Supplemental Methods

Immunophenotyping: Morphology, Histopathology, Flow cytometry, Southern blot, cytology, and immunohistochemistry. Peripheral blood was collected from the retro-orbital sinus for analysis.[1] Blood and bone marrow smears and spleen touch preps were stained using a modified Wright-Giemsa stain (Sigma Aldrich, St. Louis, MO). Formaldehyde-fixed tissues from a subset of leukemic mice were paraffin-embedded, sectioned, and stained with hematoxylin and eosin (H&E) for histopathological analysis. Leukemic cells were viably frozen for later flow cytometric analysis. All leukemias were analyzed by flow cytometry using common T-cell (CD4, CD8, and TCRß), B-cell (B220), and myeloid surface markers (Mac-1/CD11b, Gr-1). A subset of 21 leukemias was analyzed using additional markers as follows. Cryopreserved tumor cell samples were briefly thawed in a 37°C water bath and then incubated for 15 min at 37°C in DMEM containing 20% FBS, 10 U/ml Heparin (Sigma, St. Louis, MO), and 0.25 mg/ml DNAse 1 (Roche, Basel, Switzerland). Cells were pelleted and resuspended in Hank's balanced salt solution (HBSS) without Ca/Mg, 2% FBS, 2.5% cell dissociation buffer (GIBCO, Invitrogen, Carlsbad, CA), 100 U/ml Penicillin G, and 100 mg/ml streptomycin. Tumor cell suspensions were preincubated with antibodies to FcgRII/III (BD PharMingen, San Jose, CA) for 3 min to prevent nonspecific binding of labeled antibodies to the cell surface. Cells were washed and then stained with monoclonal antibodies conjugated with fluorescein-isothiocyanate (FITC), phycoerythrin (PE), or TRI-COLOR (TC) for 20 min at 4°C. Cells were stained with antibodies to CD3, CD4, CD5, CD8a, CD11b (Mac1), CD19, CD41, CD45, CD45R (B220), CD71, CD86, CD90.2, CD117 (Kit), CD138, IgD, IgM, IgK, TCR a/b, TCR g/d, I-Ab, Ly-6G (Gr1), Ly-71 (F4/80), and Ly-76 (Ter119). All antibodies were purchased from BD PharMingen, except CD11b TC, F4/80 PE, CD19 TC, CD45 TC, CD4 TC, IgK FITC, TCR a/b

Biotin, and TCR g/d FITC which were from CalTag (Invitrogen, Carlsbad, CA). Cells were analyzed with a FACSCalibur flow cytometer (Becton Dickinson, Franklin Lakes, NJ) in four-color mode using CellQuest Pro software. Cryopreserved specimens were thawed and partially depleted of dead cells with the use of Histopaque 1119 (Sigma) used according to the manufacturer's directions. Cytospins were prepared and stained with a Wright's Giemsa stain with azure B. Photographs were taken on a Nikon Eclipse 80i microscope with a Nikon Digital Sight camera using NIS-Elements F2.30 software at a resolution of 2560 ◊ 1920 (Melville, NY). Using Adobe Photoshop CS2, images were re-sized and set at a resolution of 300 pixels/inch, autocontrast was applied, and unsharp mask was used to improve image clarity. Southern blots were performed on genomic DNA from leukemias with probes detecting IgH and TCRß gene rearrangements as previously described.[2,3] For a smaller subset of animals, additional characterization of leukemia cells was performed by immunohistochemical staining of tissue sections for CD3, B220, and myeloperoxidase (MPO) using rat anti-human CD3 (Serotec, Raleigh, NC), rat anti-mouse B220 (BD Pharmingen, San Diego, CA), and rabbit anti-human MPO (Thermo Fisher Scientific, Fremont, CA) antibodies. Detection of CD3 and B220 was achieved using a Rat on Mouse HRP-Polymer detection system from Biocare Medical (Concord, CA), and of MPO using the Anti-Rabbit Envision+ System-HRP Labelled Polymer detection system from Dako (Carpinteria, CA). All leukemias were immunophenotyped with flow cytometry. The B and T cell receptor rearrangements status provided complementary information, as did a white blood cell (WBC) count for each leukemia using a hemocytometer and observation of an enlarged spleen and/or thymus. For leukemias with inconclusive data, Wright-Giemsa slides plus H&E slides were examined, with the addition of IHC in limited cases.

Note: for simplicity, T lymphoblastic leukemia/lymphoma is denoted as "leukemia" in the text although many animals exhibited only a lymphomatous component of T lymphoblastic disease.

Linker-mediated PCR of M4070-induced leukemia. In brief, 1-3 micrograms of spleen or lymph node genomic DNA from leukemic mice were digested overnight with both *SauIIIA1* and *Tsp509I* in a 50 µl reaction. A splinkerette linker sequence was made by heating equimolar amounts of the Splinklong primer (5'-CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGC TGAAT GAGACTGGTGTCGACACTAGTGG-3') and the appropriate enzyme primer (Splinkshort-SauIIIA1 5'- GATCCCACTAGTGTCGACACCAGTCTCTAATTTTTTTTTTCA AAAAAA-3' or Splinkshort-Tsp509I 5'- AATTCCACTAGTGTCGACACCAGTCTCTAAT TTTTTTTTTCAAAAAAA-3') to 95°C for 3 min and allowing them to cool to RT. The DNA fragments were ligated to the linkers in a 40 µl volume using 160 units of T4-Ligase at 16°C overnight. Another overnight digestion with EcoRV was performed before two rounds of PCR. All enzymes were purchased through New England Biolabs (Ipswich, MA). First round of PCR used a linker-specific primer (Splink1 5'-CGAAGAGTAACCGTTGCTAGGAGAGACC-3') and a primer specific to the LTR of the M4070 virus (LTR5 5'-GCGTTACTTAAGCTAGCTTGCCAAACCTAC-3'). A 1 µL aliquot of the first PCR reaction product (1:50 dilution) was the template for the second PCR (nested PCR) using a nested linker primer (Splink2 5'-GTGGCTGAATGAGACTGGTGTCGAC-3') and a nested LTR primer (LTR3 5'-GCTAGCTTGCCAAACCTACAGGTGG-3'). Purification with the QIAquick PCR purification kit (Qiagen, Valencia, CA) was performed after each step followed by elution in 30 µl deionized water.

Sequence processing and annotation. Briefly, Fasta formatted sequences were obtained from ABI trace files using phred[4] with the -alt_trim option. All raw sequences were scanned for LTR and linker recognition sequences using EMBOSS Vectorstrip[5] with successively less stringent parameters (10%, 15%, and finally 20% mismatches allowed) until the maximum number of constructs (i.e., LTR: 'GCTAGCTTGCCAAACCTA CAGGTGGGGTCTTTCA' and linker sequence: 'CCACTAGTGTCGACACCAGTCT CATTCAG') were recognized and trimmed off. Only 12,312 sequences that had a matched LTR with an insert sequence of at least 16 base pairs were carried further for mapping to the mouse genome. These trimmed insert sequences were mapped to the mouse genome (NCBI build 37, Ensembl release 55) using BLASTN (DeCypher's TeraBLASTN, Active Motif), requiring query sequences to align within 5 bp of the end of the LTR sequence that was trimmed. Additionally, the query sequence was required to match with at least 95% identity. Ambiguous sequences that mapped to multiple genomic loci were removed. 7927 uniquely mappable inserts remained. All insertion sequences from a single mouse were grouped into a single library, regardless of sequencing date or tissue of origin (e.g., thymus, lymph node, spleen). Redundant sequences from the same mouse that mapped to the same genomic position within 5 bp were coalesced into 4012 non-redundant (NR) insertions. 2267 NR insertions from different mice that mapped to the same genomic location were removed as potential artifacts (e.g., endogenous pro-viral sites or PCR contaminants). Unambiguously mapped non-redundant insertions were assigned to clusters of common integration sites (CIS) if the local density of insertions in a given window size exceeded that which would be expected by chance, as determined by Monte Carlo simulation. CIS defined by only two mice were removed. In the end, there were 700 NR insertions and 20 CIS for WT infected mice (threshold criteria: 3 or more insertions in 200,000 bp), 999 NR insertions and 37 CIS for *MLL-AF9* mice (3+ in

120,000 bp) and 1699 NR insertions and 69 CIS for the combined data set (number 4+ in 250,000 bp or 3+ in 50,000 bp).

Bioinformatic analysis. Kaplan-Meier Survival curves were generated using Genedata® Expressionist Analyst http://www.genedata.com/products/expressionist/ (Basel, Switzerland). The Log-Rank Test was used to assess significance. Visualization of genomic regions was generated using Caryoscope v_0_3_9. CIS visualization was done via the University of California Santa Cruz (UCSC) mouse genome browser created by the Genome Bioinformatics Group of UC Santa Cruz (http://genome.ucsc.edu/cgi-bin/hgGateway). CIS-associated genes were compared to the Wellcome Trust Sanger Institute's Consensus Cancer Gene database (http://www.sanger.ac.uk/genetics /CGP/Census/). For the comparison of CIS to the data contained within the Catalog of Somatic Mutations in Cancer (COSMIC) database v45_260110,[6] human orthologs of genes associated with the murine CIS were obtained by using the Ensembl biomart web application http://www.ensembl.org/biomart/martview/. A local copy of COSMIC database version was queried to generate the counts used to look for association by Fisher's exact test. Ingenuity Pathways Analyses http://www.ingenuity.com/ was used to look for Functions and Canonical Pathways overrepresented in the CIS list. Heatmaps were generated using Java Treeview version 1.1.3. Cytoscape v2.6.1 was used to generate networks based on associations. The LiftOver web application on the University of California Santa Cruz website http://genome.ucsc.edu/cgi-bin/hgLiftOver was used to define orthologous regions of the murine genome that were found to be deleted or amplified in human tumors as previously described.[7,8]

Quantitative real-time PCR. RNA was isolated from spleen, thymus and/or lymph nodes using TRIzol reagent (Invitrogen). 1 μg RNA was then reversed transcribed using Superscript III reverse transcriptase (RT) with random hexamer primer (Invitrogen) to generate cDNA templates. Quantitative real-time reverse transcription-PCR (qRT-PCR) was performed using Quantitect SYBR Green (Qiagen) on a Mastercycler® ep realplex machine (Eppendorf, Westbury, NY). Primers used to amplify each product available upon request. Program is as follows: 95ºC for 15 minutes, 40 cycles of 94ºC for 15 seconds, 60ºC for 30 seconds and 72ºC for 30 seconds followed by a melting curve. A quantitative analysis was performed using the realplex software (Eppendorf) using *beta-actin* as the endogenous control. All reactions were performed in duplicate. The ΔΔCT values were calculated for each sample and normalized to results from cDNA of hematopoietic tissues from wild-type mice.

Supplemental References

1. Kim WI, Matise I, Diers MD, Largaespada DA. RAS oncogene suppression induces apoptosis followed by more differentiated and less myelosuppressive disease upon relapse of acute myeloid leukemia. *Blood*. 2009;113(5):1086-1096.
2. Tonegawa S. The molecules of the immune system. *Sci Am*. 1985;253(4):122-131.
3. Kronenberg M, Goverman J, Haars R, et al. Rearrangement and transcription of the beta-chain genes of the T-cell antigen receptor in different types of murine lymphocytes. *Nature*. 1985;313(6004):647-653.
4. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998;8(3):175-185.
5. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 2000;16(6):276-277.
6. Forbes SA, Bhamra G, Bamford S, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet*. 2008;Chapter 10:Unit 10 11.
7. Mullighan CG, Goorha S, Radtke I, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*. 2007;446(7137):758-764.
8. Walter MJ, Payton JE, Ries RE, et al. Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci U S A*. 2009;106(31):12950-12955.
9. Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res*. 2004;32(Database issue):D523-527.

10. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer*. 2004;4(3):177-183.
11. Wolff L, Koller R, Hu X, Anver MR. A Moloney murine leukemia virus-based retrovirus with 4070A long terminal repeat sequences induces a high incidence of myeloid as well as lymphoid neoplasms. *J Virol*. 2003;77(8):4965-4971.
12. Verhaak RG, Wouters BJ, Erpelinck CA, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*. 2009;94(1):131-134.

Supplemental Tables

Table S1. Significant pair-wise comparisons of phenotype-phenotype and phenotype-candidate gene by Fisher's Exact Test. The association being tested is in the left column while the *P* value is found in the right column in significant figures. (A) Associations between phenotype variables and designations are listed with a significance of < 0.05. Lymphoid disease was further classified as L1 to indicate lymphoid disease with a high CD4+, CD8+ population, or a double positive CD4/CD8+ population. L2 lymphoid disease indicates mice with T cell lineage lymphoid disease by immunohistochemistry but also expressed the Mac1 myeloid marker on the blast cell surface. Gene receptor rearrangements by Southern blot are shown as positive if listed as JB1, JB2, or JH, or any rearrangement. Spl WT indicates the spleen weight where 0 = up to 0.2 grams (g), 1 = over 0.2 g, 2 = over 0.8 g (10x normal), and 3 = over 2 g. WBC indicates the white blood cell count where 0 = normal (up to 20), 1 = over 20 and 2 = over 100. (B) Associations between phenotype variables and CIS-associated genes are listed with a significance of < 0.05.

Table S2. CIS list from infected wild type mice. The list of CISs identified when only considering insertions from WT mice for the insertion site analysis. Columns 1-2 are the chromosome and position of each CIS according to the NCBI build 37. Column 3 shows the

range of the CIS in kilobases (KB). Column 4 shows the number of wild type mice with contributing insertions to each CIS. Columns 5-7 refer to the gene in bold in column 8. Column 5 indicates if the CIS-associated gene has previously been identified in the RTCGD,[9] column 6 indicates if the gene has been identified as a cancer gene in the Cancer Gene Census,[10] and column 7 indicates if a mutation in the human homolog of the gene has been identified as a recurring somatic mutation in cancer (COSMIC).[6] Column 8 shows all the genes in or near the CIS region that may be affected by proviral insertions. The genes in bold indicate the gene whose transcriptional start site is closest to the median of the CIS region and what were called the CIS-associated candidate genes, annotated using Ensembl release 55.

Table S3. CIS list from infected *Mll-AF9* mice. The list of CISs identified when only considering insertions from *Mll-AF9* mice for the insertion site analysis. Columns 1-2 are the chromosome and position of each CIS according to the NCBI build 37. Column 3 shows the range of the CIS in kilobases (KB). Column 4 shows the number of *Mll-AF9* mice with contributing insertions to each CIS. Columns 5-7 refer to the gene in bold in column 8. Column 5 indicates if the CIS-associated gene has previously been identified in the RTCGD,[9] column 6 indicates if the gene has been identified as a cancer gene in the Cancer Gene Census,[10] and column 7 indicates if a mutation in the human homolog of the gene has been identified as a recurring somatic mutation in cancer (COSMIC).[6] Column 8 shows all the genes in or near the CIS region that may be affected by proviral insertions. The genes in bold indicate the gene whose transcriptional start site is closest to the median of the CIS region and what were called the CIS-associated candidate genes, annotated using Ensembl release 55.

Table S4. CIS-associated candidate genes are overrepresented in Cancer Functions and Pathways according to the Ingenuity Pathway Analysis. The significant function or canonical pathway are listed under 'Function Annotation.' *P* values are listed to show significance. # genes means the number of candidate genes involved in that process. The genes involved in each function or pathway are listed under 'Gene Names.' (A) Top 10 most enriched Functions in candidate gene list. (B) Top 5 most enriched Canonical Pathways in candidate gene list.

Table S1

A. Phenotype assocation	*P* value	B. Phenotype-Candidate Gene Association	*P* value
L1__Any_Southern	1.84E-03	Any_Southern__Rasgrp1	2.51E-02
L1__JB1_Southern	7.27E-03	Any_Southern__Rras2	7.21E-03
L1__JB2_Southern	5.87E-05	JB1_Southern__Myc	7.27E-03
L1__SPL_WT_0	3.64E-03	JB1_Southern__Notch1	4.91E-03
L1__WT	1.43E-09	JB1_Southern__Rasgrp1	2.87E-02
L2__JB1_Southern	2.22E-03	JB1_Southern__Rras2	6.17E-05
L2__WBC_1	2.71E-02	JB2_Southern__Fgfr3	2.54E-02
L2__WT	2.00E-07	JB2_Southern__Galk1	9.86E-03
lymphoid__Any_Southern	5.17E-05	JB2_Southern__Rasgrp1	2.87E-02
lymphoid__JB1_Southern	4.00E-06	JB2_Southern__Rras2	1.56E-02
lymphoid__JB2_Southern	6.51E-04	SPL_WT_0__Ets1	4.21E-02
lymphoid__SPL_WT_0	3.58E-04	SPL_WT_0__Tmem173	9.25E-03
lymphoid__WT	2.20E-16	SPL_WT_2__Bcl11a	5.00E-02
WT__Any_Southern	2.93E-04	SPL_WT_2__Mn1	2.28E-02
WT__JB1_Southern	1.73E-06	SPL_WT_3__Ift57	1.47E-02
WT__JB2_Southern	9.97E-04	SPL_WT_3__Tmem49	4.30E-02
WT__SPL_WT_0	1.74E-05	WBC_0__Gimap8	4.74E-02
JB1_Southern__JB2_Southern	2.52E-03	WBC_0__Mknk2	2.93E-02
JB1_Southern__JH_Southern	8.31E-05	WBC_1__Tmem131	3.86E-02
JB1_Southern__SPL_WT_0	5.21E-02	WBC_2__Gmfg	3.79E-02
JB2_Southern__JH_Southern	3.03E-02		
WBC_0__SPL_WT_0	7.16E-05		
WBC_1__SPL_WT_1	6.12E-05		
MandL__MllAF9	1.10E-05		
MandL__SPL_WT_2	4.10E-04		
MandL__SPL_WT_3	3.21E-02		
MllAF9__SPL_WT_1	1.74E-05		
MllAF9__SPL_WT_2	1.48E-10		
MllAF9__SPL_WT_3	9.09E-11		
myeloid__MllAF9	1.82E-08		
myeloid__SPL_WT_1	2.59E-02		
myeloid__SPL_WT_2	3.07E-02		
myeloid__SPL_WT_3	3.23E-03		
myeloid__WBC_2	2.53E-04		

Table S2

Chr	CIS Position	CIS Range (KB)	wild type mice with insert	RTCGD	Cancer Gene Census	COSMIC	All genes in/near CIS
2	103601727-103784302	183	3				***Nat10***, *Lmo2, Caprin1, BX537331.1, AL928544.7 (miRNA), AL928544.5 (miRNA)*
2	117168903-117249623	81	10	YES			***Rasgrp1***
2	26315310-26400791	85	10	YES	YES	YES	***Notch1***
4	32377612-32513306	136	3	YES		YES	***Bach2***
5	108136326-108186428	50	19	YES			***Gfi1,*** *Evi5*
5	34022070-34040387	18	3	YES	YES	YES	***Fgfr3***, *Tacc3*
6	127104034-127281434	177	5	YES	YES	YES	***Ccnd2***, *AC161597.1*
7	121285480-121316277	31	11	YES		YES	***Rras2***, *Copb1*
8	10910498-11099413	189	3				***3930402G23Rik,*** *Irs2*
9	32416427-32424429	8	4	YES		YES	***Ets1***
10	20811983-20972634	161	9				***AC153556.5 (miRNA),*** *Myb*
10	41691081-41790233	99	3				***Armc2***
11	11587106-11679166	92	8	YES	YES	YES	***Ikzf1,*** *RP23-373H2*
11	115872987-115992429	119	3				***Galk1***, *Trim65, Itgb4, Mrpl38, Trim47, Uncl3d, Wbp2, Unk, H3f3b*
11	68174513-68245020	71	4	YES		YES	***Pik3r5,*** *Ntn1, AL606831.2 (miRNA)*
15	61815622-61995612	180	8	YES	YES		***Myc,*** *Pvt1*
17	29534935-29631800	97	7	YES	YES	YES	***Pim1,*** *Fgd2*
17	47649930-47669158	19	4				***Taf8***
18	35911356-36089864	179	3				***Tmem173,*** *Cxxc5*
19	37514331-37569373	55	6	YES			***Hhex***, *Exoc6*

Table S3

Chr	CIS Position	CIS Range (KB)	*Mll-AF9* mice with insert	RTCGD	Cancer Gene Census	COSMIC	All genes in/near CIS
1	173849175-173851212	2	3	YES			***Slamf6***
2	11547443-11577841	30	3	YES			***Il2ra***
2	90919486-90925879	6	3			YES	***Slc39a13,*** *Sfpi1*
2	101463219-101464091	1	4				***B230118H07Rik,*** *Rag2*
2	165781678-165837434	56	3				***RP23-108D12.5,*** *Ncoa3*
2	167750534-167785437	35	3				***Ptpn1***
3	94945270-95035953	91	3				***Tnfaip8l2 ,*** *Cdc42se1, Sema6c, Gabpb2, Mlt11, Gm128, Bnipl, Lysmd1, Scnm1*
5	108078270-108182241	104	5	YES			***Gfi1,*** *Evi5, Rpap2*
5	111861745-111980636	119	7	YES	YES		***Mn1,*** *C130026L21Rik*
6	48630644-48720382	90	5				***Gimap4,*** *Gimap cluster, Al854703, Zfp775*
7	19858212-19895233	37	5				***Fosb,*** *Rtn2, Vasp, C79127*
7	29165895-29228553	63	4				***Gmfg,*** *Plekhg2, Zfp36, Med29, Paf1, Samd4b*
7	121304579-121354575	50	5	YES		YES	***Rras2****, Copb1*
7	152231919-152233702	2	3	YES	YES		***Ccnd1,*** *Oraov1*
8	10856578-10910346	54	3				***AC116499.9****, 3930402G23Rik*
8	131049443-131116479	67	3				***Nrp1***
10	20907125-20956148	49	6	YES	YES		***Myb,*** *AC153556.2 (miRNA)*
10	76990210-77085753	96	3				***Itgb2,*** *181008A18Rik, Pttglip, Sumo3, Ube2g2*
10	79452089-79515906	64	3				***Cnn2****, Abca7, Hmha1, Gpx4, ORF61, Polr2e, Sbno2*
10	92532859-92627678	95	3				***4930485B16Rik,*** *Cdk17*
11	24098976-24156602	58	6	YES	YES	YES	***Bcl11a***
11	51713117-51817285	104	4				***Phf15,*** *Cdkn2aipn1, Ube2b, Cdkl3*
11	79467420-79566569	99	3				***Rab11fip4,*** *mmu-mir-193 (miRNA), mmu-mir-365-2 (miRNA), AL731726.1*
11	86407328-86438598	31	3				***Tmem49***
11	120491729-120493009	1	3			YES	***Mafg***
14	70133928-70229580	96	3				***Chmp7****, Tnfrsf10b, Rhobtb2, Pebp4*
15	61815640-61816583	1	3	YES	YES		***Myc***
15	62000405-62022073	22	3				***Pvt1***
15	66646986-66693812	47	3				***Tg,*** *Sla*
15	80398045-80402302	4	3				***Enthd1,*** *Grap2*
15	96373651-96486074	112	3				***AC123606.11,*** *Slc38a1, Slc38a2*
15	97443447-97559406	116	3				***Rpap3,*** *Pp11r*
16	32517604-32549358	32	3				***Zdhhc19***
16	49839166-49938369	99	3				***Cd47,*** *Gm5486, AC107830.1, AC107830.2*
17	29534871-29639166	104	6	YES	YES	YES	***Pim1,*** *Fgd2*
17	52420984-52490384	69	3				***AC121600.2,*** *AC121600.1 (miRNA)*
19	37566911-37569767	3	4				***Exoc6****, Hhex*

Table S4

A. Top 10 most enriched functions in CIS-associated candidate genes

Function Annotation	*P* value	# genes	Gene Names
transformation of cells	2.65E-11	19	*BCL11A, BCL2L1, CCND1, CCND2, ETS1, FGFR3, FOSB, HHEX, JDP2, MYB, MYC, NOTCH1, PIK3CD, PIK3R5, PIM1, RASGRP1, REL, RRAS2, STAT3*
quantity of lymphocytes	6.74E-10	16	*BCL2L1, CHST3, ETS1, FOSB, GFI1, IKZF1, IL2RA, MYC, NOTCH1, PIK3CD, PIM1, RASGRP1, REL, SLA, STAT3, TCF7*
developmental process of lymphocytes	8.04E-10	17	*BCL11A, BCL2L1, CCND1, CD74, ETS1, GFI1, IKZF1, IL2RA, MYB, MYC, NOTCH1, PIK3CD, PIM1, RASGRP1, REL, STAT3, TCF7*
developmental process of blood cells	9.34E-10	20	*BCL11A, BCL2L1, CCND1, CCND2, CD74, ETS1, GFI1, HHEX, IKZF1, IL2RA, LMO2, MYB, MYC, NOTCH1, PIK3CD, PIM1, RASGRP1, REL, STAT3, TCF7*
developmental process of T lymphocytes	1.08E-09	15	*BCL11A, CCND1, CD74, ETS1, GFI1, IKZF1, IL2RA, MYB, NOTCH1, PIK3CD, PIM1, RASGRP1, REL, STAT3, TCF7*
differentiation of cells	7.25E-09	25	*BCL11A, BCL2L1, CCND1, CCND2, CD74, ETS1, EVPL, FGFR3, GFI1, HHEX, IKZF1, IL2RA, JDP2, LMO2, MAFG, MYB, MYC, NOTCH1, PIK3CD, PIM1, RASGRP1, REL, RRAS2, STAT3, TAF8*
transformation of eukaryotic cells	7.38E-09	15	*BCL11A, CCND1, CCND2, ETS1, FGFR3, FOSB, JDP2, MYB, MYC, NOTCH1, PIK3CD, RASGRP1, REL, RRAS2, STAT3*
development of leukocytes	7.38E-09	13	*BCL2L1, CCND1, CD74, ETS1, GFI1, IL2RA, MYB, NOTCH1, PIK3CD, PIM1, REL, STAT3, TCF7*
quantity of blood cells	7.86E-09	17	*BCL2L1, CHST3, ETS1, FOSB, GFI1, IKZF1, IL2RA, MYB, MYC, NOTCH1, PIK3CD, PIM1, RASGRP1, REL, SLA, STAT3, TCF7*
hematopoiesis	7.86E-09	17	*BCL11A, BCL2L1, CCND1, CD74, ETS1, GFI1, HHEX, IKZF1, IL2RA, MYB, MYC, NOTCH1, PIK3CD, PIM1, REL, STAT3, TCF7*

B. Top 5 most enriched canonical pathways in CIS-associated candidate genes

Function Annotation	*P* value	# genes	Gene Names
GM-CSF Signaling	1.30E-08	8	*RRAS2, BCL2L1, PIK3CD, CCND1, STAT3, ETS1, PIM1, PIK3R5*
Acute Myeloid Leukemia Signaling	2.00E-08	8	*RRAS2, MYC, PIK3CD, CCND1, STAT3, PIM1, TCF7, PIK3R5*
Prolactin Signaling	1.52E-05	6	*RRAS2, MYC, PIK3CD, STAT3, TCF7, PIK3R5*
HGF Signaling	4.09E-05	6	*RRAS2, PIK3CD, CCND1, STAT3, ETS1, PIK3R5*
Chronic Myeloid Leukemia Signaling	4.11E-05	6	*RRAS2, MYC, BCL2L1, PIK3CD, CCND1, PIK3R5*

Supplemental Figures

Figure S1. Survival analyses reveal female *Mll-AF9* female mice died significantly faster than male *Mll-AF9* mice, *Mll-AF9* mice die faster than WT mice regardless of phenotype, and phenotype does not alter latency of disease. Kaplan Meier plots are shown. (A) Survival of non-infected *Mll-AF9* mice. Female mice are represented in pink whereas male mice are represented in blue ($P = 0.021$). (B) Survival of M4070 infected *Mll-AF9*. Female mice are represented in pink whereas male mice are represented in blue ($P < 0.0001$). (C) Survival of mice with lymphoid disease. *Mll-AF9* mice are shown in red and WT mice are shown in black ($P = 0.0034$). (D) Survival of mice with myeloid disease. *Mll-AF9* mice are shown in red and WT mice are shown in black ($P = 0.0002$). (E) Survival of mice with both myeloid and lymphoid disease. *Mll-AF9* mice are shown in red and WT mice are shown in black ($P = 0.038$). (F) Survival of WT mice with lymphoid (blue), myeloid (green) or both myeloid and lymphoid leukemia (pink) (not significant). (G) Survival of *Mll-AF9* mice with lymphoid (blue), myeloid (green) or both myeloid and lymphoid leukemia (pink) (not significant). (H) Survival of mice with either the L1 (brown) or L2 (blue) lymphoid phenotype (not significant).

Figure S2. Breakdown of phenotype by experimental cohort and gene receptor rearrangements by phenotype. (A) Number and percent of leukemias with each phenotype by experimental cohort. The 'other' disease group represents mice that could not be categorized as myeloid, lymphoid or both. These include one mouse that had lymphoid disease but for which myeloid leukemia could not be definitively accessed (infected wild type cohort), histiocytic sarcoma, mast cell disease, or non-definitive

phenotypes in the infected *Mll-AF9* cohort, and a mouse with myeloid disease for which lymphoid disease could not be accessed (non-infected *Mll-AF9* cohort). (B) Number and percent of leukemias with gene receptor rearrangements in each phenotype. JB1 and JB2 represent T-cell receptor rearrangements while JH represents a B-cell receptor rearrangement by Southern blot. Any Southern means at least one positive rearrangement.

Figure S3. Multiple phenotypes were observed in the experimental cohorts. (A) Cytology of myeloid and T-cell diseases in *MLL-AF9* mice infected with retrovirus. Upper left: Bone marrow, mouse 400, with myeloid neoplasm showing substantial differentiation (similar to mouse 410 shown in Figure 1Di). Upper right: Bone marrow, mouse 429, with poorly differentiated acute myeloid leukemia (histopathology shown in Figure 1Dii). Lower left: Thymus, mouse 544, composite image of thymic cells diagnosed by morphology and flow immunophenotyping as T-cell acute lymphoblastic leukemia (similar to thymus of mouse 429 shown in Figure 1Div). Lower right: Bone marrow, mouse 522, with moderately differentiated acute myeloid leukemia (histopathology shown in Figure 1Dv). Scale bar represents 20 micrometers. Note: cells in upper and lower left panels with small condensed nuclei have an appearance consistent with apoptosis. (B) Immunophenotype of leukemia from a representative mouse with both myeloid and lymphoid disease. Flow cytometry data from mouse 529 splenocytes detected myeloid leukemia with mixed myeloid differentiation in which granulocytic cells expressed both Gr1 and CD11b (i) and in which monocytic cells lack Gr1 but expressed CD11b (i) and F4/80 (iii). In addition, T-cell leukemia was detected in which

a sizable population expressed the T-cell markers CD5 (ii) and CD4 (iv), but lacked CD3 (data not shown). Morphology of this leukemia is shown in Figure 1Dii-iv. (C) Immunophenotype of representative *Mll-AF9* myeloid leukemia. Splenocytes from mouse 522 expressed Gr-1 and CD11b (i), with a subpopulation expressing F4/80 (iii); few residual T-cells were present (ii, iv). Morphology of this leukemia is shown in Figure 1Dv-vi.

Figure S4. Southern blot showing proviral insertion sites in tissues from M4070 infected animals. DNA was isolated from enlarged spleens (Spl), and thymus (Thy) of infected animals. Genomic DNA (30 μg) was cut with *PvuII* and separated on a 0.8% agarose gel. After blotting onto Nitrocellulose, hybridization was performed with a M4070-specific LTR probe.[11] There is an internal *PvuII* fragment at 2371, as seen as the dark band in all the lanes. The remaining bands represent insertions derived from the M4070 virus, which are mostly unique in each sample.

Figure S5. Human microarray analysis shows variable expression of *MN1*, *FOSB* and *BCL11A* in patients with AML. All gene expression profiles using the Affymetrix U133Plus2 gene chips of the 461 AML cases[12] are available at the Gene Expression Omnibus (accession number GSE6891). For each probe set, the geometric mean of the hybridization intensities of all patients was calculated. The level of expression of each probe set in every sample was determined relative to this geometric mean and logarithmically transformed (log2). Pearson correlation clustering was performed using the Omniviz software (version 3.6), with the probe sets that had an absolute standard

deviation larger than 4.0 among all samples (1538 probe sets). Each row represents one human AML patient. The Pearson correlation values are represented in colors corresponding to positive correlation (red) and negative correlation (blue). The expression above baseline in all AMLs for each probe set, corresponding to a CIS-associated candidate gene, is shown in a black histogram per patient sample, scaled for each probe based on the highest value. (A) The large heat map shows all AML patients stratified by karyotype and expression status. Regions defined by specific translocations or genetic abnormalities are indicated on the right. Patients with *MLL-AF9* translocations are represented with a royal blue line to the right while all other patients with t(11q23) translocations are marked with an aqua blue line. (B) Patients with *MLL* rearrangements are enlarged and shown together, separated by a horizontal black line dividing patients with *MLL-AF9* translocations (royal blue) and all other *MLL* rearrangements (aqua blue). The *MLL-AF9* AMLs were positively correlated by expression level, though they were not clustered together.

Figure S6. Highly penetrant CISs occur more frequently in infected wild type leukemias than in infected *Mll-AF9* leukemias. Highly penetrant CISs are defined as having five or more leukemias contributing to the insertions sites of the CIS. The x-axis shows the number of highly penetrant CISs found in each leukemia while the y-axis shows the number of times each CIS number occurred. Black bars represent *Mll-AF9* leukemia while gray bars represent WT leukemia.

Figure S7. shRNA for human or mouse FOSB causes a decrease in protein expression. A) Western blots showing expression of FOSB in U937 cells. The left panel represents the U937 cell line stably expressing the shRNA named F10, which the right panel shows the results of the U937 cell line stably expressing the E3 shRNA, both targeting the *FOSB* gene. The numbers above each blot correspond to the days after plating shown in Figure 5A. +/- Dox represents the presence or absence of doxycycline, which induces the short hairpin against *FOSB*. 'Scrmb' stands for scrambled, as in the U937 cell line with an shRNA that should not target any gene for degradation. β-actin is shown to demonstrate equal loading. In both cell lines, levels were constant without the presence of doxycycline, when the shRNA was not induced, comparable to the U937 line with the scrambled shRNA. When the shRNA was induced with doxycycline, FOSB levels rapidly decreased. B) Western blots showing expression of Fosb in immortalized Rosa26-rtTA-M2 mouse embryonic fibroblasts (MEFs) transduced with TRMPV vectors containing Fosb shRNAs after doxycycline induction. Tubulin is shown to demonstrate equal loading. Control shRNA included Luc and Ren while RagMEFs were cells not transduced with shRNA. 1:10 indicates a dilution of virus containing the shRNA.

Figure S1

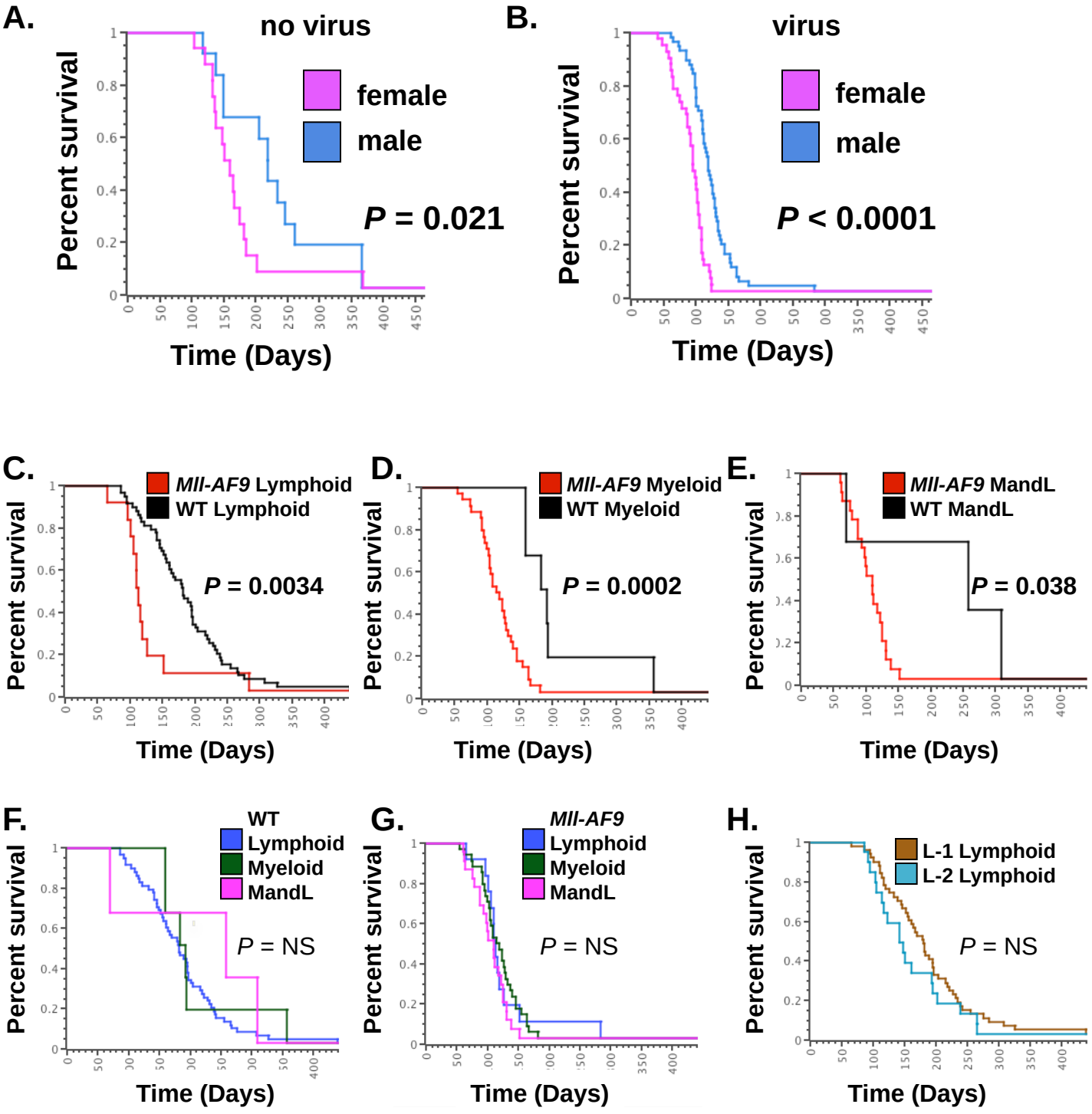

A.
no virus
female
male
P = 0.021
Percent survival
Time (Days)
B.
virus
female
male
P < 0.0001
Percent survival
Time (Days)
C.
Mll-AF9 Lymphoid
WT Lymphoid
P = 0.0034
Percent survival
Time (Days)
D.
Mll-AF9 Myeloid
WT Myeloid
P = 0.0002
Percent survival
Time (Days)
E.
Mll-AF9 MandL
WT MandL
P = 0.038
Percent survival
Time (Days)
F.
WT
Lymphoid
Myeloid
MandL
P = NS
Percent survival
Time (Days)
G.
Mll-AF9
Lymphoid
Myeloid
MandL
P = NS
Percent survival
Time (Days)
H.
L-1 Lymphoid
L-2 Lymphoid
P = NS
Percent survival
Time (Days)

Figure S2

A.

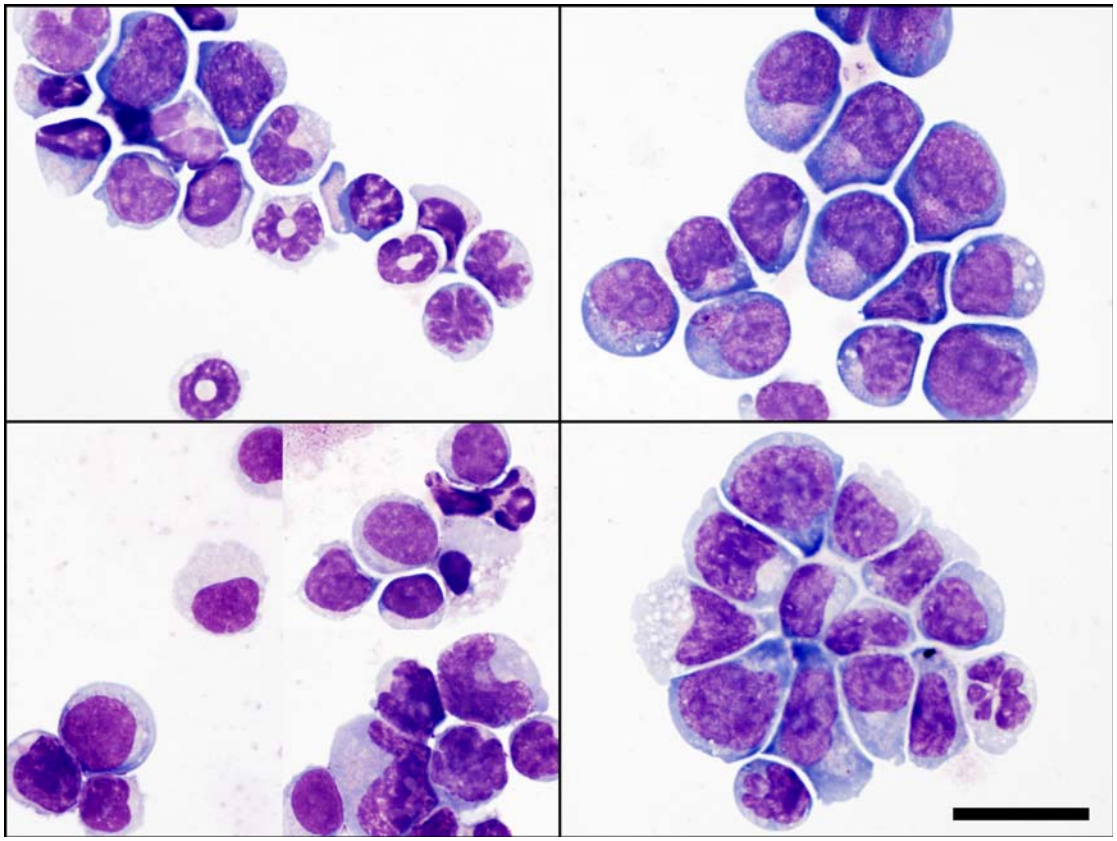
	Myeloid	Lym-phoid	L-1	L-2	MandL	Other
Infected WT	8/78 (10.3%)	64/78 (82.1%)	42/78 (53.8%)	22/78 (28.2%)	5/78 (6.4%)	1/78 (1.3%)
Infected *Mll-AF9*	37/85 (43.5%)	15/85 (17%)	13/85 (15.3%)	2/85 (2.4%)	27/85 (31.8%)	6/85 (7.1%)
Non-infected *Mll-AF9*	12/16 (75%)	0/16 (0%)	0/16 (0%)	0/16 (0%)	3/16 (18.8%)	1/16 (6.3%)

B.

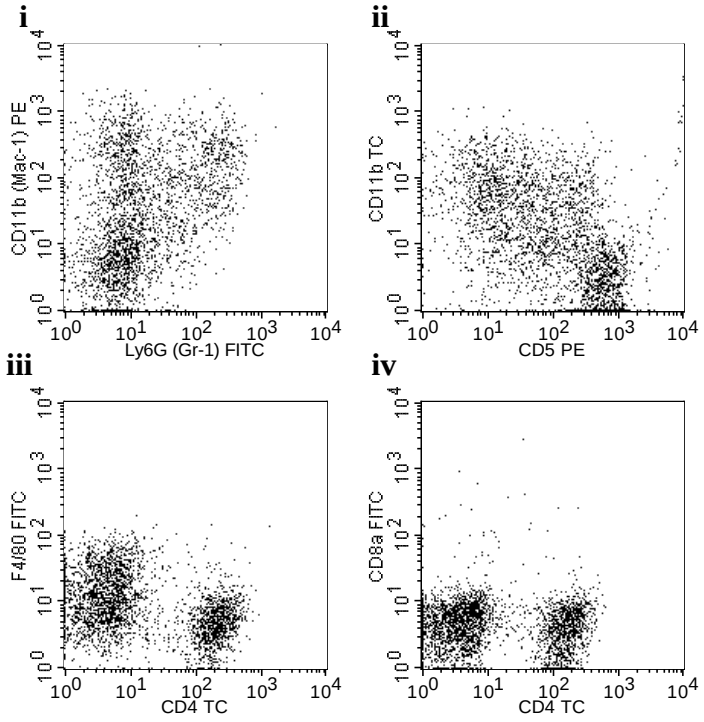
	Myeloid	Lymphoid	MandL
JB1 Southern	18/77 (23.4%)	25/48 (52.1%)	24/48 (50%)
JB2 Southern	23/62 (37.1%)	27/37 (73%)	19/33 (57.6%)
JH Southern	19/75 (25.3%)	14/47 (29.8%)	16/48 (33.3%)
Any Southern	36/77 (46.8%)	38/48 (79.2%)	40/48 (83.3%)

Figure S3

A.



B.



C.

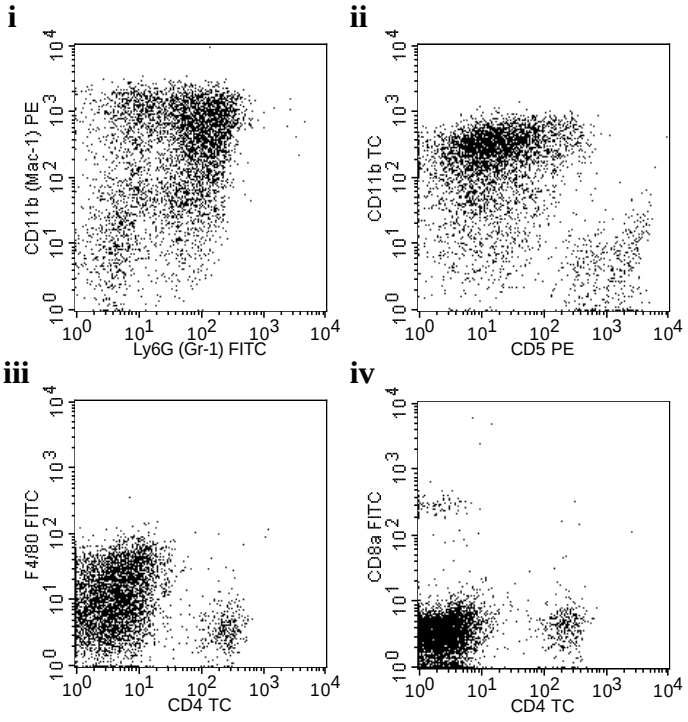


Figure S4

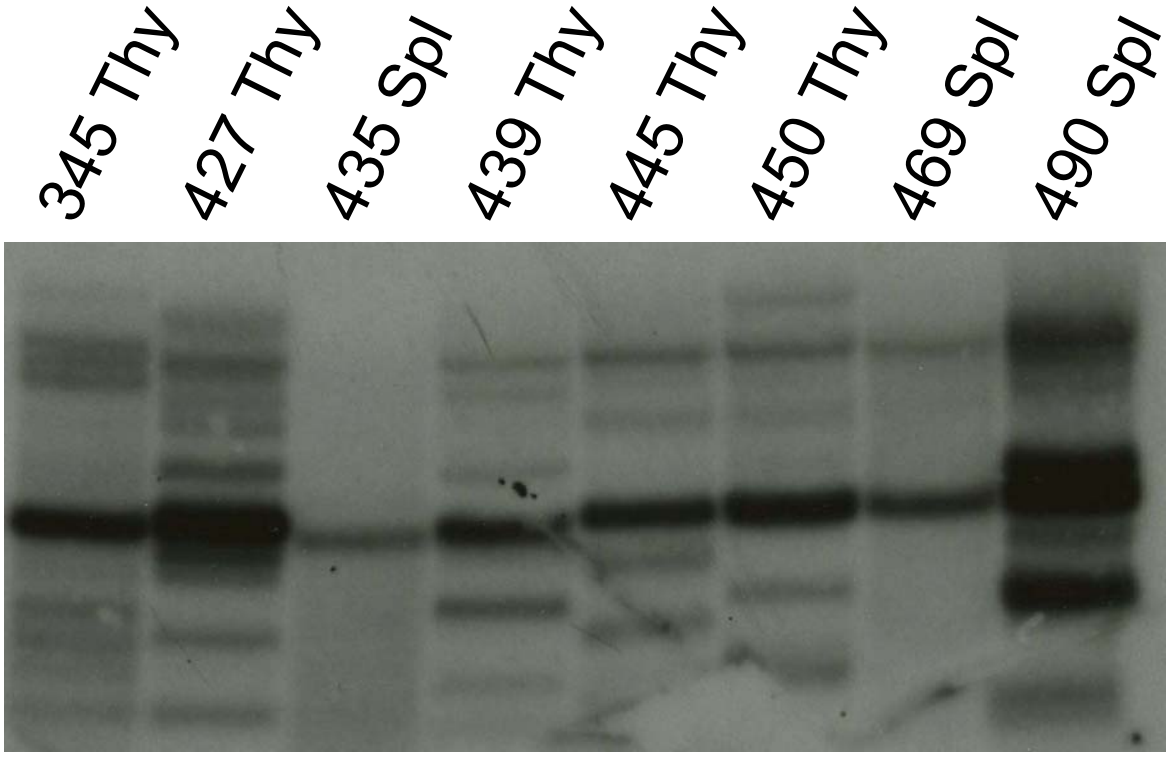


Figure S5

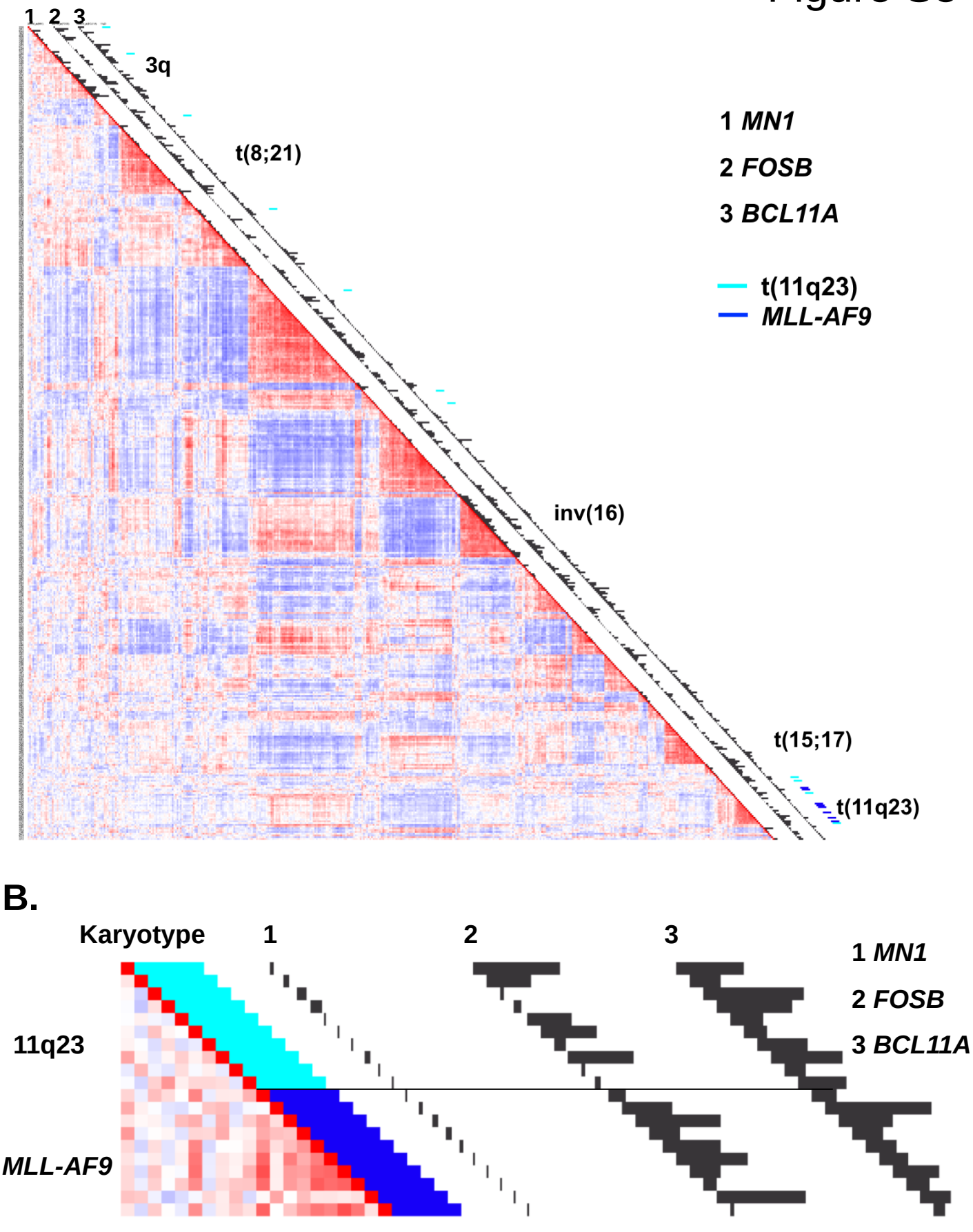

A.
1 2 3
3q
t(8;21)
inv(16)
t(15;17)
t(11q23)
1 MN1
2 FOSB
3 BCL11A
t(11q23)
MLL-AF9
B.
Karyotype
1
2
3
11q23
MLL-AF9
1 MN1
2 FOSB
3 BCL11A

Figure S6

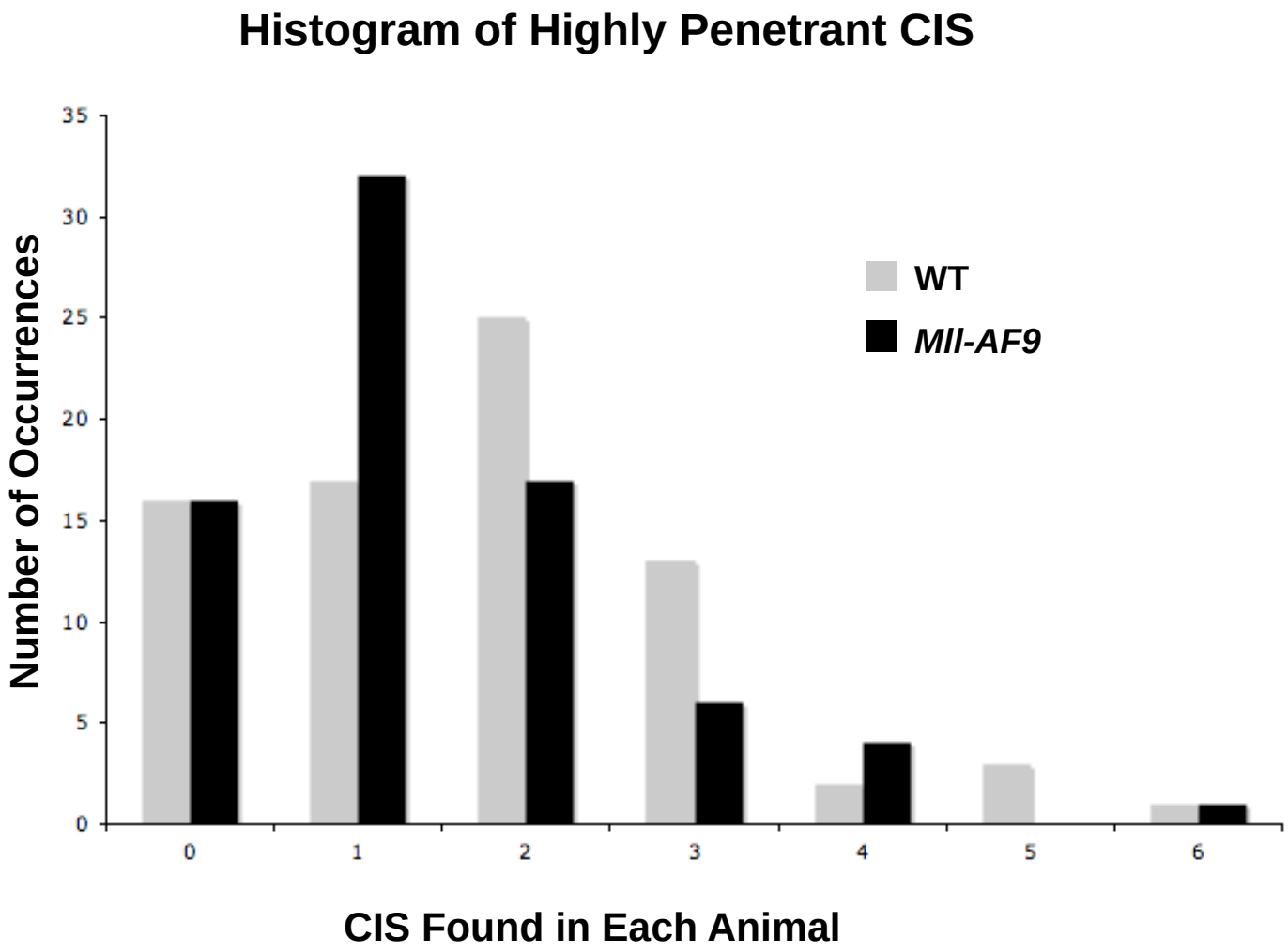
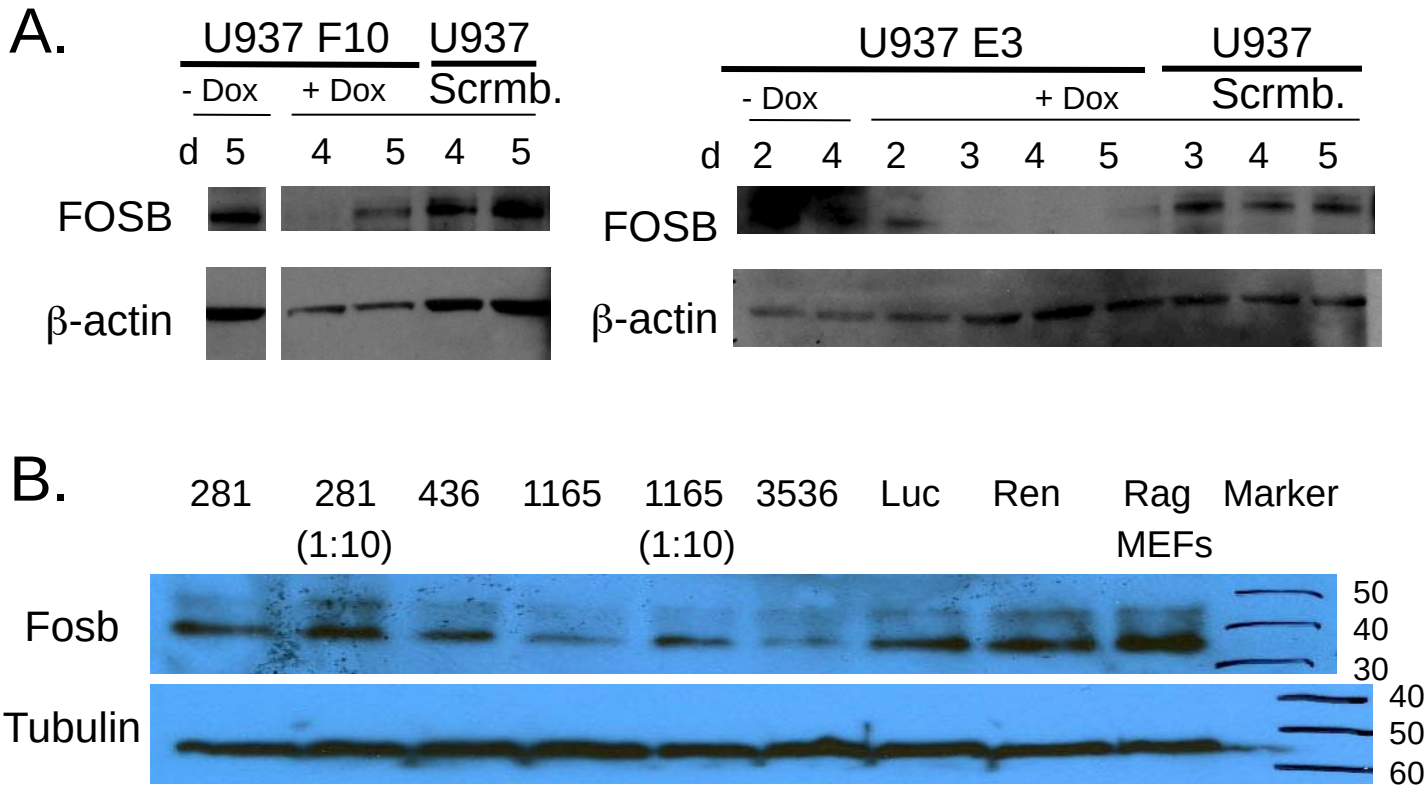

Histogram of Highly Penetrant CIS
Number of Occurrences
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
WT
Mll-AF9
CIS Found in Each Animal

Figure S7


A.
U937 F10
U937 Scrmb.
- Dox
+ Dox
d 5 4 5 4 5
FOSB
β-actin
U937 E3
U937 Scrmb.
- Dox
+ Dox
d 2 4 2 3 4 5 3 4 5
FOSB
β-actin
B.
281
281 (1:10)
436
1165
1165 (1:10)
3536
Luc
Ren
Rag MEFs
Marker
Fosb
Tubulin
50
40
30
40
50
60