Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/25/eaaz6699/DC1

Supplementary Materials for

Genome-wide kinetic properties of transcriptional bursting in mouse embryonic stem cells

Hiroshi Ochiai*, Tetsutaro Hayashi, Mana Umeda, Mika Yoshimura, Akihito Harada, Yukiko Shimizu, Kenta Nakano, Noriko Saitoh, Zhe Liu, Takashi Yamamoto, Tadashi Okamura, Yasuyuki Ohkawa, Hiroshi Kimura, Itoshi Nikaido*

*Corresponding author. Email: ochiai@hiroshima-u.ac.jp (H.O.); itoshi.nikaido@riken.jp (I.N.)

Published 17 June 2020, *Sci. Adv.* **6**, eaaz6699 (2020)
DOI: 10.1126/sciadv.aaz6699

The PDF file includes:

Figs. S1 to S6
Legends for tables S1 to S4

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/25/eaaz6699/DC1)

Tables S1 to S4

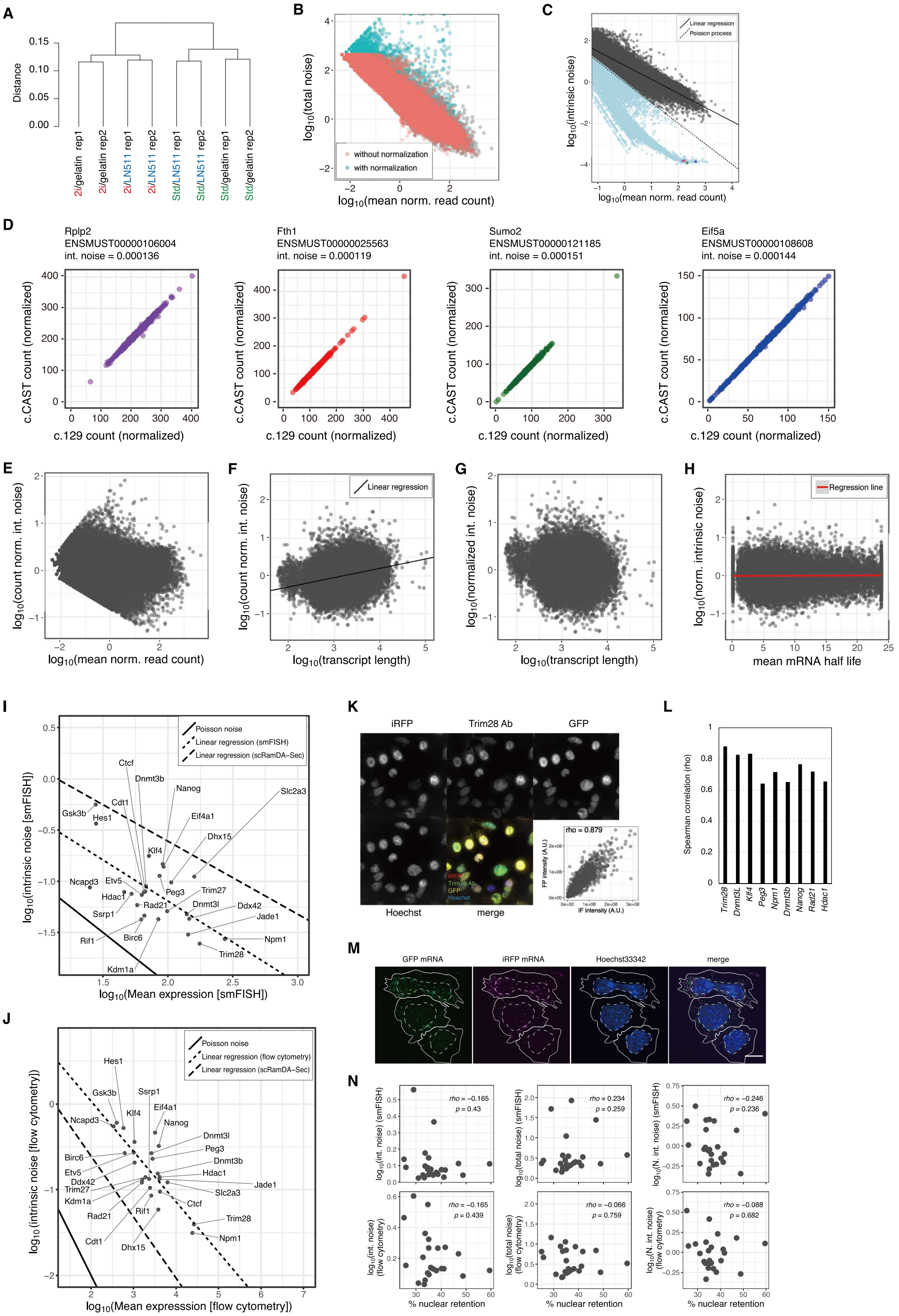

Figure S1. Intrinsic noise determination using scRNA-seq data. (A) The difference between the

transcriptomes of mESCs cultured on either Laminin-511 (LN511) or gelatin coated dishes was relatively subtle. mESCs were conditioned on either gelatin or Laminin-511 coated dishes under either Std or 2i medium. RNA was extracted from these cells and analyzed by RNA-seq. Two biological replicates were prepared. The result of clustering transcriptome data is shown. Biological replicates are similar to each other. The clusters are mainly divided into two clades depending on different medium conditions, suggesting that the influence of coating reagent difference (gelatin vs LN511) on the transcriptome is much smaller than that of the culture medium difference (Std vs 2i). **(B-G)** Processing of data obtained from scRNA-seq. **(B)** In order to calculate intrinsic noise, it was necessary to normalize so as to match the average expression level among alleles. Therefore, we first normalized the global expression level between alleles. Here, a scatter plot of the relationship between the mean read counts and total noise before and after a global normalization is shown. Even with global normalization, the shape of the distribution was not substantially changed. **(C)** Scatter plot of mean normalized read counts and intrinsic noise in data with expression levels that were normalized by quantile normalization between alleles. Theoretically, intrinsic noise should not be less than Poisson noise (1/mean read counts). Therefore, data below the Poisson noise were removed in subsequent analyses (see Materials and Methods). Intrinsic noise tends to decrease depending on the expression level. **(D)** Representative scatter plots of normalized individual allelic read counts of transcripts with intrinsic noise less than Poisson noise. The colors of spots correspond to **(C)**. These transcripts are extremely similar in expression between the alleles, resulting in very low intrinsic noise levels. **(E)** Scatter plot of read count-normalized intrinsic noise and mean normalized read counts. **(F)** Scatter plot of count-normalized intrinsic noise and transcript length. **(G)** Scatter plot of count and transcript length-normalized intrinsic noise (referred as just normalized intrinsic noise) and transcript length. **(H)** Scatter plot of mRNA half-life and normalized intrinsic noise. There was no correlation between them (*r* = 0.0104). **(I-J)** Intrinsic noise at mRNA level and protein level in KI cell line. Intrinsic noise at mRNA and protein levels was determined using data sets obtained by smFISH with allele-specific probes **(I)** and flow cytometry **(J)** in KI cell lines. Scatter plots of average expression level and intrinsic noise with regression lines are shown. The residuals from the regression line in the Y-axis direction were taken as normalized intrinsic noise. **(K-L)** Immunofluorescence of endogenous proteins in KI cell line. In some KI cell lines, the protein expressed from the target gene was fluorescently immunostained; the fluorescence image of immunofluorescence with GFP or iRFP derived from the knocked-in cassette was obtained by fluorescence microscopy. In addition, correlations of fluorescent intensity of individual cells were examined. Since the KI cassette contains 2A peptide, nuclear localization signal (NLS)-GFP (or iRFP), and PEST sequence, which induces rapid protein degradation, these endogenous proteins and knocked-in fluorescent proteins become different protein molecules after translation. **(J)** Fluorescent images of fluorescent proteins and immunostained endogenous TRIM28 proteins in the *Trim28* KI cell line. The images show maximum intensity projections of stacks. The lower right panel shows a scatter plot of the fluorescence intensities of the fluorescent proteins and the fluorescent immunostaining. **(K)** A bar graph of the Spearman's rank correlation coefficients between the fluorescence intensities of the fluorescent protein and the fluorescent immunostaining of the endogenous protein. All proteins analyzed here showed substantial correlation, suggesting that the fluorescence intensity of the fluorescent protein is indicative of endogenous protein abundance. (**M**) Representative images of smFISH using GFP and iRFP probes in *Nanog* KI cell line. Solid and dashed lines represent the plasma and nuclear membranes, respectively. Scale bar: 10 µm. (**N**) Scatter plots of nuclear retention rate of mRNA and either intrinsic noise, total noise or normalized intrinsic noise revealed by smFISH or flow cytometry.

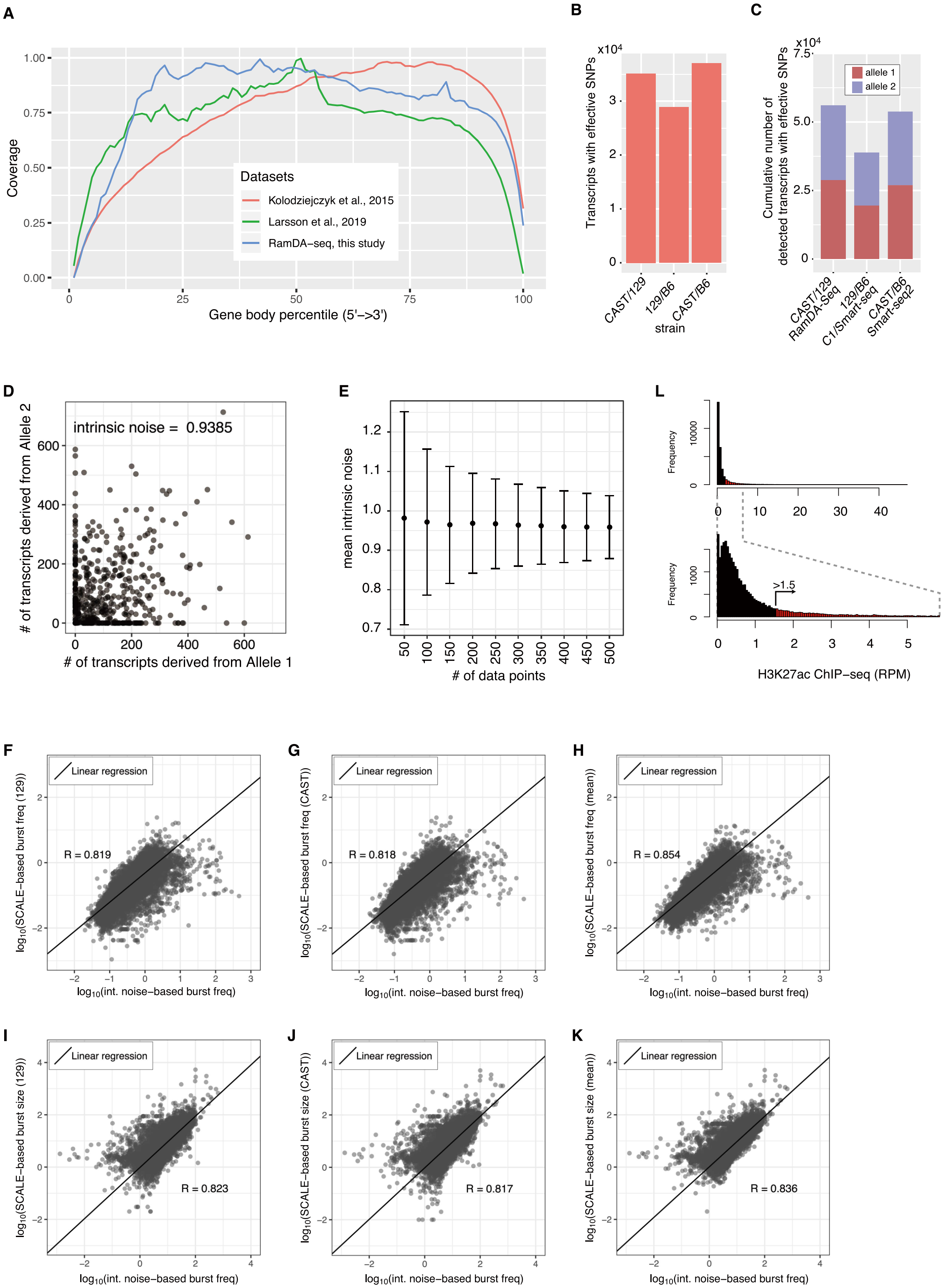

Figure S2. Comparison of the classification efficiency of each allele using SNPs with results of

single-cell (sc) RNA-seq (RamDA-seq) with that of other studies. (A) Mean read coverage over transcripts of three datasets. "RamDA-seq, this study": CAST/129 mESC scRamDA-seq; "Kolodziejczyk *et al*., 2015": 129/B6 mESC C1/Smart-seq performed in (*35*); "Larsson *et al*., 2019": CAST/B6 mESC Smart-seq2 performed in (*11*). **(B)** The theoretical numbers of transcripts with effective SNPs in three different hybrid mESC genomes. **(C)** The cumulative number of detected transcripts with effective SNPs in three different experiments. CAST/129 RamDA-seq: CAST/129 mESC scRamDA-seq in this study; 129/B6 C1/Smart-seq: 129/B6 mESC C1/Smart-seq performed in (*35*); CAST/B6 Smart-seq2: CAST/B6 mESC Smart-seq2 performed in (*11*). The classification efficiency of each allele using the SNPs of this study was comparable to that of the other studies. **(D, E)** The number of samples affects the confidence of calculated intrinsic noise. **(D)** Each of the 1,000 simulated datasets with 500 data points of the set of allele expression was generated. Mean expression of each allele and intrinsic noise in these data sets were 113.22 ± 5.74 (SD) and 0.958 ± 0.0799 (SD), respectively. One representative plot is shown. **(E)** Differences in standard deviation between datasets with different numbers of data points. We calculated the mean and standard deviation of intrinsic noise in datasets with different numbers of data points. Error bars indicate standard deviation. It is evident that the larger the number of data points, the smaller the data variation. We used 447 of 129/CAST mESCs in the G1 phase for the analysis of intrinsic noise in this study, while, Kolodziejczyk *et al*., and Larsson *et al*. used 250 of 129/B6 mESCs, and 188 of B6/CAST mESCs that were in different phases of the cell cycle and were cultured in Std medium or unknown condition in their scRNA-seq, respectively (*11*, *35*). Thus, our data are expected to be an important resource for a deeper understanding of transcriptional bursting and intrinsic noise. **(F-K)** Comparison of burst frequencies **(F-H)** and sizes **(I-K)** inferred using either intrinsic noise levels or SCALE software (see Materials and Methods). By using SCALE, the burst size and frequency parameters of individual alleles can be determined. **(F, I)** Scatter plots of intrinsic noise-based vs. SCALE-based (129 allele) parameters with regression lines. **(G, J)** Scatter plots of intrinsic noise-based vs. SCALE-based (CAST allele) parameters with regression lines. **(H, K)** Scatter plots of intrinsic noise-based vs. SCALE-based (mean of CAST and 129 allele values) parameters with regression lines. **(L)** Histogram of reads per million reads (RPM) of H3K27ac ChIP-seq of promoter-associated regions in mESCs (see Materials and Methods). Enhancers were manually defined as H3K27ac ChIP-seq RPM greater than 1.5.

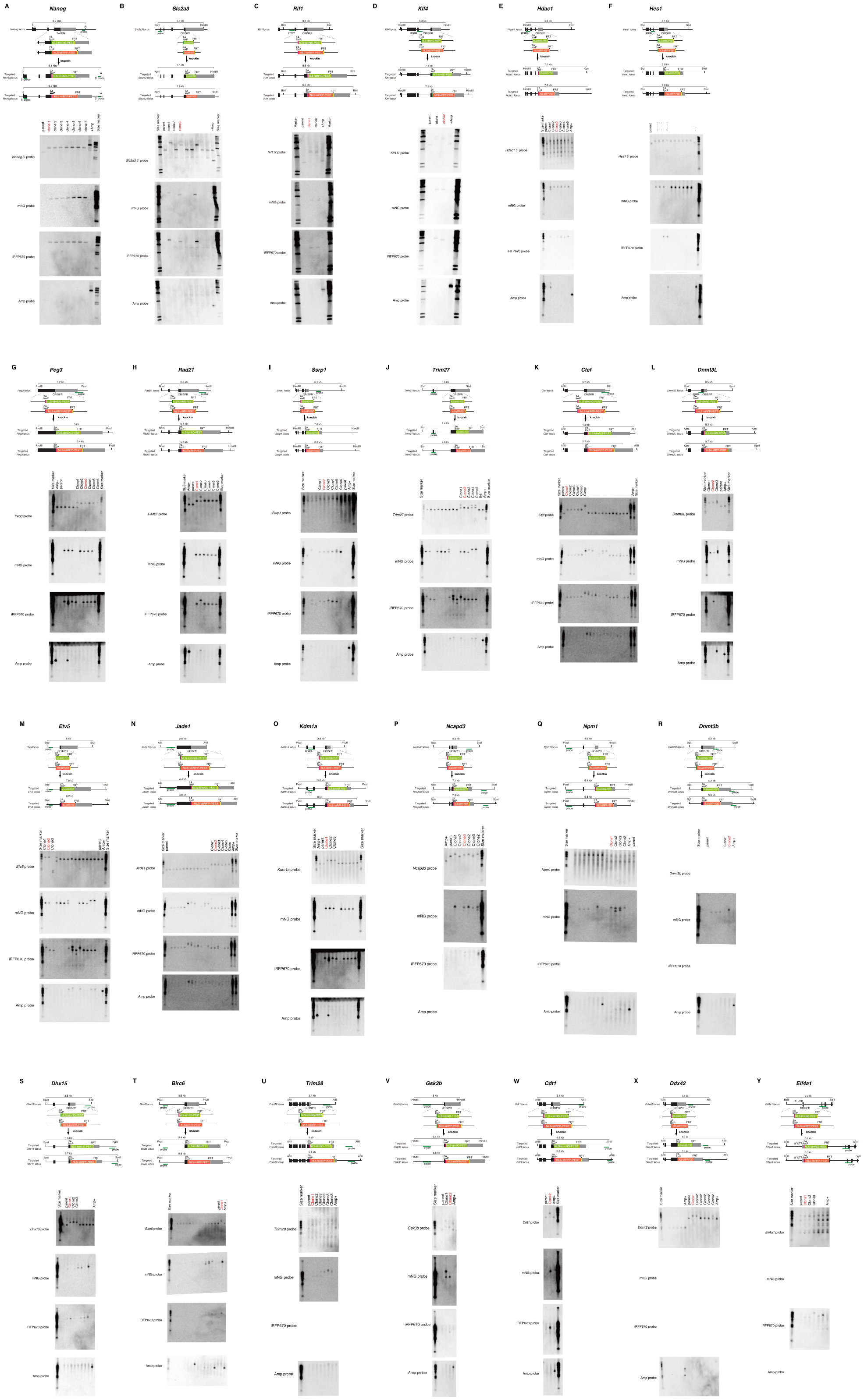
A Nanog
Nanog 5' probe
mNG probe
iRFP670 probe
Amp probe
B Slc2a3
Slc2a3 5' probe
mNG probe
iRFP670 probe
Amp probe
C Rif1
Rif1 5' probe
mNG probe
iRFP670 probe
Amp probe
D Klf4
Klf4 5' probe
mNG probe
iRFP670 probe
Amp probe
E Hdac1
Hdac1 5' probe
mNG probe
iRFP670 probe
Amp probe
F Hes1
Hes1 5' probe
mNG probe
iRFP670 probe
Amp probe
G Peg3
Peg3 probe
mNG probe
iRFP670 probe
Amp probe
H Rad21
Rad21 probe
mNG probe
iRFP670 probe
Amp probe
I Ssrp1
Ssrp1 probe
mNG probe
iRFP670 probe
Amp probe
J Trim27
Trim27 probe
mNG probe
iRFP670 probe
Amp probe
K Ctcf
Ctcf probe
mNG probe
iRFP670 probe
Amp probe
L Dnmt3L
Dnmt3L probe
mNG probe
iRFP670 probe
Amp probe
M Etv5
Etv5 probe
mNG probe
iRFP670 probe
Amp probe
N Jade1
Jade1 probe
mNG probe
iRFP670 probe
Amp probe
O Kdm1a
Kdm1a probe
mNG probe
iRFP670 probe
Amp probe
P Ncapd3
Ncapd3 probe
mNG probe
iRFP670 probe
Amp probe
Q Npm1
Npm1 probe
mNG probe
iRFP670 probe
Amp probe
R Dnmt3b
Dnmt3b probe
mNG probe
iRFP670 probe
Amp probe
S Dhx15
Dhx15 probe
mNG probe
iRFP670 probe
Amp probe
T Birc6
Birc6 probe
mNG probe
iRFP670 probe
Amp probe
U Trim28
Trim28 probe
mNG probe
iRFP670 probe
Amp probe
V Gsk3b
Gsk3b probe
mNG probe
iRFP670 probe
Amp probe
W Cdt1
Cdt1 probe
mNG probe
iRFP670 probe
Amp probe
X Ddx42
Ddx42 probe
mNG probe
iRFP670 probe
Amp probe
Y Eif4a1
Eif4a1 probe
mNG probe
iRFP670 probe
Amp probe

Figure S3. Establishment of knock-in cell lines. The upper part of the panel shows structures of the target gene near the knock-in site, the targeting vector, and the knocked-in genes. The results of Southern blotting are shown at the bottom of the panel. Parent means the parental strain (C57BL/6J, Bruce4), +Amp means a cell line in which the ampicillin resistance gene, which is contained in the backbone of the targeting vector, has been introduced into the genome. Cell lines shown in red letters were used in the downstream experiments. See Table S6 for details of knock-in cell lines used in this study. Source data are available at DOI: http://dx.doi.org/10.17632/5rchtsps3z.1.

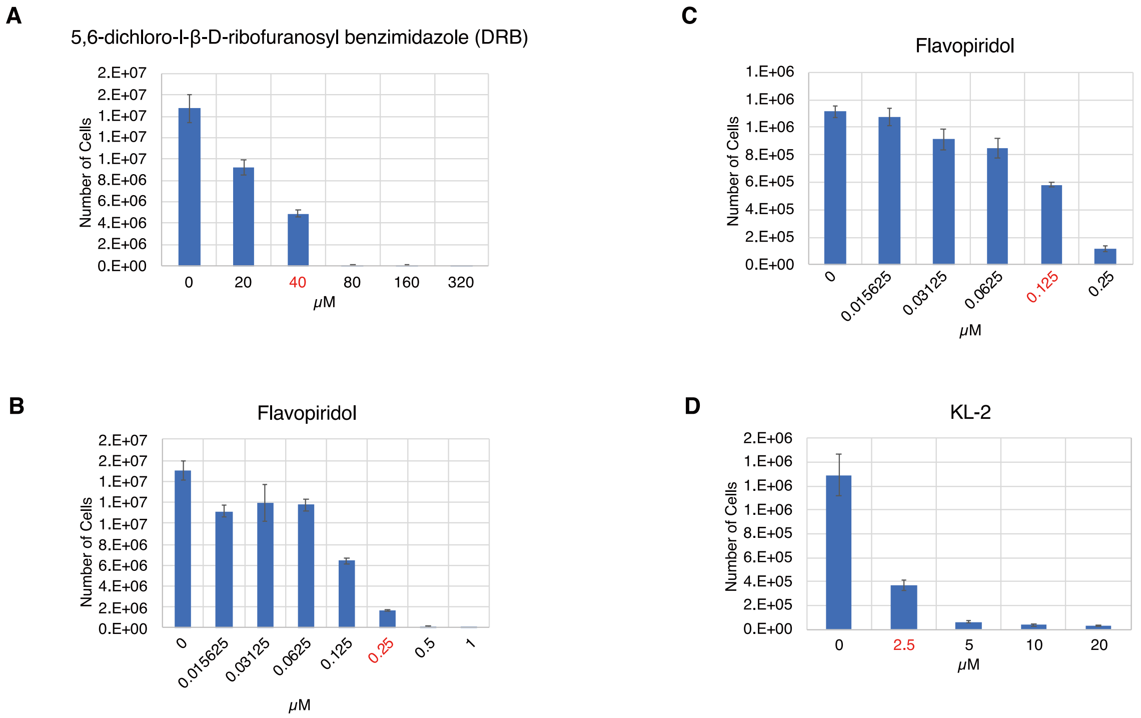

Figure S4. Effects of Pol II pause release inhibitor on cell growth. (A, B) 1 × 10^5 C57BL/6J WT mESCs conditioned to 2i medium were cultured for 2 days in the presence of 5,6-dichloro-l-β-D-ribofuranosyl benzimidazole (DRB) **(A)** or flavopiridol **(B)**, and the number of cells were counted. We decided to use the concentration (highlighted in red letters), at which the growth rate drops sufficiently, and cells were not extinct, in the following experiment. **(C, D)** Effects of Pol II pause release inhibitor on cell growth in mESCs conditioned PD-MK medium. 1 × 10^5 C57BL/6J WT mESCs conditioned to PD-MK medium were cultured for 2 days in the presence of flavopiridol **(C)** or SEC inhibitor KL-2 **(D)**, and the number of cells were counted. We decided to use the concentration (highlighted in red letters), at which the growth rate drops substantially and cells were not extinct, in the following experiment.

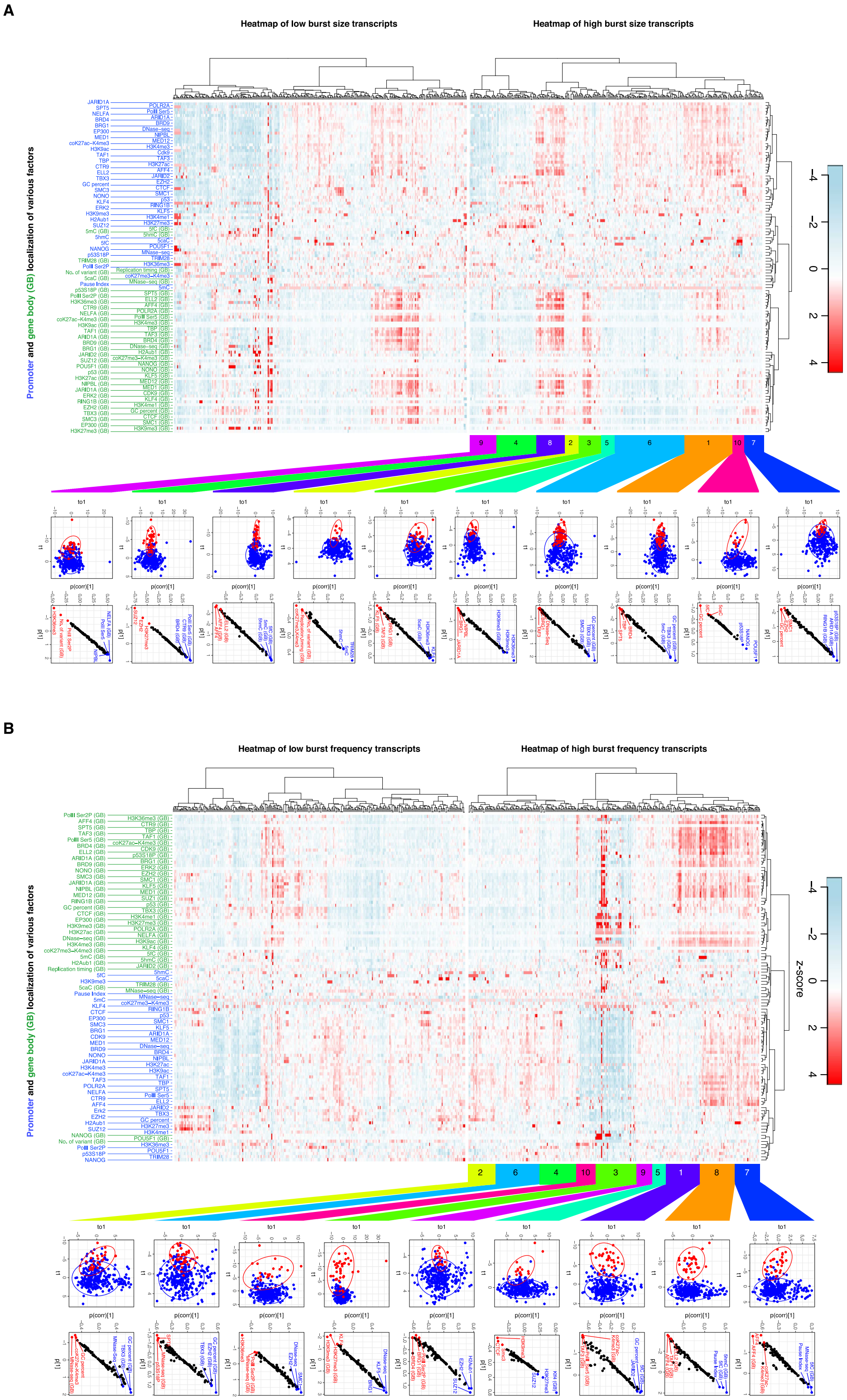
A
Heatmap of low burst size transcripts
Heatmap of high burst size transcripts
Promoter and gene body (GB) localization of various factors
−4
−2
0
2
4
9
4
8
2
3
5
6
1
10
7
to1
t1
p(corr)[1]
p[1]
B
Heatmap of low burst frequency transcripts
Heatmap of high burst frequency transcripts
Promoter and gene body (GB) localization of various factors
z-score
−4
−2
0
2
4
2
6
4
10
3
9
5
1
8
7
to1
t1
p(corr)[1]
p[1]

Figure S5. Burst size and frequency are regulated by combinations of promoter- and gene body-binding factors. (A) Burst size is regulated by combinations of promoter- and gene body-binding factors. First, the target 5,992 transcripts were ranked with the burst size, of which the upper or lower 5% (300 transcripts each) was taken as high and low burst size transcripts, respectively. The upper panel shows a heat map of promoter and gene body (GB) association of various factors in high and low burst size transcripts. The high burst size transcripts were classified into 10 clusters, and each cluster of high burst size transcripts and whole low burst size transcripts were subjected to OPLS-DA modeling. The lower panel represents score plots of OPLS-DA (the first predictive component [t_1] vs. the first orthogonal component [to_1]) and S-plots constructed by presenting the modeled covariance (p[1]) against modeled correlation {p(*corr*)[1]} in the first predictive component. **(B)** Burst frequency is regulated by combinations of promoter- and gene body-binding factors. First, the target 5,992 transcripts were ranked with the burst frequency, of which the upper or lower 5% (300 transcripts each) was taken as high and low burst frequency transcripts, respectively. The upper panel shows a heat map of promoter and GB association of various factors in high and low burst frequency transcripts. The high burst frequency transcripts were classified into 10 clusters, and each cluster of high burst frequency transcripts and whole low burst frequency transcripts were subjected to OPLS-DA modeling. The lower panel represents score plots of OPLS-DA (the first predictive component [t_1] vs. the first orthogonal component [to_1]) and S-plots constructed by presenting the modeled covariance (p[1]) against modeled correlation {p(*corr*)[1]} in the first predictive component.

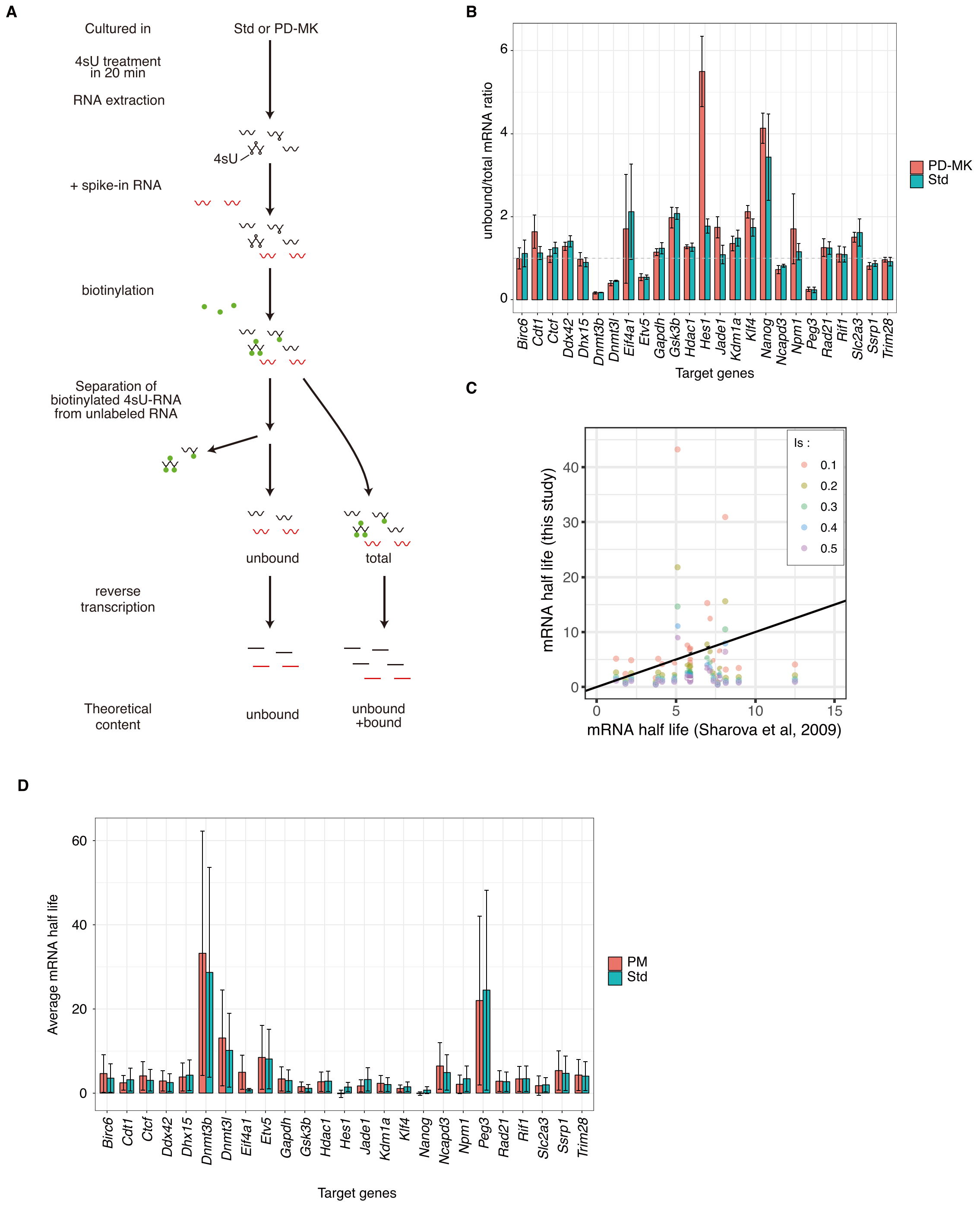

Figure S6. RNA degradation rate between PD-MK and Std conditions did not show significant difference. (A) Schematic representation of the experiment for examining RNA degradation rate. WT mESCs were transiently treated with 4-thiouridine (4sU). By this transient treatment, 4sUs were

incorporated into newly synthesized RNA. RNA was then extracted from the cells, and a known amount of spike-in RNA that does not contain 4sU was added. Thereafter, the RNA mix was biotinylated. Then, from a part of this biotinylated RNA mix, biotinylated RNA was removed using streptavidin beads, and unbiotinylated RNAs that were transcribed before the addition of 4sU were recovered. RNA samples that were not treated with streptavidin beads contained both existing RNA and newly synthesized RNA. Therefore, we refer to these RNA samples as total RNAs. These were reverse transcribed and analyzed by qPCR. **(B)** A bar graph of the ratio of unbound and total RNA. For many samples, the ratio has exceeded 1, which was theoretically impossible. It is considered that the reverse transcription efficiency of 4sU-introduced RNA could be extremely low (see Materials and Methods). **(C)** We assumed that the presence of biotinylated RNA during reverse transcription may trap reverse transcriptase, and that the efficiency of reverse transcription is further reduced globally. We assume that the global suppression effect of reverse transcriptase trapping is I_g (global inhibitory effect). Moreover, the reverse transcription inhibitory effect of biotinylated RNA itself was defined as I_s (see Materials and Methods). In order to determine the appropriate value of I_s, several values were assigned to I_s, and mRNA half-lives in the Std condition were compared with the previously reported mRNA half-lives (*23*). We found that the scaling of mRNA half-lives in the Std condition and that of previously reported mRNA half-lives were getting closer when I_s was 0.1. **(D)** Average mRNA half-life. The half-life of mRNA was calculated using the ratio obtained in **(B)** and the correction formula (see Materials and Methods). In all genes analyzed, there was no significant difference between PD-MK and Std conditions.

Table S1. Allelically normalized read count data of individual transcripts of 129 alleles. Data with intrinsic noise below Poisson noise or transcripts showing interallelic extreme expression level differences are removed (see Materials and Methods).

Table S2. Allelically normalized read count data of individual transcripts of CAST alleles. Data with intrinsic noise below Poisson noise or transcripts showing interallelic extreme expression level differences are removed (see Materials and Methods).

Table S3. smFISH count data of GFP and iRFP knocked-in allele in knock-in cell line.

Table S4. Plasmids information used for knock-in cell line establishment, cell quality and sequencing depth data of scRNA-seq, smFISH probes, sequences of oligos used in this study, and list of data sources used in this study. This Excel file contains data divided into tabs.

The “Data about scRNA-seq” tab contains multiple pieces of information (not only a table but also figures) about “cell quality and sequencing depth data of scRNA-seq”.