [21]. We therefore investigated whether the gamma -phosphate monomethylation can block the dephosphorylation activity of the CIP enzyme and consequently would misclassify 7SK TSS as pol-II TSS. Briefly, we cloned and isolated the human methyl phosphate capping enzyme (hMePCE) to add a gamma-methylation to an in vitro transcript and show that the gamma-methylated triphosphate end is resistant to CIP treatment (Supplementary text 1, Supplementary Figure 5b). Furthermore, VCE can remove the gamma phosphate for the gamma-methylated triphosphate indicating that VCE can cap a gamma-methylated triphosphate end (Supplementary text 1). Together, these results explain why 7SK TSS are classified as CIP resistant TSS and consistent with gamma-phosphate monomethylation.

This should also be the case for the U6 small nuclear RNA which is also known to possess a gamma-methyl triphosphorylated 5' end [22]. In our data, however, we found no TSS signal from the U6 gene at the annotated start, instead we found CIP sensitive TSS around 20 bp upstream of the U6 annotation (Supplementary Figures 5a, 6b) consistent with a canonical triphosphate 5' end at these positions. Interestingly, we found CIP sensitive TSS at the start of the annotated pseudogene RNAU6ATAC (Supplementary data 1).

Pre-tRNAs have been notoriously difficult to study due to rapid processing of primary transcripts relative to the exceptional stability of the mature tRNA which consequently accumulates in the cell [23]. With the ability of ReCappable-seq to capture all primary transcripts, we are now in a unique position to interrogate the TSS landscape of tRNAs. Using annotated tRNAs from GtRNAdb [24], we found 3,245 of the non-Pol-II TSS are located upstream or within tRNA genes representing the largest fraction of non-Pol-II TSS found by ReCappable-seq (Supplementary Figure 3).

TSS positions relative to the tRNA annotation highlight a large number of TSS approximately 5 bp upstream of mature tRNA 5' ends (Figure 4a) consistent with previous work using in vitro transcription which identified TSS located mostly 10 to 2 bp upstream of mature plant tRNAs [25]. Additionally, we found 3 novel TSS clustered at around -17, +7 and +40bp from the mature tRNA 5'end (Figure 4a). The downstream TSS cluster at +40bp is consistent with the notion that tRNAs might have originated from tRNA halves [26]. Further refinement relative to the type of tRNA reveals that downstream TSS are mostly found in leu tRNA and lys tRNA (Supplementary Figure 7).

ReCappable-seq also identified TSS for the vault RNAs (Supplementary Figure 6a), RMRP (Supplementary Figure 6c) and RPPH1 genes (Supplementary Figure 6d) among other genes previously shown to be transcribed by Pol-III.

Beyond TSS annotation, we examined the nucleotide composition for Pol-II and non-Pol-II TSS and found that 83.7% and 79.5% respectively of the +1 nucleotides are A or G and 70.6% and 61.7% respectively of the -1 position are C or T (Figure 3b) revealing the canonical -1[CT]+1[GA] motifs typical for TSS. Limiting TSS to uniquely mapped reads, we also interrogated the conservation profiles around Pol-II and non-Pol-II TSS. We found a peak of conservation for Pol-II TSS consistent with the fact that these sites are functional and therefore under selection (Figure 3c). Interestingly for non-Pol-II TSS, the conservation profile is very different from the Pol-II TSS indicating a very different evolutionary conservation for both types of transcripts (Figure 3c).

Genomic marks at non-Pol-II TSS can be distinct from Pol-II TSS. Consistent with the predicted transcribing polymerase, we found a higher density of Pol-II ChIP tags [17] around Pol-II TSS and a higher density of Pol-III ChIP tags around non-Pol-II TSS (Figure 3d). Interestingly, the Pol-II TSS also show a lower fraction of Pol-III tags relative to the flanking regions suggesting that genomic regions at Pol-II TSS are actively devoid of bound Pol-III. Interestingly, non-Pol-II TSS show absence of H3K36me3 and H3K79me2 signals (Supplementary Figure 4). H3K36 histone modification marks are enriched on the gene body region and play important roles in transcriptional activation [29] while H3K79me2 marks have been shown to be strongly correlated with gene activity [30]. Our data suggest that the enrichment observed for H3K36me3 and H3K79me2 are specific to Pol-II transcripts (Supplementary Figure 4). Other histone marks show distinct profiles at Pol-II and non-Pol-II TSS : H3K27ac for example, is a known histone mark for regions proximal to the TSS. While H3K27ac is observed both upstream and downstream of Pol-II TSS, H3K27ac marks are found only upstream of non-Pol-II TSS (Supplementary Figure 4). These distinct profiles between Pol-II and non-Pol-II TSS are also observed to a lesser degree for the H2AFZ, H3K4me2 and H3K4me3 histone marks (Supplementary Figure 4).

Distinctive DNase I accessibility profiles can be observed at Pol-II and non-Pol-II TSS. Pol-II transcripts show maximum DNase I accessibility a few nucleotides upstream of their TSS (Figure 3e) consistent with nucleosome depletion at pol-II TSS [31]. In contrast, non-Pol-II transcripts show minimum DNase I accessibility at TSS (Figure 3e). Distinctive nucleosome positioning has been previously observed for promoters of Pol-II and Pol-III RNA polymerases [32]. These results further extend the distinctive accessibility of DNA at TSS of Pol-II and non-Pol-II transcripts and more generally highlight a remarkably distinct epigenetic landscape for Pol-II and non-Pol-II TSS.

Identification of the human mitochondrial TSS

The property of ReCappable-seq to capture the 5' di and triphosphorylated transcripts offers the unique ability to interrogate the TSS landscape of the mitochondrial genome. The entire mitochondrial genome is known to be transcribed from both strands as long polycistronic transcripts [33]. The light strand promoter (LSP) controls the transcription of eight of the tRNAs and the MT-ND6 gene. On the heavy strand, two-promoter systems (HSP1 and HSP2) have historically been proposed to explain the higher abundance of the rRNAs. However, the two promoter model remained controversial as more recent experiments [34][35] suggest that heavy strand transcription is under the control of a single promoter (HSP1) and that the difference in abundance may be a consequence of differential turnover [33].

In agreement with previous literature, ReCappable-seq identifies LSP and HSP1 TSS at nucleotide resolution: Both TSS have a strong major peak and a few minor peaks indicative of an imprecise start of transcription (Supplementary Figure 8a and b). Interestingly, most of the transcripts starting at the major HSP1 TSS show signs of stuttering with non-templated adenosine being added to the start of the transcripts (Supplementary Figure 8b). Furthermore, we did not find TSS at the H2 position reinforcing the absence of an HSP2 promoter as indicated by previous studies [34][35]. Instead, we found three putative novel TSSs across the mitochondrial genome, two on the heavy strand (positions MT:2434 and MT:3242, human GRC38) and one on the light strand (position MT:16029, human GRCh38) (Figure 5b). Interestingly, the novel TSS on

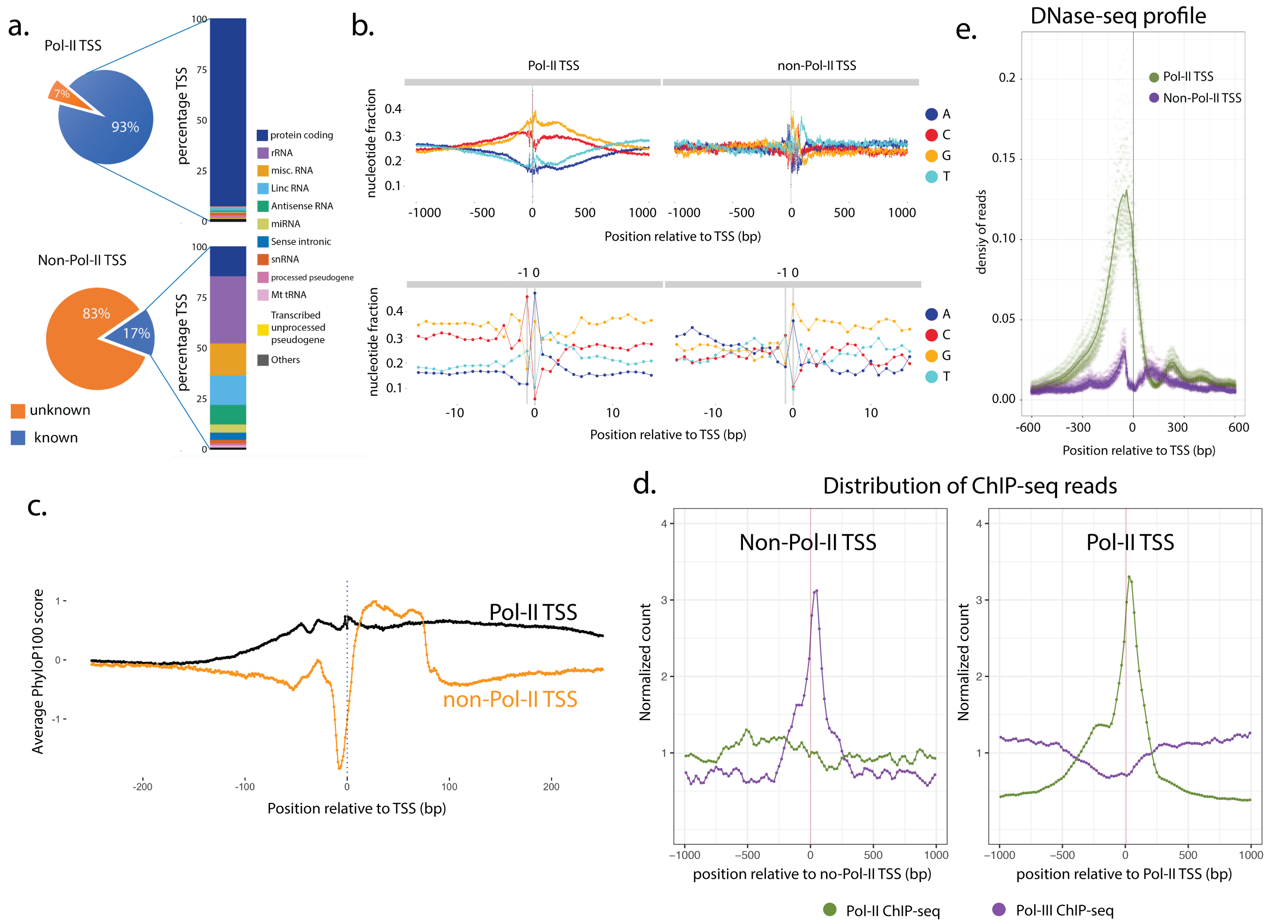


Figure 3. TSS Annotation : a. Pie chart representing the percentage of TSS that are associated with an annotated GENCODE gene (blue) or not annotated (orange) for Pol-II (upper pie chart) and non-Pol-II (lower pie chart). TSS associated with annotated genes were further broken down to gene types. b. Nucleotide composition in the 2 kb flanking regions (upper panels) and a zoom into 15 bp flanking regions (lower panels) for Pol-II TSS (left panels) and non-Pol-II TSS (right panels). 83.66% and 79.5% of Pol-II and non-Pol-II TSS respectively are starting with A or G. c. Conservation profiles using PhyloP basewise conservation score [27] derived from Multiz alignment [28] of 100 vertebrate species around Pol-II TSS (black) and non-Pol-II TSS (orange). d. Pol-II (green) and Pol-III (purple) ChIP-seq binding profiles around Pol-II TSS (right panel) and non-Pol-II TSS (left panel). e. Open chromatin profiles from ENCODE DNase-seq data for Pol-II (green) and non-Pol-II (purple) TSS.

the light strand (Supplementary Figure 8c) is located 6bp upstream for the Proline tRNA and shows similar conformation to the HSP1 with notable stuttering of non templated adenosine being added to the start of the novel transcripts. Similar nucleotide composition starting at the TSS can be found for HSP1, LSP and the predicted novel TSS with AAAGA as the common motif.

Comparison with CAGE

CAGE technology [4] relies on capturing the canonical cap structure of Pol-II transcripts using the cap-trapper method [36] and is considered the best performing method for TSS analysis [5]. We evaluated how ReCappable-seq results compare with CAGE by performing CAGE experiments on the same starting material.

Technical replicates of CAGE experiments were performed (Materials and Methods) and approximately 30 million 50 bp single end reads were obtained. CAGE reads were mapped to the reference human genome (GRCh38) and analyzed using the same pipeline as that used for ReCappable-seq described in the Methods section. Similar to ReCappable-seq, CAGE datasets show very good reproducibility between technical replicates (Pearson corr=0.99, P-value < 2.2e-16, Supplementary Figure 1d). CAGE replicates were combined to obtain the same number of mappable reads relative to ReCappable-seq.

A. Total reads

We first determined the percentage of reads mapping to

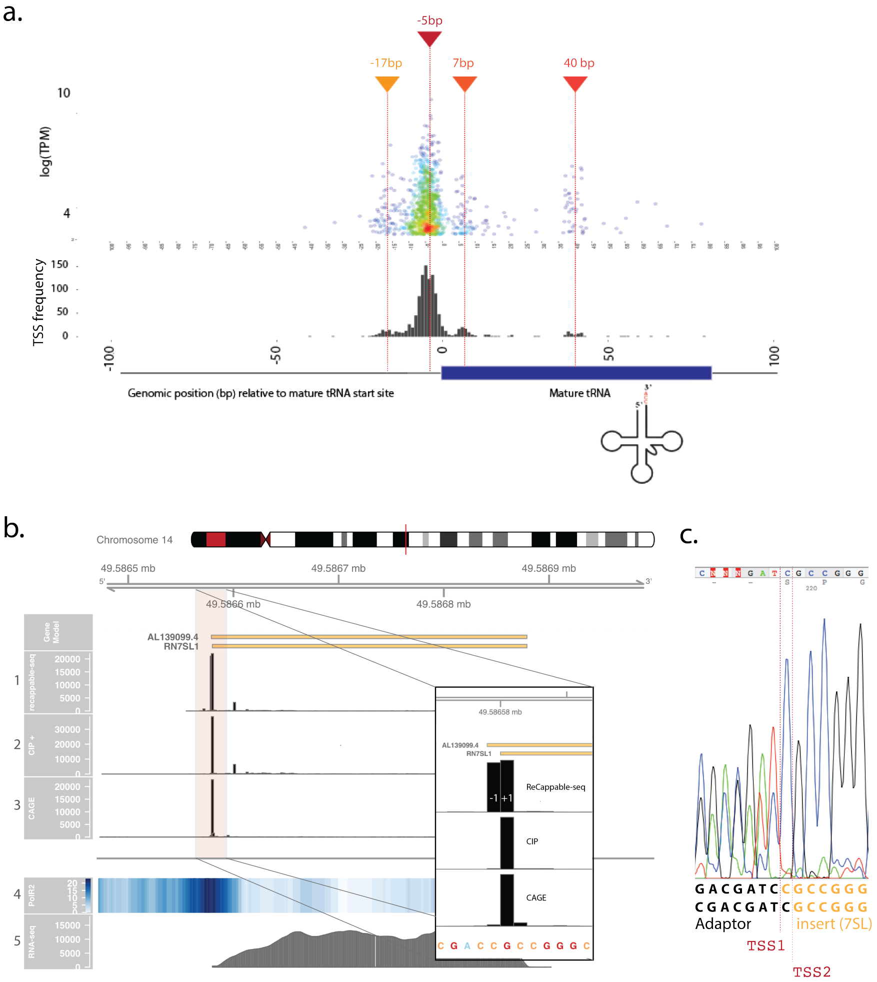


Figure 4. a. Non-Pol-II TSS flanking tRNA annotations. Top panel visualizes individual non-Pol-II TSS relative to the 5' end of the annotated mature tRNA (in bp) as a function of the TPM. Bottom panel represent the distribution of the non-Pol-II TSS relative to the start of the annotated tRNA starts (in bp). b. Example of the 7SL (RN7SL1) locus showing the distribution 5'end of mapped reads for Recappable-seq (panel 1), CIP treated Recappable-seq (panel 2) CAGE (panel 3), Pol-II ChIP-seq (Panel 4, density of reads) and RNA-seq reads (panel 5). Floating panel represents a close up of the 5'end of the RN7SL1 with essentially two TSS at -1 and +1 of the annotated RN7SL1. Only the most 5' mapped end of the reads is represented for Recappable-seq (panel 1), CIP treated Recappable-seq (panel 2) and CAGE (panel 3) to highlight the TSS positions. c. Validation of the two TSS identified for 7SL using RACE. The amplified fragments were directly sequenced using sanger sequencing with a primer located in the 7SL gene (Material and Methods). The sequencing trace reveals two products ligated to the RACE adaptor resulting from two alternative transcript starts corresponding to TSS1 and TSS2.

ribosomal RNA genes. Since processed rRNAs in this context should be considered as background contamination, similar background contamination can be expected from degraded or processed mRNA. We found that the CAGE dataset has 34% of contaminating rRNA (Supplementary Figure 2) contrasting with ReCappable-seq with only 3% remaining rRNA (Supplementary Figure 2). This difference indicates a substantially higher level of non-capped processed RNA in CAGE compared to ReCappable-seq libraries.

Another measure of contamination with non-primary 5'end reads is the fraction of genomic positions with single-read evidence only (singleton positions). These positions may be indicative of very weak TSS or, more likely, are the result of contamination with processed or degraded RNA. We found that 62% of the genomic positions found by CAGE are singleton positions compared to only 35% for ReCappable-seq. Singletons and low-coverage positions are typically removed from downstream analysis representing a notable loss of data in CAGE libraries. These results indicate an appreciable contamination with non-capped RNA in the CAGE libraries.

B. TSS positions at single nucleotide resolution :

CAGE is limited to Pol-II TSS, therefore we restricted the ReCappable-seq dataset to CIP resistant TSS consistent with transcripts derived only from Pol-II. Furthermore, CAGE libraries do not enable the equivalent of the ReCappable-seq unenriched control libraries to be obtained, preventing the identification of high confidence TSS in a similar manner. For direct comparison, we used ReCappable-seq Pol-II TSS without applying the enrichment score cutoff (Material and Methods, Figure 2b, quadrants III and IV). From this total of 42,988 candidate TSS we determined that the subset of 37,111 CIP resistant Pol-II candidate TSS detects 7,216 genes (Figure 5a). For CAGE, 92,786 TSS were identified detecting a total of 9,536 genes. The positions of both CAGE and ReCappable-seq Pol-II candidate TSS are in agreement, with 76% of ReCappable-seq positions located within 5bp of a CAGE position.

Uniquely mapped TSS are found evenly distributed across chromosomes with the notable exception of the mitochondrial genome: in total there are 18 candidate TSS positions in the mitochondrial genome (amongst them, 14 are high confidence TSS positions) defined by ReCappable-seq, from which 10 positions are located within 10 bp of the two known TSS [33] downstream of the hHSP1 (H-strand) and the hLSP (L-strand) promoters (Figure 5b). In contrast, there are 730 TSS identified by CAGE distributed across the entire mitochondrial chromosome (Figure 5b). Because the number of mitochondrial TSS is known to be limited to only 2-3 TSS in the human mitochondrial genome [33], the presence of a large number of sites in the CAGE dataset is indicative of spurious TSS. Furthermore, more than half of the CAGE mitochondrial TSS (487) are located within the ribosomal genes known to be the most highly expressed locus of the mitochondrial genome. We confirmed high level of mitochondrial rRNA expression by performing RNA-seq from total RNA (Figure 5b). The same overall trend is obtained on a nuclear rRNA locus with 2,955 CAGE TSS predicted within the rRNA locus versus only 487 Pol-II ReCappable-seq candidate TSS (among which 61 are high confidence TSS) (Supplementary Figure 9a). This large number of spurious CAGE TSS in the mitochondrial genome, specifically in the rRNA locus demonstrate a higher background level in CAGE data compared to ReCappable-seq.

To examine whether this background of CAGE also applies to nuclear derived protein coding transcripts, we used RNA-seq to identify the top 300 highly expressed protein coding genes in the human nuclear genome (Supplementary data 1). CAGE and ReCappable-seq identified TSS for 291 and 282 genes respectively. While most of the TSS are located near the 5'end of each gene for both technologies, more CAGE TSS can be found dispersed across the gene body as compared to Recappable-seq candidate TSS (Supplementary Figure 9b). Again, this pattern is indicative of lower false positive rate of TSS in ReCappable-seq data.

C. Clustered TSS:

We identified TSS clusters based on clustering of reads within a 20 bp window. Clusters containing at least one position with TPM >= 1 were compiled resulting in 11,569 TSS clusters for CAGE and 10,802 Pol-II candidate TSS clusters for ReCappable-seq. A total of 6,876 clusters (64%) are shared between CAGE and ReCappable-seq. Clusters obtained using CAGE are systematically larger with a median of 148 bp for CAGE compared to 69 bp for ReCappable-seq (Figure 5d). Using the shared clusters, we refined cluster sizes to common interval between CAGE and ReCappable-seq and compared both the total number of reads and genomic positions per common genomic interval for each dataset. While CAGE and ReCappable-seq show acceptable correlation in terms of number of reads per cluster, CAGE systematically detects more TSS positions per cluster (Figure 5c). These results together with the rRNA analysis above, indicate that there is a higher background level in CAGE data and this background is affecting the cluster boundaries and the number of individual TSS per cluster.

Collectively, our results indicate that ReCappable-seq has higher specificity for TSS compared to CAGE. The higher specificity of ReCappable-seq impacts its performance on 3 levels : first, fewer reads map to either ribosomal loci or are singletons/low coverage positions which are discarded, therefore allowing a lower sequencing depth than CAGE. Second, a smaller number of false positive TSS are detected by ReCappable-seq leading to an improved annotation of predicted isoforms and promoters. Finally, ReCappable-seq TSS clusters are narrower with fewer false positive positions than CAGE.

Discussion

In this work, we describe a novel technology, ReCappable-seq, which identifies eukaryotic TSS of all primary transcripts genome-wide at single nucleotide resolution. ReCappable-seq identifies TSS independently of the transcribing polymerases because it captures capped and triphosphorylated primary transcripts. In this respect, ReCappable-seq represents a significant departure from existing technologies that either target prokaryotic TSS [37] [9] or eukaryotic Pol-II TSS [5]. With the growing realization of the impact of non-coding RNA, the ability of ReCappable-seq to provide the global landscape of TSS is an important advance.

Similarly, ReCappable-seq can be applied to complex communities composed of both prokaryotic and eukaryotic organisms with the expected outcome of detecting all TSS regardless of the organism or transcribing polymerases while simultaneously depleting mature rRNA.

In this study we focused on the human transcriptome and demonstrate that ReCappable-seq identifies TSS from transcripts derived from Pol-I, Pol-II, Pol-III and POLRMT RNA polymerases. ReCappable-seq is in good agreement with

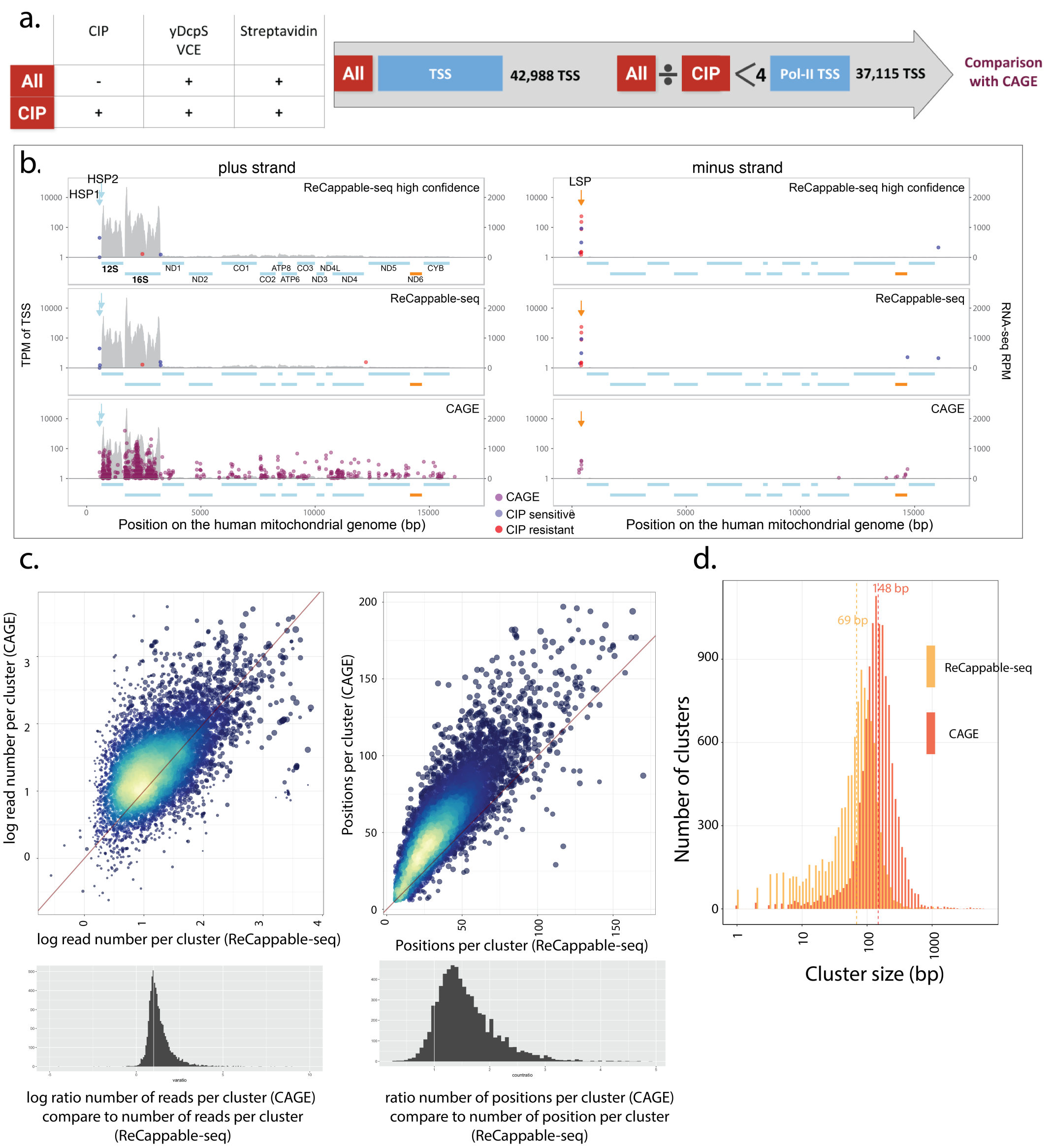


Figure 5. ReCappable-seq compared to CAGE. a. ReCappable-seq (ALL) and CIP treated ReCappable-seq (CIP) positions were compared to define the Pol-II consistent TSS fraction. The resulting 37,111 Pol-II TSS were compared with CAGE.b. ReCappable-seq high confidence TSS (top panel), ReCappable-seq candidate TSS (middle panel), and CAGE TSS (bottom panel) located on the plus strand (left panels) and minus strand (right panels) of the mitochondrial genome. RNA-seq data are shown in gray. c. Comparison between the number of TSS reads in ReCappable-seq versus CAGE (left) and TSS genomic positions in ReCappable-seq versus CAGE (Right). d. Distribution of CAGE clusters (dark orange) and ReCappable-seq clusters (light orange) as a function of their size (in bp, x-axis).

CAGE results while offering higher specificity for the 5’ end of primary RNA transcripts. We further confirm the specificity of ReCappable-seq for genuine TSS by demonstrating that the TSS obtained from highly degraded samples are comparable to TSS obtained using intact RNA.

ReCappable-seq can be complemented with an unenriched control library to better discriminate between primary and processed transcripts leading to the identification of high confidence TSS. With an estimated <10% of false positive TSS positions, this step is optional but is very useful for studying highly expressed genes. This higher specificity is particularly evident when contrasting the TSS landscape of ReCappable-seq and CAGE in the mitochondrial genome: it is reasonable to conclude that the 730 TSS identified by CAGE are false positives since none of the mitochondrial derived transcripts are known to be G capped. In contrast, the high confidence TSS of ReCappable-seq identifies the known hHSP1 and hLSP TSS and also only a handful of novel high confidence TSS.

Finally, by complementing with a library where the ReCappable protocol is preceded by a dephosphorylation (CIP) step, TSS can be classified into Pol-II and non-Pol-II consistent TSS. We have shown that this classification of TSS into two distinct groups based on the sensitivity to CIP treatment is accurate for transcripts with canonical cap structures, which correspond to the vast majority of transcripts. Interestingly, ReCappable-seq identifies some unusual Pol-II consistent transcripts that are not capped with the canonical cap structure. We show for example, that while 7SK RNA is a known Pol-III transcript, the TSS of 7SK genes appear CIP resistant. This disparity is due to the CIP resistant gamma-methylated triphosphate found in 5'ends of 7SK transcripts and consequently the 7SK TSS is classified here as Pol-II consistent. The known Pol-III transcript 7SL, is another example of interesting transcript with two adjacent 5' ends. The one corresponding to the annotated gene start is classified as Pol-II while the other one located one nucleotide upstream is classified as Pol-III. Although, theoretically, the Pol-II classification may be a result of the inability of the CIP enzyme to access the triphosphate 5' end, this Pol-II consistent TSS is confirmed by CAGE (Figure 4b). Further investigation is required to uncover the exact nature of the 7SL 5' cap structure.

Beyond the 7SL and 7SK examples, the systematic complementation of ReCappable-seq results with orthogonal datasets such as CAGE, can be used as a discovery platform to identify other interesting non-canonical capped structures. Indeed a number of non-canonical CAP structures recently reported [38] [39] [40] are expected to have distinctive outcomes. For example the trimethyl G cap is expected to be resistant to yDcpS [11] but captured by CAGE, leading to a discrepancy between ReCappable-seq and CAGE. Despite trimethyl G cap having been described on U1, U2, U4, and U5 snRNAs [41], ReCappable-seq identifies a clear pol-II consistent TSS at the start of these genes suggesting that only a fraction of the 5' end of these transcripts has been fully methylated to trimethyl G.

NAD cap is another structure that will not be decapped by yDcpS [11] but captured by CAGE because of the presence of the 2', 3' diol on the ribose. As an example, the HSP1 TSS in the mitochondria has a strong CAGE signal consistent with the NAD cap recently reported in human mitochondria [39] while ReCappable-seq shows a CIP sensitive signal consistent with a triphosphate end. Taken together, these results suggest a mixed population of RNA 5' ends at the HSP1 TSS position.

Finally, with the ability to identify the 5' ends of primary transcripts with or without a canonical cap structure, ReCappable-seq reveals a rich landscape of novel TSS that were unknown due to the lack of appropriate technologies to investigate them.

Supplementary Data

Supplementary Data 1 : List of TSS found in the human genome (GRCh38). Columns correspond to Chr (chromosome), TSS (position of the TSS, 1-coordinates), strand (+ or -), TPM (TSS tags per million), Ratio Control (TPM ReCappable-seq divided by TPM unenriched control), Ratio CIP (TPM ReCappable-seq divided by TPM CIP), Ratio CAGE ((TPM ReCappable-seq divided by TPM CAGE), feature (quadrant I, II, III or IV according to Figure 2b), source (annotation source, either GENCODE or tRNA prediction), geneType (annotation type), gene (name of the associated gene). N.A. corresponds to TSS with unknown origin.

Supplementary Data 2 : List of TSS clusters defined by ReCappable-seq. Columns represent Chr (chromosome), source, feature, Start of TSS cluster, End of TSS cluster, Total TPM of cluster, strand (+ or -), coordination, attribute. Attribute contains: number of tags of all the TSS in the cluster (nio), total TPM of all the TSS in the cluster (TPM), TSS position with the largest number of tags in the cluster (summit), number of tags of the summit (nio summit), TPM of the summit (TPM summit).

Data availability

Sequencing data and processed results were deposited and available on GEO (GSE132660).

(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132660). High confidence TSS defined by ReCappable-seq have been uploaded as a custom track in the UCSC genome browser:

(http://genome.ucsc.edu/s/rezo/ReCappable-seq). High confidence TSS defined by ReCappable-seq; bedGraph format; negative data values represent CIP sensitive TSS (Quadrant II), positive data values represents CIP resistant TSS (Quadrant IV, Figure 2).

Acknowledgments

We would like to thank the NEB sequencing core facility for NGS sequencing and New England Biolabs for supporting this research.

References

[1] R. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones, and M. Marra. Profiling the HeLa S3 transcriptome using ran-

domly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1):81–94, Jul 2008.

[2] R. E. Breitbart, A. Andreadis, and B. Nadal-Ginard. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu. Rev. Biochem.*, 56:467–495, 1987.

[3] A. Reyes and W. Huber. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.*, 46(2):582–592, Jan 2018.

[4] M. Murata, H. Nishiyori-Sueki, M. Kojima-Ishiyama, P. Carninci, Y. Hayashizaki, and M. Itoh. Detecting expressed genes using CAGE. *Methods Mol. Biol.*, 1164:67–85, 2014.

[5] X. Adiconis, A. L. Haber, S. K. Simmons, A. Levy Moonshine, Z. Ji, M. A. Busby, X. Shi, J. Jacques, M. A. Lancaster, J. Q. Pan, A. Regev, and J. Z. Levin. Comprehensive comparative analysis of 5’-end RNA-sequencing methods. *Nat. Methods*, 15(7):505–511, 07 2018.

[6] M. G. Ivanchenko and M. Megraw. NanoCAGE-XL: An Approach to High-Confidence Transcription Start Site Sequencing. *Methods Mol. Biol.*, 1830:225–237, 2018.

[7] K. Maruyama and S. Sugano. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, 138(1-2):171–174, Jan 1994.

[8] L. Marshall and R. J. White. Non-coding RNA production by RNA polymerase III is implicated in cancer. *Nat. Rev. Cancer*, 8(12):911–914, 12 2008.

[9] L. Ettwiller, J. Buswell, E. Yigit, and I. Schildkraut. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics*, 17:199, Mar 2016.

[10] H. Liu, N. D. Rodgers, X. Jiao, and M. Kiledjian. The scavenger mRNA decapping enzyme DcpS is a member of the HIT family of pyrophosphatases. *EMBO J.*, 21(17):4699–4708, Sep 2002.

[11] M. G. Wulf, J. Buswell, S. H. Chan, N. Dai, K. Marks, E. R. Martin, G. Tzertzinis, J. M. Whipple, I. R. Correa, and I. Schildkraut. The yeast scavenger decapping enzyme DcpS and its application for in vitro RNA recapping. *Sci Rep*, 9(1):8594, Jun 2019.

[12] B. Yan, M. Boitano, T. A. Clark, and L. Ettwiller. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat Commun*, 9(1):3676, 09 2018.

[13] M. Boutard, L. Ettwiller, T. Cerisy, A. Alberti, K. Labadie, M. Salanoubat, I. Schildkraut, and A. C. Tolonen. Global repositioning of transcription start sites in a plant-fermenting bacterium. *Nat Commun*, 7:13783, 12 2016.

[14] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[15] I. Financsek, K. Mizumoto, Y. Mishima, and M. Muramatsu. Human ribosomal RNA gene: nucleotide sequence of the transcription initiation region and comparison of three mammalian genes. *Proc. Natl. Acad. Sci. U.S.A.*, 79(10):3092–3096, May 1982.

[16] E. F. Fritsch Sambrook, J. and T. Maniatis. Molecular Cloning: A Laboratory Manual. 2nd ed. *Cold Spring Harbor Laboratory Press*, 1989.

[17] A. J. Oler, R. K. Alla, D. N. Roberts, A. Wong, P. C. Hollenhorst, K. J. Chandler, P. A. Cassiday, C. A. Nelson, C. H. Hagedorn, B. J. Graves, and B. R. Cairns. Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat. Struct. Mol. Biol.*, 17(5):620–628, May 2010.

[18] R. J. White. Transcription by RNA polymerase III: more complex than we thought. *Nat. Rev. Genet.*, 12(7):459–463, May 2011.

[19] D. Canella, V. Praz, J. H. Reina, P. Cousin, and N. Hernandez. Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. *Genome Res.*, 20(6):710–721, Jun 2010.

[20] M. S. Cosgrove, Y. Ding, W. A. Rennie, M. J. Lane, and S. D. Hanes. The Bin3 RNA methyltransferase targets 7SK RNA to control transcription and translation. *Wiley Interdiscip Rev RNA*, 3(5):633–647, 2012.

[21] S. Gupta, R. K. Busch, R. Singh, and R. Reddy. Characterization of U6 small nuclear RNA cap-specific antibodies. Identification of gamma-monomethyl-GTP cap structure in 7SK and several other human small RNAs. *J. Biol. Chem.*, 265(31):19137–19142, Nov 1990.

[22] A. L. Didychuk, S. E. Butcher, and D. A. Brow. The life of U6 small nuclear RNA, from cradle to grave. *RNA*, 24(4):437–460, 04 2018.

[23] J. E. Wilusz. Controlling translation via modulation of tRNA levels. *Wiley Interdiscip Rev RNA*, 6(4):453–470, 2015.

[24] P. P. Chan and T. M. Lowe. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, 44(D1):D184–189, Jan 2016.

[25] Y. Yukawa, G. Dieci, M. Alzapiedi, A. Hiraga, K. Hirai, Y. Y. Yamamoto, and M. Sugiura. A common sequence motif involved in selection of transcription start sites of Arabidopsis and budding yeast tRNA genes. *Genomics*, 97(3):166–172, Mar 2011.

[26] Y. Shen, X. Yu, L. Zhu, T. Li, Z. Yan, and J. Guo. Transfer RNA-derived fragments and tRNA halves: biogenesis,

biological functions and their roles in diseases. *J. Mol. Med.*, 96(11):1167–1176, Nov 2018.

[27] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20(1):110–121, Jan 2010.

[28] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, 14(4):708–715, Apr 2004.

[29] E. J. Wagner and P. B. Carpenter. Understanding the language of Lys36 methylation at histone H3. *Nat. Rev. Mol. Cell Biol.*, 13(2):115–126, Jan 2012.

[30] Z. Farooq, S. Banday, T. K. Pandita, and M. Altaf. The many faces of histone H3K79 methylation. *Mutat Res Rev Mutat Res*, 768:46–52, 2016.

[31] M. Radman-Livaja and O. J. Rando. Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.*, 339(2):258–266, Mar 2010.

[32] A. S. Helbo, F. D. Lay, P. A. Jones, G. Liang, and K. Gronbaek. Nucleosome Positioning and NDR Structure at RNA Polymerase III Promoters. *Sci Rep*, 7:41947, 02 2017.

[33] A. R. D'Souza and M. Minczuk. Mitochondrial transcription and translation: overview. *Essays Biochem.*, 62(3):309–320, 07 2018.

[34] M. Terzioglu, B. Ruzzenente, J. Harmel, A. Mourier, E. Jemt, M. D. Lopez, C. Kukat, J. B. Stewart, R. Wibom, C. Meharg, B. Habermann, M. Falkenberg, C. M. Gustafsson, C. B. Park, and N. G. Larsson. MTERF1 binds mtDNA to prevent transcriptional interference at the light-strand promoter but is dispensable for rRNA gene transcription regulation. *Cell Metab.*, 17(4):618–626, Apr 2013.

[35] D. Litonin, M. Sologub, Y. Shi, M. Savkina, M. Anikin, M. Falkenberg, C. M. Gustafsson, and D. Temiakov. Human mitochondrial transcription revisited: only TFAM and TFB2M are required for transcription of the mitochondrial genes in vitro. *J. Biol. Chem.*, 285(24):18129–18133, Jun 2010.

[36] P. Carninci, C. Kvam, A. Kitamura, T. Ohsumi, Y. Okazaki, M. Itoh, M. Kamiya, K. Shibata, N. Sasaki, M. Izawa, M. Muramatsu, Y. Hayashizaki, and C. Schneider. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, 37(3):327–336, Nov 1996.

[37] C. M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiss, A. Sittka, S. Chabas, K. Reiche, J. Hackermuller, R. Reinhardt, P. F. Stadler, and J. Vogel. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, 464(7286):250–255, Mar 2010.

[38] R. F. Abdelhamid, C. Plessy, Y. Yamauchi, M. Taoka, M. de Hoon, T. R. Gingeras, T. Isobe, and P. Carninci. Multiplicity of 5' cap structures present on short RNAs. *PLoS ONE*, 9(7):e102895, 2014.

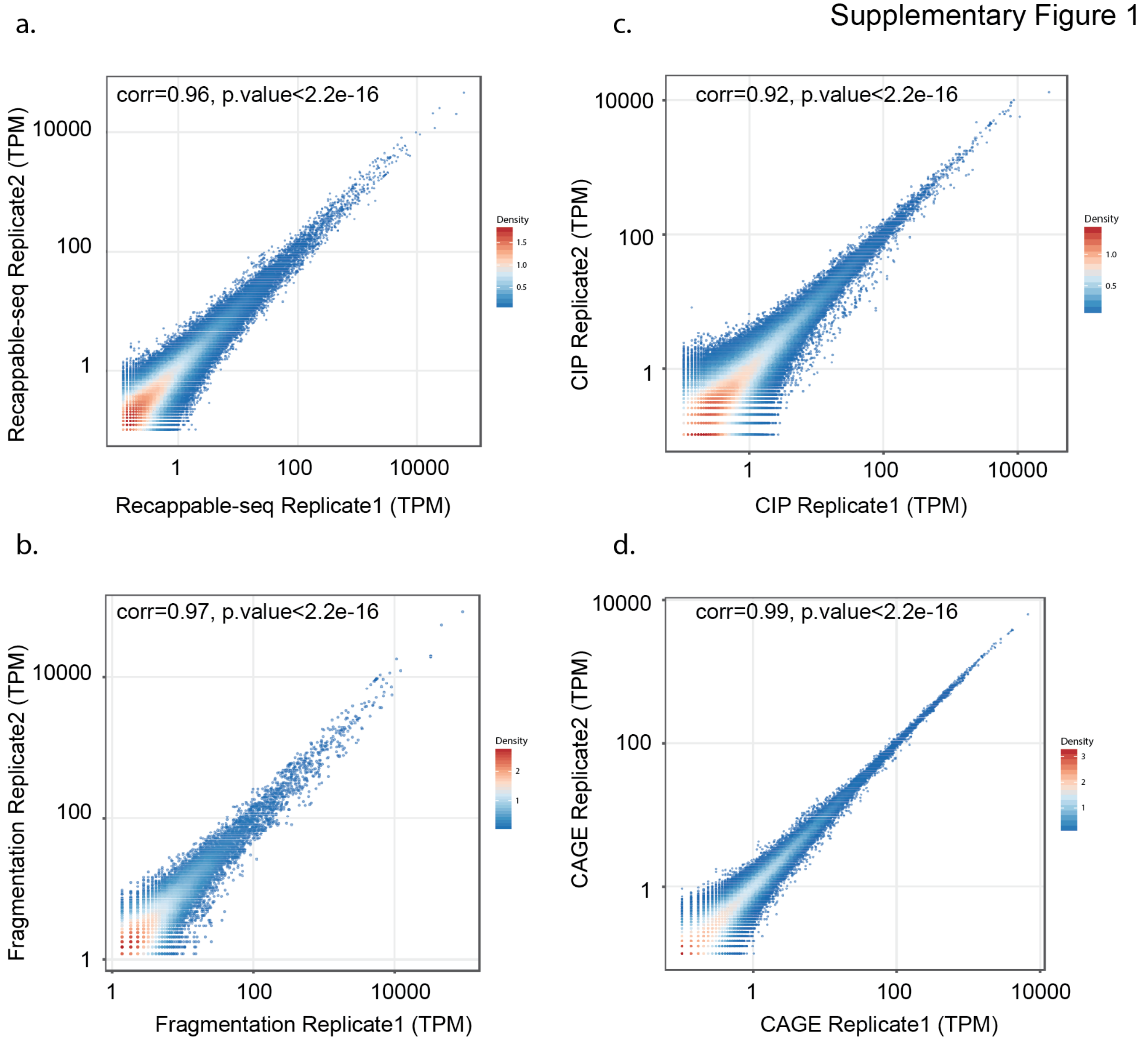
[39] J. G. Bird, U. Basu, D. Kuster, A. Ramachandran, E. Grudzien-Nogalska, A. Towheed, D. C. Wallace, M. Kiledjian, D. Temiakov, S. S. Patel, R. H. Ebright, and B. E. Nickels. Highly efficient 5' capping of mitochondrial RNA with NAD+ and NADH by yeast and human mitochondrial RNA polymerase. *Elife*, 7, 12 2018.

[40] Xu L. Balamkundu S. Cai W.M. Cui L. Liu C.F. Fu X.Y. Lin Z. Shi P.Y. Lu T.K. Luo D. Jaffrey S.R. Dedon P.C. Wang J. Chew L.A. Lai Y. Dong H. Quantifying the RNA cap epitranscriptome reveals novel caps in cellular and viral RNA. *BioRiv*, 683045, 07 2019.

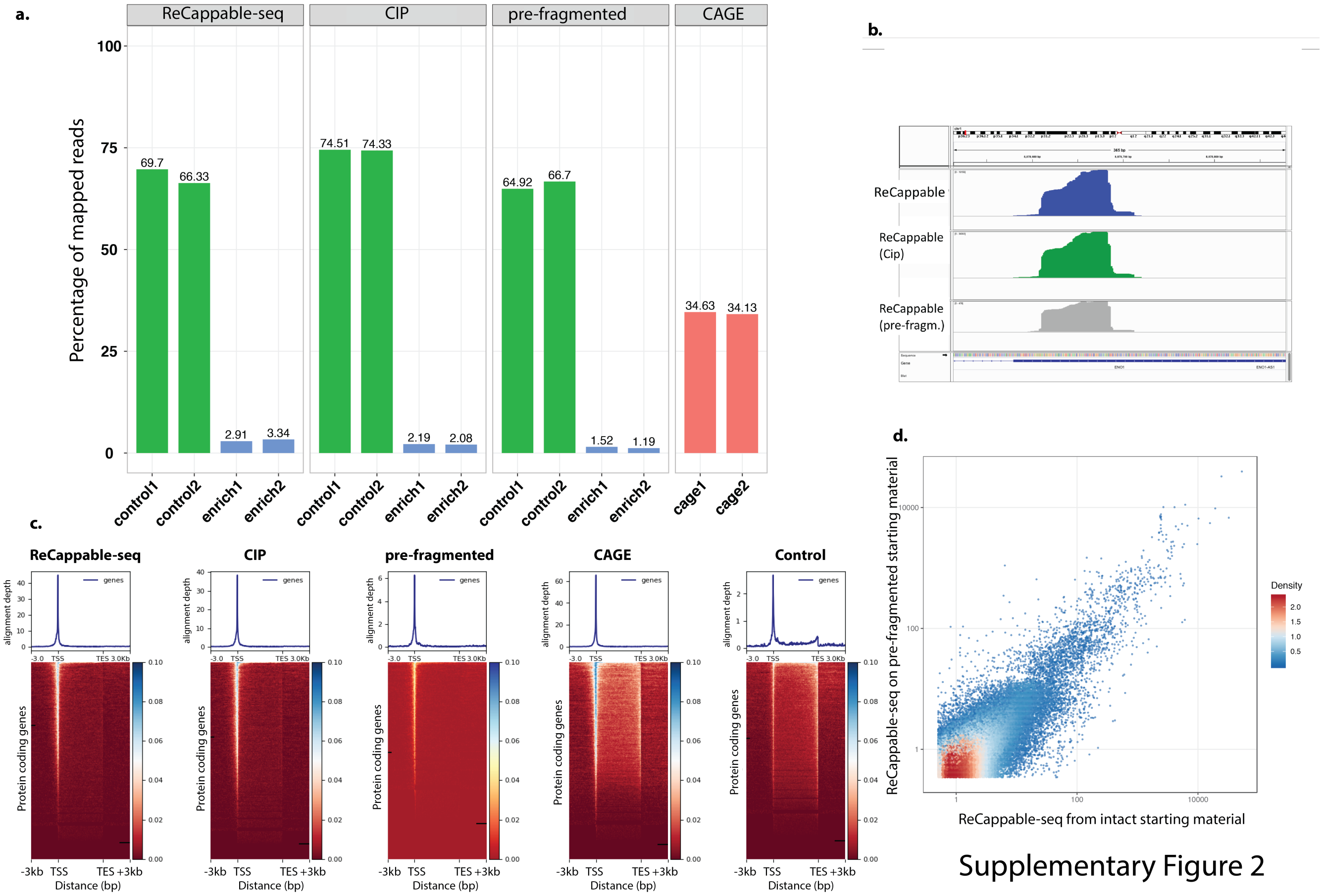
[41] S. Shuman. Transcriptional networking cap-tures the 7SK RNA 5'-gamma-methyltransferase. *Mol. Cell*, 27(4):517–519, Aug 2007.

[42] F. Ramirez, D. P. Ryan, B. Gruning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dundar, and T. Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, 44(W1):W160–165, 07 2016.

Supplementary Figure 1

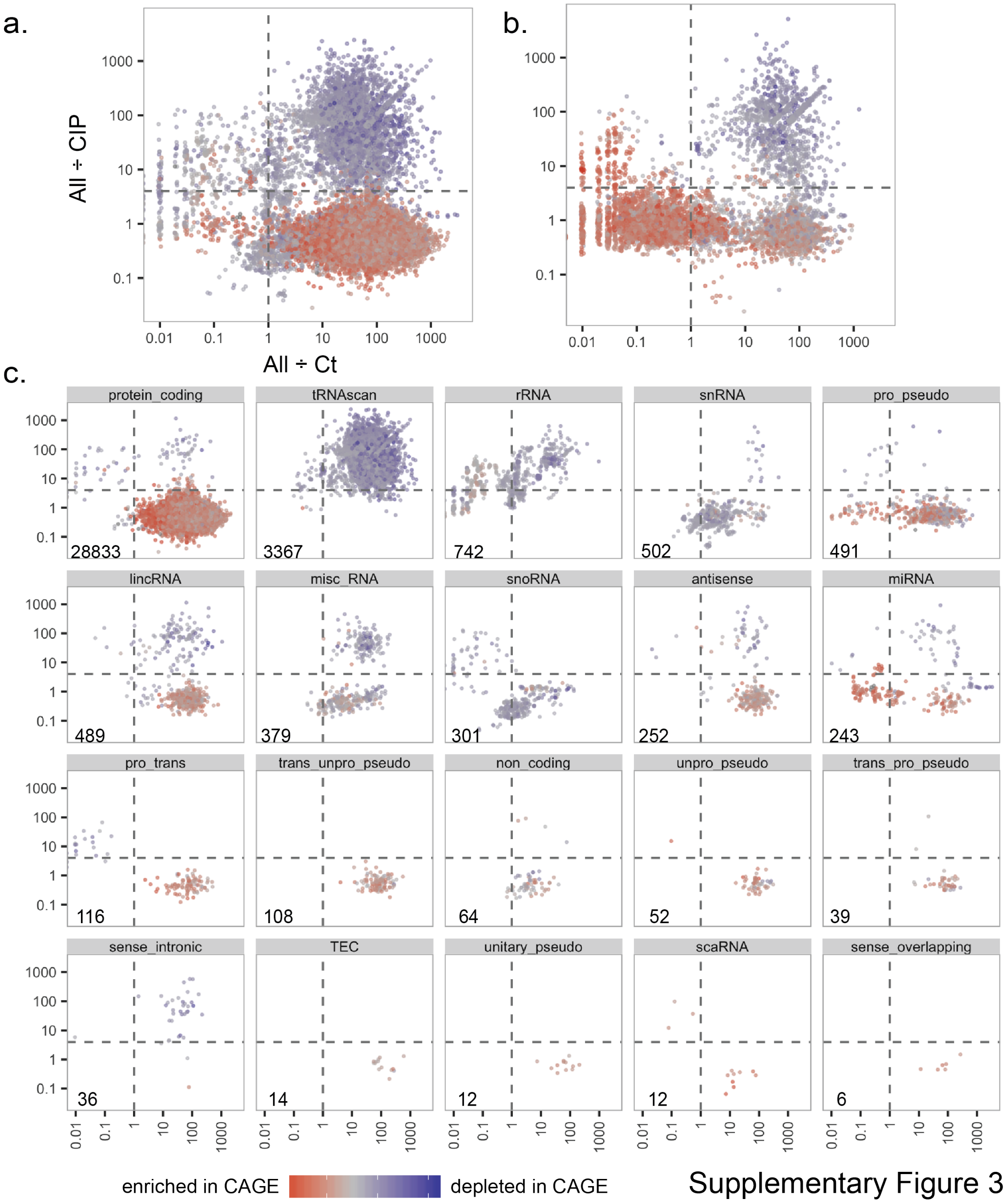


Supplementary Figure 1 : Correlation between technical replicates for a. ReCappable-seq experiments, b. ReCappable-seq performed on degraded sample, c. CIP treated ReCappable-seq (CIP) experiment and d. CAGE experiments. Only the 50 first bases of read 1 were mapped (see Material and Methods) and the replicates were subsequently combined and downsampled to 63 million mappable reads for consistency.



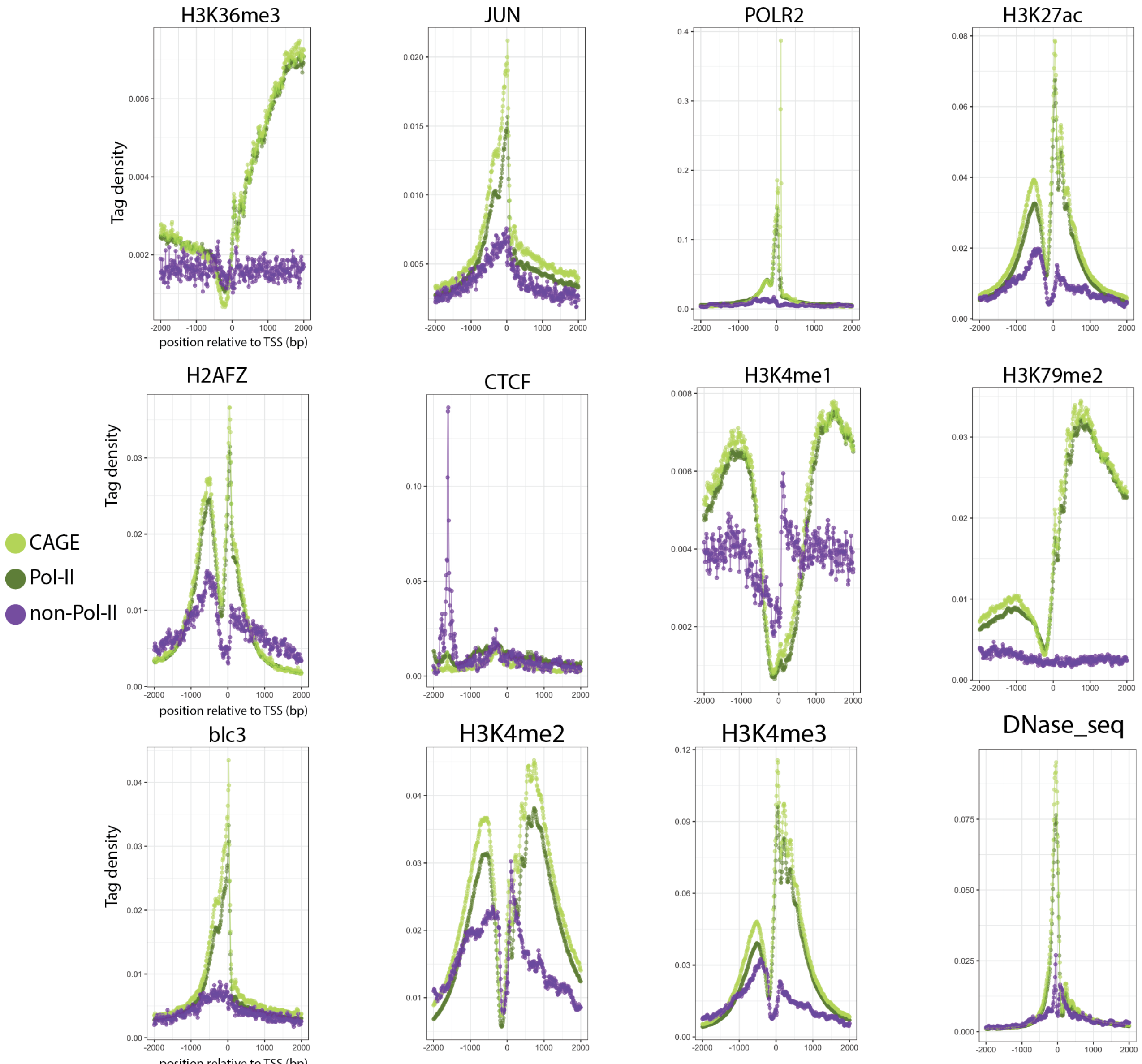
Supplementary Figure 2

Supplementary Figure 2 : a. Percentage of reads mapping to Ribosomal loci compared to all mapping reads in the ReCappable-seq and CAGE datasets. Unenriched control datasets denote Recappable-seq datasets for which the Streptavidin step has been omitted. b. Mapping profiles of ReCappable-seq (top panel), CIP (middle panel) and ReCappable-seq performed on pre-fragmented RNA (bottom panel) upstream of the ENO1 gene. c. Profiles (top panels) and density map (bottom panels) of reads relative to annotated genes for ReCappable-seq, CIP, pre-fragmented, CAGE and unenriched control (minus streptavidin) libraries using deeptools [42]. d. Correlation (corr = 0.76, P-value < 2.2e-16) between TPM from ReCappable-seq derived from intact starting material (x axis) and pre-fragmented starting material (y axis).



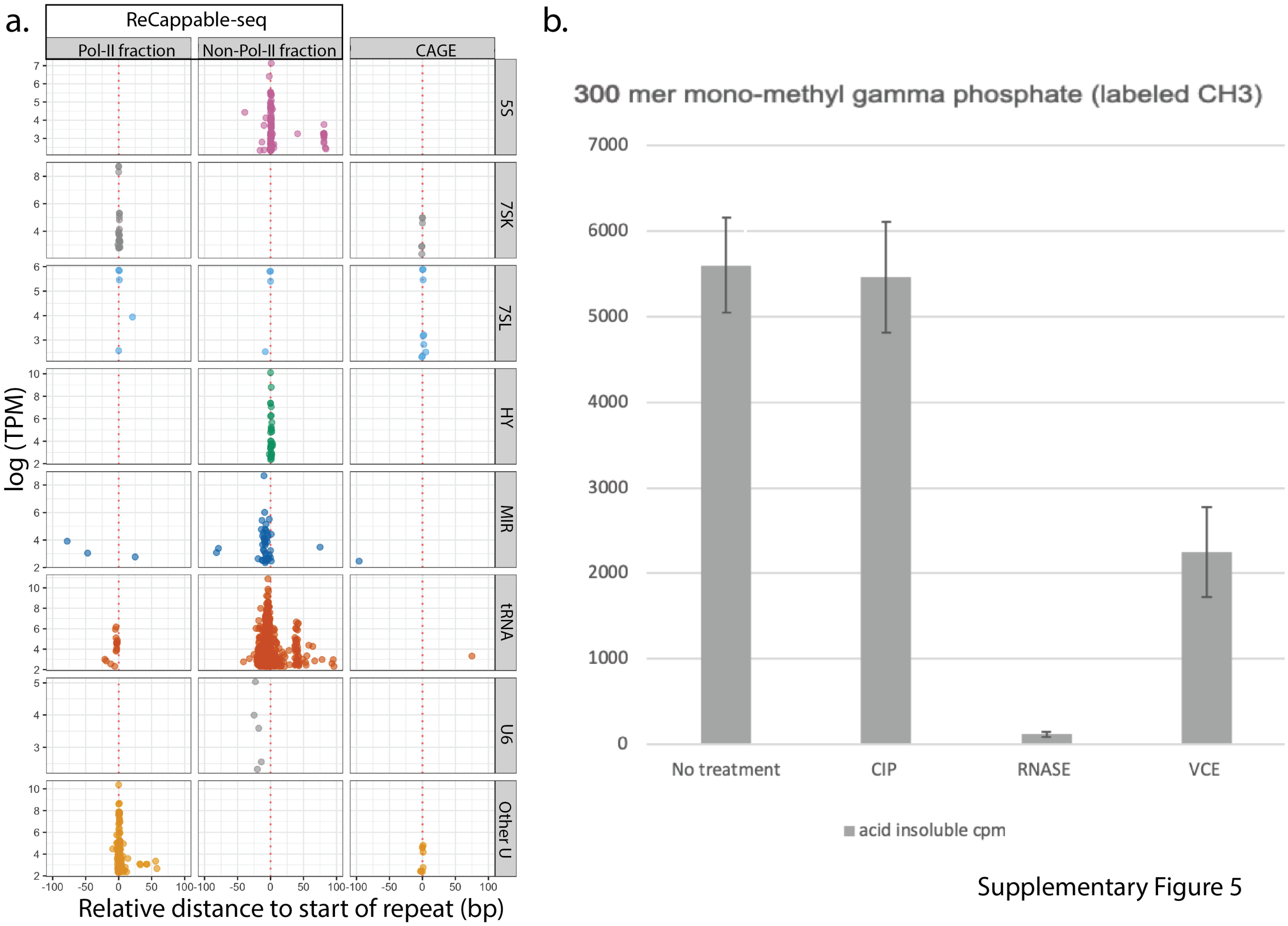
Supplementary Figure 3

Supplementary Figure 3 : Distribution of the ReCappable-seq TSS based on the ratio between TPM-ALL and TPM-Ct (x axis) and the ratio between TPM-ALL and TPM-CIP (y axis). Color of the TSS represents the ratio between TPM-ALL and TPM CAGE. a. Subset of the TSS (36,095) assigned to genes in GENCODE annotation and the tRNA prediction. b. Subset of the TSS (6,893) not assigned to any gene. c. TSS from (a) assigned to individual gene types. The number on the left-bottom corner of each quadrant represents the number of TSS assigned to the corresponding gene type.



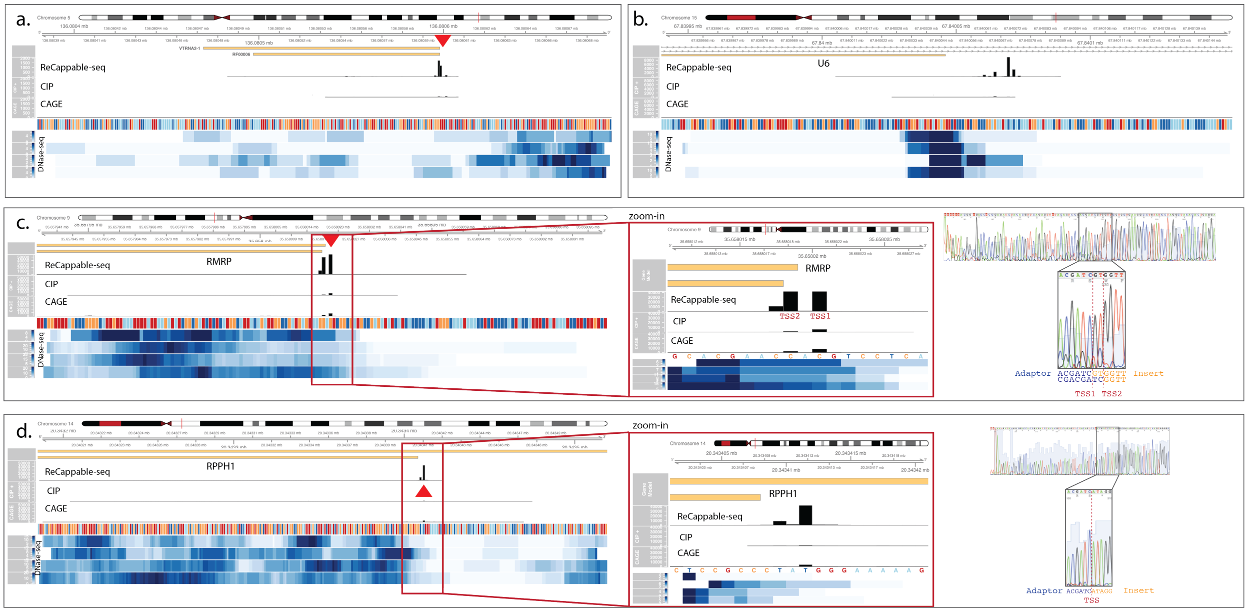
Supplementary Figure 4

Supplementary Figure 4 : Epigenetic marks in a 4 kb window centred at the Pol-II (green) and non-Pol-II (purple) TSS. Epigenetic marks were derived from publically available ChIP-seq or DNase-seq ENCODE data performed on human A549 cells.



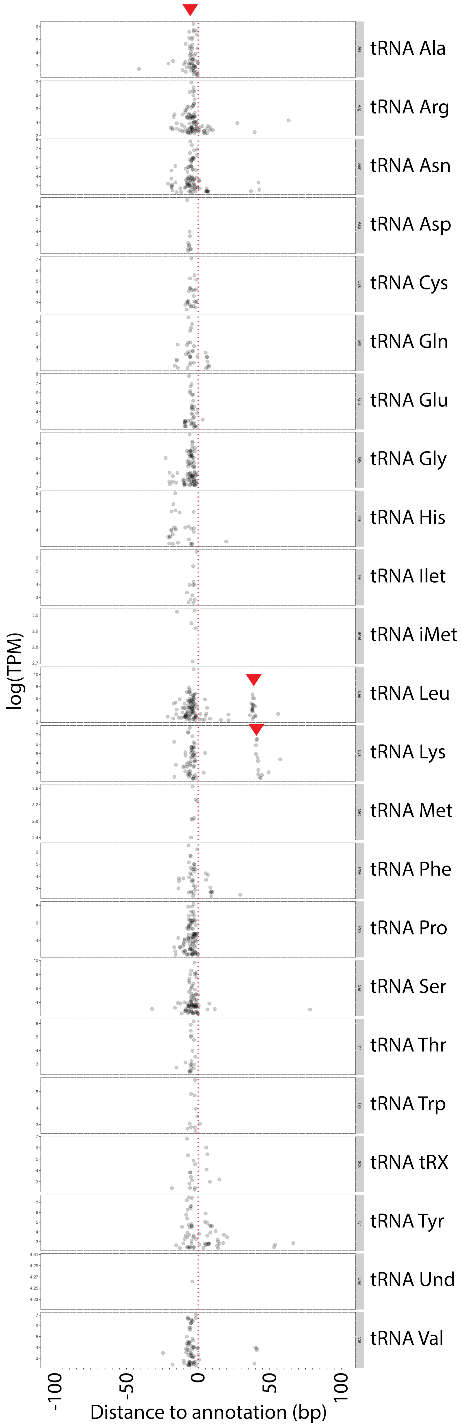
Supplementary Figure 5

Supplementary Figure 5 : a. Position of the Pol-II (left panels), non-Pol-II (middle panels) consistent TSS and CAGE TSS (right panel) relative to repeat start sites for 5S, 7SK, 7SL, HY, MIR, tRNA and U repeats. b. Remaining acid insoluble radioactivity (cpm, y-axis) after TCA precipitation of a 300 nucleotide RNA made by in vitro transcription which had been labeled with a tritiated methyl group on its 5’ gamma phosphate and incubated with either CIP, VCE or RNASE I (Supplementary Text1).



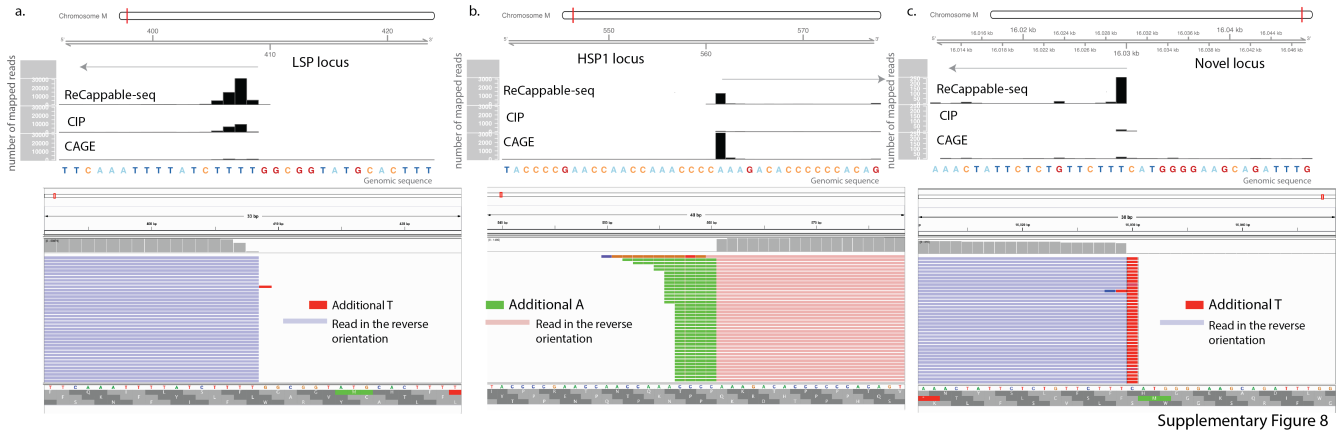
Supplementary Figure 6

Supplementary Figure 6 : a. Genomic profile of the Vault 2.1 gene. b. Genomic profile of one of the U6 gene. c. Genomic profile of the RMRP gene. d. Genomic profile of the RPPH1 gene. Red arrows denote the positions with the highest density of reads starting at these locations. RACE experiments were performed on both the RMRP and the RPPH1 transcripts revealing one transcript start for RPPH1 and two transcript starts for RMRP.

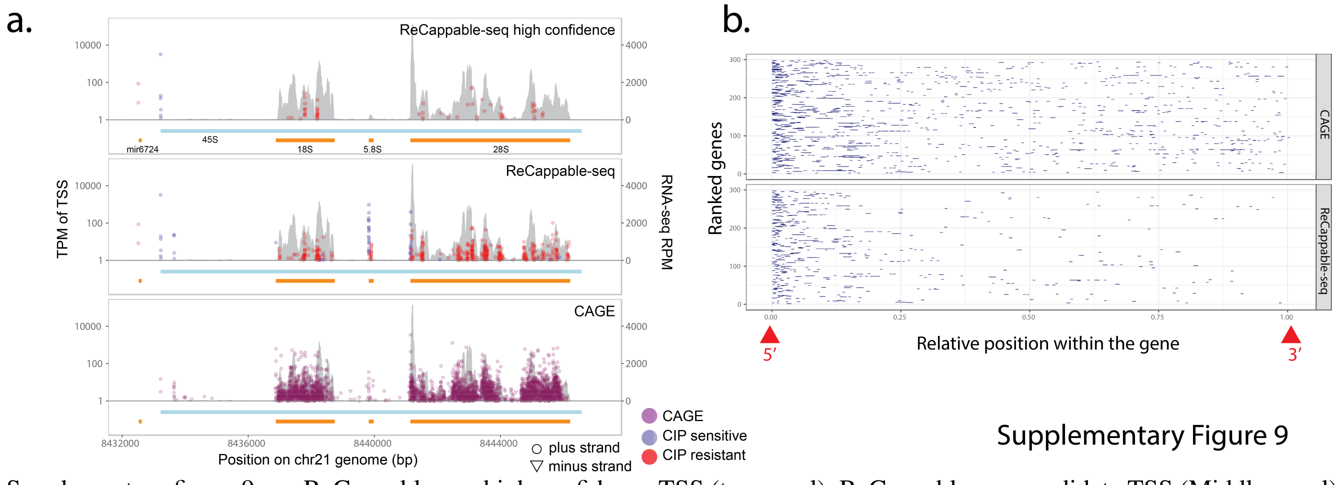


Supplementary Figure 7

Figure 6. Supplementary Figure 7 : Non-Pol-II TSS distribution around tRNA classified according to the type of tRNA. red arrows represent notable TSS clusters.



Supplementary figure 8 : Identification of mitochondrial TSS. a. LSP locus, b. HSP1 locus, c. Locus containing a putative novel light strand TSS. Upper panel describes the density of reads starting at various genomic locations in the mitochondrial genome for ReCappable-seq (top), CIP (middle) and CAGE (bottom). The lower panel displays an IGV screenshot of a subset of the mapped reads in the forward (pink) and reverse (blue) orientation. Soft-clipped sections of the reads are colored according to the nucleotide type : A (green), T (red), G (orange) and C (blue).



Supplementary figure 9 : a. ReCappable-seq high confidence TSS (top panel), ReCappable-seq candidate TSS (Middle panel) and CAGE TSS (bottom panel) located on the 45S locus on chr21. b. CAGE TSS (top panel) and ReCappable-seq candidate TSS (bottom panel) coverage in the gene body of the 300 most highly expressed protein coding genes