Cell Reports, Volume *8*

Supplemental Information

Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites

Schraga Schwartz, Maxwell R. Mumbach, Marko Jovanovic, Tim Wang, Karolina Maciag, G. Guy Bushkin, Philipp Mertins, Dmitry Ter-Ovanesyan, Naomi Habib, Davide Cacchiarelli, Neville E. Sanjana, Elizaveta Freinkman, Michael E. Pacold, Rahul Satija, Tarjei S. Mikkelsen, Nir Hacohen, Feng Zhang, Steven A. Carr, Eric S. Lander, and Aviv Regev

Supplemental Discussion

RT-dropoff one nucleotide prior to the TSS

To investigate the extent to which transcription start sites, in general, and methylated TSS, in particular, could be precisely identified based on our data, we first examined the distribution of >30,000 sites in mouse with >10 reads beginning at a specific site (across all conditions). We noted that this catalogue is enriched towards the first (annotated) position, as expected, but to an equal extent also at the second position, which was unexpected **(Fig. S4A)**. To further address this, we analyzed the proportion of sites harboring an adenosine in the first or the second position, as a function of fold change over input. At the first position, we found a strong bias towards adenosine, as anticipated (**Fig. S4B**). However, at the second position, no such bias was noted. Instead, even when reads stacked at Position 2, the bias towards adenine was still at Position 1, and the magnitude of this bias was comparable to the one obtained when the stacks were at Position 1. Thus, the stacks at Positions 1 and 2 both appear to reflect methylations occurring at Position 1. The stacks at positions 2 likely reflect drop-off of the RT polymerase one nucleotide prior to reaching the 5' terminus, possibly due to steric obstruction of the cap structure. Virtually identical results were obtained when performing the same analysis in human (data not shown).

As an independent source of validation that the sites preceding the stacks are the true TSSs, in cases in which the former lack an adenosine, note also the analysis presented in **Fig. 5C-D**, assessing correspondence with a CAGE dataset.

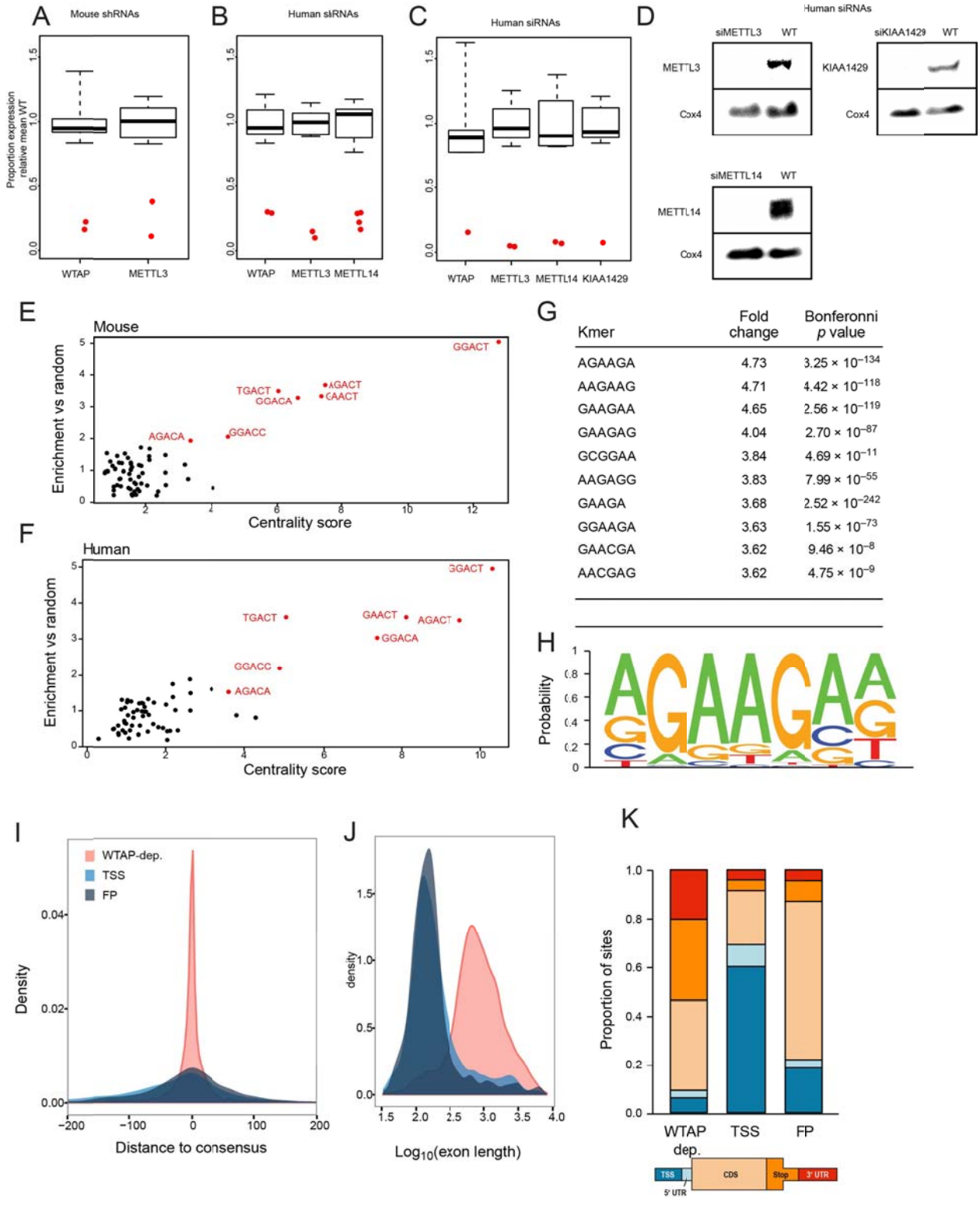
Correlation of mTSS state with transcript structure, stability and translation rates

To gain further insight into the role of mTSSs, we examined the extent to which mTSS presence in a gene was associated with gene structure, RNA stability and translational efficiency. As our ability to detect mTSSs is biased towards more highly expressed genes, we used a randomly sampled control dataset matching the expression patterns of the mTSS harboring genes (but lacking mTSSs). Specifically, we compared sets of 2,891 human genes and 3,459 mouse genes containing an mTSS in four or more of the samples, to a roughly equal number of expression-matched controls.

Across the vast majority of datasets, mTSS presence was most strongly associated with 5' UTR length, with genes harboring mTSSs having significantly shorter 5' UTRs compared to their non-methylated expression matched counterparts (**Figs. S5A-C**). This was observed using the full human and mouse datasets defined above (**Fig. S5A-C**) and when analyzing mTSSs identified in a specific condition (data not shown). Furthermore, a negative association with 5' UTR length was observed when binning genes into 5 equally sized bins based on number of samples in which an mTSS was observed (a variable capturing both our confidence in the sites, and potentially the stoichiometry of modification at site) (**Fig. S5D-E**).

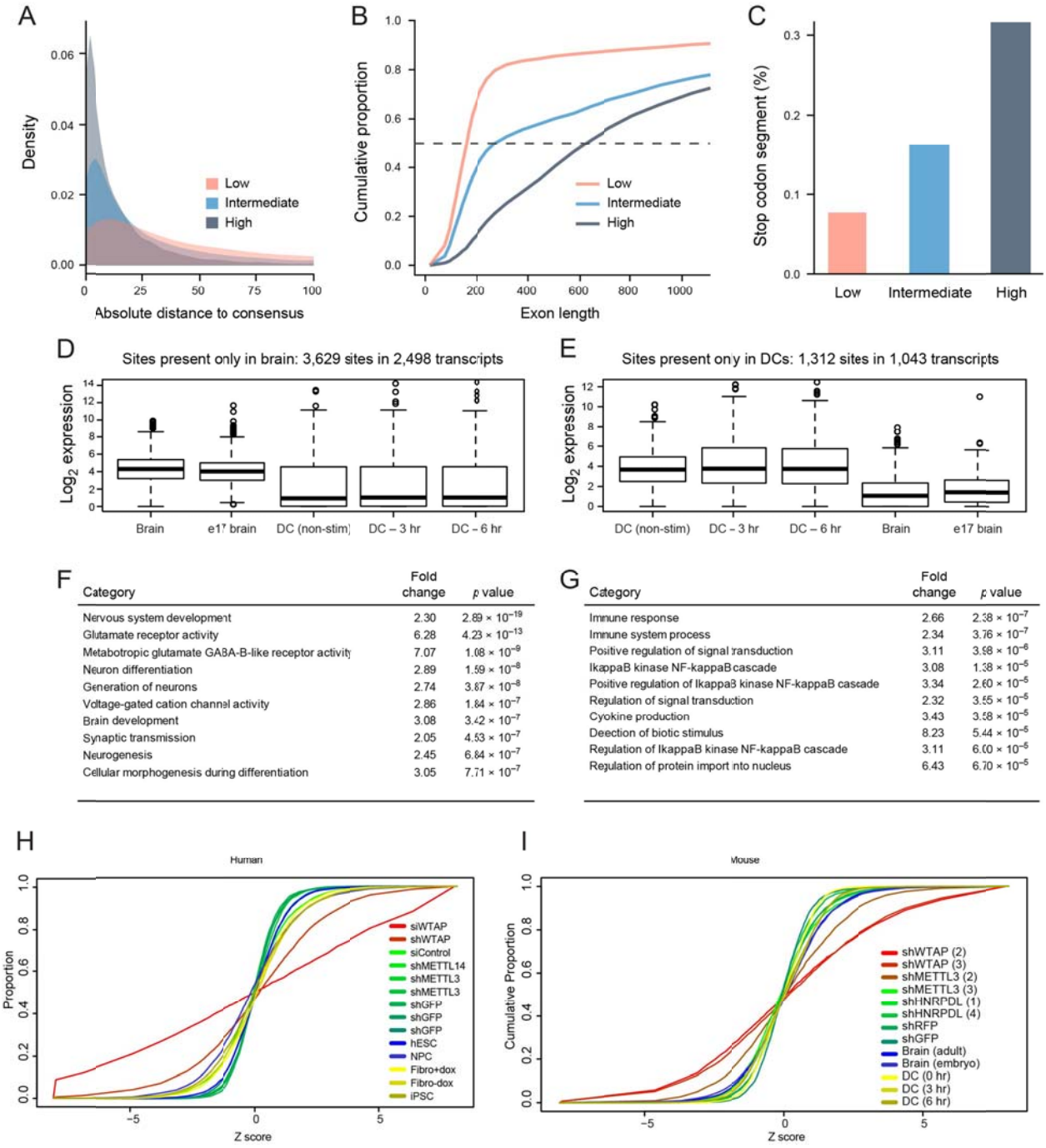
We also found a strong positive association between the presence of mTSS and translational efficiency (**Fig. S5A-B**), with genes harboring mTSSs exhibiting increased translational efficiencies compared to the expression matched controls (**Fig. S5F-G**). These effects were particularly strong when comparing our mouse dataset with translational efficiencies measured in mESC (Ingolia et al., 2011), but also observed in the human dataset when comparing with efficiencies measured in HeLa and HEK293 cells (Subtelny et al., 2014). We also observed a quantitative relationship between the number of conditions with mTSSs and translational efficiencies when using the binning approach described above (**Fig. S5H** and data not shown). As expression level, number of conditions in which an mTSS was observed, and 5' UTR length are all correlated with each other, we used a linear model to predict translational efficiencies based on these three variables. In mouse, all variables significantly contributed to the model in this joint context ($P<10^{-4}$ in all cases), whereas in human, the number of conditions in which mTSSs were observed ceased to remain significant. Finally, we note that in these analyses we noted only a negligible impact on RNA stability (**Fig. S5A-B**).

Together, our data strongly supports an inverse association between 5’ UTR length and presence of mTSS, and moreover provides indications that mTSSs in genes with shorter 5’ UTR are associated with increased translation efficiencies.



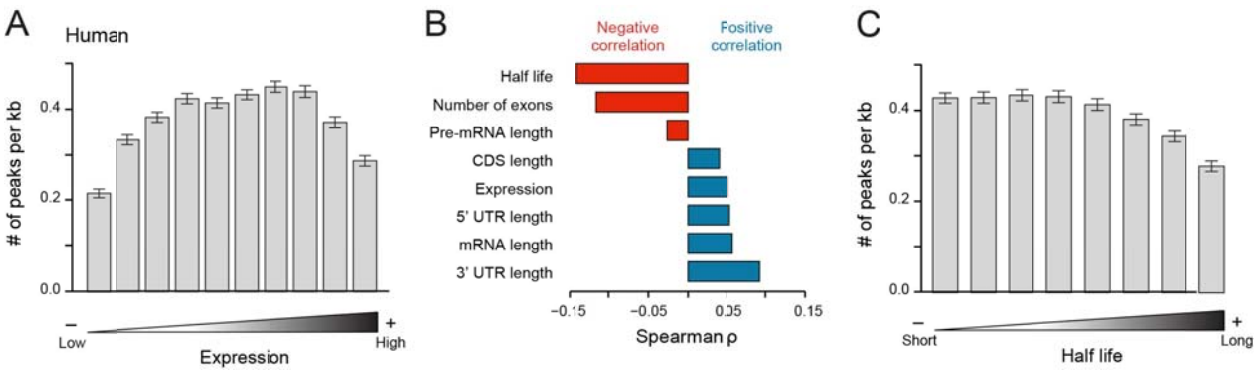
Kmer	Fold change	Bonferonni *p* value
AGAAGA	4.73	3.25×10^{-134}
AAGAAG	4.71	4.42×10^{-118}
GAAGAA	4.65	2.56×10^{-119}
GAAGAG	4.04	2.70×10^{-87}
GCGGAA	3.84	4.69×10^{-11}
AAGAGG	3.83	7.99×10^{-55}
GAAGA	3.68	2.52×10^{-242}
GGAAGA	3.63	1.55×10^{-73}
GAACGA	3.62	9.46×10^{-8}
AACGAG	3.62	4.75×10^{-9}

Supplemental Figure S1 related to Figure 2. (A-D) Knockdown efficiency. (A-C) Shown are sample transcript knockdown levels (Y axis, expressed as proportion of average gene expression among non-perturbed samples), (upon perturbation of (**A**) WTAP or METTL3 (X axis) using shRNAs in mouse embryonic fibroblasts; (**B**) WTAP, METTL3, METTL14 using shRNAs in A549 cells, or (**C**) of WTAP, METTL3, METTL14 and KIAA1429 using siRNAs in human A549 cells. Transcript levels were determined based on TMM-normalized values from the (input) RNA-seq experiments. Boxplots present the distribution of the mean-normalized expression levels across all non-perturbed samples, with whiskers extending until the minimum and maximum of the distribution. Red dots indicate levels in the perturbed samples. Some red dots are additionally labelled with information on the experiment or hairpin used **(D)** Protein knockdown levels (by Western blots) following siRNA mediated depletion of METTL3, METTL14 and KIAA1429 from A549 cells. Cox4 was used as a loading control. **(E,F)** Identification of optimal set of consensus sites. For each of 64 pentamers harboring an AC at position 3-4 (NNACA), presented is a fold change quantifying enrichment in the set of 10,000 top motifs in human **(E)** and mouse **(F)** over an equally sized shuffled dataset (Y-axis), along with a centrality score (X-axis) denoting the extent to which a motif is biased towards appearing at the very center of the detected peak (**Methods**). The motifs in red, scoring highest in both metrics and identical between human and mouse, were selected as the set of consensus motifs used throughout the manuscript. **(G,H)** Sequence features associated with false positive methylation sites. (**G**) Shown are the top scoring pentamers (first column) enriched in sites from the 'false-positive' cluster, along with their fold enrichment (second column) and a Bonferroni-corrected P-value for a chi-squared test assessing their enrichment when comparing them to a set of WTAP-dependent sequences. (**H**) Sequence logo depicting a motif obtained from clustering all enriched pentamers with fold change >2 and corrected P value < 0.05. **(I-K)** Hallmarks of methylation in human A549 cells. Distributions of distances to consensus **(I)**, lengths of internal exons harboring methylated sites **(J)**, and sites across defined gene segments **(K)**, for the sites in the three clusters (WTAP-dependent true positives: TP; False positives: FP; and TSS related: TSS) in Fig. 2F**.** Distributions and gene segments defined as in **Fig. 2B-D**.

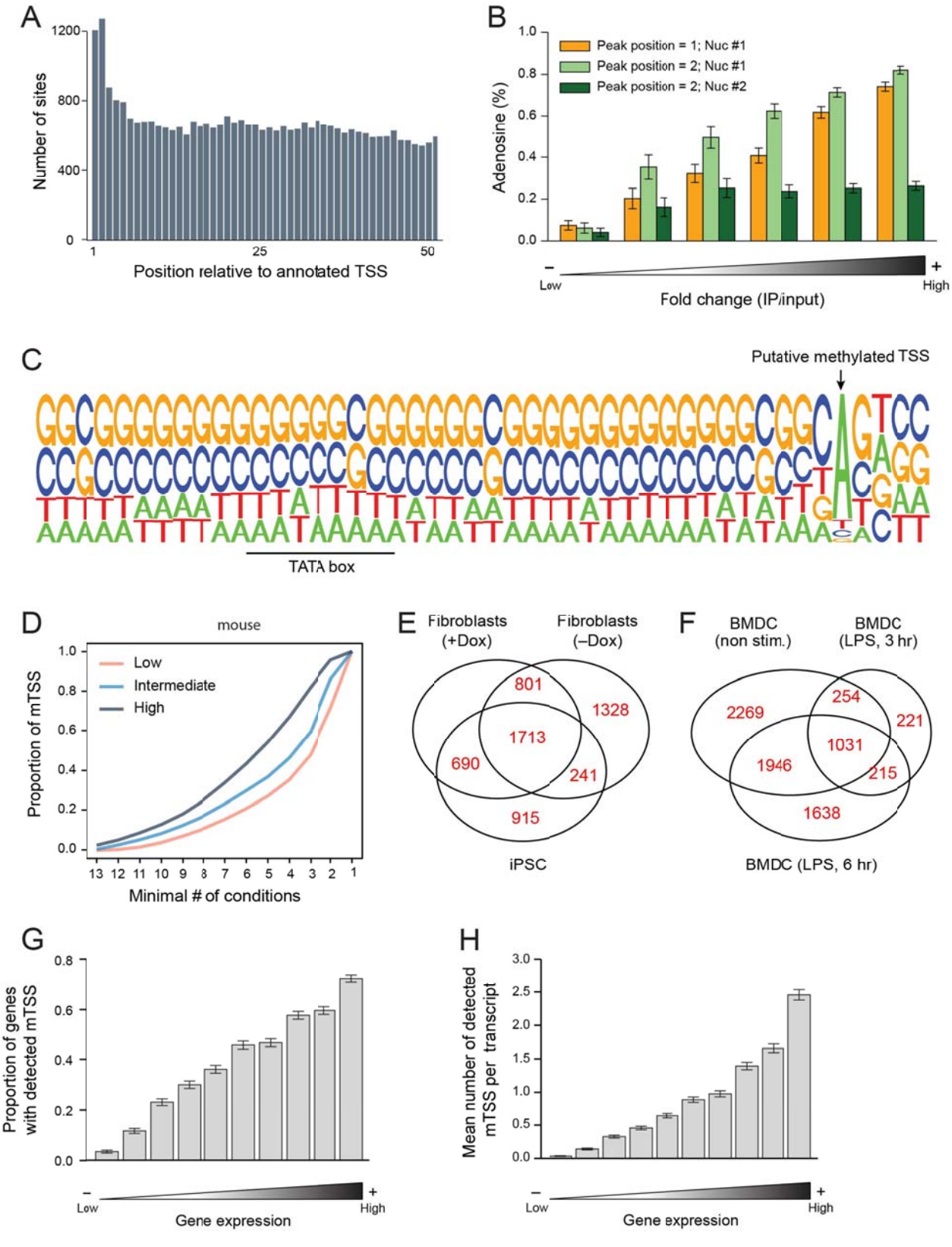


Supplemental Figure S2 related to Figure 3. Characterization of methylation sites. (A) Absolute distance from consensus for sets of sites with 'high', 'intermediate' and 'low' confidence, in human datasets, defined as described in the text. **(B)** Cumulative distributions of exon lengths in the 'high', 'intermediate' and 'low' confidence sets. The X-axis is truncated at 1000 nt. The grey dashed line indicates the median exon length across the three sets. **(C)** Proportion of sites comprised in the stop codon segments for each of the three sets. **(D,E)** Distribution of expression levels in the two brain samples and three dendritic cell samples for

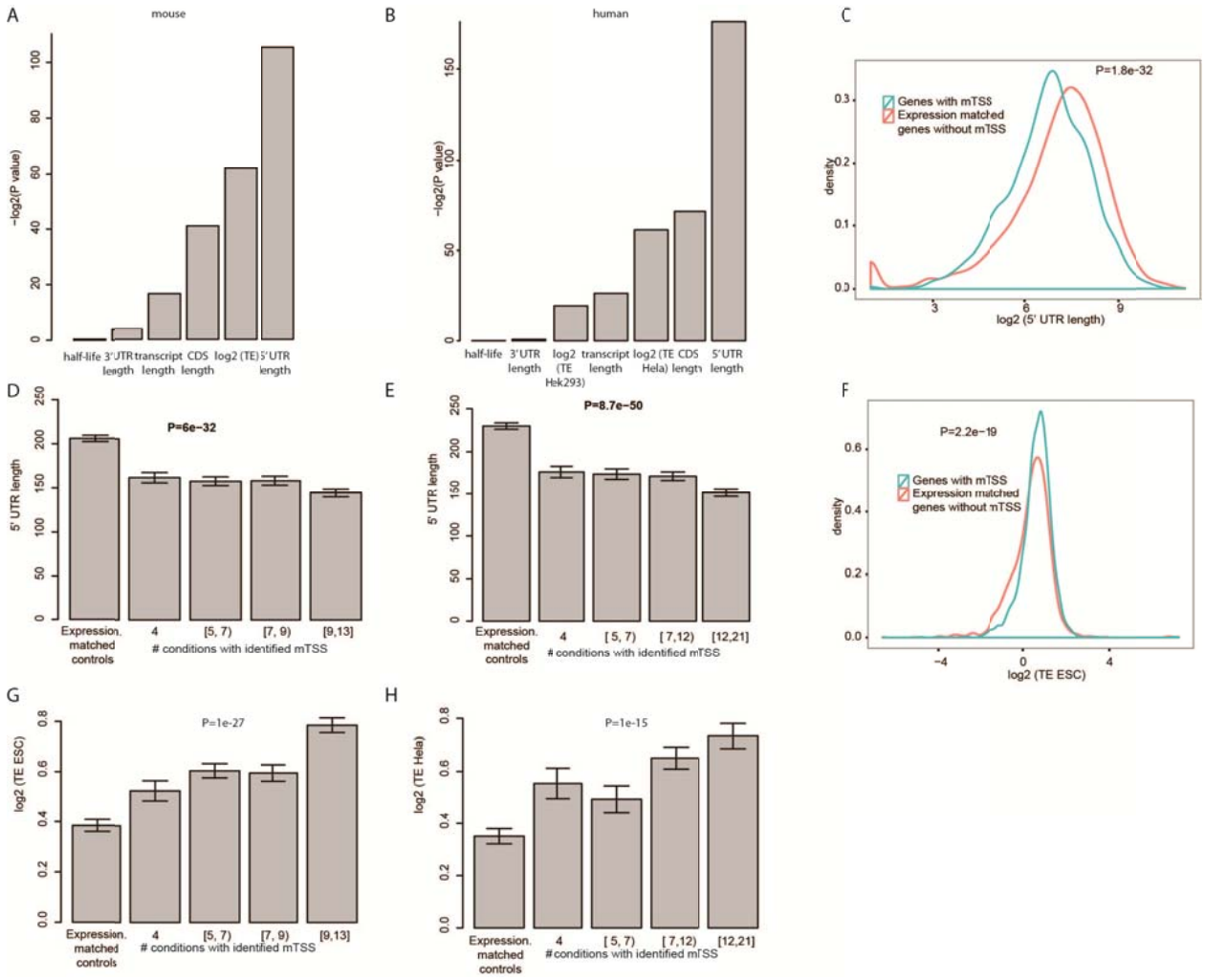
genes harboring sites present only in brain (D) or only in DCs (E). **(F,G)** GO terms enriched in genes from (D) and (E), respectively. **(H)** Variability of POI scores across the two dynamic datasets in human (reprogramming into iPSCs and differentiation into NPCs) compared to variability of these scores across different experiments performed in A549 cells. For this analysis, we designated a control cohort of perturbation experiments in A549 cells in which we failed to observe a substantial impact on methylation (replicates of the shMETTL3, shMETTL14, and shGFP replicates). For each site in each experiment in this cohort, a distribution of Z scores was calculated based on all remaining experiments in the cohort. Similarly, distributions of Z scores with respect to the control cohort were calculated for all other conditions, including the dynamic experiments and WTAP perturbation. Cumulative distributions of Z scores for a subset of these conditions are displayed. These analyses indicated the perturbations of WTAP as clear outliers, but the distributions of the remaining **(I)** Analysis as in (H) but for mouse. As cohort controls in this experiment, we made use of shMETTL(3), shHNRPDL, shRFP, shGFP, as well as the dendritic cells and the brain samples.



Supplemental Figure S3 related to Figure 4. Features associated with methylation site density in human genes. (A) Association of gene expression levels (X axis) with methylation peak density (Y axis). **(B)** Spearman's ρ (X axis) between methylation densities and different gene level features (Y axis), performed as in **Fig. 4C**. **(C)** Association of mRNA half life with methylation density. Analyses in this figure were done as presented in Figure 4, but for the human data.


A
Number of sites
1200
800
400
0
1
25
50
Position relative to annotated TSS
B
Peak position = 1; Nuc #1
Peak position = 2; Nuc #1
Peak position = 2; Nuc #2
Adenosine (%)
1.0
0.8
0.6
0.4
0.2
0.0
–
+
Low
High
Fold change (IP/input)
C
Putative methylated TSS
TATA box
D
mouse
Low
Intermediate
High
Proportion of mTSS
1.0
0.8
0.6
0.4
0.2
0.0
13 12 11 10 9 8 7 6 5 4 3 2 1
Minimal # of conditions
E
Fibroblasts (+Dox)
Fibroblasts (–Dox)
801
1328
1713
690
241
915
iPSC
F
BMDC (non stim.)
BMDC (LPS, 3 hr)
2269
254
221
1031
1946
215
1638
BMDC (LPS, 6 hr)
G
Proportion of genes with detected mTSS
0.8
0.6
0.4
0.2
0.0
–
+
Low
High
Gene expression
H
Mean number of detected mTSS per transcript
3.0
2.5
2.0
1.5
1.0
0.5
0.0
–
+
Low
High
Gene expression

Supplemental Figure 4 related to Figure 5. Anal ysis of methylated TSSs in human and mouse. (A) Distribution of relative positions of all mouse methylated sites harboring >10 reads (across all IP samples) with respect to annotated transcription start sites. **(B)** Proportion of positions containing an adenosine (Y axis), as a function of fold change between number of reads in the IP sample compared to the input sample (X axis), for 1,247 sites occurring at the very first annotated position (orange bars), and for 1,292 sites occurring at the second annotated position, showing separately for the latter the first position (light green bars) and the second (dark green bars). Error bars indicate the standard error of the proportion. **(C)** Sequence logo surrounding the human internal mTSSs (position >4) derived from 3,552 sites with log2 fold changes >4.5, as shown for mouse in **Fig. 5b. (D)** Proportion of methylated TSS sites (Y axis) in each of three equally sized fold-change bins (low, intermediate and high) detected in at least a given number of experiments among 13 experiments in mouse (X axis), including both the perturbation experiments (displayed in **Fig. 2**) and the experiments examining the dynamics of the methylations (displayed in **Fig. 3**). **(E,F)** Venn-diagrams showing the extent of overlap in detection of mTSSs across pluripotency reprogramming (human, **E**), and dendritic cell stimulation (mouse, **F**) and. Only sites with fold- changes >4 were considered as present in this analysis. Gene expression levels were not controlled for in this analysis. **(G)** Evaluation of saturation in mTSS detection in MEFs. Genes expressed above the 60^{th} percentile were binned into 10 equally sized bins based on their expression (X axis). The proportion of genes for which an mTSS was identified is plotted for each bin (Y axis). Error bars – standard error of proportion. **(H)** As in (G), but quantifying the average number of unique mTSSs identified per transcript, as a function of expression. Error bars – SEM.



Supplemental Figure 5 related to Figure 5. Features correlating with mTSS presence. **(A-B)** Log-transformed P values (Y axis), obtained for a non-parametric Mann-Whitney test comparing distributions of each indicated variable (X axis) in genes harboring an mTSSs and in expression matched controls in mouse (A) and human (B) datasets. **(C)** Distributions of 5' UTR lengths within genes harboring an mTSS (blue) and matched controls (red) in mouse. **(D-E)** Quantitative relationship between binned number of samples in which an mTSS was identified (X axis) and mean 5' UTR length (Y axis) in mouse (D) and human (E). Lengths are also shown for expression matched controls (left most bar). Error bars – standard error of the mean (SEM). P values derived using the Kruskal-Wallis test are indicated; Post-hoc paired wilcox-test were significant for all pairwise comparisons involving the two extreme bins. (F) Distribution of translational efficiencies within genes harboring an mTSS (blue) and matched controls (red), in mouse, measured in mES cells. **(G-H)** Quantitative relationship between binned number of samples in which an mTSS was identified (X axis) and mean translational efficiencies (Y axis) in mouse (G) and human (H). Efficiencies are also shown for expression matched controls (leftmost bar). Error bars – standard error of the mean (SEM). P values derived using the Kruskal-Wallis test are indicated; In human, post-hoc paired Mann-Whitney tests were significant across all

pairwise comparison in human with the exception of the second and third bins, whereas in mouse only the pairwise interactions involving either of the extreme bins were significant.