Sequencing of First-strand cDNA Library Reveals Full-length Transcriptomes

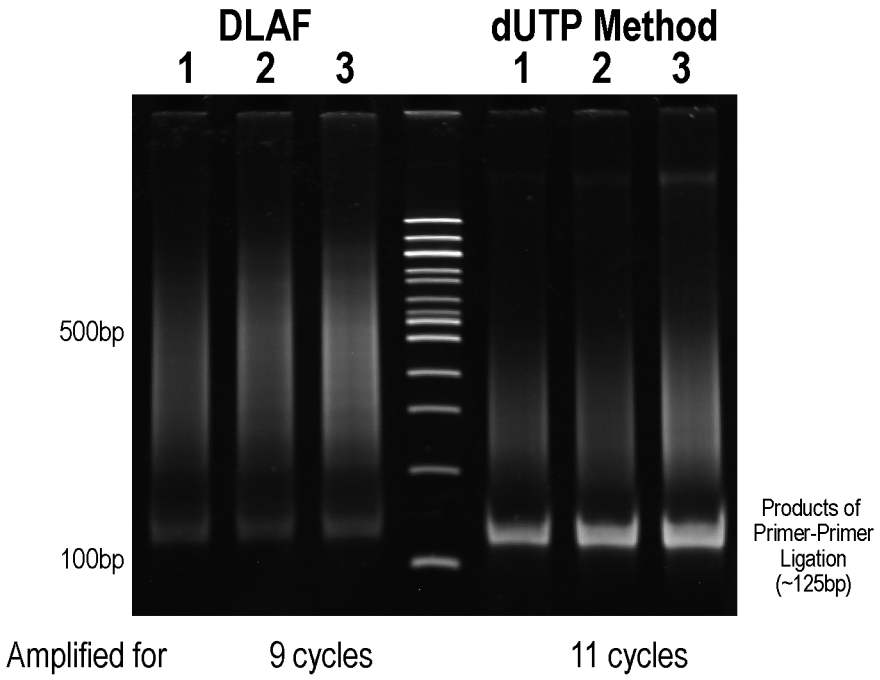
Saurabh Agarwal, Todd S. Macfarlan , Maureen A. Sartor & Shigeki Iwase

Supplementary Figure 1

a

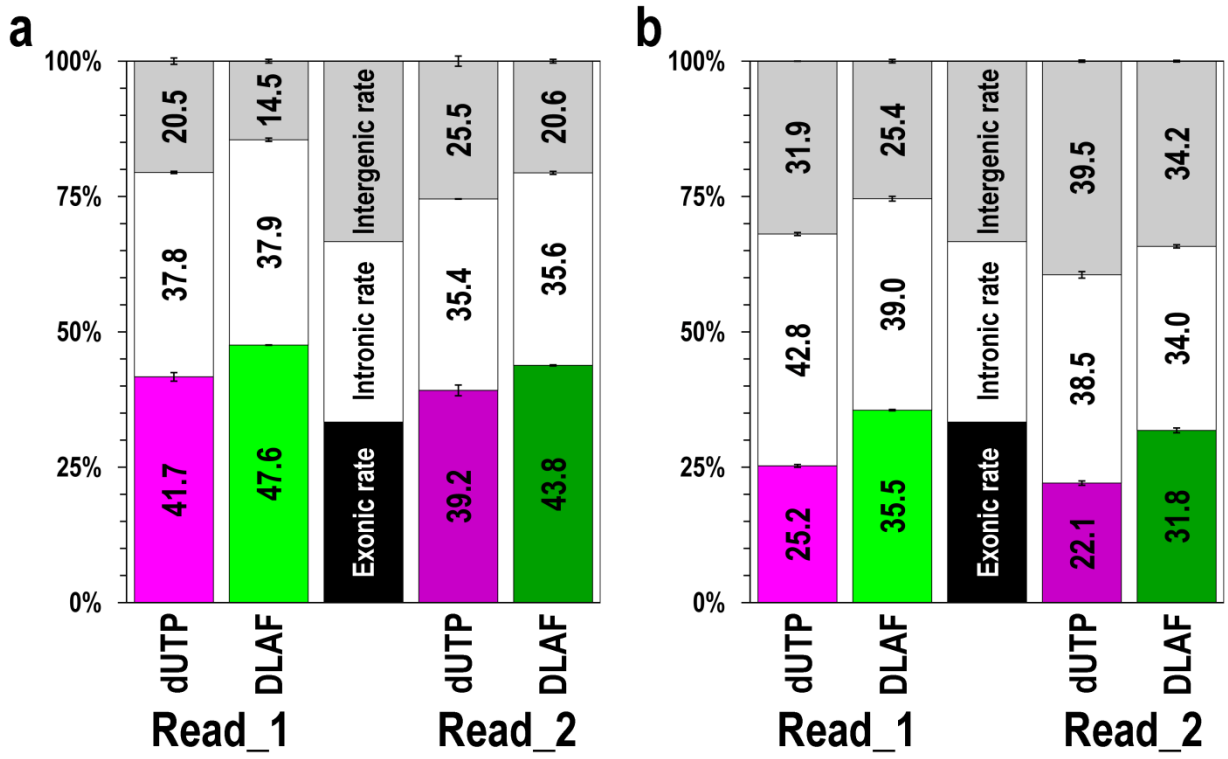
RNA Sample	qPCR threshold cycle (C_t)			
	DLAF (*n*=3)		dUTP (*n*=3)	
	Mean	Stdev	Mean	Stdev
#1	8.591	0.0876	12.065	0.1757
#2	8.624	0.0563	10.909	0.0702
#3	8.390	0.0414	10.218	0.0167
Mean	8.535 [A]	0.1268	11.064 [B]	0.9333
Mean ΔC_t $^{(B-A)}$ = 2.529 (*p*-value = 0.00961)				

b



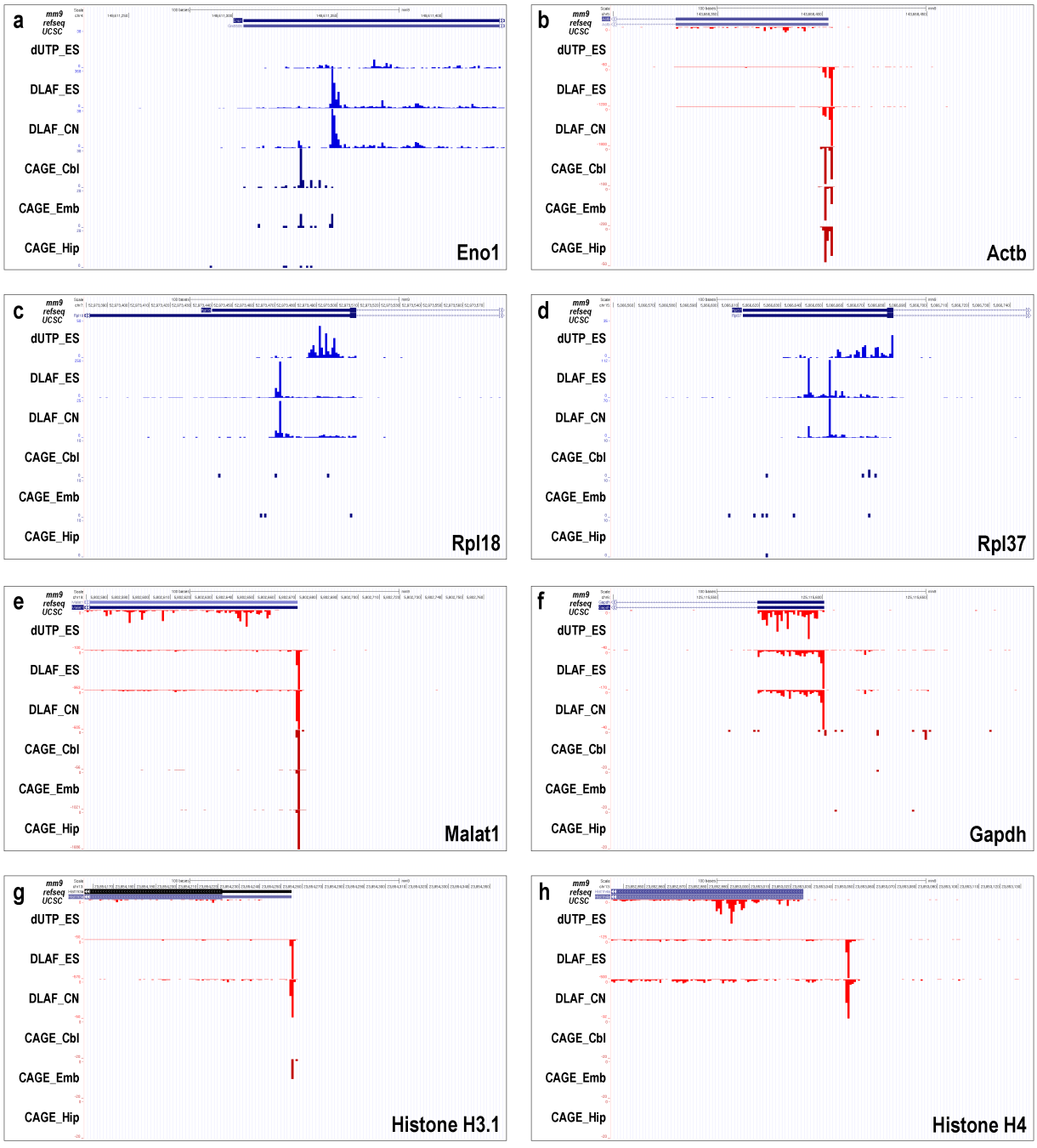
Supplementary Figure 1. Comparison of cDNA yield between the dUTP and DALF methods. (a) Quantitative PCR to compare the yields of the two methods. 2 percent of the libraries (before the final PCR step) were amplified in triplicates using 24–25 bp oligonucleotides matching the Illumina sequencing primers. DLAF libraries gave significantly lower Ct values compare to dUTP libraries (*P* = 0.00961, two-sided, unpaired-samples Student's *t*-test). **(b)** Final PCR products for the libraries. Due to lower yields, the dUTP method libraries were amplified for 2 additional cycles to give approximately similar yields to the DLAF libraries.

Supplementary Figure 2



Supplementary Figure 2. Relative coverage of intragenic (exonic and intronic) and intergenic regions. Comparison between dUTP and DLAF libraries using WT mES cells **(a)** or *Kdm1a* deficient mES cells **(b)** are shown. DLAF libraries show a higher rate of mapping to exonic regions compared to dUTP libraries. Average of two biological replicates is shown and error-bars indicate the range of data.

Supplementary Figure 3



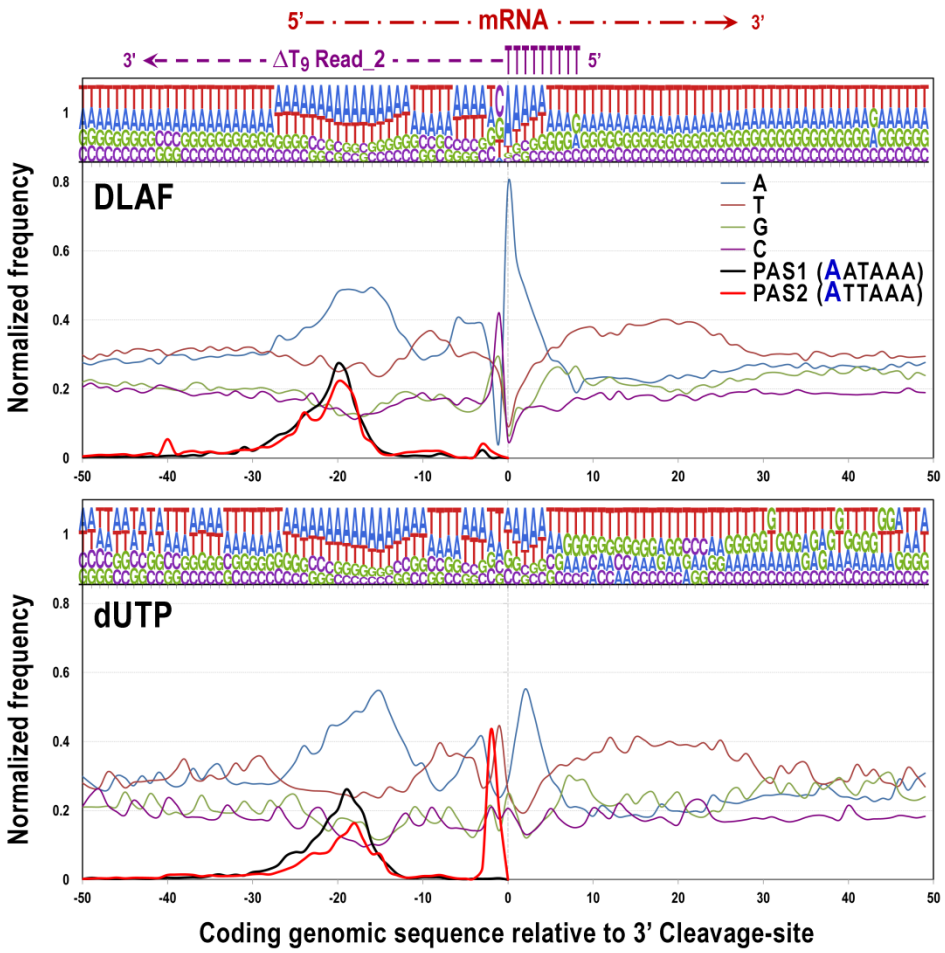
Supplementary Figure 3. Transcription start sites (TSS) detected by Cap Analysis of Gene Expression (CAGE) and DLAF and dUTP libraries. Only the first sequenced nucleotides of read_1 are represented along a 201-bp window around the annotated TSS. CN: mouse cortical neurons. ES: mouse ES cells. Cbl: mouse cerebellum. Hip: mouse hippocampus. For many genes, including *Actb* **(b)** and *Malat1* **(e)**, DLAF read_1 peaks coincide with the CAGE peaks. For some TSSs, CAGE did not give a discrete signal, whereas DLAF could identify the 5' ends **(c, d, and f)**. Non-polyadenylated mRNA, such as histones H3.1 **(g)** and H4 **(h),** may not be detected by CAGE likely due to exclusion by oligo(dT)-mediated enrichment of the mRNA. Blue and red represent transcription on sense and antisense strands, respectively.

Supplementary Figure 4

a

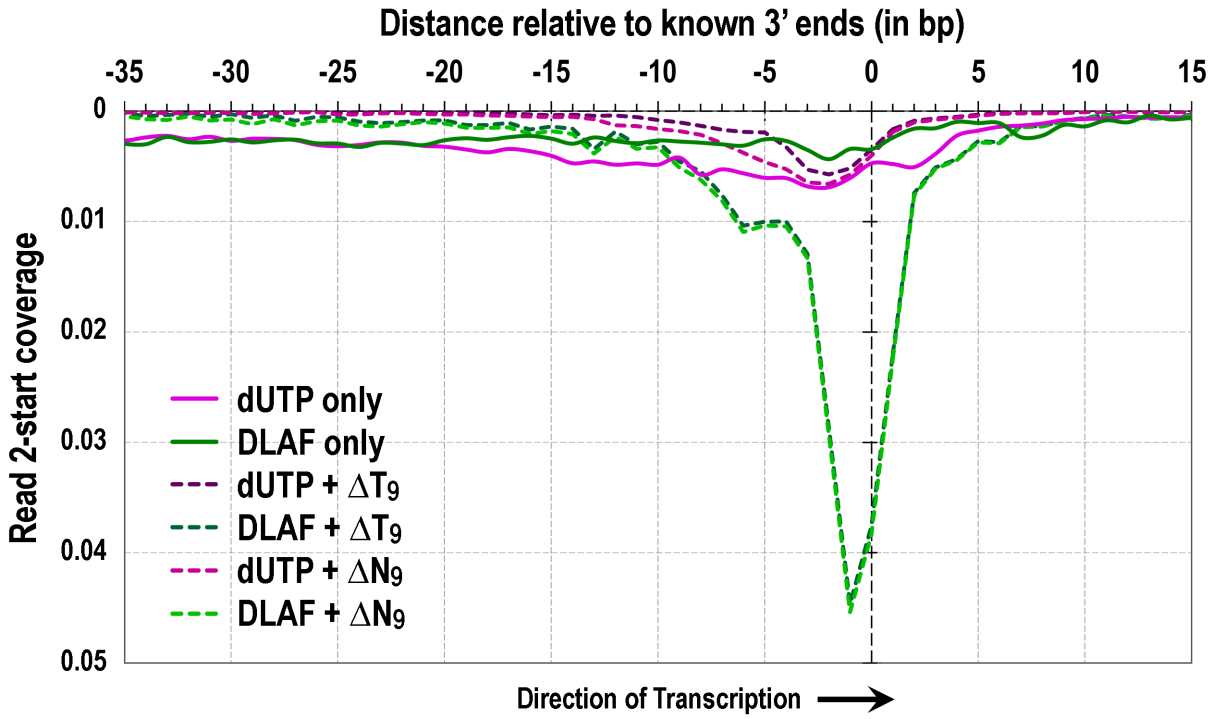
Read_2	Mapped reads ^{M} (in millions)	Mapped ΔT_9 reads ΔT_9 (in millions)	Yield of ΔT_9 reads ΔT_9 / M (as % of M)	Number of ΔT_9 reads containing PAS1 (AATAAA) ^{n1} (as % of ΔT_9)	Number of ΔT_9 reads containing PAS2 (AATAA) ^{n2} (as % of ΔT_9)	Fraction of ΔT_9 reads containing PAS $\frac{n1+n2}{\Delta T_9}$ (as % of ΔT_9)	Yield of PAS containing ΔT_9 reads in each library $\frac{n1+n2}{M}$ (as % M)	Average yield of PAS containing ΔT_9 reads $\text{Average}\ \frac{n1+n2}{M}$ (as % of M) ^{Ave}	Ratio of average yields of PAS containing ΔT_9 Reads $\frac{\text{Ave}^{DLAF}}{\text{Ave}^{dUTP}}$
DLAF_1	17.16	0.379	2.21%	186,899 (**49.31%**)	45,620 (**12.04%**)	**61.35%**	1.355%	**1.329%** ± 0.0368%	**6.74** (*p*=0.0014)
DLAF_2	45.22	1.011	2.24%	473,476 (**46.83%**)	115,750 (**11.45%**)	**58.28%**	1.303%		
dUTP_1	23.49	0.058	0.25%	31,249 (**54.01%**)	7,175 (**12.40%**)	**66.42%**	0.164%	**0.197%** ± 0.0475%	
dUTP_2	23.55	0.081	0.34%	43,858 (**54.01%**)	10,484 (**12.91%**)	**66.93%**	0.231%		

b



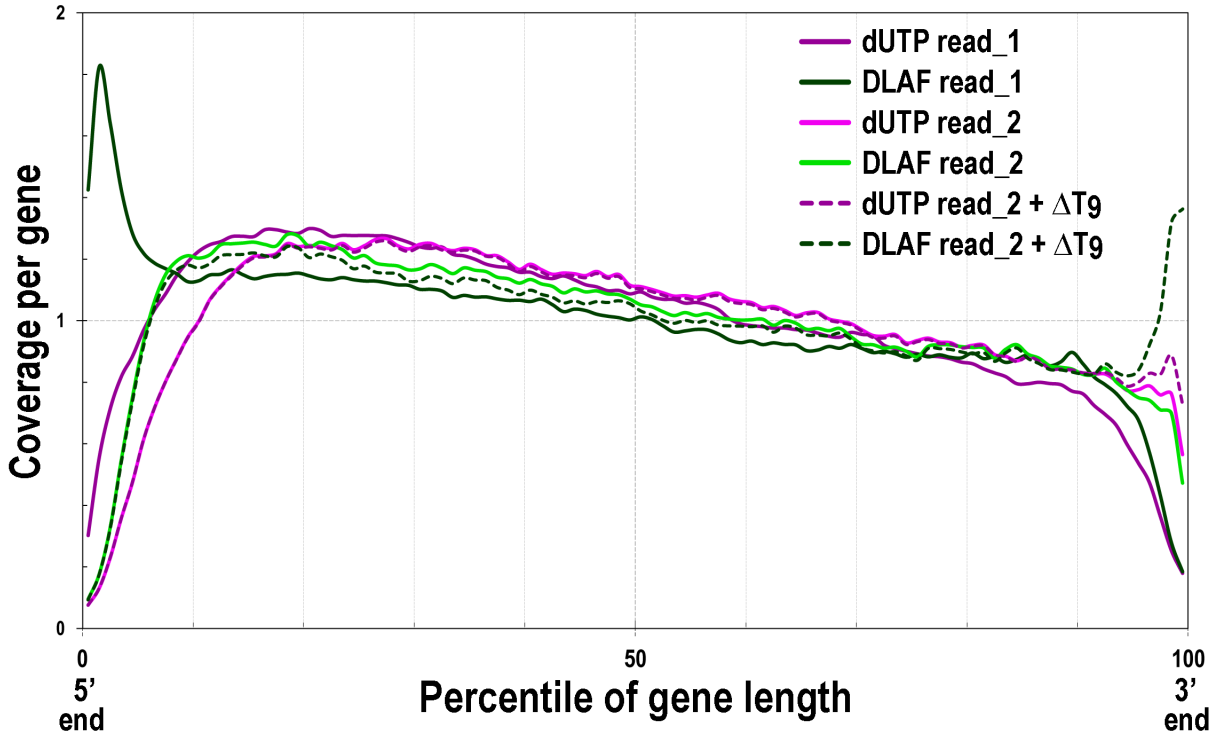
Supplementary Figure 4. Polyadenylation signal (PAS) analysis of ΔT_9 read_2 from DLAF and dUTP libraries. ΔT_9 read_2 from mES cells were analyzed**. (a)** The numbers and frequency of ΔT_9 read_2 sequences that contained either AATAAA(PAS1) or ATTAAA(PAS2). DLAF results in a higher yield of PAS-containing ΔT_9 read_2. *p*-value from two-sided, unpaired-samples Student's *t*-test is shown. **(b)** Base frequency in the coding genomic sequences upstream and downstream of ΔT_9 read_2. Direction of mRNA transcription and position and direction of T_9 and ΔT_9 read_2 are shown at the top. X-axis indicates the genomic position relative to the last base of T_9 (shown as 0). Positve X-axis shows downstream genomic sequence. Y-axis indicates the frequency of a nucleotide at a given position and the frequency distribution of PAS1 or PAS2 respectively. DLAF ΔT_9 read_2 show an enrichment of upstream and downstream U-rich genomic sequences, which are known to be associated with PAS. Both libraries showed peak of PAS frequency at around -20 base positions.

Supplementary Figure 5



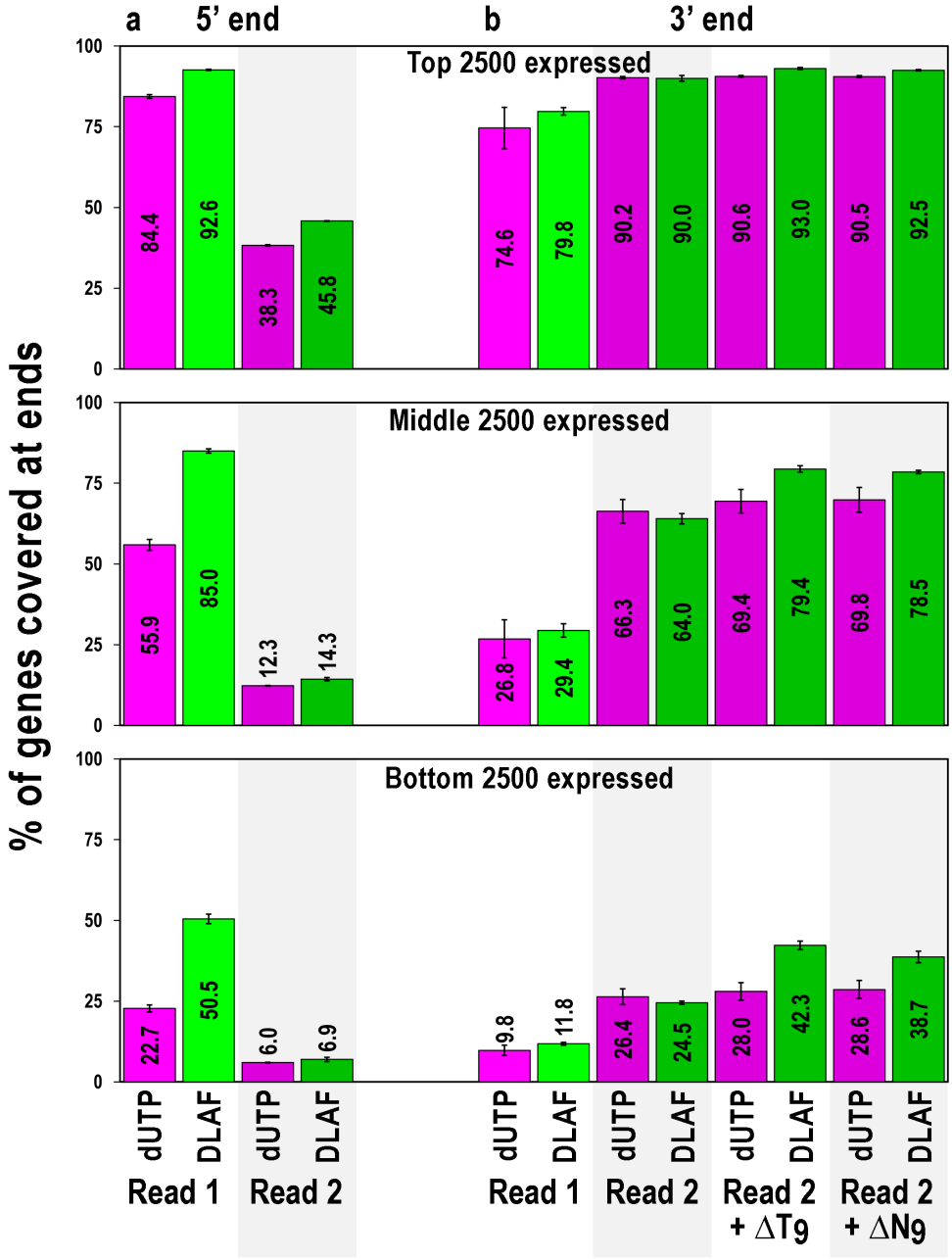
Supplementary Figure 5. Read-start coverage at 3' ends of genes and identification of polyadenylation sites. Read-start coverage represents only the first nucleotide of read_2 near the 3' ends of 5,000 middle-expressed refseq (mm9) genes. Solid lines: coverage in the first alignment. Dashed lines: coverage with combined ΔT_9 or ΔN_9 reads. The data are normalized to million total non-rRNA reads per gene. Average of two biological replicates is shown. The inclusion of ΔT_9 or ΔN_9 to DLAF read_2 leads to a strikingly increased signal, which culminates at the -1 base position relative to the known 3' ends. RNA cleavage sites prior to polyadenylation carry a consensus sequence of 5'-CA-3' [1,2]. The maximum signal at the -1 position is likely because of the exclusion of the last A by the computational trimming of T_9. dUTP read_2 after the incorporation of the ΔT_9 and ΔN_9 reads show some enrichment, but the signal is approximately 8 times lower compared to DLAF (averaged across -5 to +5 bases of 3' ends).

Supplementary Figure 6



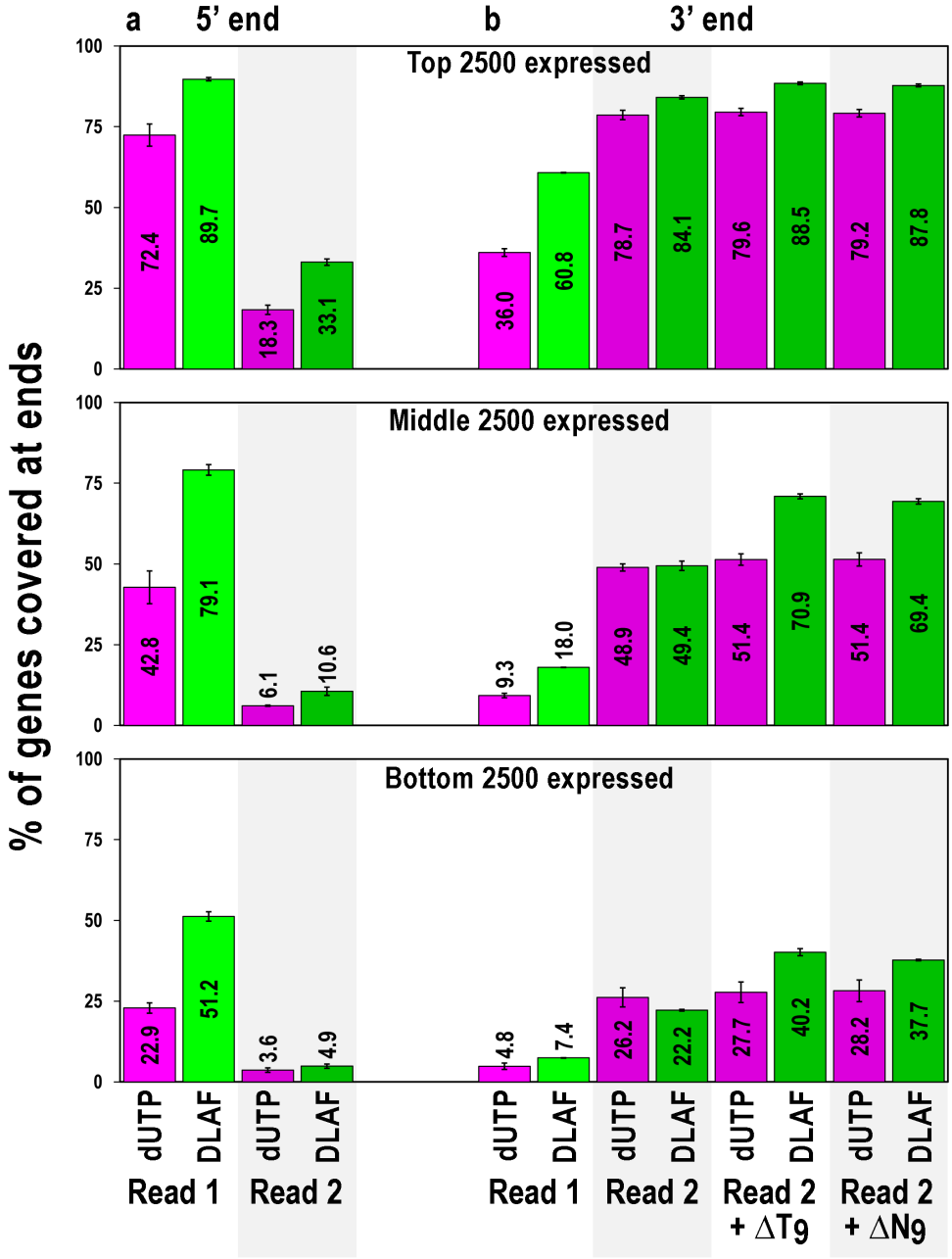
Supplementary Figure 6. End-to-end gene coverage with the DLAF method. Read coverage along gene length is shown for the 5,000 middle-expressed genes using RNA-SeQC. As already shown in Fig. 2, DLAF read_1 shows a distinct enrichment of 5' ends of the genes, whereas dUTP read_1 shows depletion. Dashed lines denote read_2 coverage after inclusion of ΔT_9 reads. When ΔT_9 reads are merged with read_2, DLAF read_2 shows a distinct increase in 3' end coverage. The dUTP method libraries show only a subtle improvement. The signal is normalized to the total number of reads mapping to the 5,000 middle-expressed genes in each library. Average of two biological replicates is shown.

Supplementary Figure 7



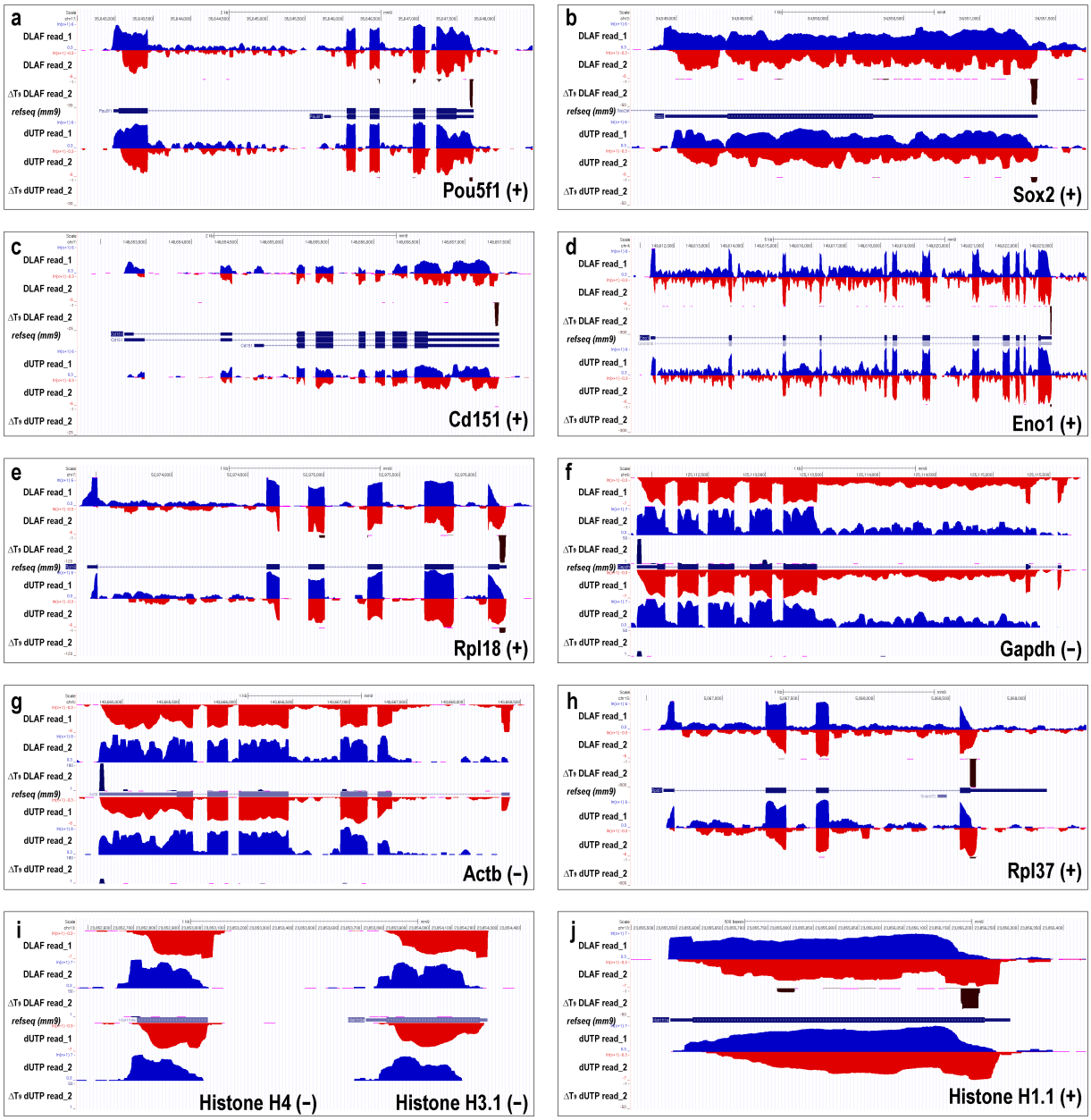
Supplementary Figure 7. Percentage of genes covered at 5' and 3' ends in WT mES cells. A gene end was defined as the terminal 50 bases at either the 5' end or the 3' end of a known transcript. Coverage of an end is counted when it is covered by least 5 reads out of 12.5 million randomly sampled non-rRNA non-mtRNA reads. Genes were categorized into the top, middle, and bottom 2,500 genes based on their expression levels. **(a)** 5' end coverage. DLAF read_1 show a markedly higher number of genes covered at the 5' end than the dUTP method. The differences between the DLAF and dUTP methods are more pronounced for bottom-expressed genes, indicating the high sensitivity of the DLAF method. **(b)** 3' end coverage. Read_1 and read_2 for both methods show comparable coverage of the 3' ends after initial mapping. When remapped ΔT_9 or ΔN_9 reads are included, the DLAF libraries show a higher number of genes that are covered at 3' ends, whereas the improvement is minimal for the dUTP method libraries. Average of two biological replicates is shown and error-bars indicate the range of data.

Supplementary Figure 8



Supplementary Figure 8. Percentage of genes covered at 5' and 3' ends in *Kdm1a* deficient mES cells. (a) 5' end coverage. DLAF read_1 show a markedly higher number of genes covered at the 5' end than the dUTP method. The differences between the DLAF and dUTP methods are more pronounced for bottom-expressed genes, indicating the high sensitivity of the DLAF method. **(b)** 3' end coverage. Read_1 and read_2 for both methods show comparable coverage of the 3' ends after initial mapping. When remapped ΔT_9 or ΔN_9 reads are included, the DLAF libraries show a substantially higher number of genes that are covered at 3' ends. Average of two biological replicates is shown and error-bars indicate the range of data.

Supplementary Figure 9



Supplementary Figure 9. Identification of polyadenylation sites using a novel analysis. UCSC genome browser images showing the read coverage. ES-cell specific genes are *Pou5f1* **(a)**, *Sox2* **(b),** and *Cd151* **(c)**. Ubiquitously expressed house-keeping genes are *Eno1* **(d),** *Rpl18* **(e),** *Gapdh* **(f),** and *Actb* **(g)**. Histone genes are H3.1 and H4 **(i),** and H1.1 **(j)**. Annotated refseq (mm9) genes are shown in the middle of each panel. The gene-encoding strands are indicated in parentheses. Only coding-sense transcripts are shown. Blue and red indicate the strand to which the reads align. The RNA-seq signal is shown as $\log_e(1+x)$, where x represents the number of reads normalized to 10 million non-rRNA , non-mtRNA reads. ΔT_9 reads are shown on a linear scale. Most of the genes, *Pou5f1*, *Sox2*, *Cd151*, *Eno1*, *Rpl18*, *Gapdh*, and *Actb* **(a-g)**, show ΔT_9 peaks near the annotated 3' ends. *Rpl37* **(h)** shows a peak approximately 500 bases upstream of the annotated 3' end, indicating a novel polyadenylation site. Histone genes H3.1 and H4 **(i)** that are known to be non-polyadenylated show an absence of the ΔT_9 signal. It should be noted that some noise can emanate from adenine-rich sequences, which is exemplified by histone H1.1 **(j)**. The DLAF library shows a stronger ΔT_9 signal at polyadenylation sites than the corresponding dUTP library.

Supplementary Figure 10

a

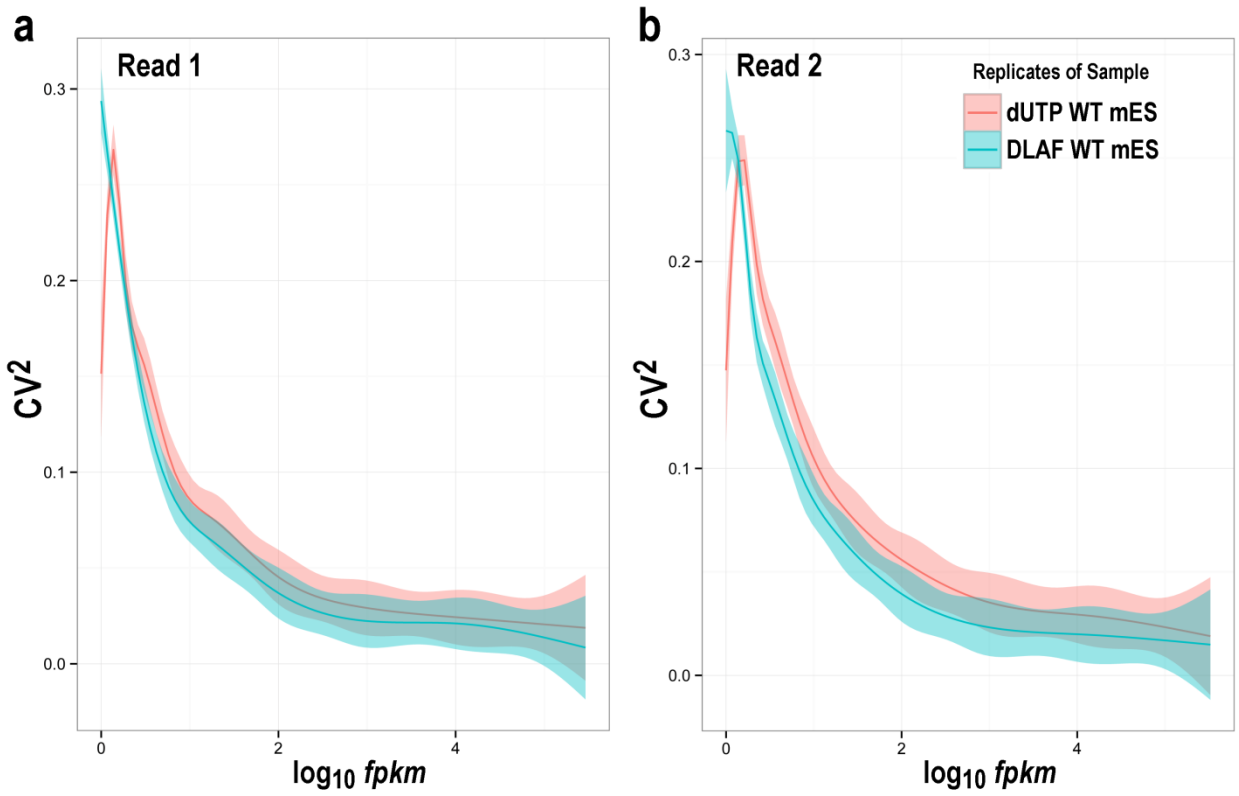
WT mESCs	Pearson Correlation								
Spearman Correlation	Sample	DLAF_1_r1	DLAF_2_r1	dUTP_1_r1	dUTP_2_r1	DLAF_1_r2	DLAF_2_r2	dUTP_1_r2	dUTP_2_r2
	DLAF_1_r1	1	**0.973**	0.964	0.968	0.972	0.971	0.963	0.966
	DLAF_2_r1	**0.98**	1	0.963	0.968	0.969	0.972	0.961	0.965
	dUTP_1_r1	0.97	0.968	1	**0.97**	0.961	0.962	0.968	0.966
	dUTP_2_r1	0.974	0.973	**0.976**	1	0.966	0.966	0.968	0.971
	DLAF_1_r2	0.978	0.976	0.967	0.972	1	**0.974**	0.966	0.968
	DLAF_2_r2	0.977	0.977	0.967	0.971	**0.98**	1	0.964	0.968
	dUTP_1_r2	0.967	0.965	0.975	0.974	0.97	0.968	1	**0.97**
	dUTP_2_r2	0.971	0.97	0.973	0.977	0.973	0.973	**0.976**	1

b

KO mESCs	Pearson Correlation								
Spearman Correlation	Sample	DLAF_1_r1	DLAF_2_r1	dUTP_1_r1	dUTP_2_r1	DLAF_1_r2	DLAF_2_r2	dUTP_1_r2	dUTP_2_r2
	DLAF_1_r1	1	**0.965**	0.949	0.949	0.961	0.959	0.946	0.948
	DLAF_2_r1	**0.975**	1	0.949	0.95	0.96	0.961	0.946	0.948
	dUTP_1_r1	0.956	0.957	1	**0.956**	0.947	0.946	0.955	0.952
	dUTP_2_r1	0.956	0.958	**0.961**	1	0.947	0.948	0.955	0.956
	DLAF_1_r2	0.97	0.969	0.954	0.953	1	**0.965**	0.949	0.949
	DLAF_2_r2	0.968	0.97	0.954	0.955	**0.974**	1	0.95	0.95
	dUTP_1_r2	0.953	0.954	0.959	0.96	0.956	0.958	1	**0.956**
	dUTP_2_r2	0.955	0.955	0.958	0.962	0.955	0.957	**0.961**	1

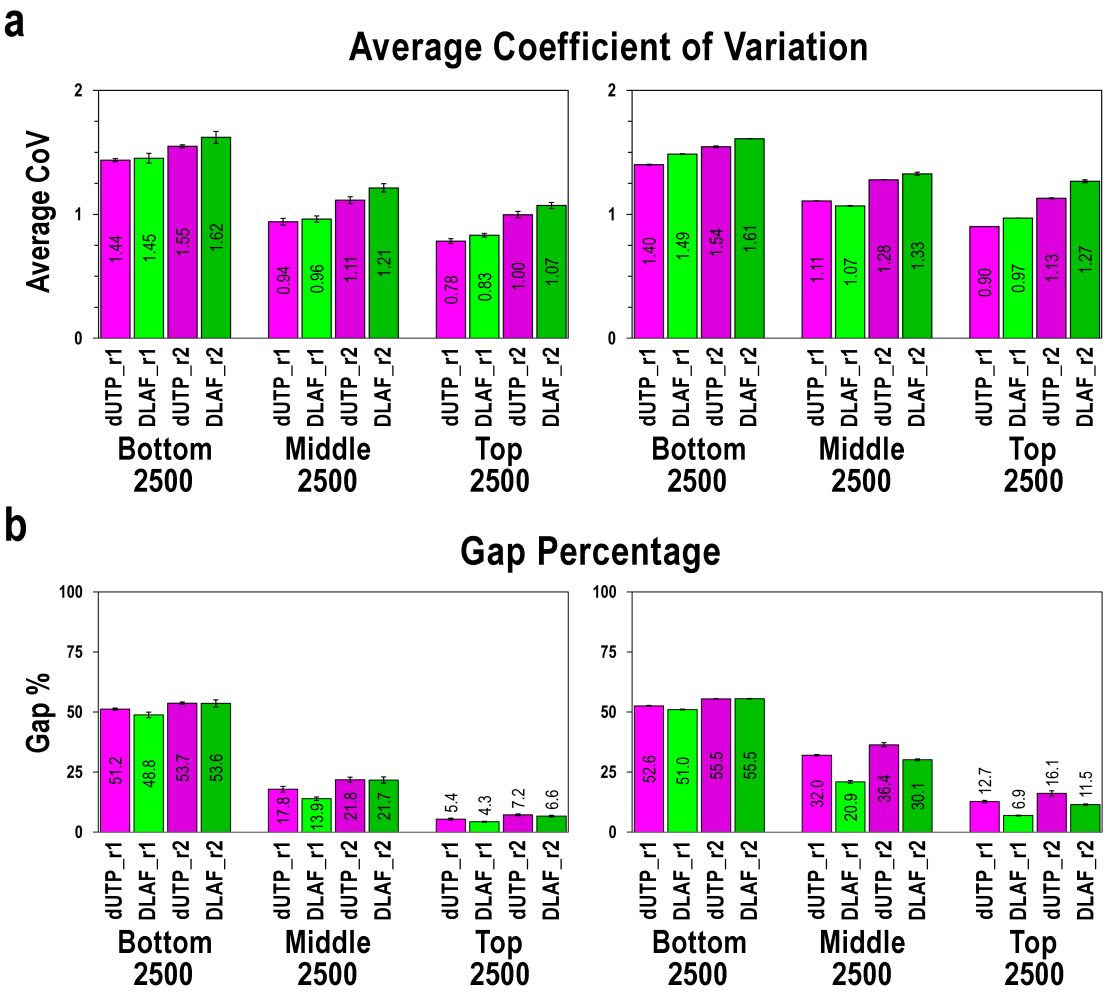
Supplementary Figure 10. Correlation of gene expression of DLAF and dUTP libraries. RNA-SeQC was used to generate the matrix for pairwise Pearson's (dark red) and Spearman's (purple) coefficients of the correlation of FPKM values from WT **(a)** and *Kdm1a* deficient mES cells **(b)**. Analysis was done for 12.5 million randomly sampled non-rRNA and non-mtRNA reads. The correlation between biological replicates is shown in bold.

Supplementary Figure 11



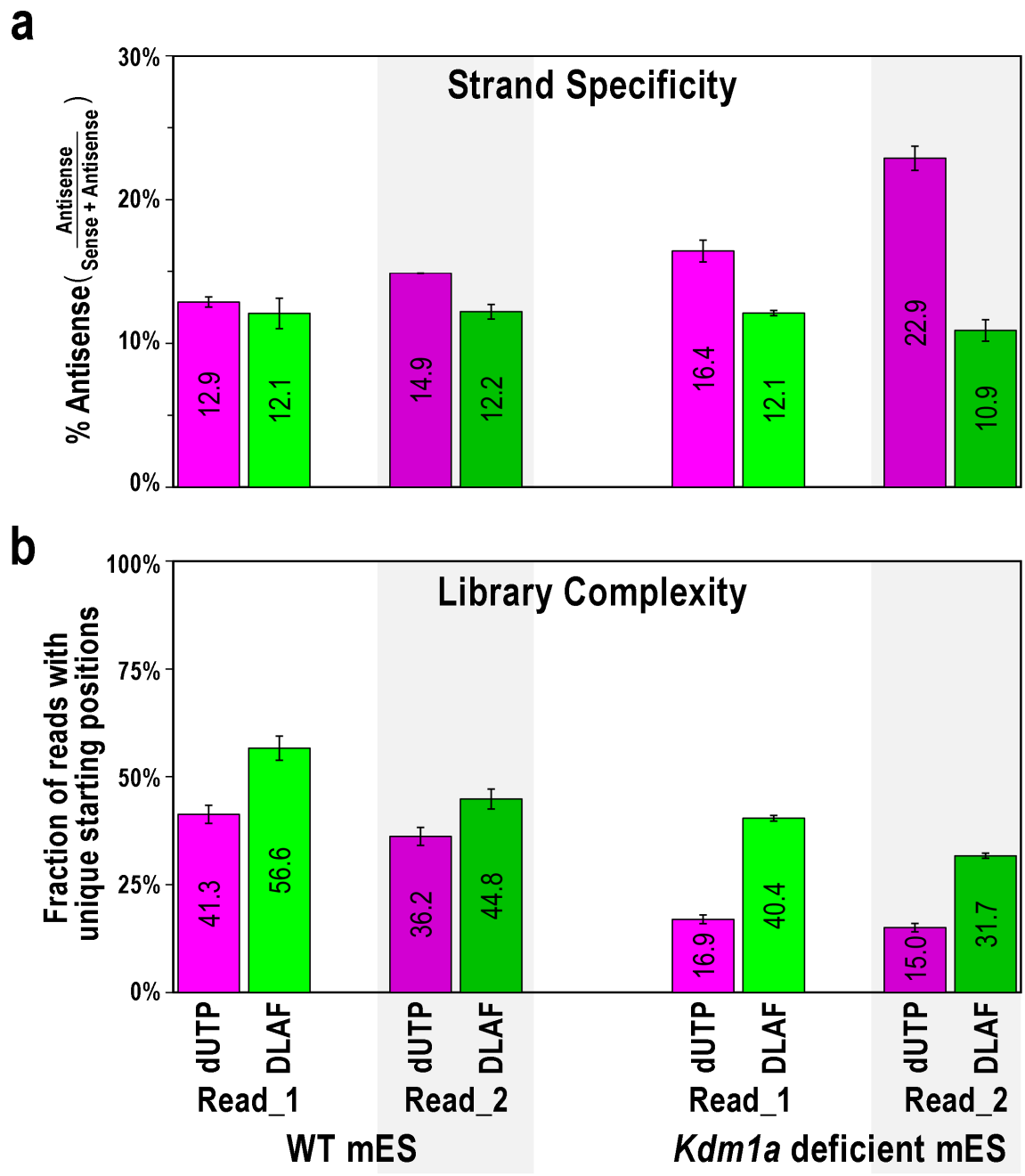
Supplementary Figure 11. Coefficient of variation of gene expression. Squared coefficient of variation of gene expression (CV^2), a measure of cross-replicate variability, was calculated by Cuffdiff and CummeRbund. The DLAF libraries from WT mES cells show a slightly lower CV^2 than the dUTP libraries, indicating a slightly higher reproducibility. Data from two biological replicates are shown.

Supplementary Figure 12



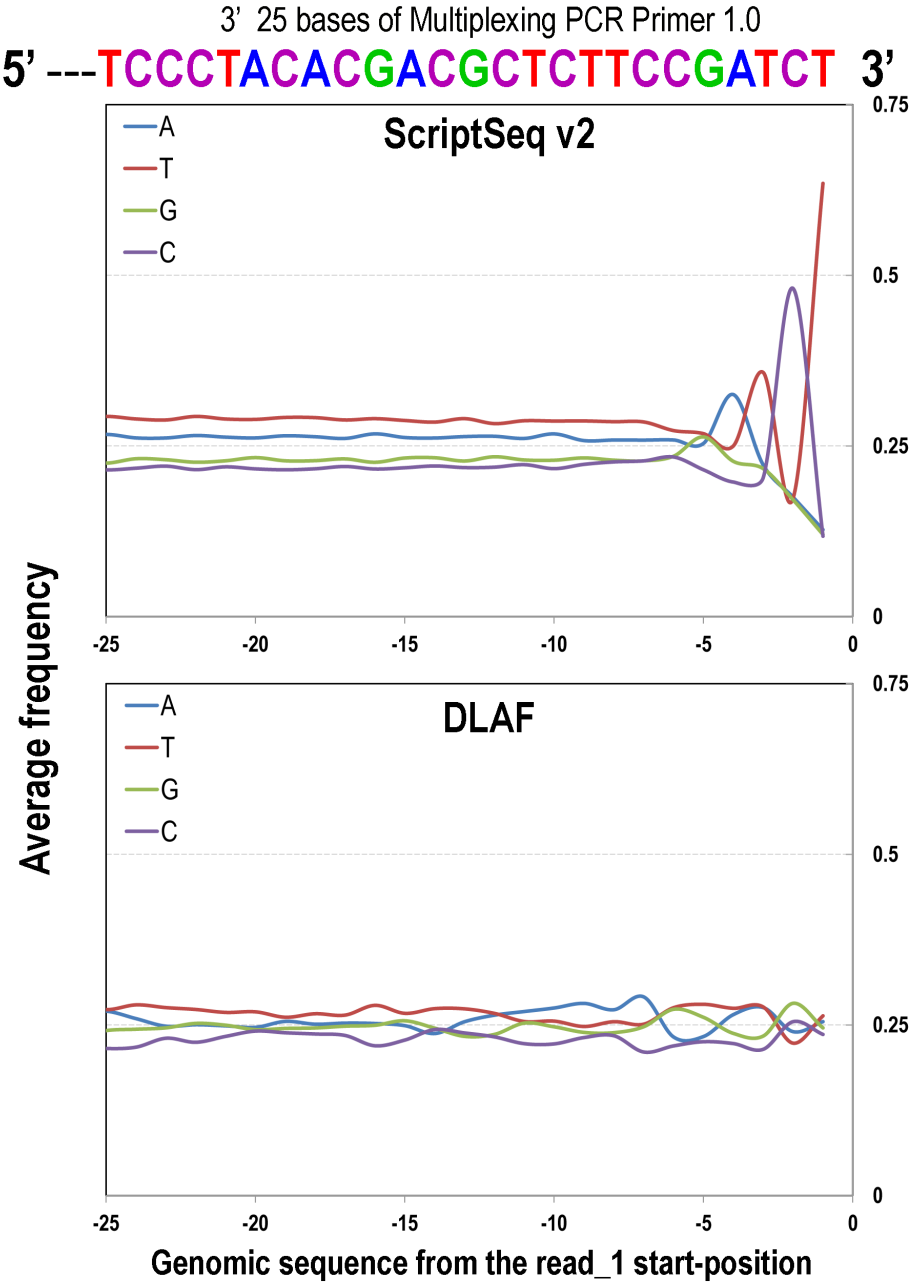
Supplementary Figure 12. Evenness and continuity of coverage. (a) The average coefficient of variation in 'evenness of coverage' (average CoV) was calculated by RNA-SeQC for the 2,500 bottom-, middle and top-expressed genes using 12.5 million randomly sampled non-rRNA and non-mtRNA reads. For both reads, the dUTP method libraries show a slightly lower average CoV, indicating a slightly more uniform coverage in both WT and *Kdm1a* deficient mES cells. **(b)** Continuity of coverage. Total cumulative gap length over the total cumulative transcript lengths (gap %) was calculated by RNA-SeQC. DLAF libraries show a consistently lower gap % in both WT and *Kdm1a* deficient mES cells. However, the improvement in *Kdm1a* deficient mES cells could also be attributed to the lower coverage of the genic regions (Supplementary Fig. 2). Average of two biological replicates is shown and error-bars indicate the range of data.

Supplementary Figure 13



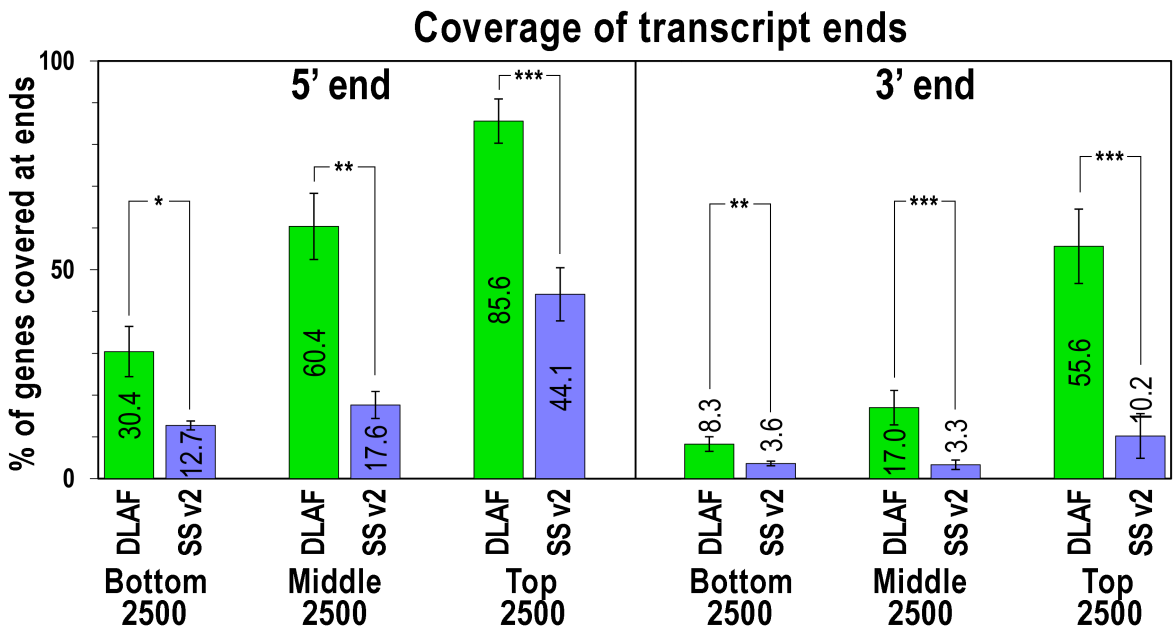
Supplementary Figure 13. Strand specificity and complexity of the DLAF and dUTP libraries. (a) Antisense rate was calculated as the fraction of reads that map to known transcripts in an antisense direction out of the total reads for the transcripts (antisense %). **(b)** The complexity of the libraries was estimated as the fraction of 12.5 million randomly-sampled non-ribosomal and non-mitochondrial reads with unique starting positions using the rmdup utility of samtools. For each cell line, DLAF libraries showed a markedly higher complexity for both read_1 and read_2. Average of two biological replicates is shown and error-bars indicate the range of data.

Supplementary Figure 14



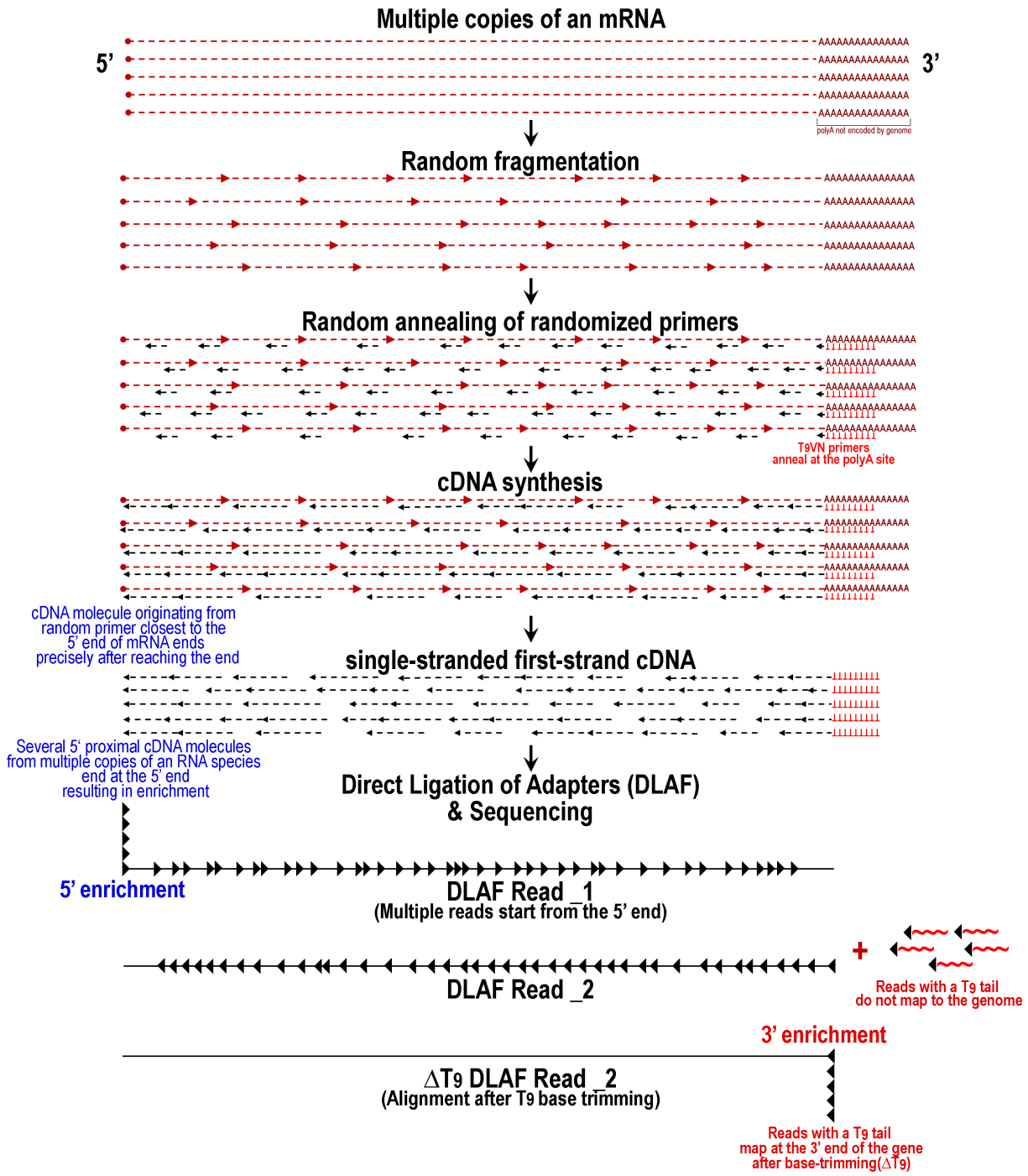
Supplementary Figure 14. Sequence bias in ScriptSeq v2 libraries. Genomic sequences upstream of read_1 were extracted using getFasta utility of the BEDTools[3] and checked for their sequence content. ScriptSeq libraries show an enrichment of RNA fragments originating from downstream of regions that contain GATCT sequence. Last 25 bases of Illumina's multiplexing PCR primer 1.0 are shown in color. DLAF libraries do not show such enrichment. Average of three biological replicates is shown.

Supplementary Figure 15



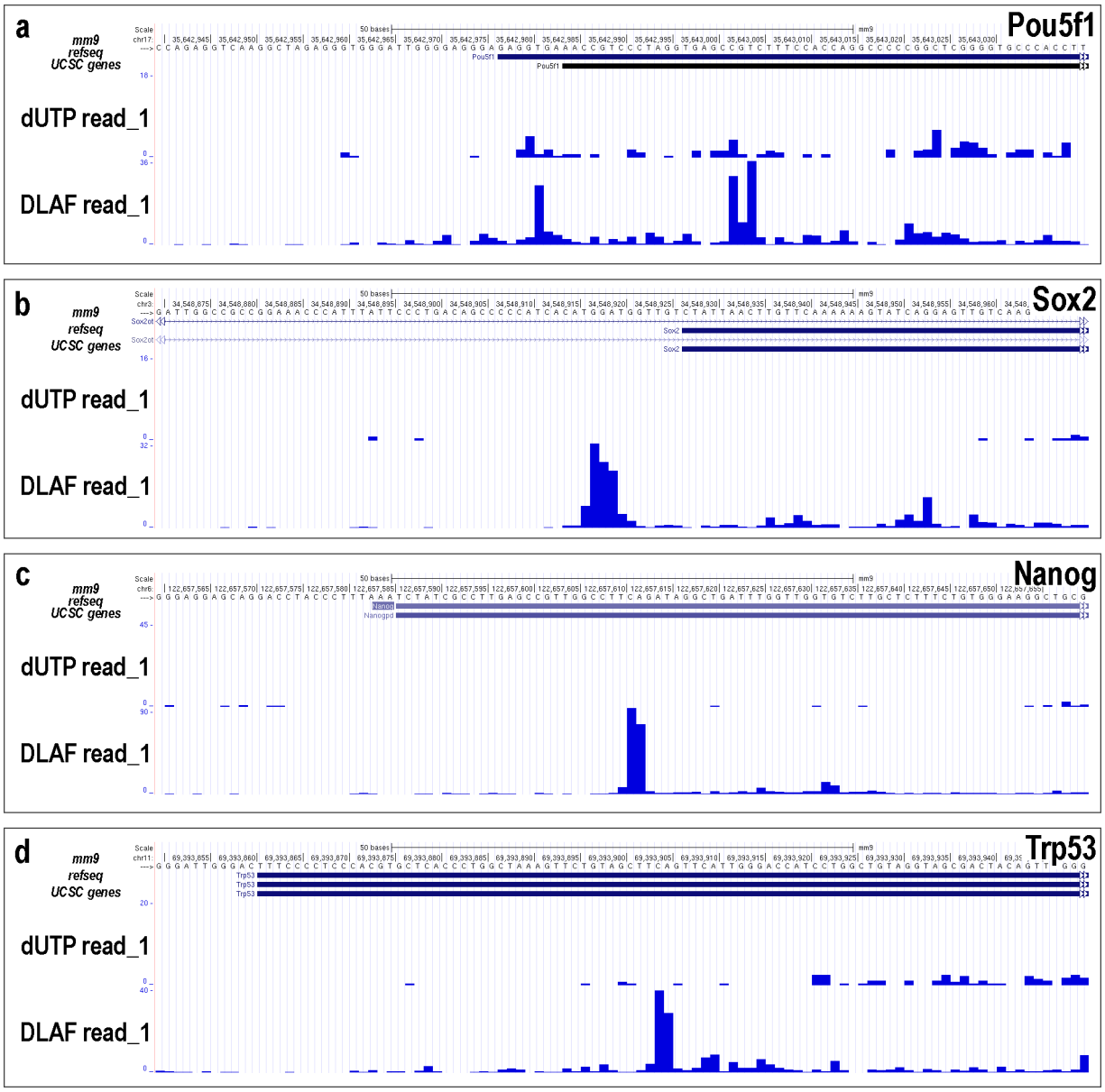
Supplementary Figure 15. Coverage of transcript ends by DLAF and ScriptSeq libraries. Coverage was calculated by RNA-SeQc using 15 million randomly sampled non-rRNA non-mtRNA reads. ScriptSeq read_1 show a markedly lower number of genes covered at both the 5'- and 3'-ends than the DLAF libraries. Average of three biological replicates is shown and error-bars indicate the standard deviation. ***$P < 0.01$, **$P < 0.05$, and *$P < 0.1$ in two-sided, unpaired-samples Student's *t*-tests.

Supplementary Figure 16



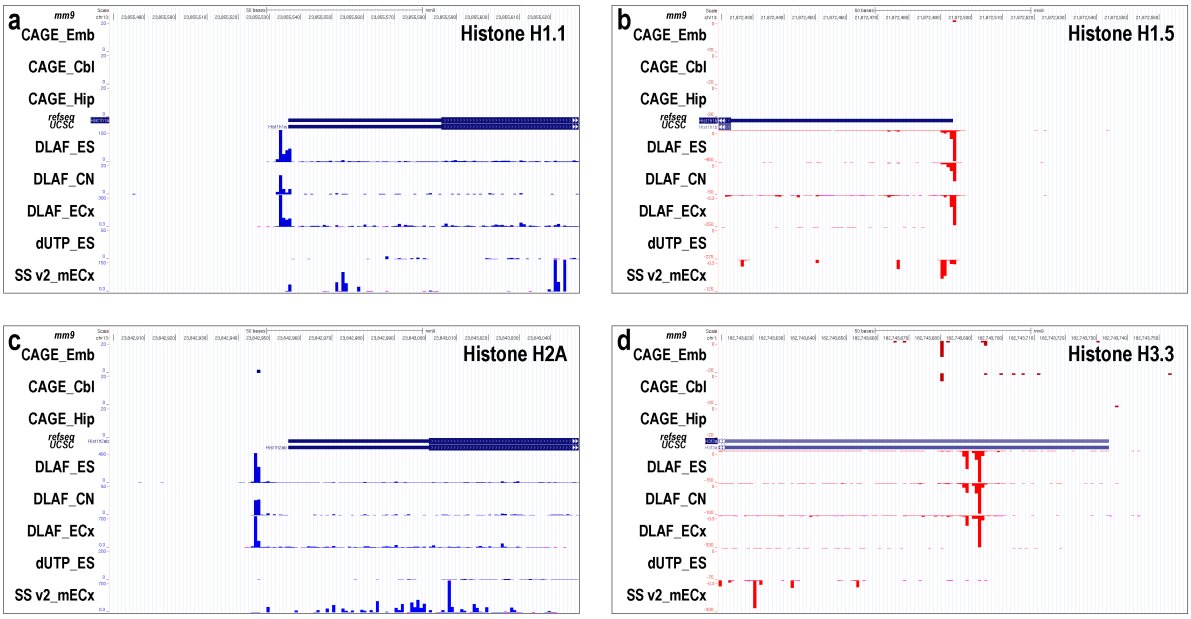
Supplementary Figure 16. Schematic explanation of the enrichment of 5' and 3' ends with DLAF. Enrichment of 5' ends is likely a result of multiple cDNA molecules that end precisely at the 5' end. After random fragmentation, random primers anneal to initiate cDNA synthesis from multiple positions. In the middle of transcripts, RT results in randomly distributed ends. However, at the 5' end of an mRNA, cDNA synthesis stops when reverse transcriptase reaches the last nucleotide on the RNA (5' end). During sequencing, read_1 start precisely from the 3' ends of the cDNA, thus resulting in enrichment at the 5' end of the genes. Similar enrichment is also observed at the 3' ends of polyadenylated RNA because T_9VN random primers bind predominantly at the sites of polyadenylation.

Supplementary Figure 17



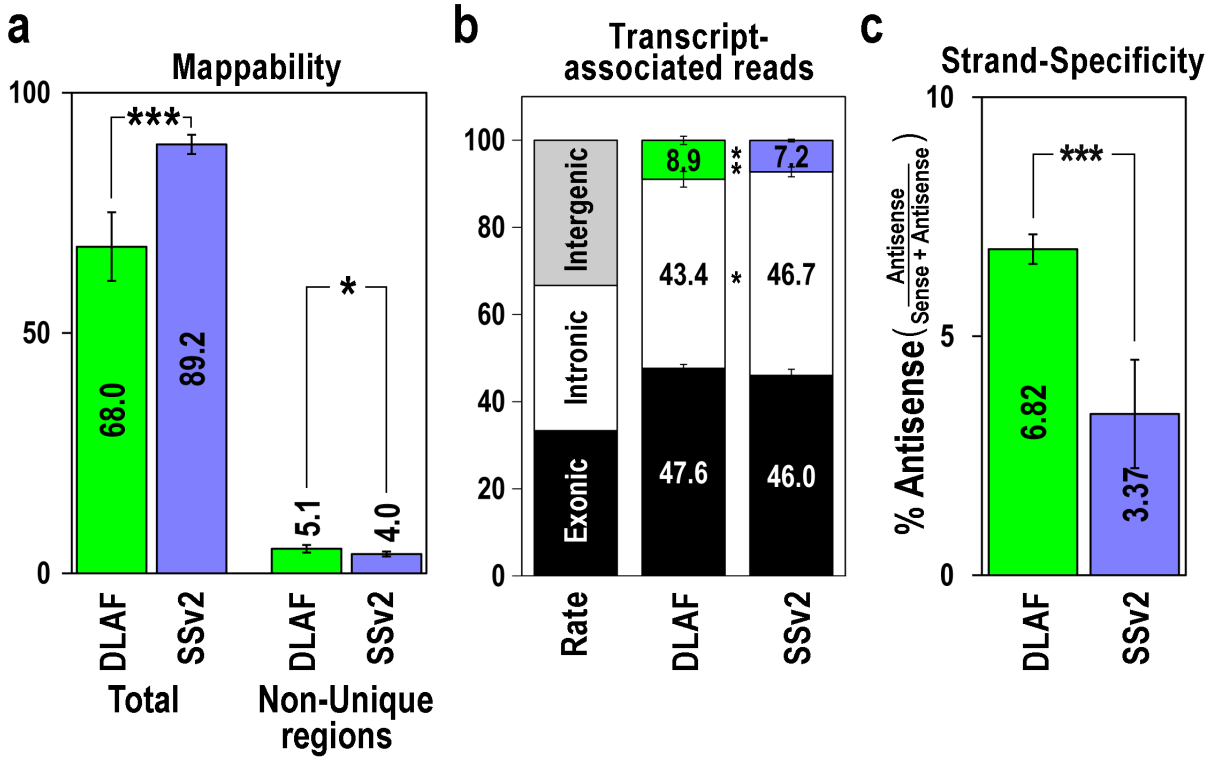
Supplementary Figure 17. Utility of DLAF in identification of novel TSSs of mRNA. The UCSC genome browser showing the read_1 start positions near the promoters of four genes, namely *Pou5f1* (**a**), *Sox2* **(b)**, *Nanog* **(c),** and *Trp53* **(d)** from libraries from mES cells. At these loci, the start positions of DLAF read_1 did not match their annotated TSSs. Thus, the data can be used to identify novel tissue/cell type-specific TSSs at near-base resolution. A signal at the -1 and -2 base positions relative to the highest peak is likely a result of the non-templated incorporation of nucleotides by reverse transcriptase past the 5' end of RNA[4](See supplementary Table 1).

Supplementary Figure 18



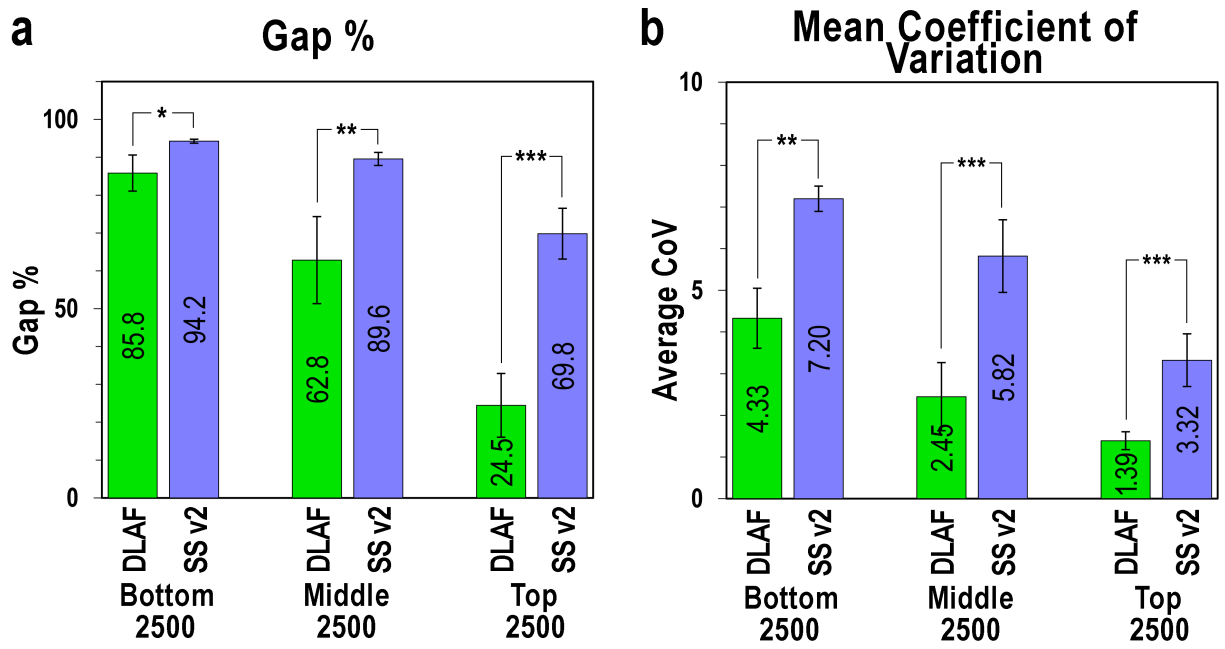
Supplementary Figure 18. Utility of DLAF in identification of novel TSSs of non-polyadenylated genes. The UCSC genome browser showing the read_1 start positions near the promoters of four histone genes, namely *H1.1* **(a)**, *H1.5* **(b)**, *H2A* **(c)** and *H3.3* **(d)**. At these loci, the start positions of DLAF read_1 from 3 different cell lines namely WT mES cells, cortical neurons and embryonic cortex show sharp peaks around the annotated TSSs, whereas such signal was not detected in dUTP or ScripSeq v2 libraries. DeepCAGE data from a previous study shows little or no signal at the TSS for these genes, which are known to be devoid of a poly(A) tail.

Supplementary Figure 19



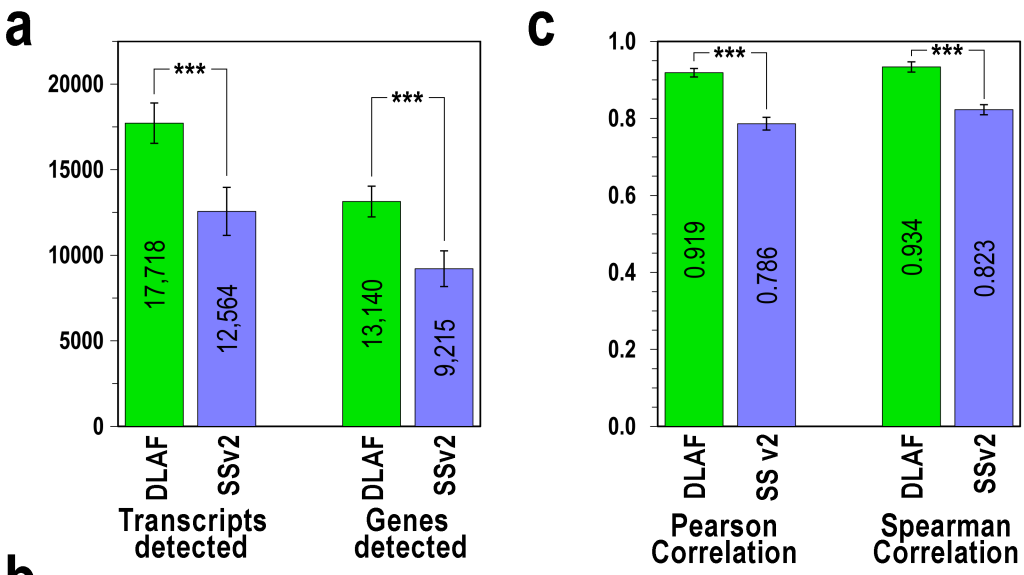
Supplementary Figure 19. Mappability of reads, relative coverage of genic/ intergenic regions, and strand-specificity of DLAF and ScriptSeq libraries. (a) ScriptSeq libraries show a significantly higher total mappability of reads to the genome and a slightly lower mappability to the non-unique regions of the genome. **(b)** ScriptSeq libraries show a slightly but significantly lower mapping rate to intergenic regions. **(c)** ScriptSeq libraries also show a significantly higher strand-specificity compared to DLAF libraries. Antisense rate was calculated as the fraction of reads that map to known transcripts in an antisense direction out of the total reads for the transcripts (antisense %). Average of three biological replicates is shown and error-bars indicate the standard deviation. ***$P < 0.01$, **$P < 0.05$, and *$P < 0.1$ in two-sided, unpaired-samples Student's *t*-tests.

Supplementary Figure 20



Supplementary Figure 20. Continuity/evenness of coverage of DLAF and ScriptSeq libraries. Analysis was performed using 15 million randomly sampled non-rRNA and non-mtRNA reads using RNA-SeQC **(a)** Continuity of coverage. DLAF read_1 show a significantly lower gap% for each of the three categories of the genes indicating significantly higher continuity of coverage for all three categories than the ScriptSeq libraries. **(b)** Evenness of coverage. DLAF read_1 show a significantly lower mean coefficient of variation in evenness for each of the three categories of the genes indicating significantly higher uniformity of coverage for all three categories than the ScriptSeq libraries. Average of three biological replicates is shown and error-bars indicate the standard deviation. ***$P < 0.01$, **$P < 0.05$, and *$P < 0.1$ in two-sided, unpaired-samples Student's *t*-tests.

Supplementary Figure 21



b

		Pearson Correlation					
Spearman Correlation	Sample	DLAF_1	DLAF_2	DLAF_3	SSv2_1	SSv2_2	SSv2_3
	DLAF_1	1	**0.918**	**0.908**	0.774	0.771	0.830
	DLAF_2	**0.937**	1	**0.930**	0.772	0.758	0.831
	DLAF_3	**0.919**	**0.945**	1	0.787	0.770	0.842
	SSv2_1	0.815	0.831	0.833	1	**0.768**	**0.800**
	SSv2_2	0.810	0.815	0.813	**0.808**	1	**0.791**
	SSv2_3	0.847	0.863	0.862	**0.833**	**0.827**	1

Supplementary Figure 21. Number of transcripts/genes detected and correlation of gene expression between the DLAF and ScriptSeq libraries. Analysis was carried out using15 million randomly sampled non-rRNA and non-mtRNA reads using RNA-SeQC. **(a)** For the same number of reads, DLAF read_1 detected a significantly higher number genes and transcripts. **(b)** RNA-SeQC was used to generate the matrix for pairwise Pearson's (dark red) and Spearman's (purple) coefficients of the correlation of FPKM values for ScriptSeq v2 and DLAF libraries from mECx. Inter-replicate correlation is shown in bold. **(c)** DLAF libraries show a significantly lower *r* values in both Pearson's and Spearman's correlations between replicates, indicating a higher reproducibility. Average of three biological replicates is shown and error-bars indicate the standard deviation. ***$P < 0.01$ in two-sided, unpaired-samples Student's *t*-tests.

Supplementary Table 1

Top-2500 expressed gene	Annotation used to define +1 position	Signal at +1 position A (as % of T)	Signal at 0 position B (as % of T)	Signal at -1 position C (as % of T)	Signal at -2 position D (as % of T)	Total signal T (A+B+C+D)
Rmrp	Refseq TSS	10618.0 (95.29%)	476.63 (4.28%)	26.20 (0.24%)	22.43 (0.20%)	11143.26
Malat1	Refseq TSS	372.04 (26.86%)	962.70 (69.51%)	43.81 (3.16%)	6.50 (0.47%)	1385.04
Actb	DeepCAGE Peak	1190.32 (93.54%)	81.12 (6.37%)	1.05 (0.08%)	0.00 (0.00%)	1272.48
Hist1H4d	DeepCAGE Peak	112.77 (13.63%)	679.11 (82.09%)	29.76 (3.60%)	5.66 (0.68%)	827.29
Eef2	DeepCAGE Peak	92.85 (16.51%)	434.08 (77.19%)	23.06 (4.10%)	12.37 (2.20%)	562.36
Rpl8	DeepCAGE Peak	49.68 (10.67%)	380.85 (81.81%)	31.44 (6.75%)	3.56 (0.77%)	465.52
Hist1h3d	Refseq TSS	125.55 (28.99%)	268.71 (62.05%)	21.80 (5.03%)	16.98 (3.92%)	433.04
Rps21	DeepCAGE Peak	22.64 (7.91%)	186.34 (65.13%)	73.57 (25.71%)	3.56 (1.25%)	286.11
Gapdh	Refseq TSS	165.79 (98.51%)	1.47 (0.87%)	1.05 (0.62%)	0.00 (0.00%)	168.31
Rps27	Refseq TSS	81.12 (75.44%)	15.51 (14.42%)	9.22 (8.58%)	1.68 (1.56%)	107.53
Fraction of total signal Average ± S.D.		**46.74% ± 38.84%**	**46.37% ± 35.08%**	**5.79% ± 7.53%**	**1.10% ± 1.22%**	100%

Supplementary Table 1: Fraction of reads mapping to bases immediately upstream of the TSSs. Data is shown for genes from the top-2500 expressed genes in WT mES cells in a DLAF library. Ten highly expressed genes that show a single DLAF peak either at a Refseq annotated TSS or at the same position as a DeepCAGE peak [5] were selected for this analysis. +1 position indicates the first base on a transcript. 0 indicates the base immediately upstream of the TSS. DLAF read_1 show ~46% reads mapping to positions +1 and 0, whereas positions -1 and -2 show greatly reduced signals.

Supplementary Table 2

Sample Name and replicate	Cell Line	Suffix for Single-End runs	Suffix for Paired-End runs			
			Run 2		Run 1	
DLAF_WT_mES_17_rep1 DLAF_WT_mES_17_rep2	WT mES	SR_RUN_read1 (Yes)	PE_RUN_read1 (No) [β]	PE_RUN_read2 (Yes)	PE_RUN_small_read1 (No) [α]	PE_RUN_small_read2 (No) [α]
DLAF_KDM1A_KD_mES_03_rep1 DLAF_KDM1A_KD_mES_03_rep2	Kdm1a -/- mES					
dUTP_WT_mES_17_rep1 dUTP_WT_mES_17_rep2	WT mES		PE_RUN_read1 (No)	PE_RUN_read2 (Yes)		
dUTP_KDM1A_KD_mES_03_rep1 dUTP_KDM1A_KD_mES_03_rep2	Kdm1a -/- mES					
miRNA_small_RNA_WT_mES	WT mES	read1_no_barcode (Yes)				
DLAF_WT_mCN_rep1	mCN	SR_RUN_1_read1 (Yes) SR_RUN_2_read1 (Yes)				
DLAF_mECx_rep1 DLAF_mECx_rep2 DLAF_mECx_rep3	mECx	SR_RUN (Yes)				
DLAF_mECx_Klenow_low_rep1 DLAF_mECx_Klenow_low_rep2						
DLAF_mECx_Klenow_high_rep1 DLAF_mECx_Klenow_high_rep2						
SSv2_mECx_rep1 SSv2_mECx_rep2 SSv2_mECx_rep3						

Supplementary Table 2. File and sequencing run information for the samples in this study. 'Yes' or 'No' in parentheses indicate whether the results from these runs are presented in this study. α indicates a low number of reads from sequencing. β indicates that read_1 base 3 had a high 'N' content and lower quality. Reads from one or more replicates of the same sample (shown as in the same box in column 1) were merged in a strand-specific manner to generate the bigwig files.

Supplementary Note 1. RNA ligation and 3' split-adaptor method.

One strategy to omit second-strand cDNA synthesis is to directly ligate sequencing adaptors to RNA molecules (RNA ligation), which has been used for 5'-Rapid amplification of cDNA ends (5' RACE) and ssRNA-seq [6]. However, T4 RNA ligase, and its derivatives, suffer from moderate enzymatic activity compared to T4 DNA ligase due to its high *K*m value [7]. In addition, T4 RNA ligase reportedly exhibits a sequence bias to terminal nucleotides [8,9]. The lower enzymatic activity limits the utility of the RNA ligation method to relatively abundant RNA samples. In addition, the lengthy process of RNA ligation involves a risk of contamination with RNases. The extensive RNA handling processes include phosphorylation and dephosphorylation prior to the two ligation steps for each end of RNA and gel purification to remove unligated adaptors. Indeed, the loss of RNA due to circularization has been reported.

Alternatively, RT using a 3'-split adaptor can be used to omit second-strand synthesis [10]. A 3'-split DNA adaptor [10] consists of an oligo(dT) sequence at the 3' end and a defined sequence (for sequencing) at the 5' end. The oligo(dT) sequence anneals to poly(A) sequences that are artificially attached to the 3' end of fragmented mRNA species, thereby priming the RT reaction. Resulting single-strand cDNA is circularized by ligation, cleaved at a modified nucleotide between the adaptor sequences, PCR-amplified, and subjected to sequencing. Importantly, the oligo(dT) sequence in the adaptor will not differentiate artificially added poly(A) tails from endogenous poly(A) tails of mRNA. Therefore, in the current form, the 3'-split DNA adaptor method is not suitable for the analysis of the polyadenylation of mRNAs, which often has important biological implications.

Supplementary Note 2. Modifications made during the preparation of dUTP method libraries.

We made minor adjustments to the dUTP method to compare it to DLAF accurately. We used 5'-phosphorylated random primers for RT to simplify the process. We also omitted *E. coli* DNA ligase

during second-strand synthesis because the ligase might have ligated 5'-phosphorylated fragments, which would have likely resulted in chimeric cDNA species. It has been shown that *E. coli* ligase is not required during second-strand synthesis reaction [11], and, in fact, several commercially available kits do not contain this enzyme. We excluded library DNAs with less than 125-bp inserts to minimize the contamination of concatenated adaptors and cDNAs generated from small rRNA fragments and tRNA.

Supplementary Note 3. Decreasing coverage in the 5' → 3' direction for DLAF and dUTP libraries.

We postulate that the moderately increasing coverage in the 3' → 5' direction could be attributed to RT. During the RT reaction, random oligonucleotides likely anneal to the middle of fragmented RNA molecules rather than precisely at the 3'-termini of RNAs. The RT reaction then proceeds until the reverse transcriptase reaches the 5' end of the RNA fragment or encounters the 5' end of a downstream cDNA. Therefore, in a given population of randomly fragmented RNA molecules, 3'-ends are less likely to be reverse transcribed than are the 5' ends. This might explain the gradual decrease of coverage in the 5' → 3' direction.

Supplementary Note 4. ΔT_9 read_2 contain known features of polyadenylation sites.

We sought to verify the presence of polyadenylation signal (PAS) on the ΔT_9 reads. Canonical mammalian PAS consists of a either the highly conserved AAUAAA (PAS1) or the less prevalent variant AUUAAA (PAS2) sequence [12]. PAS hexamers are usually embedded within U-rich sequences upstream and downstream of 3'-cleavage sites (CSs) [12,13]. The first bases of PAS hexamers are distributed 10~30 bases upstream of the RNA 3' CSs with a peak at 20~21st bases [12]. The canonical CS sequence has been shown to be 5'-CA-3' [2], after which the RNA is cleaved and a poly(A) tail is added [12,14]. We found that 48.07% of the ΔT_9 reads in the DLAF libraries from the WT mES cells contained PAS1, whereas 11.74% of the ΔT_9 reads contained PAS2 (Supplementary Fig. 4a). This result is consistent with the previous

studies showing that PAS1 and PAS2 were present on ~55% and ~16% of transcripts in human cells respectively [12,13]. The first bases of PAS1 and PAS2 in the DLAF libraries were distributed between 30 and 15 bases from the site of T_9 trimming, with a peak at the 20th base (Supplementary Fig. 4b). This is almost identical to the reported distribution pattern of PAS hexamers. Interestingly, cytosine was observed as the most frequent base at the first nucleotide of ΔT_9 reads in DLAF libraries (Supplementary Fig. 4b). We obtained similar results regarding the distribution of PAS hexamers in the dUTP libraries; however, the preference of cytosine near the CSs was not observed (Supplementary Fig. 4b).

Supplementary Note 5. Comparative analysis of DLAF and ScriptSeq libraries.

To determine if the differences in the mean values for various metrics for DLAF- and ScriptSeq -libraries were statistically significant, we employed two-sided, unpaired-samples Student's *t*-tests to calculate the *p*-values (*P*), unless otherwise mentioned**.** We first compared the yields of libraries by semi-quantitative PCR. ScriptSeq libraries showed a lower yield than that of DLAF (data not shown), and needed ~ 4 additional cycles of PCR to achieve visibility on a gel. Multiplexed libraries were subjected to single-end sequencing and were mapped as described earlier. Interestingly, ScriptSeq libraries showed a significantly higher overall mapping rate (89.22% for ScriptSeq vs. 67.96% for DLAF, $P < 0.01$, Supplementary Fig. 19a). Mapping rates to the exonic, intronic, and intergenic regions were largely similar between the two methods, with only slight differences (Supplementary Fig. 19b). The ScriptSeq libraries also showed exceptionally low rates of antisense transcripts (3.37%) than did the DLAF libraries (6.82%, $P < 0.01$, Supplementary Fig. 19c).

We then evaluated the quality of the libraries by examining the continuity/evenness of coverage, reproducibility in expression profiling, and strand specificity. The ScriptSeq libraries showed a significantly higher gap percentage (ScriptSeq: 69.8% vs. DLAF: 24.5% in top-2500 expressed genes, $P < 0.05$, Supplementary Fig. 20a) as well as higher mean coefficients of variation in evenness (ScriptSeq:

3.32 vs. DLAF: 1.39 for top-2500 expressed genes, $P < 0.05$, Supplementary Fig. 20b). Results were similar regardless of gene expression levels (Supplementary Figs. 20a and b) indicating a highly discontinuous coverage of transcripts in the ScriptSeq libraries. Consistently, significantly lower numbers of genes or transcripts were detected in the ScriptSeq libraries (ScriptSeq: 9,215 genes vs. DLAF: 13,140 genes, $P < 0.05$, Supplementary Fig. 21a). The reproducibility in expression profiling in the ScriptSeq was significantly lower than that in the DLAF libraries (ScriptSeq: $r = 0.786$ vs. DLAF: $r = 0.919$, Pearson's correlation, $P < 0.01$, Supplementary Fig. 21b and c). These data demonstrate that, though the ScriptSeq had a higher mapping rate than did the DLAF, the mapped ScriptSeq reads represent lower reproducibility and a biased population of the transcripts.

Despite the overall similarity, there are two key differences between DLAF and ScriptSeq. First, DLAF employs randomized oligonucleotides for RT, whereas ScriptSeq uses random oligonucleotides attached to Illumina reverse primer. Second, during the tagging step, i.e. ligation for DLAF, cDNA ends hybridize to DLAF adaptors only through complementarity to random oligos (Fig. 1), whereas TSO in ScriptSeq can potentially hybridize to cDNAs that are similar to the tagging sequence. This is because, unlike the DLAF adaptors, the tagging sequence in the TSO is single stranded. The use of oligonucleotides with some fixed DNA sequences in both the RT and tagging steps might possibly give rise to a sequence bias in ScripSeq libraries and, therefore, a biased representation of transcripts.

Supplementary Note 6. Storage and usage of actinomycin D.

Actinomycin D is an important reagent for the preparation of ssRNA-seq libraries using the DLAF or dUTP method [15]. It increases the strand specificity of the libraries by inhibiting DNA-dependent DNA polymerase activity during RT [16,17]. We recommend dissolving actinomycin D (Sigma) in 100% DMSO (Sigma) at a concentration greater than 2 mg/ml and storing it at -70°C or lower in small aliquots to avoid repeated freeze-thaws. We have found that stocks made in ethanol and stored at -20°C show a degradation

over time, which could be attributed to the fact that ethanol does not freeze at -20°C. We do not recommend using actinomycin D at a concentration greater than 5μM to maintain efficient RT.

References for supplementary information

1. Sheets, M.D., C.Ogg, S. & P.Wickens, M. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* **18**, 5799–5805 (1990).
2. Chen, F., MacDonald, C.C. & Wilusz, J. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res.* **23**, 2614-2620 (1995).
3. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
4. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques* **30**, 892-7 (2001).
5. Valen, E. *et al.* Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* **19**, 255-65 (2009).
6. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-36 (2008).
7. Uhlenbeck, O.C. & Cameron, V. Equimolar addition of oligoribonucleotides with T4 RNA ligase. *Nucleic Acids Res.* **4**, 85-98 (1977).
8. England, T.E. & Uhlenbeck, O.C. 3'-terminal labelling of RNA with T4 RNA ligase. *Nature* **275**, 560-1 (1978).
9. Romaniuk, E., McLaughlin, L.W., Neilson, T. & Romaniuk, P.J. The effect of acceptor oligoribonucleotide sequence on the T4 RNA ligase reaction. *Eur. J. Biochem.* **125**, 639-43 (1982).
10. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-23 (2009).
11. D'Alessio, J.M. & Gerard, G.F. Second-strand cDNA synthesis with E. coli DNA polymerase I and RNase H: the fate of information at the mRNA 5' terminus and the effect of E. coli DNA ligase. *Nucleic Acids Res.* **16**, 1999-2014 (1988).
12. Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10**, 1001-10 (2000).
13. Tian, B. & Graber, J.H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **3**, 385-96 (2012).
14. Proudfoot, N.J. Ending the message: poly(A) signals then and now. *Genes Dev.* **25**, 1770-82 (2011).
15. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
16. Perocchi, F., Xu, Z., Clauder-Munster, S. & Steinmetz, L.M. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.* **35**, e128 (2007).
17. Rill, R.L. & Hecker, K.H. Sequence-specific actinomycin D binding to single-stranded DNA inhibits HIV reverse transcriptase and other polymerases. *Biochemistry* **35**, 3525-33 (1996).