Supplemental Information for:

H2B monoubiquitylation is a 5′-enriched active transcription mark and correlates with exon-intron structure in human cells

Inkyung Jung[1,4], Seung-Kyoon Kim[2,4], Mirang Kim[3,4], Yong-Mahn Han[2], Yong Sung Kim[3], Dongsup Kim[1,*], and Daeyoup Lee[2,*]

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea

[2] Department of Biological Sciences, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea

[3]Medical Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, South Korea

[4] These authors contributed equally to this work.

[*]To whom correspondence should be addressed:

Tel: +82-42-350-4317

Fax: +82-42-350-4310

E-mail: kds@kaist.ac.kr

Tel: +82-42-350-2623

Fax: +82-42-350-2610

E-mail: daeyoup@kaist.ac.kr

Supplementary Figures

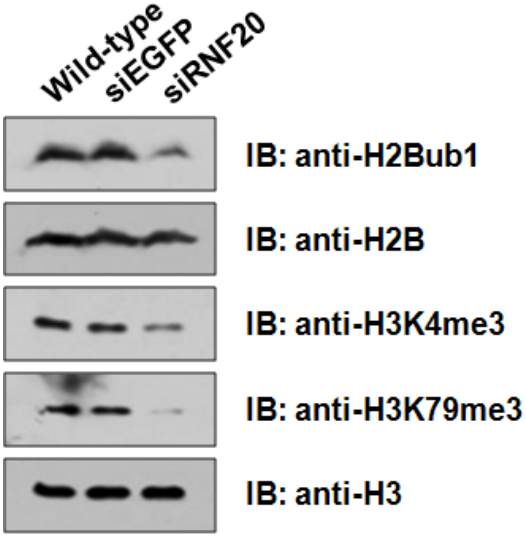
A

PANTHER biological processes	Benjamini P value			
	H2Bub1	H2Bub1-H3K79me1	H2Bub1-H3K79me2	H2Bub1-H3K79me3
Protein modification	1.7×10^{-17}	3.8×10^{-19}	8.8×10^{-16}	5.6×10^{-13}
Nucleoside, nucleotide and nucleic acid metabolism	1.4×10^{-10}	1.4×10^{-13}	2.2×10^{-20}	2.1×10^{-13}
Pre-mRNA processing	1.2×10^{-10}	4.9×10^{-12}	2.4×10^{-14}	2.6×10^{-15}
Protein phosphorylation	1.0×10^{-10}	6.0×10^{-12}	5.5×10^{-9}	1.1×10^{-9}
Intracellular protein traffic	8.9×10^{-11}	1.6×10^{-8}	7.9×10^{-11}	1.2×10^{-8}
Protein metabolism and modification	2.5×10^{-9}	1.4×10^{-11}	1.1×10^{-9}	1.7×10^{-6}
mRNA splicing	8.2×10^{-9}	1.4×10^{-10}	7.8×10^{-12}	2.1×10^{-13}
Cell cycle	2.9×10^{-5}	2.6×10^{-6}	2.2×10^{-7}	1.4×10^{-4}

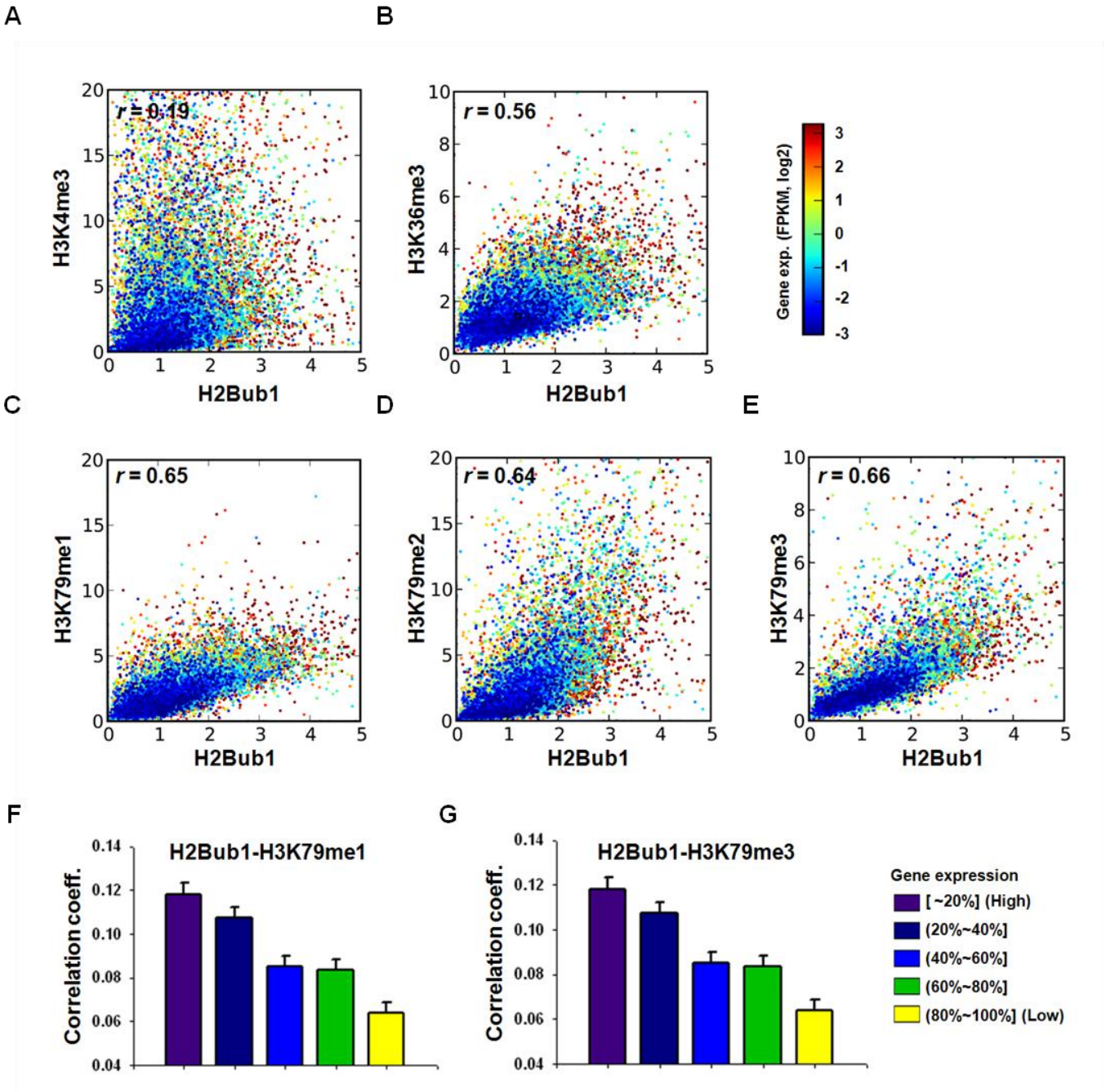
B

GO biological processes enrichment among 2-fold up- or downregulated genes (Benjamini P value)			
2-fold down (4,654)		2-fold up (5,304)	
Cell cycle	1.5×10^{-4}	Neuron development	5.3×10^{-3}
Cell cycle process	2.2×10^{-4}	Negative regulation of macromolecule metabolic process	3.5×10^{-3}
Cell cycle phase	1.4×10^{-3}	Neuron differentiation	3.4×10^{-3}
Phosphorus metabolic process	1.2×10^{-3}	Regulation of transcription from RNA polymerase II promoter	3.3×10^{-3}
Phosphate metabolic process	1.2×10^{-3}	Blood vessel development	2.7×10^{-3}
Protein amino acid phosphorylation	5.0×10^{-3}	Embryonic development ending in birth or egg hatching	2.7×10^{-3}
Intracellular signaling cascade	5.4×10^{-3}	Chordate embryonic development	2.8×10^{-3}
Mitotic cell cycle	4.8×10^{-3}	Blood vessel morphogenesis	2.7×10^{-3}
Cellular protein catabolic process	7.9×10^{-3}	Positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.7×10^{-3}
Proteolysis involved in cellular protein catabolic process	8.2×10^{-3}	Response to extracellular stimulus	2.5×10^{-3}
Cell division	1.0×10^{-2}	Negative regulation of gene expression	2.6×10^{-3}
Modification-dependent macromolecule catabolic process	9.5×10^{-3}	Positive regulation of gene expression	2.5×10^{-3}
Modification-dependent protein catabolic process	9.5×10^{-3}	Vasculature development	2.8×10^{-3}
Protein catabolic process	1.8×10^{-2}	Neuron projection development	2.9×10^{-3}
M phase	1.6×10^{-2}	Positive regulation of nitrogen compound metabolic process	3.1×10^{-3}
Death	1.6×10^{-2}	Cell projection organization	3.0×10^{-3}
Cell death	1.6×10^{-2}	Positive regulation of transcription	3.1×10^{-3}
M phase of mitotic cell cycle	1.7×10^{-2}	Response to nutrient levels	3.6×10^{-3}
Organelle fission	1.7×10^{-2}	Negative regulation of macromolecule biosynthetic process	4.9×10^{-3}
Nuclear division	1.6×10^{-2}	Negative regulation of cellular biosynthetic process	4.9×10^{-3}

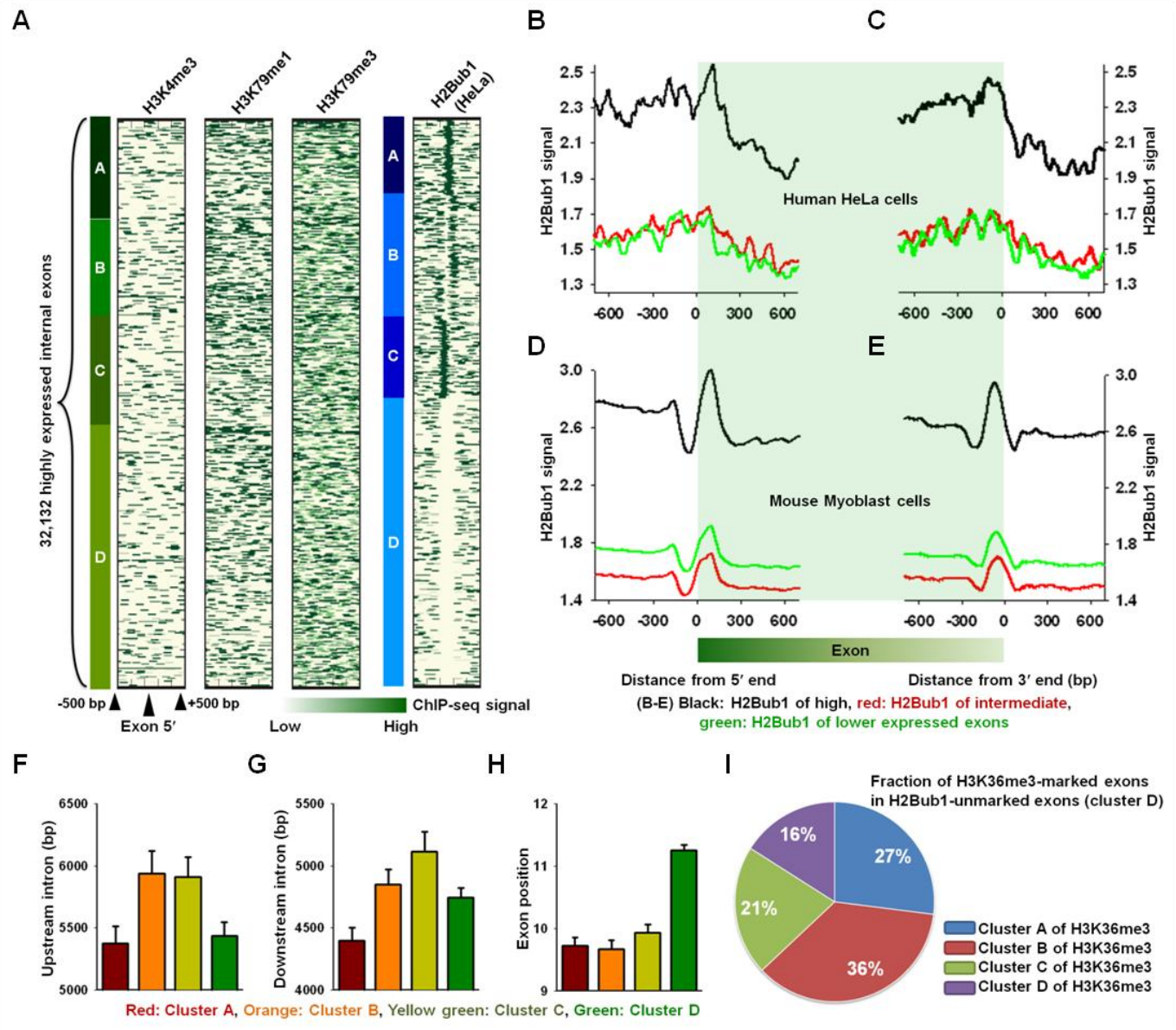
Supplementary Figure 1 Gene set analysis for biological processes. (A) Enrichment of PANTHER biological processes among the modification peak-targeted genes. Among 25,133 genes defined by Refseq in the UCSC Genome Browser, 4,173 genes contained at least one H2Bub1 peak. Among these, 3,362, 3,362, and 2,765 genes were targeted by H3K79me1, me2, and me3, respectively. Gene set analysis was conducted for these groups (H2Bub1 targeted, H2Bub1-H3K79me1 doubly targeted, H2Bub1-H3K79me2 doubly targeted, and H2Bub1-H3K79me3 doubly targeted). Our results revealed that eight of the top 10 enriched PANTHER biological processes in each group were commonly recognized. Protein modification, metabolism and mRNA processing-related genes were commonly enriched. The Benjamini P values for eight biological processes are shown. (B) GO biological process enrichment for genes showing 2-fold up- or downregulation following the siRNA-mediated knock-down of *RNF20*. In siRNF20-transfected NCCIT cells, 4,654 and 5,304 genes were down- and upregulated, respectively, more than 2-fold compared to wild-type cells. The top 20 enriched GO biological processes are shown for each group, along with p values.



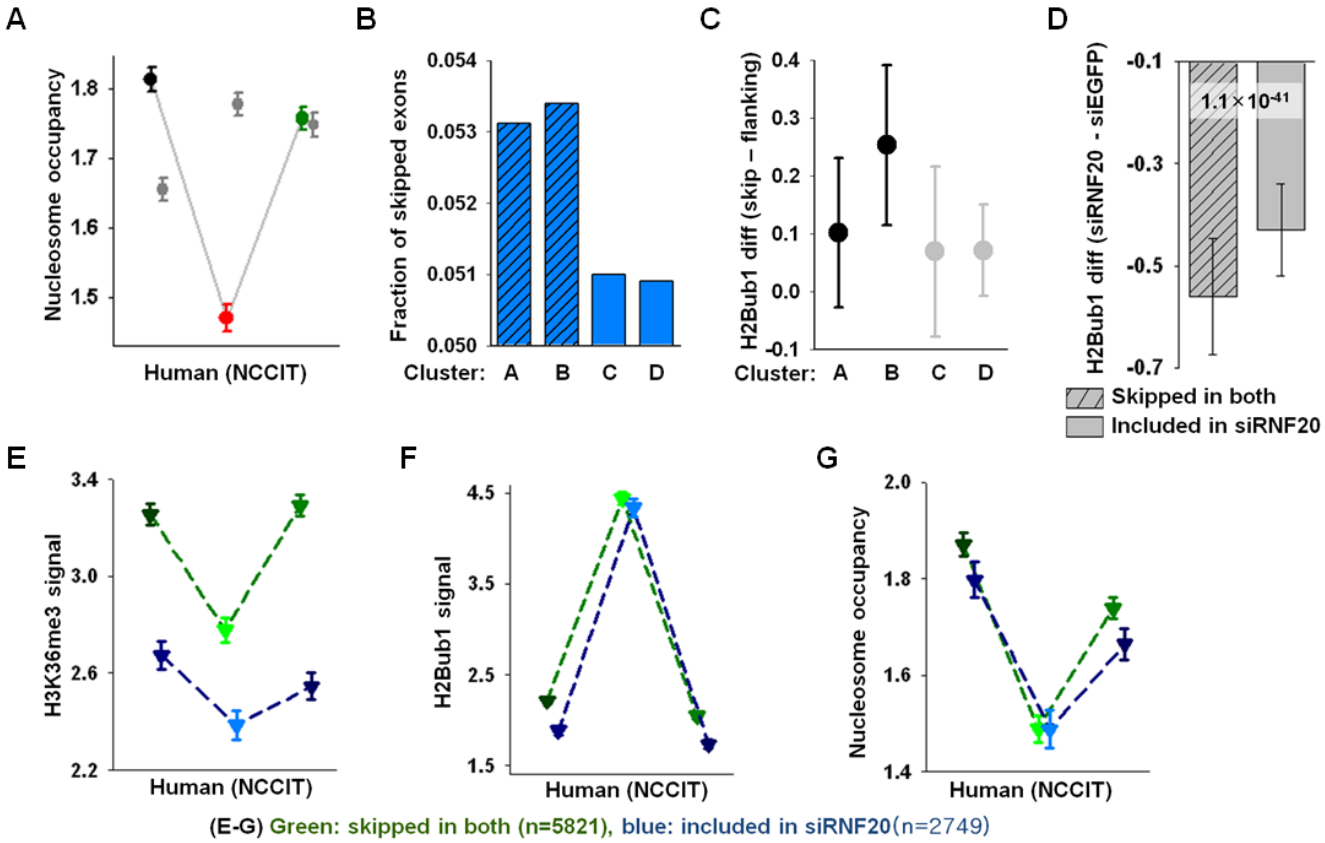
Supplementary Figure 2 Effects of siRNA-mediated knockdown of *RNF20* (human *BRE1*) on histone modifications. EGFP- and *RNF20*-siRNA (60 nM each) were transiently transfected into NCCIT cells. Whole-cell extracts were analyzed by SDS-PAGE and immunoblot (IB) analysis was performed using the indicated antibodies. Histone H3 was used as a loading control.



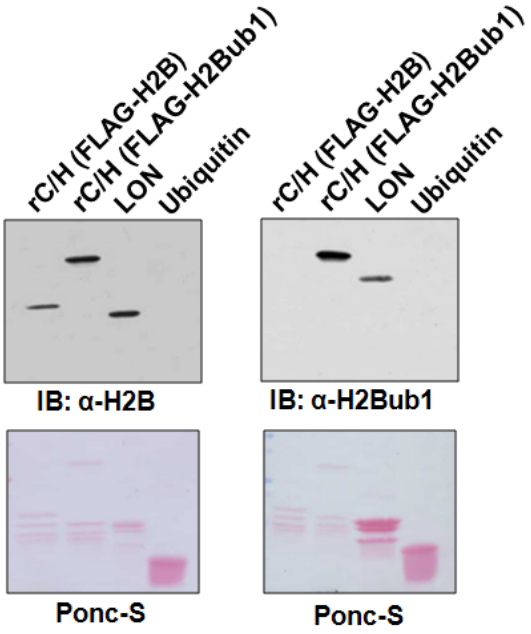
Supplementary Figure 3 Genome-wide co-occurrence patterns between H2Bub1 and other modifications in gene body regions. (A-E) Each spot indicates the average ChIP-seq signal in the gene body region of each gene, and its color indicates the gene expression level defined by Cufflinks. The Pearson correlation coefficients between H2Bub1 and the other modifications are shown (r =). H3K79 methylations were strongly correlated with H2Bub1, but H3K4me3 was not (r = 0.19). (F and G) Average Pearson correlation coefficients between H2Bub1 and H3K79 methylations are shown according to gene expression levels. Correlation coefficients were calculated using the ChIP-seq signals of intronic and exonic regions. Genes were equally divided into five groups according to expression levels (purple for highly expressed genes and yellow for lower expressed genes). Highly expressed genes tended to have stronger correlation coefficients between H2Bub1 and H3K79 methylations compared to the lower expressed genes.



Supplementary Figure 4 Modification patterns at the exon boundaries and the cluster-specific properties of H2Bub1. (A) The modification patterns at the exon boundaries (500 bp upstream and downstream of the 5′ end of exons) were sorted based on the clustered results described in Figure 3A. H3K4me3, H3K79me1, and H3K79me3 do not show enriched signals at the exon boundaries. H2Bub1 was enriched at the exon boundaries in HeLa cells. (B-E) The average H2Bub1 signals at exon boundaries are shown according to exon expression levels for human HeLa cells and mouse myoblast cells. For highly expressed exons (top 30% of exons, black lines in B-E), H2Bub1 was strongly enriched at both the 5′ and 3′ ends of the exon boundaries. For intermediately (30–70% of exons, red lines in B-E) or lower (bottom 30%, green lines in B-E) expressed exons, no specific patterns were observed. (F and G) H2Bub1 cluster-dependent properties are shown. The y-axis indicates the average values with standard deviation error bars. (H) In terms of exon position, H2Bub1 cluster D tended to contain more exons near the 3′ end. The y-axis indicates the average values of the exon positions with standard deviation error bars. (I) Fraction of matched exons between H2Bub1 cluster D and each H3K36me3 cluster showed that 84% of the exons in H2Bub1 cluster D (H2Bub1 depleted) were marked by H3K36me3.



Supplementary Figure 5 Skipped exon-specific modification patterns. (A) Nucleosome occupancy in the skipped exons and their flanking exons as described in Figure 4A-F. (B) Fraction of skipped exons for each cluster defined in Figure 3A. (C) Differences of H2Bub1 signals between skipped exons and their flanking exons. The y-axis indicates the average value of H2Bub1 signals with standard deviation error bars. (D) H2Bub1 differences among persistently skipped exons and included exons in siEGFP-transfected and siRNF20-transfected NCCIT cells. P value was calculated using the KS-test. (E-G) Modification signals in wild-type cells for persistently skipped exons (green) and wild-type specific skipped exons (blue).



Supplementary Figure 6 The specificities of the anti-H2B and -H2Bub1 antibodies. Immunoblot (IB) analysis with anti-H2B and -H2Bub1 antibodies on rC/H (FLAG-H2B) (recombinant core histone; FLAG tagged H2B), rC/H (FLAG-H2Bub1) (recombinant core histone; FLAG tagged H2Bub1), LON (long oligonucleosomes), and ubiquitin. Ponc-S indicates the results from Ponceau S staining. The purification of histone molecules and the reconstitution of histone octamers were performed as previously described (Oh et al. 2010). Reconstituted core histones were subjected to Western blot analyses with the indicated antibodies.

Supplementary Methods

Modification peak detection. For peak detection, we used MACS-1.3.7.1 (Zhang et al. 2008) running under the default parameters with two exceptions: the –m-fold option was changed to five and the –pvalue option was changed to 1.0×10^{-4}. The results from the no treatment input sequencing (IgG) was used as a control. The numbers of detected peaks were as follows: H2Bub1 (14,679), H3K79me1 (52,721), H3K79me2 (63,656), H3K79me3 (25,232), H3K4me3 (28,911), and H3Ac (54,660). Overlapping peaks were counted if the middle of the H2Bub1 peak was located within 2 kb upstream or downstream of the other modification peaks. For the modification peaks of mouse myoblasts cells we downloaded pre-calculated detected peaks using MACS from ArrayExpress with accession number E-GEOD-25308 (Asp et al. 2011).

***K*-means clustering in the regions surrounding transcriptional start sites.** Diverse modification patterns in the regions surrounding the transcriptional start sites (TSS) were recognized by utilizing the *k*-means clustering method. Genes were first divided according to their gene expression levels: highly expressed (top 30%), lower expressed (bottom 30%), and intermediately expressed (middle 40%). Each group was then clustered according to their H2Bub1 patterns within 5 kb upstream and downstream of the TSS. We performed *k*-means clustering for each group with Euclidian distance, which avoided the generation of gene expression level-dependent clusters. Since *k*-means clustering generates clusters according to pre-defined cluster numbers, different numbers of clusters were tested (from two to 10). Three distinct patterns were consistently recognized. Other modification patterns are presented according to the clustering results of H2Bub1.

The effect of gene expression levels on the correlation coefficients between H2Bub1 and H3K79 methylations. The correlation patterns between H3K79 methylations and H2Bub1 affected gene activation. Pearson correlation coefficients were calculated between H3K79 methylations and H2Bub1 within each gene containing at least three exons (19,240 genes). For every gene, each modification signal was calculated by units of exons and introns; therefore, the modification pattern of each gene could be presented as 2*n*-1 *by* 1 vector, where *n* represents the number of exons. These vectors were then used to calculate Pearson correlation coefficients for the modification patterns. All genes were then equally divided into five groups according to their expression levels, and the average Pearson correlation coefficient between modifications was calculated for each group.

***K*-means clustering in the regions surrounding the intron-exon boundaries.** The *k*-means method was used to determine clustering modification patterns for H2Bub1 and H3K36me3 in the regions surrounding 500 bp upstream and downstream of the intron-exon boundaries. From among all exons, we selected 107,107 internal exons separated by more than 500 bp of flanking exons. The exons were divided into three groups according to their expression levels: highly expressed exons (top 30%), lower expressed exons (bottom 30%), and intermediately expressed exons (middle 40%). From these we selected exons that demonstrated less than one standard deviation between 110 bp upstream and downstream of the 5′ end of the exon. These exons showed very low H2Bub1 signals and could create errors during *k*-means clustering because the distance option was set as a correlation coefficient. These exons were therefore defined as the modification-depleted cluster (Cluster D). During trials performed using the different pre-defined cluster numbers, three exon expression level independent patterns were consistently recognized. The H3K4me3 and H3K79 methylation patterns were examined according to the clustering results of H2Bub1, in order to investigate possible crosstalk between H2Bub1 and H3 methylations in the enrichment of exon boundaries.

H2Bub1 signal normalization by nucleosome occupancy. The H2Bub1 signal was normalized with respect to nucleosome occupancy, in order to remove the effects of nucleosome enrichment at the exon boundaries. A relative signal was defined by dividing the (H2Bub1 signal+λ) by (nucleosome occupancy+λ) and taking log2, where λ, which was set to one, was a pseudo-count value used to avoid zero division. After the H2Bub1 signal was normalized with respect to nucleosome occupancy, H2Bub1 enrichment was still observed at the exon boundaries.

Calculating nucleosome occupancy in the modification peaked regions. The modification peaked regions were defined using the MACS software, and the average nucleosome occupancy was calculated for each peak as the ratio of target read count/target size divided by the total read count/genome size.

REFERENCES

Asp, P., R. Blum, V. Vethantham, F. Parisi, M. Micsinai, J. Cheng, C. Bowman, Y. Kluger, and B.D. Dynlacht. 2011. PNAS Plus: Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proc Natl Acad Sci U S A* **108:** E149-158.

Oh, S., K. Jeong, H. Kim, C.S. Kwon, and D. Lee. 2010. A lysine rich region in Dot1p is crucial for direct interaction with H2B ubiquitylation and high level methylation of H3K79. *Biochem Biophys Res Commun* **399:** 512-517.

Zhang, Y., T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, and X.S. Liu. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9:** R137.