**Supplementary Information**

*** Supplementary Methods**

*** Supplementary Figures**

*** Inventory of Supplementary Tables**

## Supplementary Methods

### Quantitative reverse transcription PCR (qRT-PCR)

Total RNAs were extracted from mouse nucleated BM cells with Rneasy Mini kit (QIAGEN), and subject to reverse transcriptase reaction using ReverTra Ace qPCR RT Master Mix with gDNA Remover (TOYOBO). cDNA was amplified with SYBR Premix Ex Taq II (TaKaRa) using LightCycler480 system (Roche Diagnostics). The expression of 18s rRNA was used for normalization of the results. The primer sets for qPCR are as follows: 5'-TGGAAGATGATGAAGAGCCAAT-3' and 5'-TCGGGCTTCAGTTCTGTTCT-3' for *Stag2*, 5'-GGACACCGTAATTTCCCTTTTG-3' and 5'-TCATTGGCTCTCTTCCCAATC-3' for *Stag1*, 5'-GCTCTGGTAAGTCAAATCTCATGGA-3' and 5'-CCCTCAGGTTGCTGGTCTTTT-3' for *Smc1*, 5'-GCTGGCGGGCAACAGTGAAC-3' and 5'-AGCCAACCTCGCAATTCCTCGC-3' for *Smc3*, 5'-CCTCAGCAGGTAGAGCAAATGG-3' and 5'-GCATCTGCTGAGTGCGTTTGTT-3' for *Rad21*, and 5'-GCAATTATTCCCCATGAACG-3' and 5'-GGGACTTAATCAACGCAAGC-3' for 18s.

### Western blotting

Nucleated BM cells were isolated using a density gradient solution (Histopaque-1077, Sigma-Aldrich) and were lysed in RIPA lysis buffer. SDS-PAGE and western blotting were performed following the standard protocol. Antibodies used are as follows: SMC1 (Abcam, ab9262), SMC3 (Abcam, ab9263), RAD21 (Abcam, ab992), STAG2 (Novus, NBP1-30472 (for mouse), or Santa Cruz, sc-81852 (for human)), and RUNX1 (Active Motif, 39000 (for mouse), or Santa Cruz, sc-365644 (for human)), and Actin (Santa Cruz, sc1616).

### Co-immunoprecipitation

Mouse 32Dcl3 cell line or human K562 cell line were used for co-immunoprecipitation analysis as previously described (1) with minor modifications. Nuclei were prepared using NE-PER Nuclear and Cytoplasmic Extraction Kit (Thermo Scientific) according to the manufacturer's protocol. Immunoprecipitation was performed using NHS Mag Sepharose (GE Healthcare) magnetic beads conjugated with SMC1 (Abcam, ab9262) or SMC3 (Abcam, ab9263) antibody and incubated with the cell extracts overnight at 4℃. After washing with lysis buffer, the beads were suspended with SDS-PAGE sample buffer.

**Histology and cytology**

For histological analysis, tissue samples were fixed in 4% paraformaldehyde, embedded in paraffin, sectioned and stained with hematoxylin and eosin. For cytological analysis, cytospin preparations of BM samples (Thermo Scientific Cytospin 4) or PB smears were stained using the May–Grünwald–Giemsa staining method.

**Immunostaining**

Purified mouse c-Kit$^+$ HSPCs and K562 cell lines were transferred onto Poly-D-Lysine coated cover glass and fixed for 20 min in 4% formaldehyde/PBS. Cells were then permeabilized in 0.5% Triton X-100/PBS for 10 min and incubated for 30 min in 5% skim milk/PBST (0.1% Tween-20 in PBS) for blocking. For STAG2 and RUNX1 staining, cells were incubated overnight at 4℃ at an antibody dilution of 1:50 for mouse monoclonal anti-STAG2 (Santa Cruz, sc-81852) and 1:200 for rabbit monoclonal anti-Runx1 (Abcam, ab92336). Subsequent staining with Alexa Fluor 488 donkey anti-mouse IgG (H+L) (Invitrogen A-21202) and Alexa Fluor 594 goat anti-rabbit IgG (H+L) (Invitrogen A-11037) was performed for 60 min at room temperature at a dilution of 1:1000. Stained cells were treated in DAPI solution (1μg/ml) for 30 min and were mounted with ProLong Gold antifade reagents (Invitrogen).

**Microscopy and data analysis**

Super-resolution images were obtained using LSM880 Airy scan (Zeiss) with a 100x oil objective lens (NA 1.46, alpha Plan-Apochromat 100x/1.46 Oil Ph3 M27). For colocalization analysis, random 4 μm x 4 μm squares in DAPI positive regions were cropped from central 5 images in z-stack images. Spots segmentation was performed using auto local threshold (MidGrey method). For quantification of the random colocalization as negative control, each square image was flipped horizontally and vertically. These steps were performed using ImageJ. Appropriate sample size was checked by G*Power 3.1 (2,3).

**Single-cell differentiation assay**

Single-cell differentiation assay was performed as previously described (4) with minor modifications. c-Kit$^+$ HSPCs were transduced by FLAG-tagged Hoxa9 or mock in the pGCDNsam-IRES-EGFP vector, and GFP-positive cells were sorted at one cell per well into a 96-well plate of which each well contains IMDM with 10% FBS, 2-β-mercaptoethanol, 10 ng/ml mouse SCF, TPO and IL-3, and 40 ng/ml human EPO. After 14 day-culture, each generated colony was subjected to FACS analysis and was classified to granulocyte-, monocyte-, and/or erythroid-containing colony if it contained > 10% of corresponding cells.

**RNA-sequencing**

RNA was extracted using RNeasy Mini Kit (QIAGEN) or NucleoSpin RNA XS (Macherey-Nagel). Libraries for RNA-seq were prepared using the NEBNext Ultra RNA Library Prep kit for Illumina (New England BioLabs) and were subjected to sequencing using HiSeq 2500 or NovaSeq 6000 instrument (Illumina) with a standard 100-150-bp paired-end protocol as previously described (5). RNA-seq experiments were performed in two or more biological replicates. The sequencing reads were aligned to the reference genome (hg19 or mm9) using STAR (v2.5.3) (6). Reads on each refSeq gene were counted with featureCounts (v1.5.3) (7) from Subread package, and edgeR package in R (8) was used to identify the differentially expressed genes with FDR threshold of 0.05 and to generate the multidimensional scaling (MDS) plot. The analysis was performed in genes expressed at >1 count per million (CPM) in two or more samples, and generalized linear models were used to compare gene expression data. Differentially expressed genes between WT- and SKO/RKO/DKO-transplanted LSK cells (FDR < 0.05) were grouped into 5 clusters using k-means clustering. Motif analysis was performed using the HOMER findMotifs.pl program (9). For the gene promoters, enrichment of known transcription factor motifs was analyzed from -2,000 to +1,000 bp from the transcription start site (TSS), and genes without significant expression changes were used as backgrounds. Gene ontology (GO) analysis was performed using Database for Annotation, Visualization, and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov). MSigDB overlap analysis was performed using MSigDB database and hallmark gene sets (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp). Tissue Specific Expression Analysis (TSEA) was performed with TSEA tool (http://genetics.wustl.edu/jdlab/tsea/) (10). RNA expression analysis in the hematopoietic system was carried out using Haemopedia RNA-seq datasets (11), and averages of log2 (TPM + 1) values of each gene set were calculated for each cell type. GSEA (v2.2.4) (12) was used to determine the sets of genes that are significantly different between groups. Gene sets characteristic of GMP and B-lymphocytes were generated using datasets from previous reports (13,14). For network analysis, GSEA analysis was performed with gene sets downloaded from https://www.baderlab.org/Software/EnrichmentMap (Human_GOBP_AllPathways_no_GO_iea_March_01_2018_symbol.gmt file). Enriched pathways were visualized using Enrichment Map (15) with the q-value < 0.05 and overlap similarity coefficient parameters > 0.5. For RNA-seq of human MDS/AML (16-18), reads aligned to the human reference genome (hg19) or counted read data were obtained and analyzed as described above. Human AML cases (17) were grouped into 2 clusters using genes with high pausing levels (pausing index >20) and k-means clustering. RNA-seq datasets used are described in **Supplementary Table S7**. Differentially expressed genes shown in **Fig. 2I** are described in **Supplementary Tables S8-9**. Differentially expressed genes shown in **Fig. 3J** are described in **Supplementary Tables S10-15**.

**ChIP-sequencing**

ChIP-seq experiments were performed using c-Kit$^+$ HSPCs or HL-60 cell lines. Cells were fixed in PBS with 1% formaldehyde (Thermo Fisher Scientific) for 10 min at room temperature with gentle mixing. The reaction was stopped by adding glycine solution (10x) (Cell Signaling Technology) and incubating for 5 minutes at room temperature, and the cells were washed in cold PBS twice. The cells were then processed with SimpleChIP Plus Sonication Chromatin IP Kit (Cell Signaling Technology) and Covaris E220 (Covaris) according to the manufacturer's protocol. The antibodies used for ChIP are as follows: STAG1 (Protein Tech, 14015-1-AP), STAG2 (Novus, NBP1-30472), SMC1 (Abcam, ab9262), CTCF (Cell Signaling Technology, D31H2), RUNX1 (Abcam, 23980), total Pol II (CST, D8L4Y), Ser5-P Pol II (Abcam, ab5408), H3K27ac (Cell Signaling Technology, D5E4), H3K27me3 (Cell Signaling Technology, C36B11), H3K4me1 (Cell Signaling Technology, D1A9), or H3K4me3 (Cell Signaling Technology, C42D8). After purification of ChIPed DNA, ChIP-seq libraries were constructed using ThruPLEX DNA-seq kit (Takara) according to the manufacturer's protocol, and then subjected to sequencing using HiSeq 2500 or NoveSeq 6000 (Illumina). ChIP-seq experiments were performed in two or more biological replicates with input controls. The sequencing reads were aligned to the reference genome (hg19 or mm9) using bowtie (v1.2.2) (19) following trimming of adapters and read tails to a total length of 50 base pairs using cutadapt. Duplicates and reads on blacklisted regions (ENCODE) were removed by Picard and bedtools, respectively. Peaks were called using MACS (v2.1.1) for each replicate individually with a *P*-value threshold of $1 \times 10^{-3}$ unless otherwise specified, and overlapped peaks among replicates were regarded as consensus peak sets. Motif analysis and peak annotation were performed with HOMER. Super-enhancers were identified with H3K27ac ChIP-seq data in WT HSPCs using ROSE (20) with default parameters. Identified super-enhancers are described in **Supplementary Table S20**. Super-enhancers in human HSCs were previously described (21), and we used the dataset of BI_CD34_Primary_RO01536 for assignment of super-enhancer-associated genes. Calculation of ChIP signal intensities around peaks and generation of read density profile plots and heatmaps were performed using deeptools (22). In metaplot analysis, statistical significance was assessed with one-sided Wilcoxon rank-sum test at each bin. Visualization of sequence data was performed using IGV. For clustering of cohesin binding sites, we calculated logarithm of H3K27ac and Ctcf ChIP signal intensities summed up around ± 200 bp from centers of cohesin binding sites (Stag1 and/or Stag2 peaks) by deeptools, performed clustering using flowPeaks (23), and regarded the H3K27ac high clusters as cohesin cluster II and the others as cluster I. Binding profiles of cohesin components, Pol II, Mediator, and ten hematopoietic TFs were similarly calculated around ± 200 bp from centers of cohesin binding sites (**Fig. 5K**). For analysis of combinatorial binding of ten transcription factors (Asxl1, Fli1, Gata2, Gfi1b, Lmo2, Lyl1, Meis1, Pu1, Runx1, and Scl) (**Fig. 5L**), each peak was called using MACS (v1.4.2) with a *P*-value threshold of $1 \times 10^{-5}$, and number of transcription factors whose peaks overlapped with regions ± 500 bp around CC-II sites annotated to each group gene was counted.

We also used ChIP-seq datasets (10 TFs: Asxl1, Fli1, Gata2, Gfi1b, Lmo2, Lyl1, Meis1, Pu1, Runx1, Scl; Pol II, Med12 in mouse, Pol II in human (24-27)) in previous studies.

**Hi-C**

Hi-C experiments were performed using MboI restriction enzyme as previously described (28). Briefly, two million mouse c-Kit$^+$ HSPCs or HL-60 cells were crosslinked with 1% formaldehyde for 10 min at room temperature. Cells were permeabilized and chromatin was digested with MboI restriction enzyme, and the ends of restriction fragments were labeled with biotinylated nucleotides and ligated. After crosslink reversal, DNA was purified and sheared with Covaris M220 (Covaris). Then point ligation junctions were pulled down with streptavidin beads. Then libraries were constructed with Nextera Mate Pair Sample Preparation Kit (Illumina) according to the manufacturer's protocol, and subject to sequencing using NovaSeq 6000 (Illumina) with a standard 100- or 150-bp paired-end protocol. Hi-C experiments were performed in biological duplicates. The sequencing reads were processed using Juicer (28) and hg19 or mm10 reference genome. After filtering of reads, the average valid interactions per genotype resulted in 1.79 billion for mouse HSPCs and 1.66 billion for HL-60 cells. For comparative analysis, the valid interactions after filtering were randomly resampled and arranged in the number of the lowest sample. Contact matrices used for further analysis were created for each replicate as well as merged one by genotype and Knight-Ruiz (KR)-normalized with Juicer. Genomic compartmentalization (A or B compartments) was analyzed using Eigenvector (28) at 25kb resolution, and A-compartments were assigned to the genomic bin with positive eigenvector values as well as higher gene density and B-compartments were the opposite. The insulation score was calculated as previously described (29) at 5kb resolution, and visualized by deeptools. Loops were called at 5kb and 10kb resolutions using HICCUPS (28) and then merged to construct loop sets. Loops were classified into CC-I loops (whose anchors overlapped with at least one CC-I sites but not with CC-II) and CC-II loops (whose anchors overlapped with at least one CC-II but not CC-I sites). Loops whose anchors corresponded to the pairs of Ctcf (cohesin cluster-I sites), enhancers (H3K4me1 peaks overlapped with H3K27ac peaks in mouse or H3K27ac peaks excluding peaks overlapped with TSSs (± 2 kb) in human), and promoters (TSSs overlapped with H3K4me3 peaks) were counted, and plotted by igraph package in R software. Aggregated intensities of "peaks" (pixels corresponding to pairs of loop anchors in the contact matrices) were calculated using aggregate peak analysis (APA) (28) with -r 5000 -n 15 parameters, which calculates the sum of a series of submatrices around peaks derived from the contact matrix. Each of these submatrices is a pixel square centered at a single peak in the upper triangle of the contact matrix. Topologically associating domains (TADs) were called at 5kb resolution using Arrowhead (28). TAD boundaries were defined as +- 5kb from the 5'- or 3'- ends of TADs, and insides were regions insides of both boundaries. For aggregated TAD analysis, we selected TADs which did not enclose other TADs, and

were located in compartment A and in the size range 100-300 kb, got submatrices corresponding to TAD regions derived from the contact matrix, resized each of them into a 100 x 100 submatrix, and calculated the sum of size-normalized submatrices. We also performed hierarchical TADs analysis using rGMAP (30) at 5kb resolution with dom_order = 3 parameter, which identifies hierarchical TADs structures such as TADs (level 1) and sub-TADs (level 2/3), and performed aggregated TAD analysis separately according to TAD levels as described above without any additional filters to select TADs. Hi-C contact matrices were visualized by Juicebox (28) or HiCExplorer (31). Annotations on the mm9 reference genome were converted to those on mm10 and vice versa using Lift Genome Annotations (UCSC).

**Splicing analysis**

RNA-seq datasets in *Sf3b1* K700E mutant cells (GSE85712) and *Srsf2* P95H mutant cells (DRA006224) were previously described (32,33). We took annotation-free approach for alternative splicing analysis using JUM (34). Junctions that have more than 5 reads in 5 (for *Srsf2*) or 2 (for others) replicates of one condition were filtered for downstream analysis. According to inclusion vs exclusion criteria (shown in **Supplementary Fig. S15F**), percent spliced in (PSI) values were adjusted for each junction using a custom script. For alternative first exon (AFE), alternative last exon (ALE), and tandem UTR events, PSI values were calculated using MISO (35). Differential PSI was assessed using moderated t-test and Benjamini-Hochberg correction. For splicing events except for composite events, minimum q-value for each event was considered as a representative statistics. Events which passed a q-value threshold of 0.20 were considered as altered splicing events.

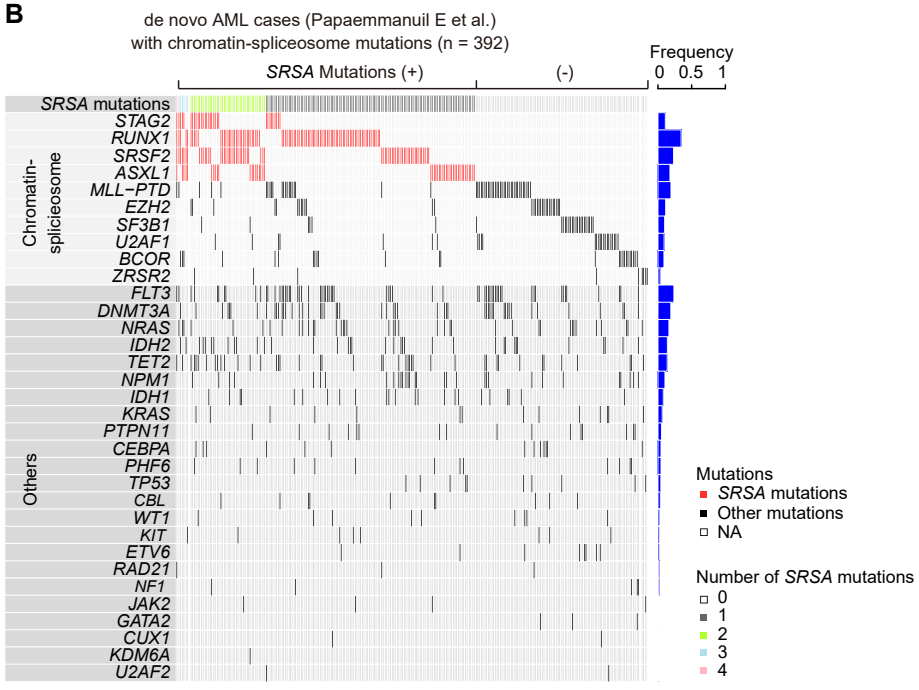
**Supplementary References**


1. Deardorff MA, Bando M, Nakato R, Watrin E, Itoh T, Minamino M*, et al.* HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle. Nature **2012**;489:313-7.
2. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods **2009**;41:1149-60.
3. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods **2007**;39:175-91.
4. Ema H, Morita Y, Yamazaki S, Matsubara A, Seita J, Tadokoro Y*, et al.* Adult mouse hematopoietic stem cells: purification and single-cell assays. Nat Protoc **2006**;1:2979-87.
5. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R*, et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. Nature **2011**;478:64-9.
6. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics **2013**;29:15-21.

7. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **2014**;30:923-30.
8. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **2010**;26:139-40.
9. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P*, et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell **2010**;38:576-89.
10. Dougherty JD, Schmidt EF, Nakajima M, Heintz N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. Nucleic Acids Res **2010**;38:4218-30.
11. Choi J, Baldwin TM, Wong M, Bolden JE, Fairfax KA, Lucas EC*, et al.* Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans. Nucleic Acids Res **2019**;47:D780-D5.
12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **2005**;102:15545-50.
13. Jojic V, Shay T, Sylvia K, Zuk O, Sun X, Kang J*, et al.* Identification of transcriptional regulators in the mouse immune system. Nat Immunol **2013**;14:633-43.
14. Mullenders J, Aranda-Orgilles B, Lhoumaud P, Keller M, Pae J, Wang K*, et al.* Cohesin loss alters adult hematopoietic stem cell homeostasis, leading to myeloproliferative neoplasms. J Exp Med **2015**;212:1833-50.
15. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One **2010**;5:e13984.
16. Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, Robertson A*, et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med **2013**;368:2059-74.
17. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL*, et al.* Functional genomic landscape of acute myeloid leukaemia. Nature **2018**;562:526-31.
18. Shiozawa Y, Malcovati L, Galli A, Pellagatti A, Karimi M, Sato-Otsubo A*, et al.* Gene expression and risk of leukemic transformation in myelodysplasia. Blood **2017**;130:2642-53.
19. Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics **2010**;Chapter 11:Unit 11 7.
20. Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR*, et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell **2013**;153:320-34.
21. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA*, et al.* Super-enhancers in the control of cell identity and disease. Cell **2013**;155:934-47.

22. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res **2014**;42:W187-91.
23. Ge Y, Sealfon SC. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Bioinformatics **2012**;28:2052-8.
24. Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P*, et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell **2010**;7:532-44.
25. Li Z, Zhang P, Yan A, Guo Z, Ban Y, Li J*, et al.* ASXL1 interacts with the cohesin complex to maintain chromatid separation and gene expression for normal hematopoiesis. Sci Adv **2017**;3:e1601602.
26. Aranda-Orgilles B, Saldana-Meyer R, Wang E, Trompouki E, Fassl A, Lau S*, et al.* MED12 Regulates HSC-Specific Enhancers Independently of Mediator Kinase Activity to Control Hematopoiesis. Cell Stem Cell **2016**;19:784-99.
27. Cui K, Zang C, Roh TY, Schones DE, Childs RW, Peng W*, et al.* Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell **2009**;4:80-93.
28. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT*, et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **2014**;159:1665-80.
29. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ*, et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature **2015**;523:240-4.
30. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. Nat Commun **2017**;8:535.
31. Ramirez F, Bhardwaj V, Arrigoni L, Lam KC, Gruning BA, Villaveces J*, et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun **2018**;9:189.
32. Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ*, et al.* Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. Cancer Cell **2016**;30:404-17.
33. Kon A, Yamazaki S, Nannya Y, Kataoka K, Ota Y, Nakagawa MM*, et al.* Physiological Srsf2 P95H expression causes impaired hematopoietic stem cell functions and aberrant RNA splicing in mice. Blood **2018**;131:621-35.
34. Wang Q, Rio DC. JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. Proc Natl Acad Sci U S A **2018**;115:E8181-E90.
35. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods **2010**;7:1009-15.

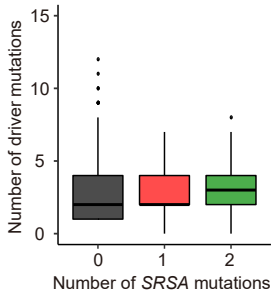
**Supplementary Figure 1** **Ochi et al.**

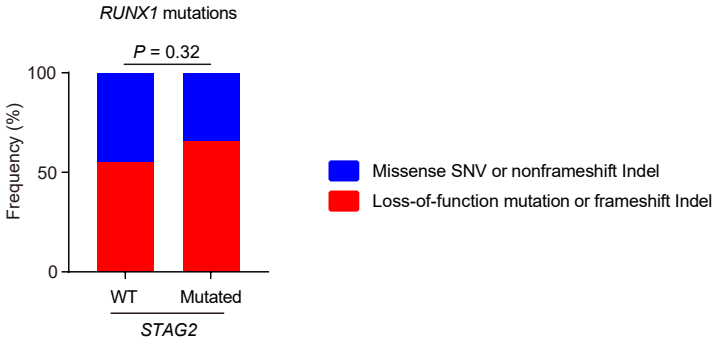
**Supplementary Figure 1. Characteristics of human MDS/AML cases with '*SRSA*' mutations (*STAG2*, *RUNX1*, *SRSF2*, and/or *ASXL1*).**

**A**, Mutational profile of MDS/AML cases with *SRSA* mutations. **B**, Mutational profile of de novo AML cases (Papaemmanuil et al., 2016) with chromatin-spliceosome mutations (*STAG2*, *RUNX1*, *SRSF2*, *ASXL1*, *EZH2*, *SF3B1*, *U2AF1*, *BCOR*, *ZRSR2* mutations, or *MLL*-partial tandem duplication (PTD)).

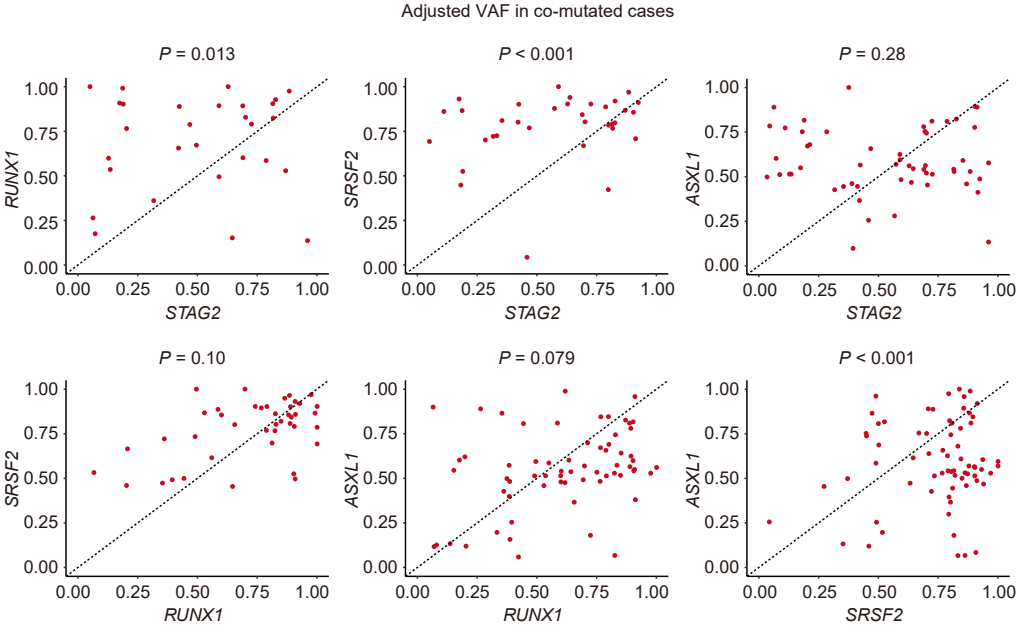
**A**


Number of driver mutations
15
10
5
0
0
1
2
Number of *SRSA* mutations

**B**

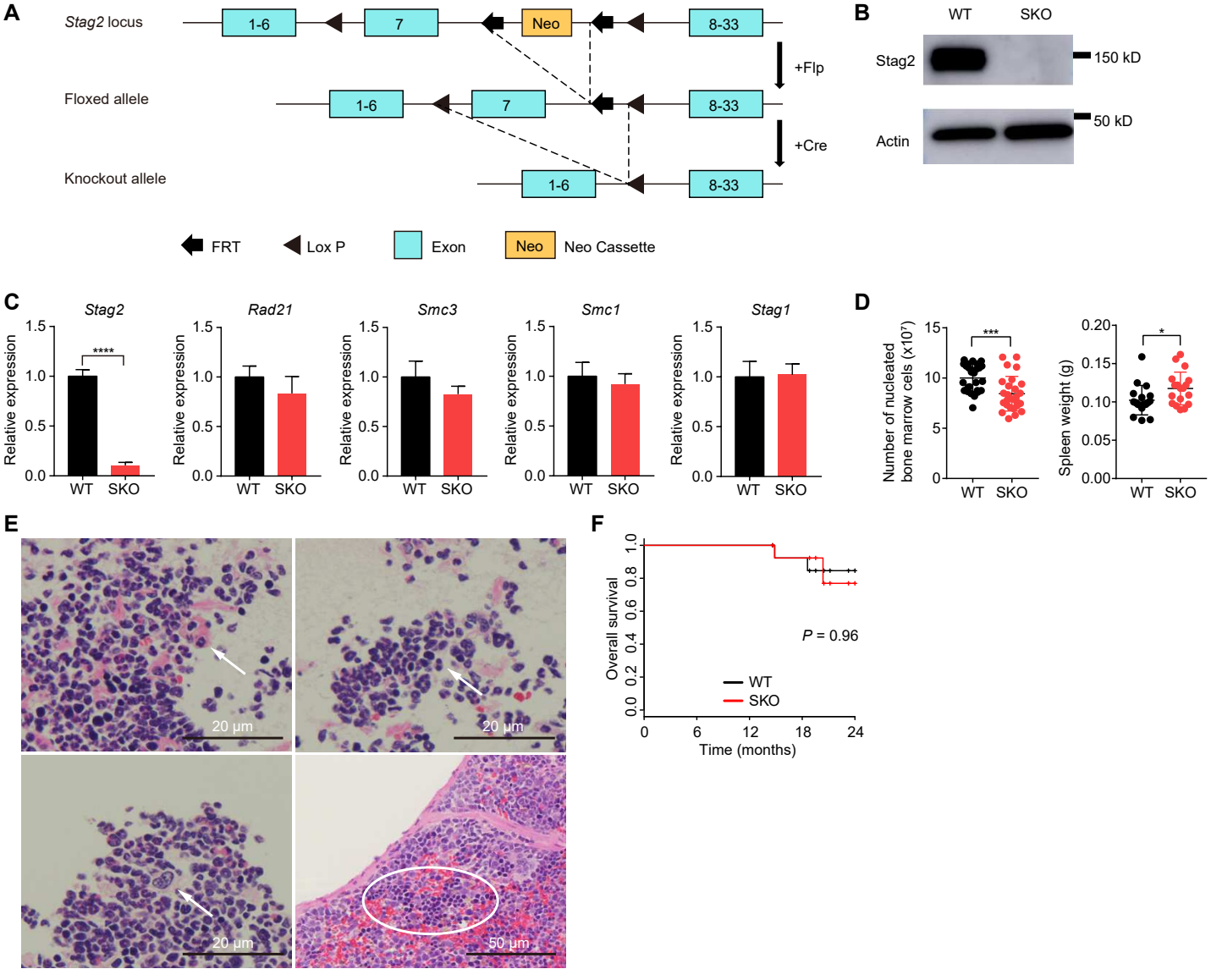

*RUNX1* mutations
*P* = 0.32
Frequency (%)
100
50
0
WT
Mutated
*STAG2*
Missense SNV or nonframeshift Indel
Loss-of-function mutation or frameshift Indel

**C**


Adjusted VAF in co-mutated cases
*P* = 0.013
*P* < 0.001
*P* = 0.28
*P* = 0.10
*P* = 0.079
*P* < 0.001
*RUNX1*
*SRSF2*
*ASXL1*
*STAG2*
*RUNX1*
*SRSF2*
0.00
0.25
0.50
0.75
1.00

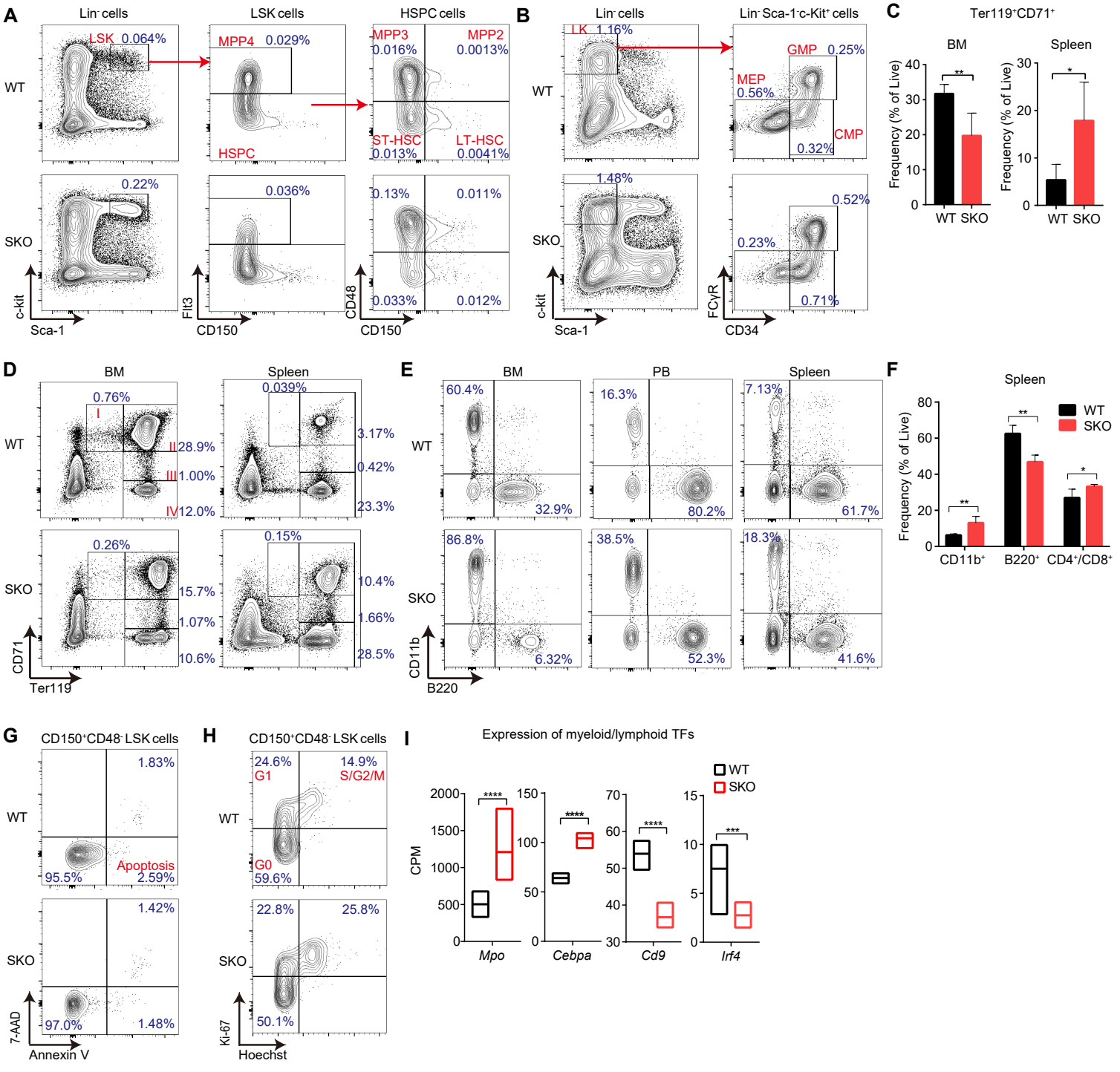
**Supplementary Figure 2. Characteristics of *SRSA* mutations.**

**A**, Number of driver mutations (in addition to *SRSA* mutations) according to the number of *SRSA* mutations. **B**, Proportion of loss-of-function or other *RUNX1* mutations in *STAG2*-WT or mutated cases. *P*-value was calculated by Fisher's exact test. *RUNX1* mutations have a slightly higher frequency of loss-of-function mutations (nonsense, frameshift, or splicing mutations) in *STAG2*-mutated cases than WT, although the difference was not significant ($P = 0.32$). **C**, Scatter plots of adjusted VAF values for each combination of *SRSA* mutations. *P*-values were calculated using distances to diagonal lines and Student's t-test.


A
Stag2 locus
1-6
7
Neo
8-33
+Flp
Floxed allele
+Cre
Knockout allele
FRT
Lox P
Exon
Neo Cassette
B
WT
SKO
Stag2
150 kD
Actin
50 kD
C
Stag2
Rad21
Smc3
Smc1
Stag1
Relative expression
****
D
Number of nucleated bone marrow cels (x10^7)
***
Spleen weight (g)
*
E
20 μm
50 μm
F
Overall survival
P = 0.96
Time (months)

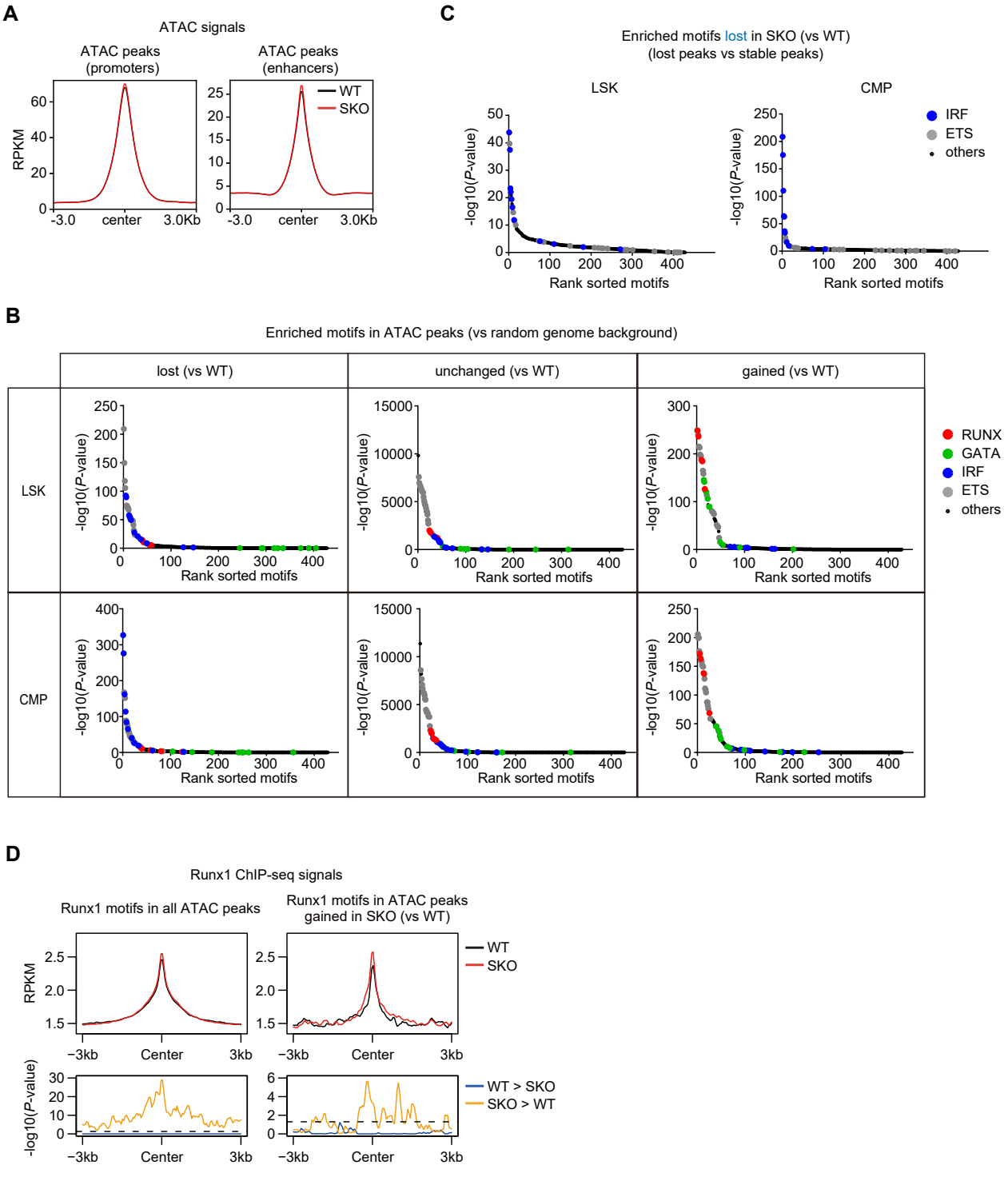
**Supplementary Figure 3. Development of *Stag2* conditional knockout mice and examination of hematological phenotypes.**

**A**, Schematic depiction of the targeted *Stag2* allele. FRT, flippase recognition target. **B**, Representative western blot analysis of Stag2 expression in the BM nucleated cells of WT and SKO mice. **C**, Real-Time qRT-PCR of indicated genes (relative expression, normalized by expression of 18s rRNA, mean ± SD, n = 3). **D**, Absolute number of nucleated BM cells in bilateral femurs and tibias (n = 26), and spleen weight of WT and SKO littermate male mice are plotted as dots (n = 17, mean ± SD). **E**, Section of BM and spleen stained with hematoxylin and eosin. Arrows indicate dysplastic cells in the BM and circle shows the erythroblastic islet in the spleen suggesting the extramedullary hematopoiesis. **F**, Kaplan-Meier plots for overall survival of WT and SKO mice (n = 14 per genotype). *P*-value was calculated by log-rank test. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$. Two-tailed unpaired Student's t-test in (**C-D**).


A
Lin- cells
LSK cells
HSPC cells
LSK 0.064%
MPP4 0.029%
MPP3 0.016%
MPP2 0.0013%
ST-HSC 0.013%
LT-HSC 0.0041%
HSPC
WT
SKO
0.22%
0.036%
0.13%
0.011%
0.033%
0.012%
c-kit
Sca-1
Flt3
CD150
CD48
CD150
B
Lin- cells
Lin- Sca-1- c-Kit+ cells
LK 1.16%
GMP 0.25%
MEP 0.56%
CMP 0.32%
1.48%
0.52%
0.23%
0.71%
FCγR
CD34
C
Ter119+CD71+
BM
Spleen
Frequency (% of Live)
WT SKO
D
BM
Spleen
0.76%
I
II 28.9%
III 1.00%
IV 12.0%
0.039%
3.17%
0.42%
23.3%
0.26%
15.7%
1.07%
10.6%
0.15%
10.4%
1.66%
28.5%
CD71
Ter119
E
BM
PB
Spleen
60.4%
32.9%
16.3%
80.2%
7.13%
61.7%
86.8%
6.32%
38.5%
52.3%
18.3%
41.6%
CD11b
B220
F
Spleen
WT
SKO
Frequency (% of Live)
CD11b+
B220+
CD4+/CD8+
G
CD150+CD48- LSK cells
1.83%
95.5%
Apoptosis
2.59%
1.42%
97.0%
1.48%
7-AAD
Annexin V
H
CD150+CD48- LSK cells
24.6%
G1
14.9%
S/G2/M
G0
59.6%
22.8%
25.8%
50.1%
Ki-67
Hoechst
I
Expression of myeloid/lymphoid TFs
WT
SKO
CPM
Mpo
Cebpa
Cd9
Irf4

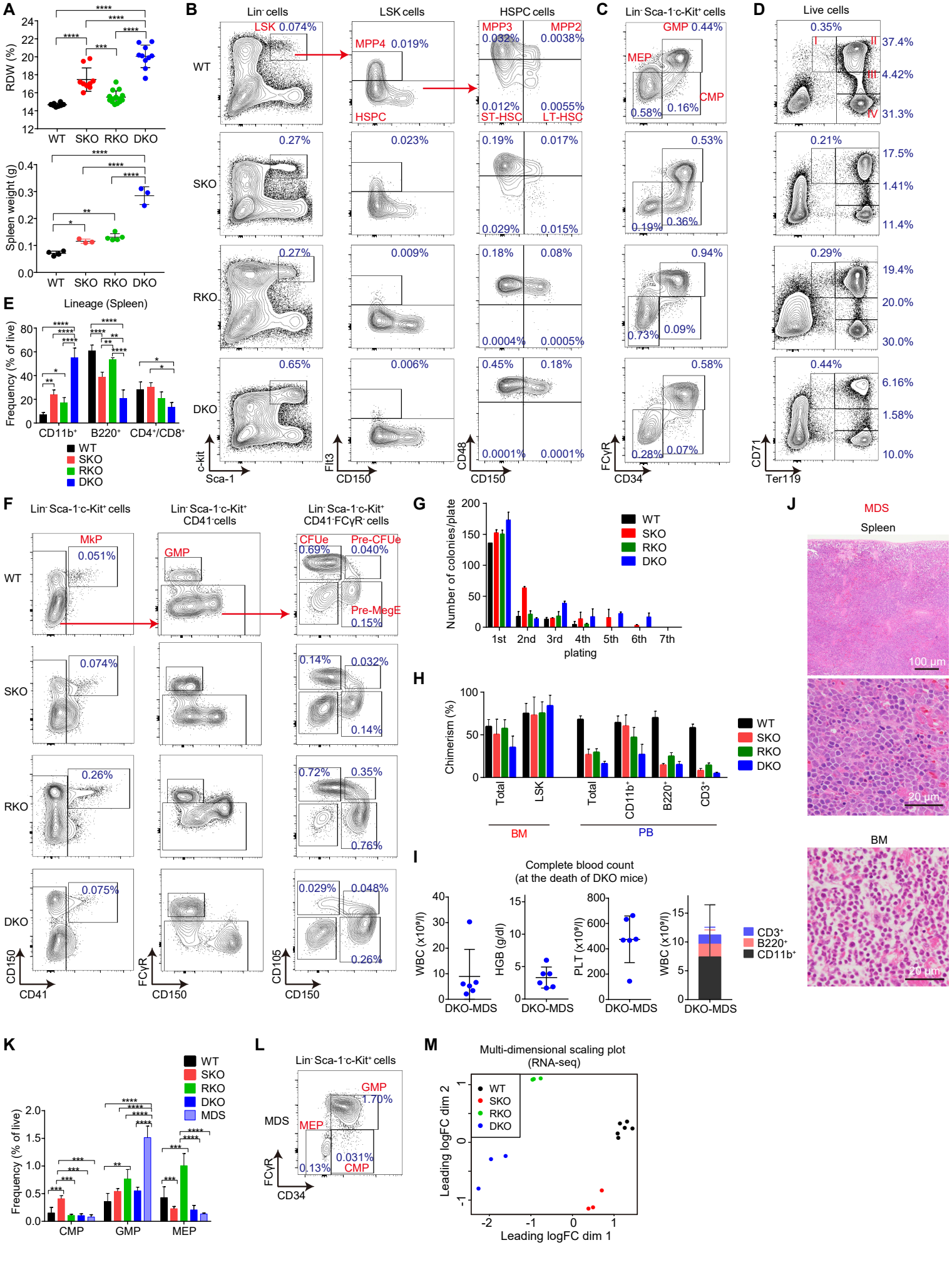
**Supplementary Figure 4. Flow cytometry and transcriptome analysis of *Stag2* conditional knockout mice.**

**A-B**, Representative flow cytometry analysis of the BM LSK (**A**) and Lin-negative/Sca1$^-$/c-Kit$^+$ (LK) populations (**B**) of WT and SKO mice. **C**, Frequency of erythroblasts (Ter119$^+$CD71$^+$) in the BM and spleen (n =5, mean ± SD). **D-E**, Representative flow cytometry analysis of erythroid maturation in the BM and spleen (**D**) and lineage-committed cells in the BM, PB and spleen (**E**). **F**, Frequency of lineage-committed cells in the spleen (n = 4, mean ± SD). **G-H**, Representative flow cytometry analysis of apoptosis (**G**) and cell-cycle (**H**). **I**, Expression levels of myeloid/lymphoid TFs in LSK cells indicated by CPM (min to max values with mean, n = 3). *P*-values were calculated using edgeR package. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$. Two-tailed unpaired Student's t-test in (**C, F**).


A
ATAC signals
ATAC peaks (promoters)
ATAC peaks (enhancers)
RPKM
WT
SKO
-3.0
center
3.0Kb
C
Enriched motifs lost in SKO (vs WT)
(lost peaks vs stable peaks)
LSK
CMP
-log10(P-value)
Rank sorted motifs
IRF
ETS
others
B
Enriched motifs in ATAC peaks (vs random genome background)
lost (vs WT)
unchanged (vs WT)
gained (vs WT)
LSK
CMP
RUNX
GATA
IRF
ETS
others
D
Runx1 ChIP-seq signals
Runx1 motifs in all ATAC peaks
Runx1 motifs in ATAC peaks gained in SKO (vs WT)
WT
SKO
WT > SKO
SKO > WT
-3kb
Center
3kb

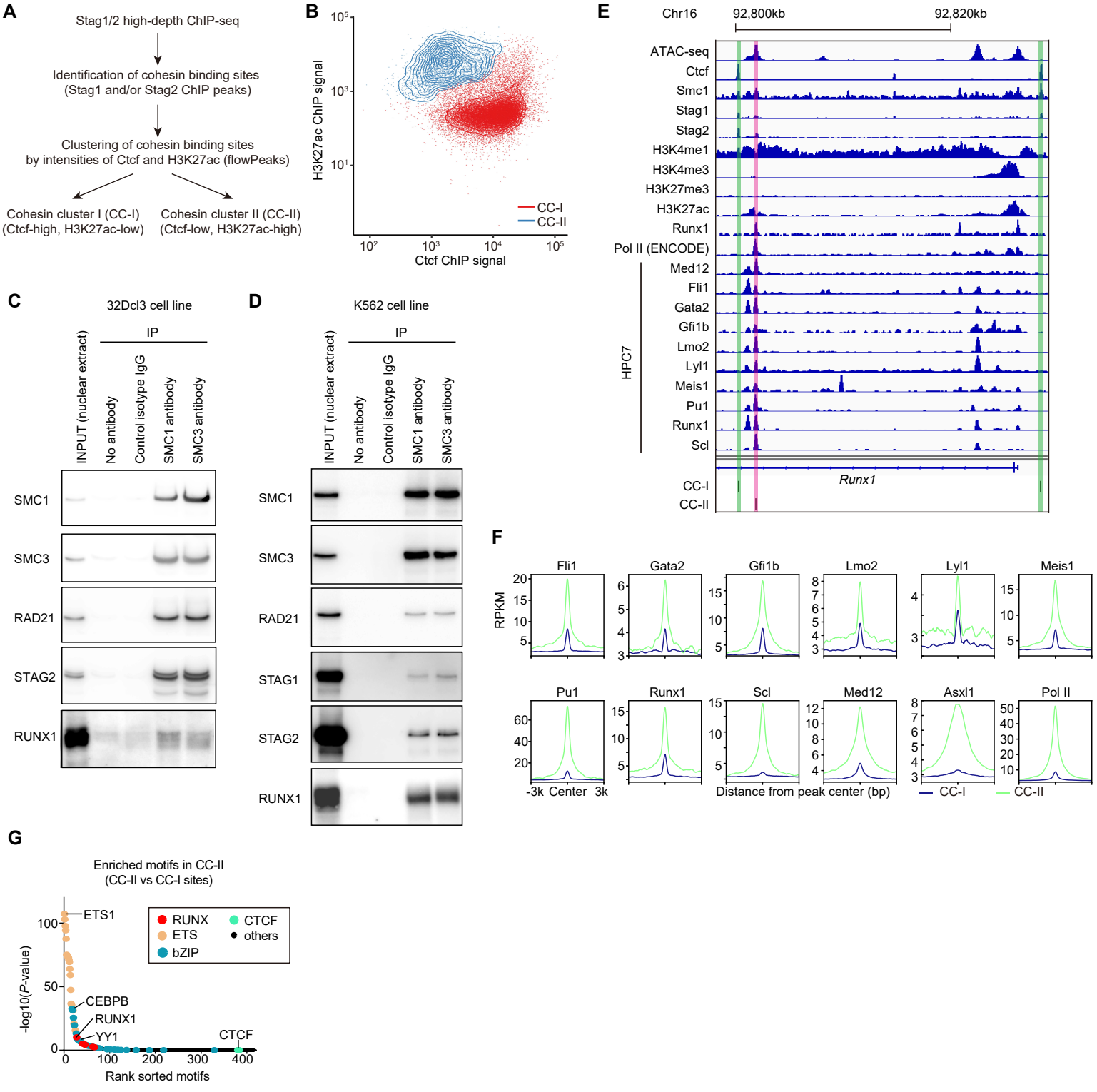
**Supplementary Figure 5. Epigenome analysis of *Stag2* conditional knockout mice.**

**A**, Average signal intensities of ATAC-seq around the ATAC-peaks in promoters (left) or enhancers (right) in WT- and SKO-derived LSK cells. **B**, Enrichment of known TF motifs in the ATAC-seq peaks with gained, lost, or unchanged accessibility in SKO-derived LSK or CMP cells compared with WT cells. The sorted motif rank and –log10(*P*-value) of a motif enrichment test using random genome backgrounds are indicated in horizontal and vertical axis, respectively. **C**, Enrichment of known transcription factor motifs in the ATAC-seq peaks that lost accessibility in SKO-derived LSK (left panel) and CMP cells (right panel) compared with WT. Stable peaks are used as backgrounds. **D**, Average Runx1 ChIP-seq signals of WT- and SKO-derived c-Kit$^+$ HSPCs around Runx1 motifs in all ATAC peaks (left panel), or in gained ATAC peaks in SKO-derived LSK cells compared with WT (right panel). *P*-values were calculated by one-sided Wilcoxon rank-sum test comparing the ChIP-intensities in each bin. Horizontal dashed lines indicate $P = 0.05$.


A
RDW (%)
WT SKO RKO DKO
Spleen weight (g)
WT SKO RKO DKO
B
Lin- cells
LSK cells
HSPC cells
WT
SKO
RKO
DKO
LSK 0.074%
MPP4 0.019%
HSPC
MPP3 0.032%
MPP2 0.0038%
0.012% ST-HSC
0.0055% LT-HSC
0.27%
0.023%
0.19% 0.017%
0.029% 0.015%
0.27%
0.009%
0.18% 0.08%
0.0004% 0.0005%
0.65%
0.006%
0.45% 0.18%
0.0001% 0.0001%
c-kit
Sca-1
Flt3
CD150
CD48
CD150
C
Lin- Sca-1- c-Kit+ cells
GMP 0.44%
MEP
CMP
0.58% 0.16%
0.53%
0.19% 0.36%
0.94%
0.73% 0.09%
0.58%
0.28% 0.07%
FCγR
CD34
D
Live cells
0.35%
I
II
III
IV
37.4%
4.42%
31.3%
0.21%
17.5%
1.41%
11.4%
0.29%
19.4%
20.0%
30.0%
0.44%
6.16%
1.58%
10.0%
CD71
Ter119
E
Lineage (Spleen)
Frequency (% of live)
CD11b+ B220+ CD4+/CD8+
WT
SKO
RKO
DKO
F
Lin- Sca-1- c-Kit+ cells
Lin- Sca-1- c-Kit+ CD41- cells
Lin- Sca-1- c-Kit+ CD41- FCγR- cells
MkP 0.051%
GMP
CFUe 0.69%
Pre-CFUe 0.040%
Pre-MegE 0.15%
0.074%
0.14%
0.032%
0.14%
0.26%
0.72%
0.35%
0.76%
0.075%
0.029%
0.048%
0.26%
CD150
CD41
FCγR
CD150
CD105
CD150
G
Number of colonies/plate
1st 2nd 3rd 4th 5th 6th 7th
plating
WT
SKO
RKO
DKO
H
Chimerism (%)
Total LSK Total CD11b+ B220+ CD3+
BM
PB
WT
SKO
RKO
DKO
I
Complete blood count
(at the death of DKO mice)
WBC (x10⁹/l)
HGB (g/dl)
PLT (x10⁹/l)
WBC (x10⁹/l)
DKO-MDS
CD3+
B220+
CD11b+
J
MDS
Spleen
100 μm
20 μm
BM
20 μm
K
Frequency (% of live)
CMP GMP MEP
WT
SKO
RKO
DKO
MDS
L
Lin- Sca-1- c-Kit+ cells
MDS
GMP 1.70%
MEP
0.031%
0.13%
CMP
FCγR
CD34
M
Multi-dimensional scaling plot
(RNA-seq)
Leading logFC dim 2
Leading logFC dim 1
WT
SKO
RKO
DKO

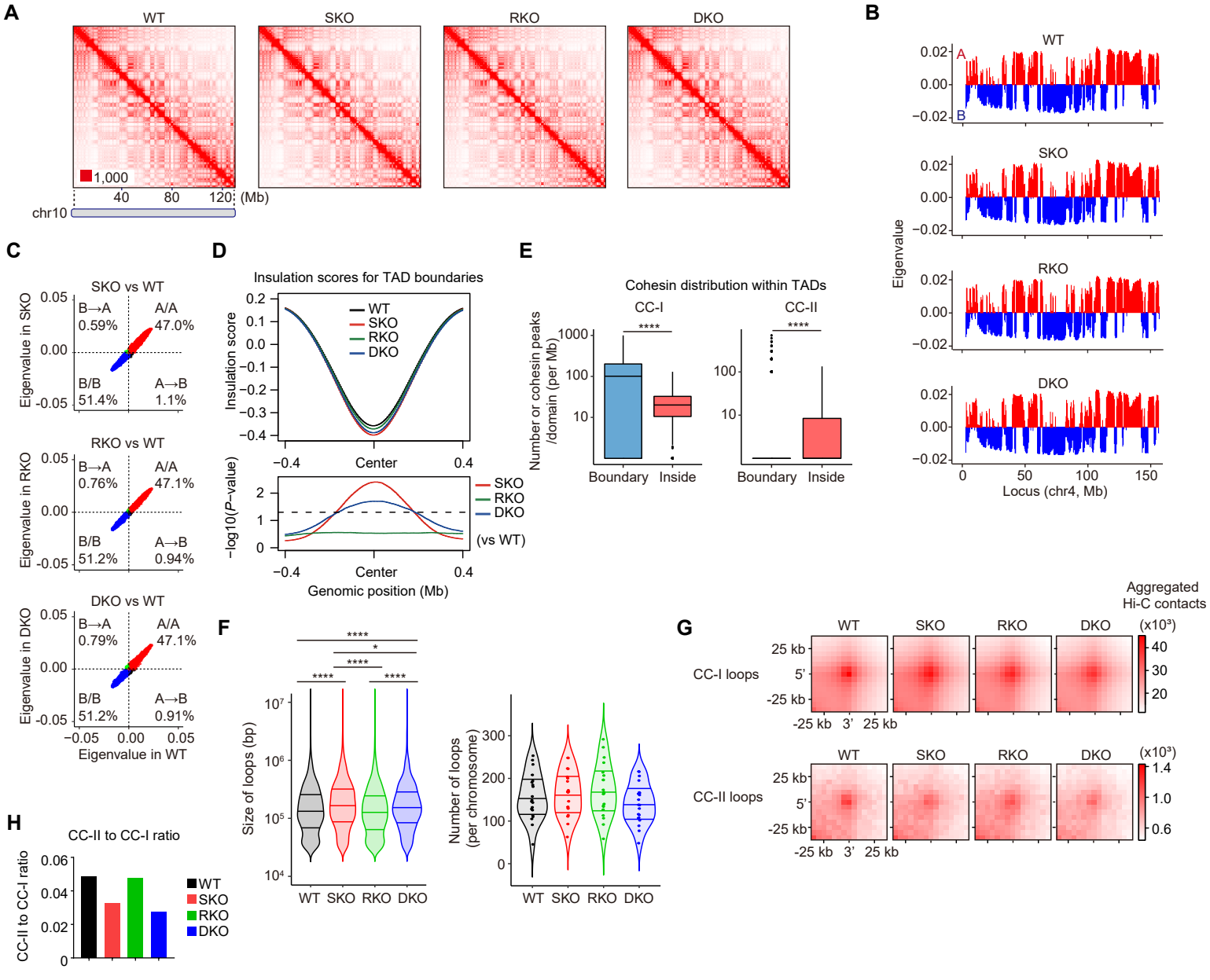
**Supplementary Figure 6. Phenotypes of *Stag2*/*Runx1* conditional knockout mice.**

**A**, RDW and spleen weight are plotted as dots (n = 8 for WT, 9 for SKO, 14 for RKO, and 10 for DKO in RDW and n = 5 for WT and RKO, and 3 for SKO and DKO in spleen weight, mean ± SD). **B-D**, Representative flow cytometry analysis of BM LSK cells (**B**), Lin-negative/Sca1$^-$/c-Kit$^+$ cells (**C**), and erythroid precursors (**D**). **E**, Frequency of each lineage-committed cells in the spleen (n = 5 for WT and RKO, and 3 for SKO and DKO, mean ± SD). **F**, Representative flow cytometry analysis of the megakaryocytic and erythroid progenitors in the BM. **G**, Colony counts in methylcellulose replating experiments (mean ± SD, n = 2) of BM cells. **H**, Percentages of CD45.2$^+$ donor cells within each fraction of BM or PB after competitive BM transplantation (16 weeks after pIpC injection) are shown (n = 4, mean ± SD). **I**, WBC, HGB, PLT counts, and total cell number of granulocytes/monocytes (CD11b$^+$), B lymphoid (B220$^+$) and T lymphoid (CD4$^+$/CD8$^+$) cells in the PB of mice that developed MDS (n = 6, mean ± SD). **J**, Section of the spleen (upper panels) and BM (lower panel) stained with hematoxylin and eosin, showing the infiltrating dysplastic myeloid cells in the spleen and BM. **K**, Frequencies of myeloid progenitors in the BM of WT, SKO, RKO or DKO-transplanted mice and MDS mice (DKO mice that developed MDS) (n = 5 for WT and RKO, and 3 for SKO, DKO and MDS, mean ± SD). **L**, Representative flow cytometry analysis of the myeloid progenitors in the BM of DKO mice that developed MDS, showing the expansion of the GMP fractions. **M**, Multi-dimensional scaling plot in which distances correspond to leading logFC between each pair of RNA-seq sample in WT/SKO/RKO/DKO-derived LSK cells. The leading logFC is the average of the largest absolute logFC between each pair of samples. The horizontal and vertical axis show the leading logFC of dimension 1 and 2, respectively. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$. *P*-values were calculated by ordinary one-way ANOVA with Bonferroni analysis in (**A, E, K**).


A
Stag1/2 high-depth ChIP-seq
Identification of cohesin binding sites
(Stag1 and/or Stag2 ChIP peaks)
Clustering of cohesin binding sites
by intensities of Ctcf and H3K27ac (flowPeaks)
Cohesin cluster I (CC-I)
(Ctcf-high, H3K27ac-low)
Cohesin cluster II (CC-II)
(Ctcf-low, H3K27ac-high)
B
H3K27ac ChIP signal
Ctcf ChIP signal
CC-I
CC-II
C
32Dcl3 cell line
INPUT (nuclear extract)
IP
No antibody
Control isotype IgG
SMC1 antibody
SMC3 antibody
SMC1
SMC3
RAD21
STAG2
RUNX1
D
K562 cell line
STAG1
E
Chr16
92,800kb
92,820kb
ATAC-seq
Ctcf
Smc1
Stag1
Stag2
H3K4me1
H3K4me3
H3K27me3
H3K27ac
Runx1
Pol II (ENCODE)
HPC7
Med12
Fli1
Gata2
Gfi1b
Lmo2
Lyl1
Meis1
Pu1
Runx1
Scl
Runx1
CC-I
CC-II
F
Fli1
Gata2
Gfi1b
Lmo2
Lyl1
Meis1
RPKM
Pu1
Runx1
Scl
Med12
Asxl1
Pol II
-3k Center 3k
Distance from peak center (bp)
CC-I
CC-II
G
Enriched motifs in CC-II
(CC-II vs CC-I sites)
RUNX
ETS
bZIP
CTCF
others
ETS1
CEBPB
RUNX1
YY1
CTCF
-log10(P-value)
Rank sorted motifs

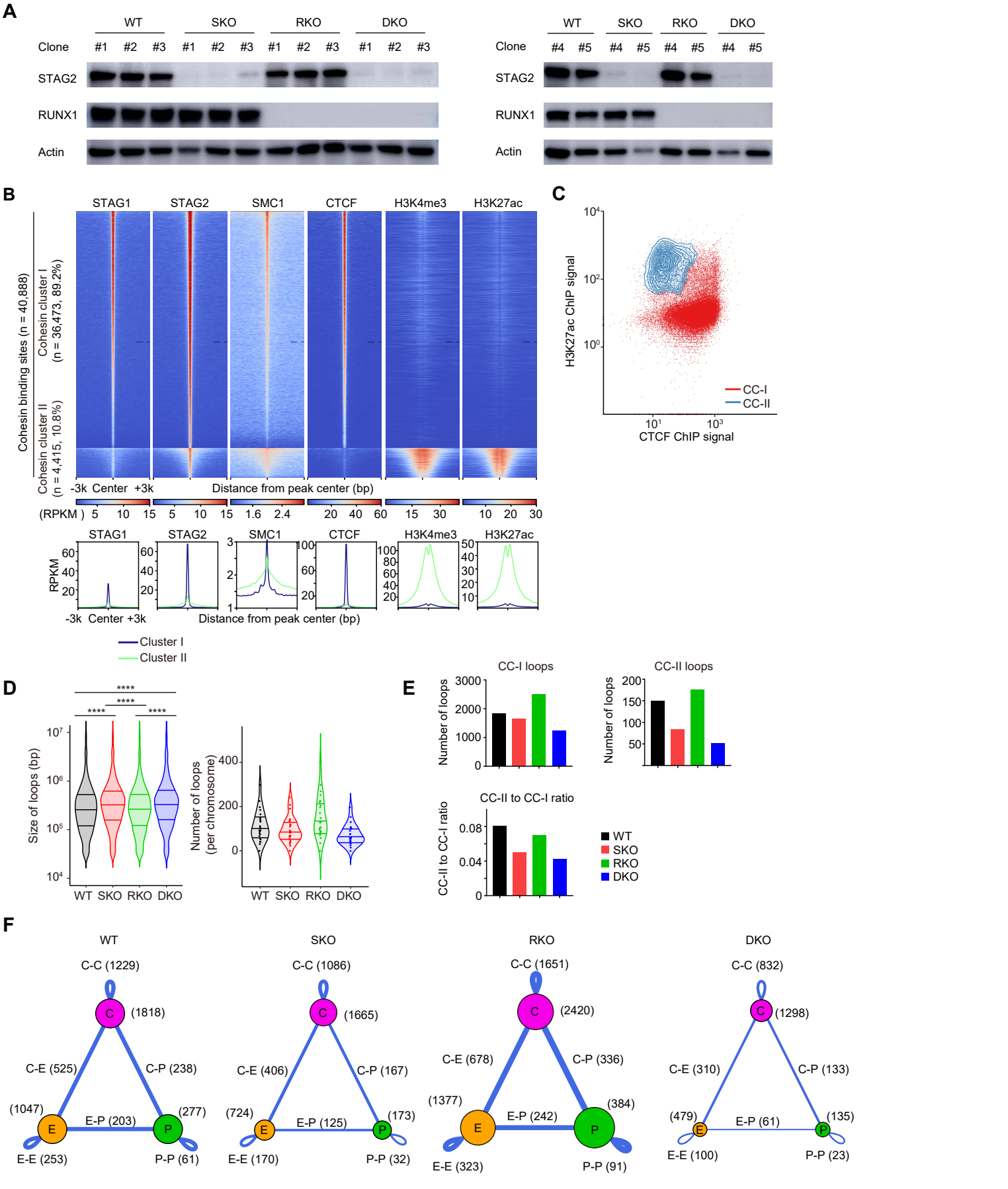
**Supplementary Figure 7. ChIP-seq analysis and identification of CC-I and CC-II sites.**

**A**, Summary of the methods used to identify the two types of cohesin binding sites in ChIP-seq analysis. **B**, Scatterplot and density plot of Ctcf and H3K27ac ChIP intensities for each cohesin binding site, indicated as RPKM values summed up around ± 200 bp from the center of each peak, according to the clusters of cohesin binding sites. CC-I, cohesin-cluster I; CC-II, cohesin-cluster II. **C-D**, Co-immunoprecipitation and western blotting experiments showing the physical interactions of cohesin complex with Runx1/RUNX1 in mouse 32Dcl3 (**C**) or human K562 (**D**) leukemia cell lines. Nuclear extractions were subjected to immunoprecipitation using indicated antibodies above the photos, followed by western blotting using antibodies indicated on the left. **E**, Genome browser snapshot demonstrating the co-localization of various transcriptional regulators at CC-II site at *Runx1* gene locus. **F**, Distribution of indicated proteins around cohesin binding sites were analyzed using published ChIP-seq data of HPC7 and others (Aranda-Orgilles et al., 2016; Li et al., 2017; Wilson et al., 2010), and average ChIP-seq read intensities around CC-I (blue) and CC-II (green) sites are depicted. **G**, Enrichment of known transcription factor motifs in the ChIP-seq peaks of CC-II sites compared with CC-I sites. The sorted motif rank and –log10(*P*-value) of a motif enrichment test are indicated in horizontal and vertical axis, respectively.

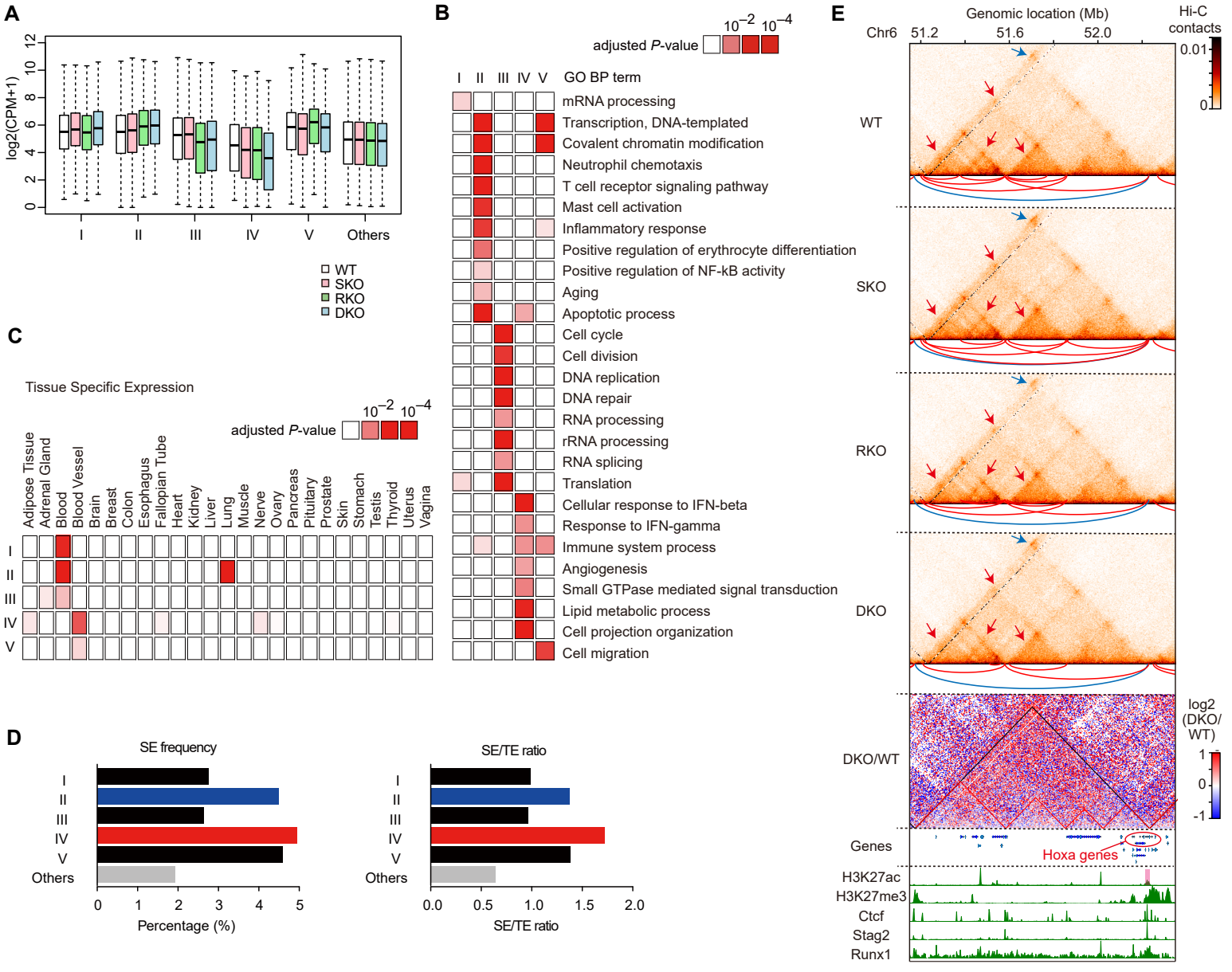

A
WT
SKO
RKO
DKO
1,000
40
80
120
(Mb)
chr10
B
WT
SKO
RKO
DKO
A
B
0.02
0.00
−0.02
Eigenvalue
0
50
100
150
Locus (chr4, Mb)
C
SKO vs WT
RKO vs WT
DKO vs WT
B→A
A/A
B/B
A→B
0.59%
47.0%
51.4%
1.1%
0.76%
47.1%
51.2%
0.94%
0.79%
47.1%
51.2%
0.91%
Eigenvalue in SKO
Eigenvalue in RKO
Eigenvalue in DKO
Eigenvalue in WT
0.05
0.00
−0.05
D
Insulation scores for TAD boundaries
Insulation score
−log10(P−value)
WT
SKO
RKO
DKO
(vs WT)
−0.4
Center
0.4
Genomic position (Mb)
E
Cohesin distribution within TADs
CC-I
CC-II
****
Number or cohesin peaks /domain (per Mb)
1000
100
10
Boundary
Inside
F
****
*
Size of loops (bp)
Number of loops (per chromosome)
WT SKO RKO DKO
G
Aggregated Hi-C contacts
CC-I loops
CC-II loops
25 kb
5'
-25 kb
3'
(x10³)
40
30
20
1.4
1.0
0.6
H
CC-II to CC-I ratio
0.06
0.04
0.02
0

**Supplementary Figure 8. Hi-C analysis in *Stag2*/*Runx1* conditional knockout mice.**

**A**, Knight-Ruiz (KR)-normalized Hi-C contact matrices in whole chromosome 10, generated by Juicebox. The intensity of each pixel represents the normalized number of contacts between a pair of loci, and maximum intensity of Hi-C contact is indicated in the lower left of the panel. **B**, First eigenvalues for each genotype at each genomic bin in chromosome 4 indicated as snapshot showing the genomic locus and corresponding values. A-compartments were assigned to the genomic bin with positive eigenvector values as well as higher gene density and B-compartments were the opposite. **C**, Scatterplot of the first eigenvalues for SKO, RKO, or DKO vs WT. Numbers within the plots indicate the percentage of bins, in which assignments to A- or B-compartments were changed or unchanged in SKO, RKO, or DKO compared with WT. Colors of dots represent the changed (green, B to A; black, A to B) or unchanged (red, A to A; blue, B to B) bins. **D**, Average insulation scores at the center of all TAD boundaries. Distance from the boundary and average insulation scores are indicated in horizontal and vertical axis, respectively. *P*-values were calculated by bin-wise one-sided Wilcoxon rank-sum test. **E** Number of cohesin peaks (CC-I or CC-II) insides or at the boundaries of TADs. *P*-values were calculated by two-sided Wilcoxon rank-sum test. A horizontal dashed line indicates $P = 0.05$. **F**, Violin plots showing the size distribution (left) and numbers of all loops (right). *P*-values were calculated by pairwise comparisons using two-sided Wilcoxon rank-sum test with Bonferroni correction. **G**, Aggregate peak analysis (Rao et al., 2014) to measure the aggregate strength of loops anchored at CC-I or CC-II loops, showing the diminishment of CC-II loops particularly in DKO. The number of aggregated Hi-C contacts for each type of loops is indicated in the color bars. **H**, Ratio of CC-II loops to CC1 loops in each genotype. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$.

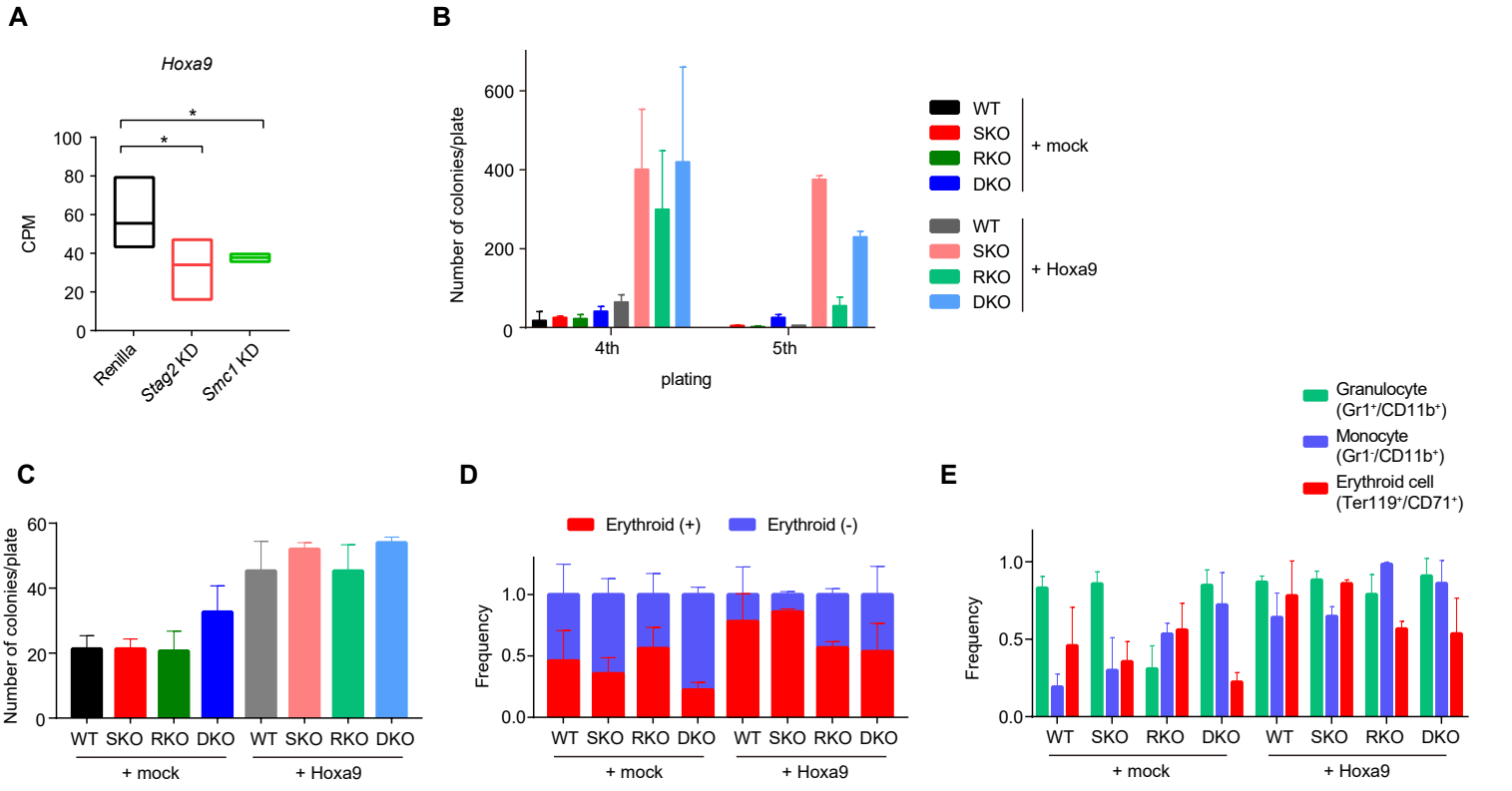

A
WT
SKO
RKO
DKO
Clone
#1 #2 #3 #1 #2 #3 #1 #2 #3 #1 #2 #3
STAG2
RUNX1
Actin
#4 #5 #4 #5 #4 #5 #4 #5
B
STAG1 STAG2 SMC1 CTCF H3K4me3 H3K27ac
Cohesin binding sites (n = 40,888)
Cohesin cluster I (n = 36,473, 89.2%)
Cohesin cluster II (n = 4,415, 10.8%)
-3k Center +3k
Distance from peak center (bp)
(RPKM)
RPKM
Cluster I
Cluster II
C
H3K27ac ChIP signal
CTCF ChIP signal
CC-I
CC-II
D
Size of loops (bp)
Number of loops (per chromosome)
WT SKO RKO DKO
E
CC-I loops
CC-II loops
Number of loops
CC-II to CC-I ratio
F
WT
SKO
RKO
DKO
C-C (1229)
C-C (1086)
C-C (1651)
C-C (832)
(1818)
(1665)
(2420)
(1298)
C-E (525)
C-P (238)
C-E (406)
C-P (167)
C-E (678)
C-P (336)
C-E (310)
C-P (133)
(1047)
(277)
(724)
(173)
(1377)
(384)
(479)
(135)
E-P (203)
E-P (125)
E-P (242)
E-P (61)
E-E (253)
P-P (61)
E-E (170)
P-P (32)
E-E (323)
P-P (91)
E-E (100)
P-P (23)

**Supplementary Figure 9. Hi-C analysis of *STAG2*/*RUNX1* knockout HL-60 human leukemia cell lines.** **A**, Representative western blots of STAG2 and RUNX1 expression in HL-60 cell lines with *STAG2*/*RUNX1* KO. **B**, Upper panels: ChIP-seq density heatmap in parent (WT) HL-60 cell lines centered on STAG1- and/or STAG2-cohesin binding sites, in which cohesin binding sites were divided into CC-I and CC-II according to the ChIP signals for CTCF and H3K27ac (see also panel (**C**) and **Supplementary Fig. S7A**). Lower panels: Average ChIP-seq read intensity plot for CC-I (blue) and CC-II (green) distribution around the cohesin binding sites. **C**, Scatter plot and density plot of CTCF and H3K27ac ChIP intensities for each cohesin binding site, indicated as RPKM values summed up around ± 200 bp from the center of each peak, according to the clusters of cohesin binding sites in HL-60 cell lines. **D**, Violin plots showing the size distribution (left) and numbers of all loops (right). *P*-values were calculated by pairwise comparisons using two-sided Wilcoxon rank-sum test with Bonferroni correction. **E**, Number of CC-I or CC-II loops and ratio of number of CC-II loops to CC-I loops. **F**, Summary of major types of loops identified in each genotype of HL-60 cell lines. CTCF sites (CC-I sites) and active enhancers/promoters in which loops were anchored are displayed as purple, orange, and green circles, respectively. The loops between two sites are displayed as blue lines, and the width of the lines is proportional to the number of loops relative to WT. E, Enhancer; P, Promoter; C, CTCF; C-C, CTCF-CTCF; C-E, CTCF-Enhancer: C-P, CTCF-Promoter; E-E, Enhancer-Enhancer; E-P, Enhancer-Promoter; P-P, Promoter-Promoter. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$.


A
log2(CPM+1)
I
II
III
IV
V
Others
WT
SKO
RKO
DKO
B
adjusted P-value
GO BP term
mRNA processing
Transcription, DNA-templated
Covalent chromatin modification
Neutrophil chemotaxis
T cell receptor signaling pathway
Mast cell activation
Inflammatory response
Positive regulation of erythrocyte differentiation
Positive regulation of NF-kB activity
Aging
Apoptotic process
Cell cycle
Cell division
DNA replication
DNA repair
RNA processing
rRNA processing
RNA splicing
Translation
Cellular response to IFN-beta
Response to IFN-gamma
Immune system process
Angiogenesis
Small GTPase mediated signal transduction
Lipid metabolic process
Cell projection organization
Cell migration
C
Tissue Specific Expression
adjusted P-value
Adipose Tissue
Adrenal Gland
Blood
Blood Vessel
Brain
Breast
Colon
Esophagus
Fallopian Tube
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Skin
Stomach
Testis
Thyroid
Uterus
Vagina
D
SE frequency
Percentage (%)
SE/TE ratio
E
Genomic location (Mb)
Chr6
Hi-C contacts
WT
SKO
RKO
DKO
DKO/WT
log2 (DKO/WT)
Genes
Hoxa genes
H3K27ac
H3K27me3
Ctcf
Stag2
Runx1

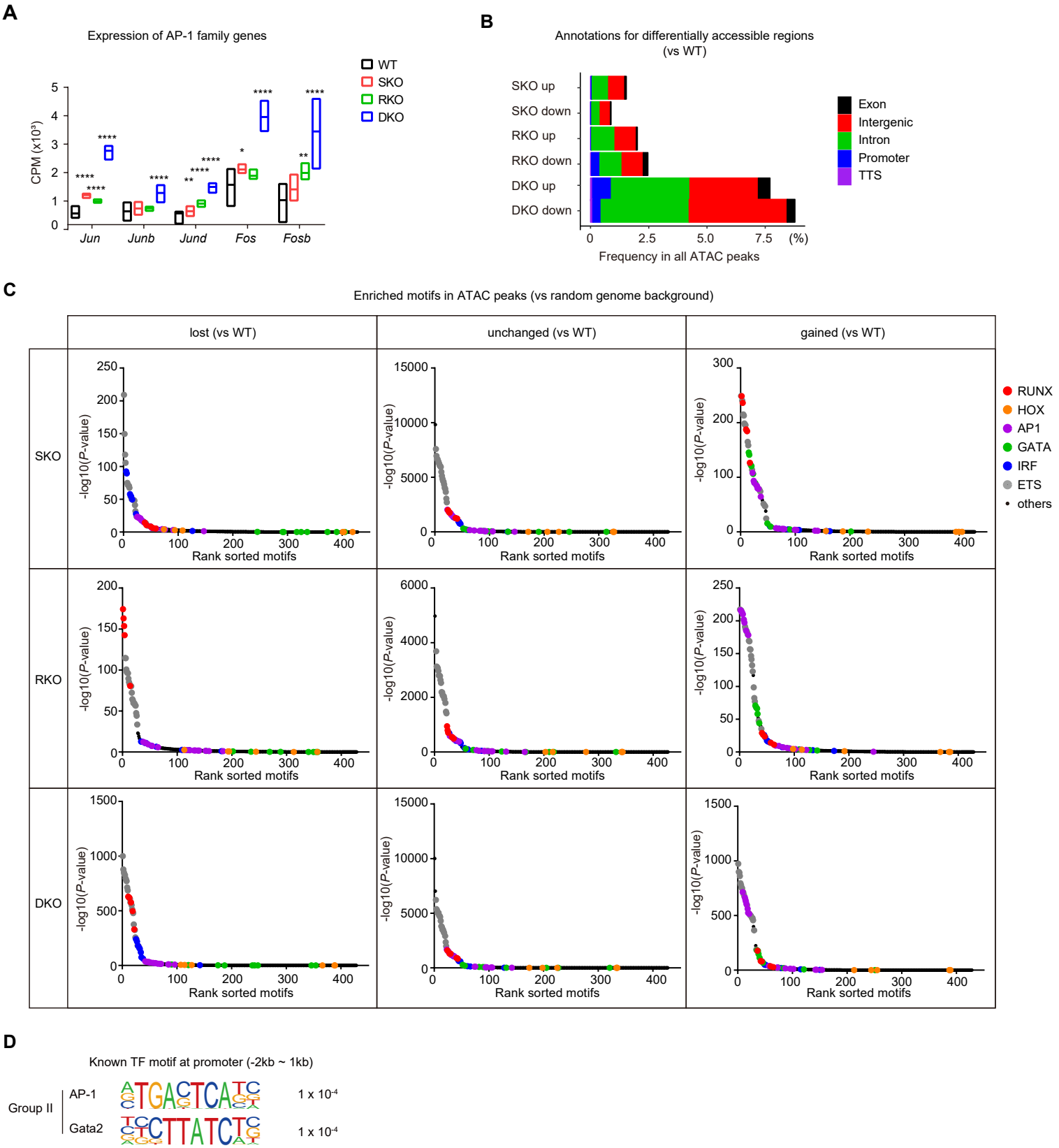
**Supplementary Figure 10. Analysis of transcriptomes, super-enhancers, and Hi-C datasets in Stag2/Runx1 deficient HSPCs.**

**A**, Box plots showing expression levels of each DEG group in WT/SKO/RKO/DKO-derived LSK cells. The vertical axis represents the log2(CPM+1) in the indicated genotype and DEG group. **B**, Summary of representative gene ontology (GO) terms associated with the indicated DEG groups with corresponding adjusted *P*-values, determined by DAVID. **C**, Summary of enrichment of genes of the indicated DEG groups in tissue-specific gene sets in various tissues, determined by Tissue Specific Expression Analysis (TSEA). Adjusted *P*-values are displayed as heatmap. **D**, Frequency of SE-associated gene (upper) and ratio of frequency of SE-associated genes to that of TE-associated genes in each DEG group (bottom). **E**, Genome browser snapshot demonstrating the Hi-C contacts, chromatin loops, and ChIP-seq profiles at the Hoxa gene cluster including *Hoxa9* gene. The black and red triangles in the DKO/WT Hi-C contact map shows the primary TAD and sub-TADs called in WT, respectively. The arcs below each Hi-C contact map show the loops identified in corresponding Hi-C data. Note that smaller loops (red arrows) and Hi-C contacts within sub-TADs (red triangle) were weakened in DKO, while a larger Ctcf-mediated loop (blue arrow) was rather enhanced.


A
Hoxa9
CPM
Renilla
Stag2 KD
Smc1 KD
B
Number of colonies/plate
plating
4th
5th
WT
SKO
RKO
DKO
+ mock
+ Hoxa9
C
Number of colonies/plate
D
Erythroid (+)
Erythroid (-)
Frequency
E
Granulocyte (Gr1+/CD11b+)
Monocyte (Gr1-/CD11b+)
Erythroid cell (Ter119+/CD71+)
Frequency

**Supplementary Figure 11. Effects of Hoxa9 overexpression on Stag2/Runx1 deficient HSPCs.**

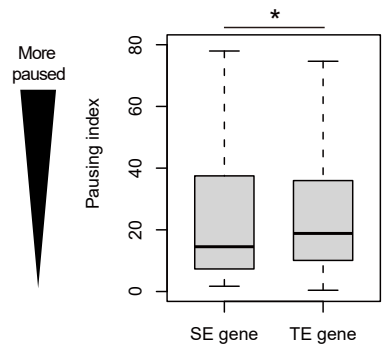
**A**, Expression levels of *Hoxa9* in *Stag2* or *Smc1* knockdown (KD) LSK cells (Mullenders et al., 2015), indicated by CPM (min to max values with mean, n = 4 for Renilla and 3 for KD groups). *P*-values were calculated using edgeR package in R software. **B**, Colony counts at 4th and 5th plating in methylcellulose replating experiments (mean ± SD, n = 2) of c-Kit$^+$ cells transduced with mock- or Hoxa9-expressing retroviral vector. Transduced cells were selected by G418 at the first plating. **C**, Colony counts per 96-well plate in single-cell liquid culture assay (mean ± SD, n = 3) of c-Kit$^+$ cells transduced with mock- or Hoxa9-expressing retroviral vector. **D**, Frequencies of colonies containing or not containing erythroid cells (mean ± SD, n = 3). **E**, Frequencies of granulocyte-, monocyte-, and erythroid-containing colonies (mean ± SD, n = 3).


A
Expression of AP-1 family genes
WT
SKO
RKO
DKO
CPM (x10³)
Jun
Junb
Jund
Fos
Fosb
B
Annotations for differentially accessible regions (vs WT)
SKO up
SKO down
RKO up
RKO down
DKO up
DKO down
Exon
Intergenic
Intron
Promoter
TTS
Frequency in all ATAC peaks
C
Enriched motifs in ATAC peaks (vs random genome background)
lost (vs WT)
unchanged (vs WT)
gained (vs WT)
SKO
RKO
DKO
-log10(P-value)
Rank sorted motifs
RUNX
HOX
AP1
GATA
IRF
ETS
others
D
Known TF motif at promoter (-2kb ~ 1kb)
Group II
AP-1
1 x 10⁻⁴
Gata2
1 x 10⁻⁴

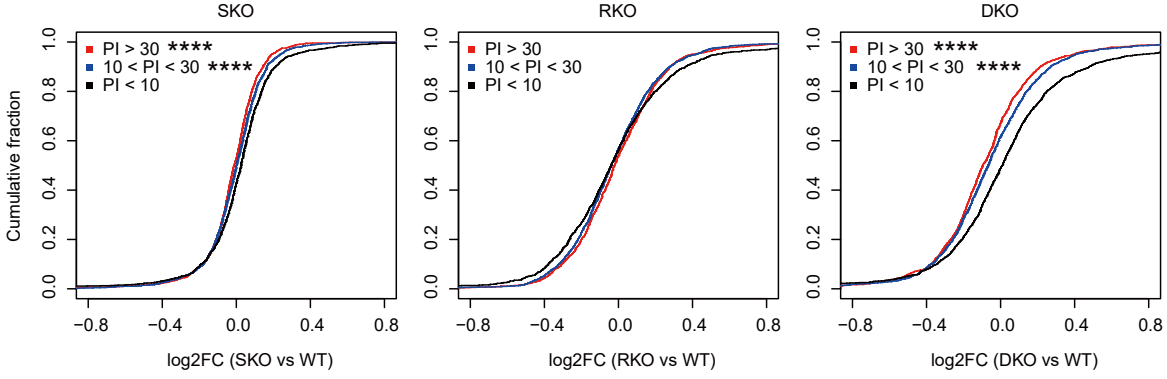
**Supplementary Figure 12. Analysis of ATAC-seq in Stag2/Runx1 deficient HSPCs.**

**A**, Expression of AP-1 family genes in LSK cells as indicated by CPM (min to max values with mean, n = 6 for WT and 3 for the others), in which *P*-values (vs WT) were calculated with edgeR package. **B**, Genomic annotation of differentially accessible ATAC peaks in SKO-, RKO- and DKO-derived LSK cells compared with WT. **C**, Enrichment of known TF motifs in the ATAC-seq peaks with gained, lost, or unchanged accessibility in SKO/RKO/DKO-derived LSK cells compared with WT. The sorted motif rank and –log10(*P*-value) of a motif enrichment test using random genome backgrounds is indicated in horizontal and vertical axis, respectively. **D**, Motifs and corresponding *P*-values identified by known TF motif search in the HOMER software in the promoter regions of genes in DEG group II. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$.

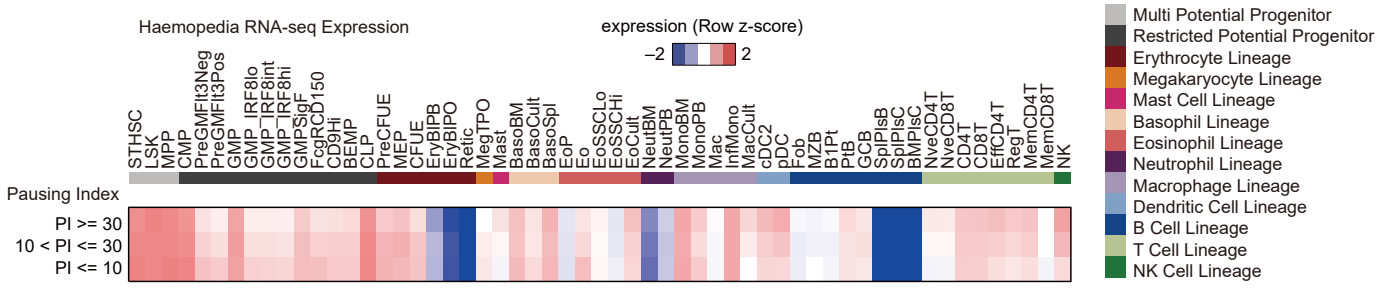
**A**


More paused
Pausing index
0
20
40
60
80
*
SE gene
TE gene

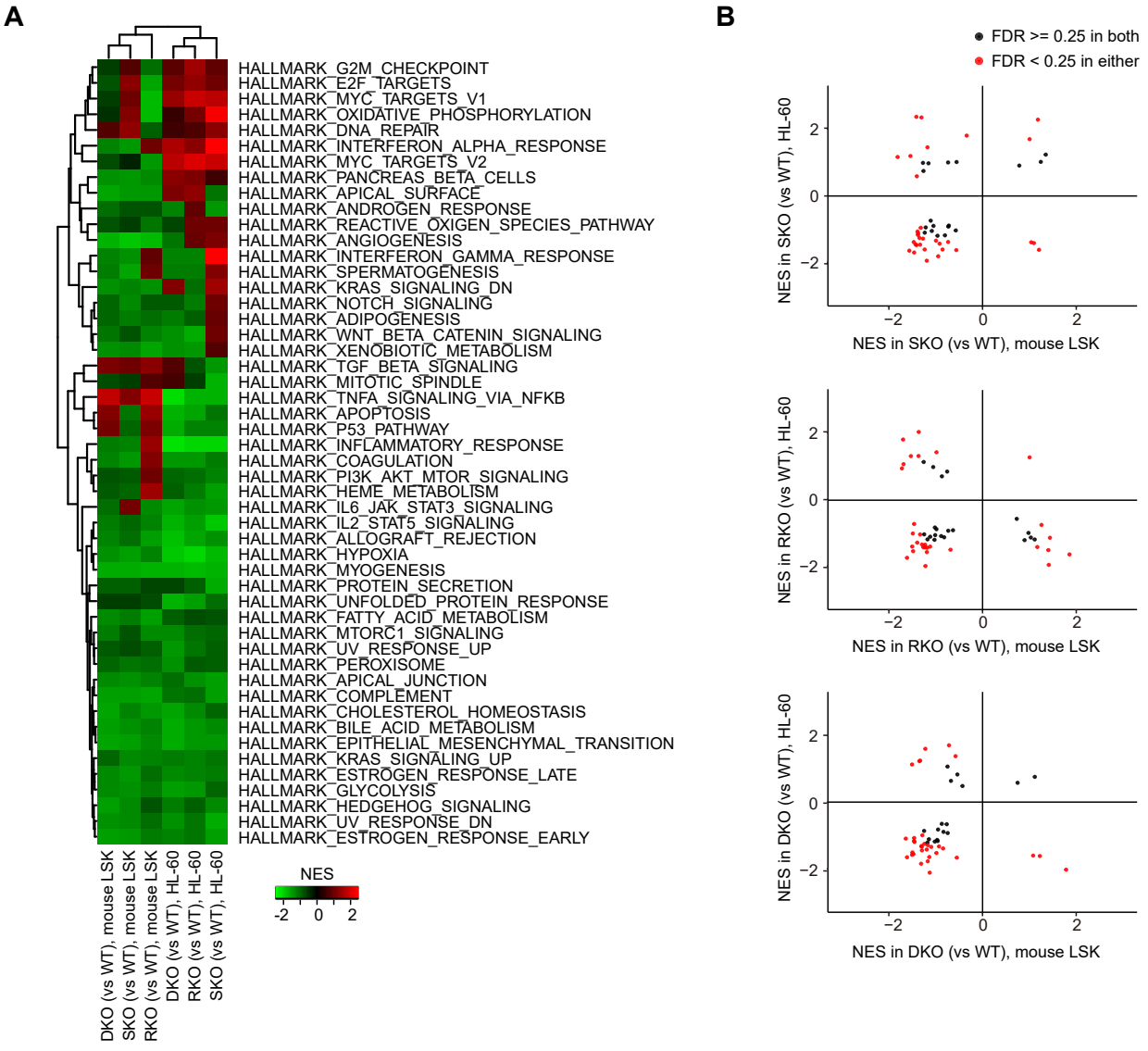
**B**


SKO
RKO
DKO
PI > 30 ****
10 < PI < 30 ****
PI < 10
Cumulative fraction
log2FC (SKO vs WT)
log2FC (RKO vs WT)
log2FC (DKO vs WT)

**C**

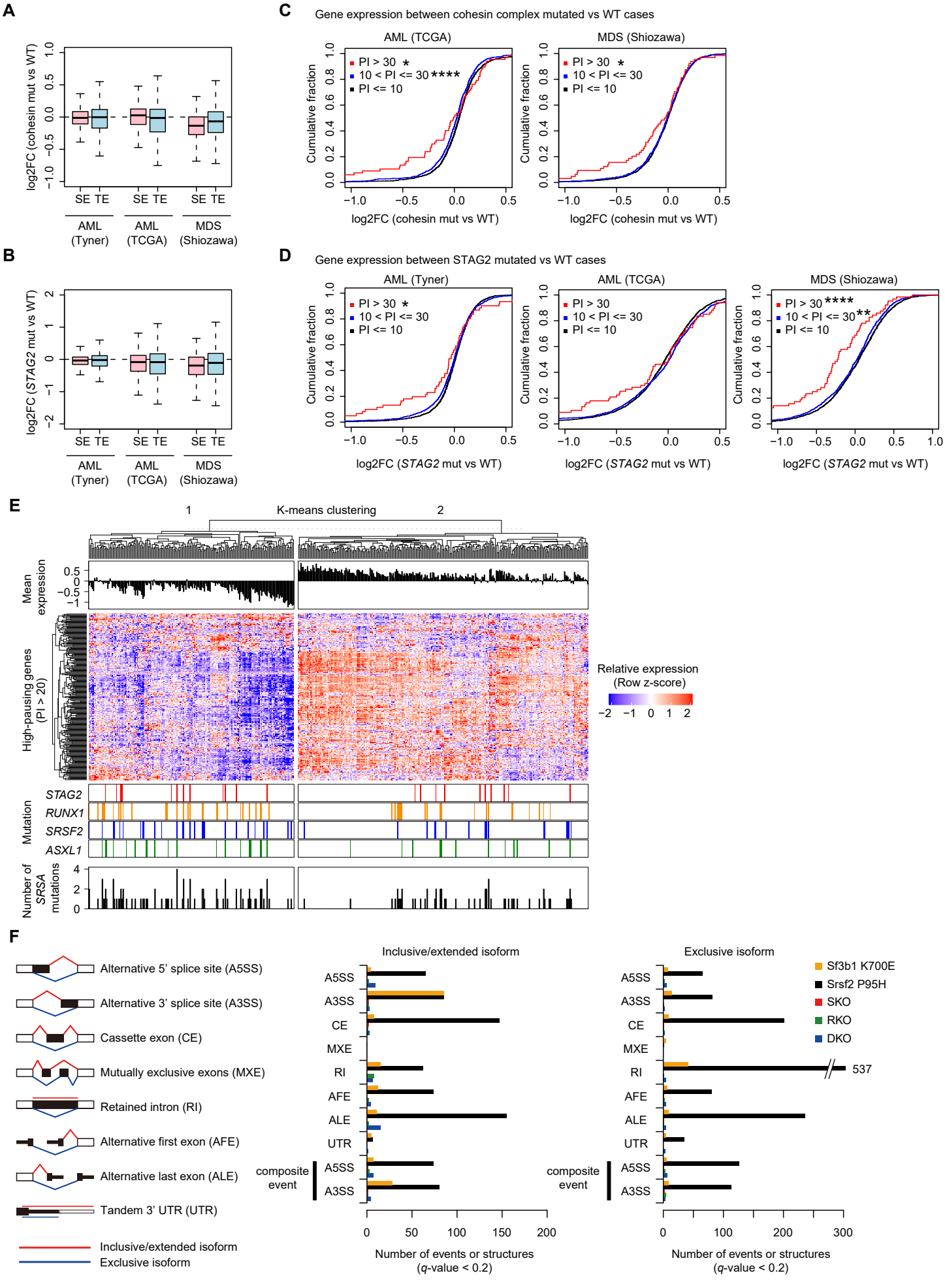

Haemopedia RNA-seq Expression
expression (Row z-score)
−2
2
STHSC
LSK
MPP
CMP
PreGMFlt3Neg
PreGMFlt3Pos
GMP
GMP_IRF8lo
GMP_IRF8int
GMP_IRF8hi
GMPSigF
FcgRCD150
CD9Hi
BEMP
CLP
PreCFUE
MEP
CFUE
EryBlPB
EryBlPO
Retic
MegTPO
Mast
BasoBM
BasoCult
BasoSpl
EoP
Eo
EoSSCLo
EoSSCHi
EoCult
NeutBM
NeutPB
MonoBM
MonoPB
Mac
InfMono
MacCult
cDC2
pDC
Fob
MZB
B1Pt
PtB
GCB
SplPlsB
SplPlsC
BMPlsC
NveCD4T
NveCD8T
CD4T
CD8T
EffCD4T
RegT
MemCD4T
MemCD8T
NK
Pausing Index
PI >= 30
10 < PI <= 30
PI <= 10
Multi Potential Progenitor
Restricted Potential Progenitor
Erythrocyte Lineage
Megakaryocyte Lineage
Mast Cell Lineage
Basophil Lineage
Eosinophil Lineage
Neutrophil Lineage
Macrophage Lineage
Dendritic Cell Lineage
B Cell Lineage
T Cell Lineage
NK Cell Lineage

**Supplementary Figure 13. Analysis of Pol II pausing and expression in Stag2/Runx1 deficient HSPCs.** **A**, Pausing indices of SE-associated genes and TE-associated genes. Note that SE-associated genes show lower degrees of promoter-proximal pausing consistent with the highly active status of transcription. *P*-value was calculated by one-sided Wilcoxon rank-sum test. **B,** Cumulative probability distributions of expression changes (log2FC) of genes grouped by Pol II pausing indices in SKO/RKO/DKO compared with WT. *P*-values (vs genes with PI no more than 10) were calculated by one-sided Wilcoxon rank-sum test. **C**, Expression specificity of genes classified by Pol II pausing indices across diverse hematopoietic lineages. Average expression levels of genes in the indicated groups in each hematopoietic lineage are shown. Mouse expression datasets of diverse hematopoietic lineages are from Haemopedia RNA-seq datasets. Color scales are normalized along each row.


A
HALLMARK_G2M_CHECKPOINT
HALLMARK_E2F_TARGETS
HALLMARK_MYC_TARGETS_V1
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_DNA_REPAIR
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_MYC_TARGETS_V2
HALLMARK_PANCREAS_BETA_CELLS
HALLMARK_APICAL_SURFACE
HALLMARK_ANDROGEN_RESPONSE
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY
HALLMARK_ANGIOGENESIS
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_SPERMATOGENESIS
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_NOTCH_SIGNALING
HALLMARK_ADIPOGENESIS
HALLMARK_WNT_BETA_CATENIN_SIGNALING
HALLMARK_XENOBIOTIC_METABOLISM
HALLMARK_TGF_BETA_SIGNALING
HALLMARK_MITOTIC_SPINDLE
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_APOPTOSIS
HALLMARK_P53_PATHWAY
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_COAGULATION
HALLMARK_PI3K_AKT_MTOR_SIGNALING
HALLMARK_HEME_METABOLISM
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_ALLOGRAFT_REJECTION
HALLMARK_HYPOXIA
HALLMARK_MYOGENESIS
HALLMARK_PROTEIN_SECRETION
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_FATTY_ACID_METABOLISM
HALLMARK_MTORC1_SIGNALING
HALLMARK_UV_RESPONSE_UP
HALLMARK_PEROXISOME
HALLMARK_APICAL_JUNCTION
HALLMARK_COMPLEMENT
HALLMARK_CHOLESTEROL_HOMEOSTASIS
HALLMARK_BILE_ACID_METABOLISM
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_KRAS_SIGNALING_UP
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_GLYCOLYSIS
HALLMARK_HEDGEHOG_SIGNALING
HALLMARK_UV_RESPONSE_DN
HALLMARK_ESTROGEN_RESPONSE_EARLY
DKO (vs WT), mouse LSK
SKO (vs WT), mouse LSK
RKO (vs WT), mouse LSK
DKO (vs WT), HL-60
RKO (vs WT), HL-60
SKO (vs WT), HL-60
NES
-2 0 2
B
FDR >= 0.25 in both
FDR < 0.25 in either
NES in SKO (vs WT), HL-60
NES in SKO (vs WT), mouse LSK
NES in RKO (vs WT), HL-60
NES in RKO (vs WT), mouse LSK
NES in DKO (vs WT), HL-60
NES in DKO (vs WT), mouse LSK

**Supplementary Figure 14. Transcriptome analysis in HL-60 cell lines and mouse LSK cells.**

**A**, Heatmap of NES values in GSEA analysis of SKO/RKO/DKO-mouse LSK cells or HL-60 cell lines compared with WT using hallmark gene sets. **B**, Scatterplots of NES values comparing SKO/RKO/DKO with WT in HL-60 cell lines and mouse LSK cells. Gene sets with FDR < 0.25 in either HL-60 cell lines or mouse LSK cells are indicated in red color.


A
log2FC (cohesin mut vs WT)
SE TE SE TE SE TE
AML (Tyner)
AML (TCGA)
MDS (Shiozawa)
B
log2FC (STAG2 mut vs WT)
C
Gene expression between cohesin complex mutated vs WT cases
AML (TCGA)
MDS (Shiozawa)
PI > 30 *
10 < PI <= 30 ****
PI <= 10
Cumulative fraction
log2FC (cohesin mut vs WT)
D
Gene expression between STAG2 mutated vs WT cases
AML (Tyner)
AML (TCGA)
MDS (Shiozawa)
PI > 30 ****
10 < PI <= 30 **
log2FC (STAG2 mut vs WT)
E
K-means clustering
1
2
Mean expression
High-pausing genes (PI > 20)
Relative expression (Row z-score)
Mutation
STAG2
RUNX1
SRSF2
ASXL1
Number of SRSA mutations
F
Alternative 5' splice site (A5SS)
Alternative 3' splice site (A3SS)
Cassette exon (CE)
Mutually exclusive exons (MXE)
Retained intron (RI)
Alternative first exon (AFE)
Alternative last exon (ALE)
Tandem 3' UTR (UTR)
Inclusive/extended isoform
Exclusive isoform
Inclusive/extended isoform
Exclusive isoform
A5SS
A3SS
CE
MXE
RI
AFE
ALE
UTR
composite event
Number of events or structures (q-value < 0.2)
537
Sf3b1 K700E
Srsf2 P95H
SKO
RKO
DKO

**Supplementary Figure 15. An association between cohesin mutation and Pol II pausing in human MDS/AML and analysis of alternative splicing events in Stag2/Runx1 deficient LSK cells.**

**A-B**, Box plots showing expression changes of SE- and TE-associated genes identified in human CD34-positive HSPCs in cohesin (**A**) or *STAG2*-mutated (**B**) cases compared with WT cases in RNA-seq datasets from three independent MDS/AML cohorts. The vertical axis represents the log2(FC) in the indicated gene sets. **C-D**, Cumulative probability distributions of expression changes (log2FC) of genes grouped by pausing indices in cohesin-mutated cases (vs WT) (**C**) or *STAG2*-mutated cases (vs WT) (**D**) in RNA-seq datasets of MDS (Shiozawa et al., 2017) and AML (Ley et al., 2013; Tyner et al., 2018). *P*-values (vs genes with PI no more than 10) were calculated by one-sided Wilcoxon rank-sum test. **E**, K-means clustering analysis of RNA-seq dataset of AML (Tyner et al., 2018) using expression of genes with PI >20. Each row and column represent each gene and case, respectively. The Color scales are normalized along each row. Mean expression of genes (PI >20) is shown in the above of the heatmap, and presence or absence of each mutation and number of SRSA mutations are shown in the below. **F**, Numbers of alternative splicing events identified between SKO/RKO/DKO-derived LSK cells and WT cells. Numbers of alternative splicing events identified in cells having Sf3b1 K700E and Srsf2 P95H, major mutations in splicing factors, are also shown. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$.

**Inventory of Supplementary Tables**

**Supplementary Table S1. Diagnosis of patients with MDS/AML in correlation analysis.**

**Supplementary Table S2. Characteristics of MDS patients with STAG2, RUNX1, SRSF2, and/or ASXL1 mutations.**

**Supplementary Table S3. Correlations between mutations in patients with MDS/AML.**

**Supplementary Table S4. The antibodies used in the FACS experiments.**

**Supplementary Table S5. sgRNA sequences used for gene knockout and PCR primers for confirmation.**

**Supplementary Table S6. Genotypes of CRISPR-KO HL-60 clones.**

**Supplementary Table S7. Mutations in human MDS/AML RNA-seq datasets.**

**Supplementary Table S8. Up-regulated genes in RNA-seq of SKO LSK cells (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S9. Down-regulated genes in RNA-seq of SKO LSK cells (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S10. Up-regulated genes in RNA-seq of SKO LSK cells in BMT experiments (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S11. Down-regulated genes in RNA-seq of SKO LSK cells in BMT experiments (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S12. Up-regulated genes in RNA-seq of RKO LSK cells in BMT experiments (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S13. Down-regulated genes in RNA-seq of RKO LSK cells in BMT experiments (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S14. Up-regulated genes in RNA-seq of DKO LSK cells in BMT experiments (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S15. Down-regulated genes in RNA-seq of DKO LSK cells in BMT experiments (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S16. Gained peaks in ATAC-seq of SKO LSK cells (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S17. Lost peaks in ATAC-seq of SKO LSK cells (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S18. Gained peaks in ATAC-seq of SKO CMP cells (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S19. Lost peaks in ATAC-seq of SKO CMP cells (vs WT, edgeR; FDR < 0.05).**

**Supplementary Table S20. Super-enhancers of mouse HSPCs identified using H3K27ac ChIP-seq of WT HSPCs.**

**Supplementary Table S21. Information about the external ChIP-seq datasets used in this study.**