Supplemental Material

Alu distance orders CpG-island promoters from evenly expressed genes to developmental regulators across human hematopoiesis

Contents

Supplemental Figures (this file, after the tables)

Supplemental Fig. S1. Distances from the TSS to the nearest retroelements

Supplemental Fig. S2. Per-type distances, Alu subfamilies, length-matched deviations and island-length quintiles

Supplemental Fig. S3. Evenness of expression by distance on each side and by promoter class

Supplemental Fig. S4. Promoter-class shares and gene-level differences across GTEx tissues

Supplemental Fig. S5. Proliferation and sample-set sensitivity

Supplemental Fig. S6. Associations with differentiation stage with counts-per-million normalization

Supplemental Fig. S7. Promoter classes in a mammary epithelial hierarchy

Supplemental Fig. S8. Histone marks by expression and methylation measured from the TSS and from the island edge

Supplemental Tables (this file)

Supplemental Table S1. Gene sets and filtering steps

Supplemental Table S2. Sensitivity to the TSS definition

Supplemental Table S3. Island length, Alu distance and the island track in the full gene set and the share set

Supplemental Table S4. Promoter classes

Supplemental Table S5. Expression datasets

Supplemental Table S6. ENCODE datasets

Supplemental Table S7. Sensitivity of the Alu-far results

Supplemental Table S8. Histone mark and methylation boundary by promoter group

Supplemental Methods and Supplemental References (this file, at the end)

Provided as separate files

Supplemental Data S1 (xlsx). Promoter class of the 31,962 genes, per-sample shares of transcript output by promoter group, and the gene contributions cited in the text.

Supplemental Data S2 (xlsx). Registration record of the 42 recorded predictions that bear on the reported results, with decision rules, verdicts and registration commits.

Supplemental Code (zip). Code that reproduces the figures and tables, with the stored result tables.

Supplemental Table S1. Gene sets and filtering steps.

Step	Description	Genes
1	Ensembl genes matched (UCSC Genome Browser, GRCh38)	63,512
2	Valid biotype annotation	46,303
3	UCSC Known Gene category, MIR host and LINC genes removed	43,111
4	GTEx v8 expression across 54 tissues (classification set)	31,962
5	Protein-coding genes in GENCODE v26	18,730
6	Present in the compiled hematopoietic expression table	18,357
7	Matched to gene symbols in Corces et al. 2016 (share set)	17,832
8	Ribosomal-protein genes analyzed separately (81)	17,751

Genes and coordinates from the UCSC Genome Browser (GRCh38), expression from GTEx v8 and protein-coding status from GENCODE v26. The share set was matched to gene symbols in Corces et al. 2016.

Supplemental Table S2. Sensitivity to the TSS definition.

	All genes (31,962)	Single-transcript genes (15,896)
Median Gini, CpG-island / non-CpG-island genes	0.355 / 0.794	0.438 / 0.862
Median CV, CpG-island / non-CpG-island genes	0.724 / 2.485	0.956 / 2.970
Median Gini by distance to the nearest retroelement (≤ 2 / 2–4 / 4–10 / > 10 kb)	0.330 / 0.462 / 0.613 / 0.763	0.388 / 0.547 / 0.683 / 0.787
Length-matched deviation in median Gini by distance to the nearest Alu (≤ 2 / 2–4 / 4–10 kb)	−0.022 / +0.048 / +0.189	−0.041 / +0.059 / +0.179

Deviation, observed minus permuted median within island-length quintiles (10,000 permutations, $P \leq 1 \times 10^{-4}$). Fig. 1D shows the same deviation by distance to the nearest retroelement of any type (−0.017 / +0.080 / +0.214 in all genes).

Supplemental Table S3. Island length, Alu distance and the island track in the full gene set and the share set.

	Full set, masked	Full set, unmasked	Share set, masked	Share set, unmasked
(a) Island length + Alu distance, partial R^2 of Alu distance	0.0556	0.0556	0.0545	0.0552
(a) Partial R^2 of island length	< 0.0001	0.0003	0.0006	0.0013
(a) + regional terms, partial R^2 of Alu distance	0.0134	—	—	—
(b) (a) + Alu count within 10 kb, partial R^2 of Alu distance	0.0111	0.0109	0.0112	0.0112
(b) Partial R^2 of island length	0.0001	0.0005	0.0010	0.0018
(b) + regional terms, partial R^2 of Alu distance	0.0094	—	—	—
(c) Eight-term model, partial R^2 of Alu distance	0.0075	0.0077	0.0081	0.0086
(c) Partial R^2 of island length	0.0024	0.0036	0.0045	0.0061
Model R^2, (a) / (b) / (c)	0.064 / 0.093 / 0.351	0.064 / 0.093 / 0.351	0.067 / 0.094 / 0.330	0.067 / 0.094 / 0.332
Genes in the models	13,409	13,403	10,869	10,864
Median island length, Alu-near 1 / Alu-mid 1 / Alu-far (bp)	820 / 1,340 / 1,923	900 / 1,406 / 2,011	812 / 1,316 / 1,927	891 / 1,364 / 2,052
CpG-island status unchanged, classes 1–9	—	99.96%	—	99.95%

Full set, 31,962 genes with RP genes in their classes. Share set, 17,832 protein-coding genes without the 81 RP genes (17,751). Partial R^2, linear models of the Gini coefficient among CpG-island genes of classes 1–9 under the current classification, with log distance to the nearest Alu and log island length (a), with the number of Alu elements within 10 kb of the TSS added (b), and with promoter configuration, GC content and CpG observed/expected ratio within 500 bp of the TSS, log mean expression across tissues and log gene length also added (c). Continuous terms were standardized. The partial R^2 of a term is the reduction in residual sum of squares when the term is added to the model with all other terms, divided by the residual sum of squares of that model. Alone, Alu distance gave R^2 0.064 (full set) and 0.066 (share set) and island length 0.009 and 0.013. With island length as a natural spline (4 df), the partial R^2 of Alu distance in the full set was 0.047 (a), 0.009 (b) and 0.006 (c), and with Alu distance in three bins (≤ 2, 2–4 and 4–10 kb, 13,176 genes) it was 0.057, 0.021 and 0.014. For the coefficient of variation, the corresponding values with the masked track were 0.020,

0.002 and 0.001. Regional terms (pre-specified, full set, masked track), Alu base fraction and GC fraction within 50 and 500 kb of the TSS and the number of protein-coding genes within 500 kb (13,409 genes). With the 50-kb terms alone the partial R^2 of Alu distance was 0.016 (a) and 0.012 (b), with the 500-kb terms alone 0.034 and 0.010, and for the coefficient of variation with all regional terms 0.003 and 0.002. Among the 13,320 CpG-island genes with the nearest retroelement within 10 kb, the median Gini rose from ≤ 2 to 2–4 and 4–10 kb within quintiles 1–4 of the 50-kb Alu fraction (0.448, 0.565 and 0.661 in quintile 1 and 0.306, 0.375 and 0.461 in quintile 4) and within all five quintiles of the 500-kb GC fraction, and the fifth Alu quintile, with 23 genes at 4–10 kb, was not evaluable (0.298, 0.351 and 0.380). Alu-near 1 genes made up 14% of the lowest and 70% of the highest Alu quintile, and Alu-far genes 5.6% and 0.2%. Island status, fraction of genes in classes 1–9 that keep a TSS-overlapping island with the unmasked track.

Supplemental Table S4. Promoter classes.

Group	Genes	Protein-coding genes	Island length (bp)	Gini	HKG H1 (%)	HKG H2 (%)
Alu-near 1	5,793	5,060	820	0.32	33.4	20.2
Alu-mid 1	1,122	929	1,340	0.43	19.6	11.9
Alu-far	323	224	1,923	0.69	8.9	5.4
Alu-near 2	3,702	3,166	798	0.34	28.1	16.4
Alu-mid 2	902	760	1,232	0.51	14.5	8.2
L1/LTR	1,572	1,253	831	0.46	20.4	8.9
Non-CpG island	18,548	7,338	—	0.79	8.4	4.6
All genes	31,962	18,730	—	—	20.1	11.6

Genes, island length (median) and Gini (median) refer to the 31,962-gene classification set, with RP genes in their classes. HKG, housekeeping genes as a percentage of the group's protein-coding genes (GENCODE v26) in the lists of Eisenberg and Levanon 2013 (H1) and Hounkpe et al. 2021 (H2); both lists consist of protein-coding genes. Alu-near 1 contains 44.9% (H1) and 47.0% (H2) of all listed housekeeping genes and 27.0% of protein-coding genes. In the share set without RP genes (17,751), the groups in the order of the table contain 4,834, 890, 217, 3,005, 741, 1,182 and 6,882 genes. Transcription factors (Lambert et al. 2018) made up 95 of 323 Alu-far genes (29.4%) and 383 of 5,793 Alu-near 1 genes (6.6%), Fisher odds ratio 5.88 (95% CI 4.48–7.68) and Mantel–Haenszel odds ratio over island-length quintiles 4.42 (3.31–5.90).

Supplemental Table S5. Expression datasets.

Dataset	Accession	Samples	Groups
GTEx v8	GTEx Portal	54 tissues	—
Compiled	GSE107011, GSE69239, GSE65996, GSE89322, GSE89778, GSE73211, GSE53094	146	pluripotent 8, HSPC 12, lineage progenitor 14, mature 112
Corces et al. 2016	GSE74246	49	HSC/MPP 8, progenitor 18, mature 23
E-MTAB-13862	ArrayExpress	4 donors	HSC, $CD34^{+}CD38^{+}$
Granja et al. 2019	GSE139369	3 donors	author cell labels; hematopoietic stem cells and progenitors (≥ 30 cells per donor and label)
Ainciburu et al. 2023	GSE180298	8 donors	author cell labels; hematopoietic stem cells, LMPP and lineage-restricted progenitors (≥ 30 cells per donor and label)
Pellacani et al. 2016	Supplementary RPKM tables of the article (raw data EGA EGAS00001000552)	1 pooled sample (6 donors) per population	basal, luminal progenitor and mature luminal cells; stromal cells as reference
Human Cell Atlas bone marrow (Hay et al. 2018)	Census of Immune Cells (HCAData, ExperimentHub EH2047)	8 donors	hematopoietic stem cells (4 donors), CMP, GMP, MEP, B, T, NK, monocytes

Supplemental Table S6. ENCODE datasets.

Cell	Assay	Experiment (files)	Donor
HPC	WGBS	ENCSR388RMS (ENCFF494SXJ, ENCFF851IIS)	ENCDO412QUR
HPC	H3K27me3	ENCSR355PUX (ENCFF804VSQ)	ENCDO412QUR
HPC	H3K4me3	ENCSR045QJH (ENCFF563ICK)	ENCDO930ZLO, ENCDO026AEL (cultured cord-blood $CD34^+$ cells)
T cell	WGBS	ENCSR663MXB (ENCFF434DHW, ENCFF534WDB)	ENCDO661BYS
B cell	WGBS	ENCSR284TCU (ENCFF701UYQ, ENCFF412HWN)	ENCDO661BYS
T cell	H3K4me3	ENCSR852FRR (ENCFF020QFY)	ENCDO190HXC
T cell	H3K27me3	ENCSR733QOZ (ENCFF506LCB)	ENCDO190HXC
B cell	H3K4me3	ENCSR000DQR (ENCFF192GXE)	ENCDO115AAA
B cell	H3K27me3	ENCSR404MOX (ENCFF480VDY)	ENCDO190HXC

GRCh38. WGBS, methylation files for the plus and minus strands. ChIP-seq, signal p-value bigWig files (HPC H3K4me3 and H3K27me3 pooled from two replicates, others single replicates). H3K27me3 peak files used for marked promoters, ENCFF716GWM (HPC), ENCFF407VKP (T cell) and ENCFF169OGM (B cell).

Supplemental Table S7. Sensitivity of the Alu-far results.

Analysis	Alu-far genes, all / share set	Corces AUC, TPM / CPM	ρ with proliferation, compiled, TPM / CPM	Partial ρ with stage, compiled / Corces (TPM)
Thresholds in use	323 / 217	0.969 / 1.000	+0.04 / −0.04	−0.588 / −0.579
Alu–Alu thresholds applied to Alu-mixed genes (pre-specified)	173 / 115	0.917 / 0.917	+0.43 / +0.34	−0.298 / −0.634
CCND2 removed	322 / 216	1.000 / 1.000	—	—
171 genes that are TFs or carry H3K27me3 in HPC	— / 171	0.969 / 0.958	—	—
Stress-response score added to the adjustment	323 / 217	—	—	−0.569 / −0.303
The same, *CCND2* removed	322 / 216	—	—	−0.576 / −0.038
Island-length-matched genes, registered match	— / 217	1.000 / 1.000	—	−0.202 / −0.757
Island-length-matched genes, 1,000 random matches (post hoc)	— / 217	0.46 / 0.50	—	—

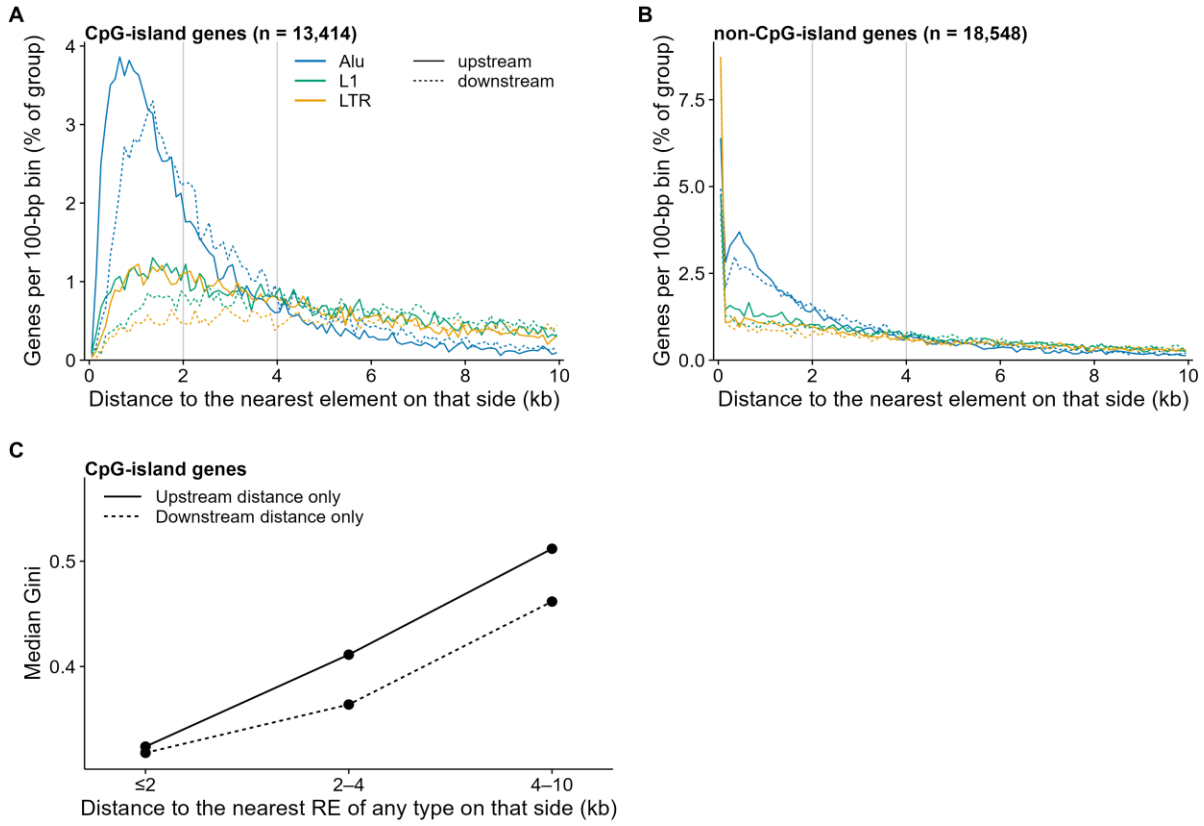
Thresholds in use, 7 kb upstream and 8 kb downstream for Alu–Alu genes and 5 and 4 kb for Alu-mixed genes. Under the alternative thresholds, the longer islands (median 2,247 bp), less even expression (median Gini 0.738), transcription-factor enrichment (Mantel–Haenszel odds ratio, MH OR, 8.15) and H3K27me3 enrichment (MH OR 6.84–9.40) kept their direction, and only the pre-specified limit of $|\rho| \leq 0.20$ with the proliferation score was not met (Supplemental Data S2). AUC, probability that an HSC/MPP sample has a higher share than a CMP, GMP or MEP sample (Corces, donor-centered). Partial ρ, Spearman correlation with differentiation stage after adjustment for the proliferation score. Stress-response score, mean z-score of 20 dissociation-response genes (van den Brink et al. 2017), none of which is in Alu-far. Alu-far genes, in the 31,962-gene classification set and in the 17,832-gene share set, in which AUCs and correlations were computed. TF, transcription factor in Lambert et al. 2018 (version 1.01), H3K27me3, ENCODE HPC peak within 1 kb of the TSS (ENCFF716GWM, Supplemental Table S6). Island-length-matched genes, CpG-island genes of classes 1–3, 5, 6, 8 and 9 in the share set matched one to one to the Alu-far

genes on masked island length (median 1,927 bp in both), with the registered prediction that their Corces AUC would be below 0.75 with TPM or CPM (not supported). In this match *HSP90AB1*, *NPR3* and *DUSP1* supplied 80% of the difference with TPM, and 16% of the 217 matched genes were TFs, against 43% of the 217 Alu-far genes. Random matches, the same number of genes sampled within 20 island-length bins, with medians shown, Corces AUC ≥ 0.75 in 25% (TPM) and 28% (CPM) of draws and compiled AUC medians 0.26 and 0.25. Genes matched to Alu-near 1 genes on island length (4,785) gave Corces AUC 0.813 and 0.938, opposite to Alu-near 1 (0.010 and 0.000), and medians of 0.75 and 0.91 over 1,000 random matches. —, not computed.

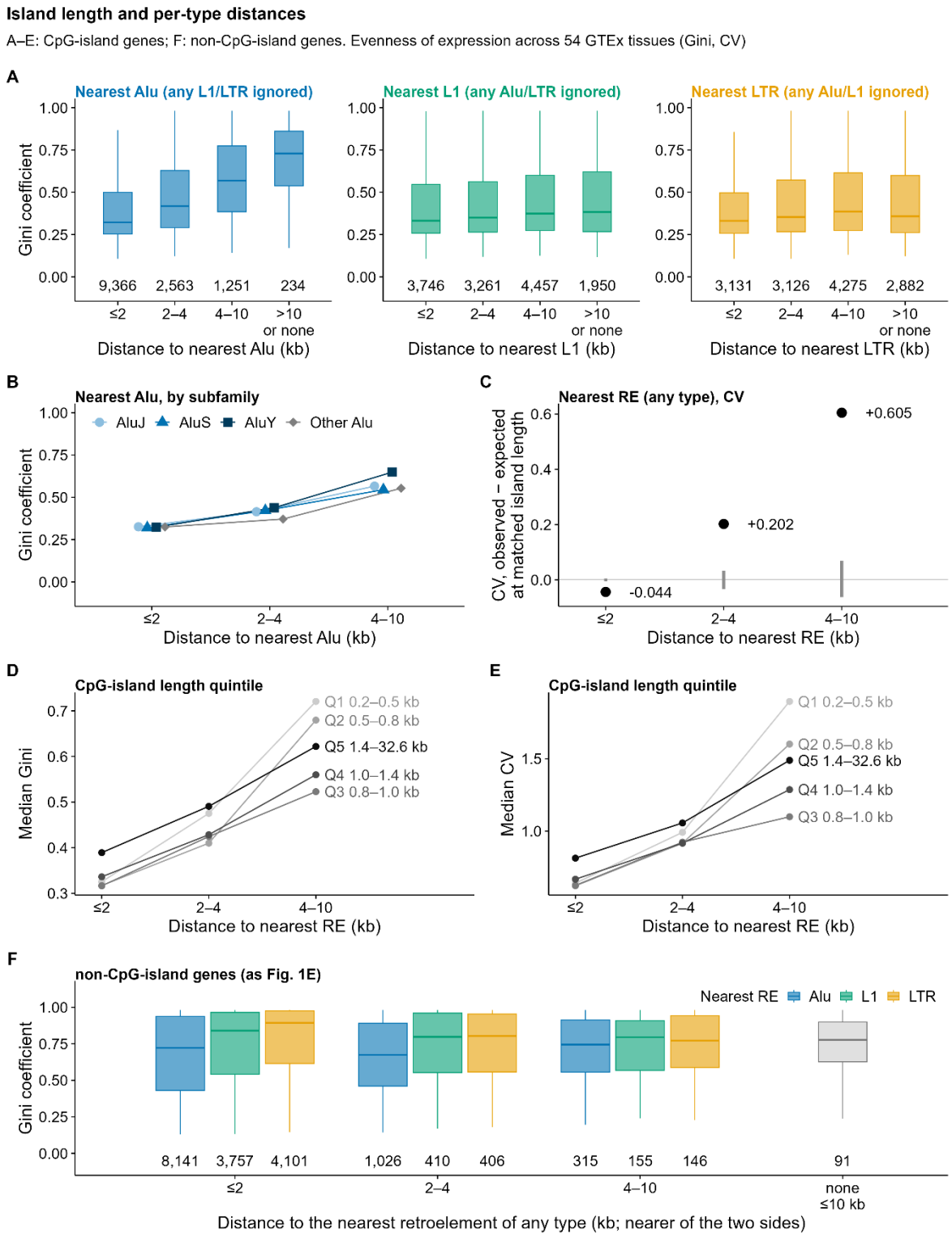
Supplemental Table S8. Histone mark and methylation boundary by promoter group.

Group	H3K27me3-marked promoters, MH odds ratio	Boundary of the top minus the bottom expression tertile (bp; positive, farther from the TSS)	Cox hazard ratio per 50 expression-percentile points
Alu-near	0.33–0.36	+500 to +700	0.65–0.73
Alu-mid	1.93–1.98	+400 to +500	0.85–0.88
Alu-far	5.99–6.99	+100 to +900	1.40–1.61
L1/LTR	1.83–1.93	—	—

Values are ranges over HPC, T cells and B cells. MH, Mantel–Haenszel. Groups, Alu-near (classes 1 and 5), Alu-mid (2, 3 and 6), Alu-far (4 and 7) and L1/LTR (8 and 9), RP genes excluded. Odds ratios against all other CpG-island genes over island-length quintiles (all autosomes), all intervals excluding 1. Boundary, first 100-bp bin from the TSS whose mean methylation of CpGs outside repeats is at least 50% (Methods), with Kaplan–Meier medians and Cox models on genes measured by Corces et al. (all autosomes). Bootstrap intervals of the boundary shift exclude 0 for Alu-near and Alu-mid genes and include 0 for Alu-far genes in all cell types. Hazard ratios below 1 indicate a boundary farther from the TSS with higher expression, and all intervals exclude 1. Interaction ratios against Alu-near were 1.21–1.31 for Alu-mid and 2.18–2.26 for Alu-far genes ($P < 0.001$), and adjustment for the TSS-to-edge distance changed hazard ratios by at most 0.11. Histone signal over TSS ± 2.5 kb relative to the mean of the four groups (Fig. 6A, even-numbered autosomes) was 1.07–1.09 (Alu-near), 0.94–1.01 (Alu-mid), 0.36–0.43 (Alu-far) and 0.70–0.71 (L1/LTR) for H3K4me3 and 0.57–0.70, 1.56–1.83, 3.30–4.25 and 1.38–1.57 in the same order for H3K27me3. The divisors, the mean signal p-value over TSS ± 2.5 kb of the 5,783 pooled genes, were 26.3 (HPC), 51.2 (T cells) and 35.3 (B cells) for H3K4me3 and 2.67, 1.02 and 0.943 for H3K27me3. —, not computed.

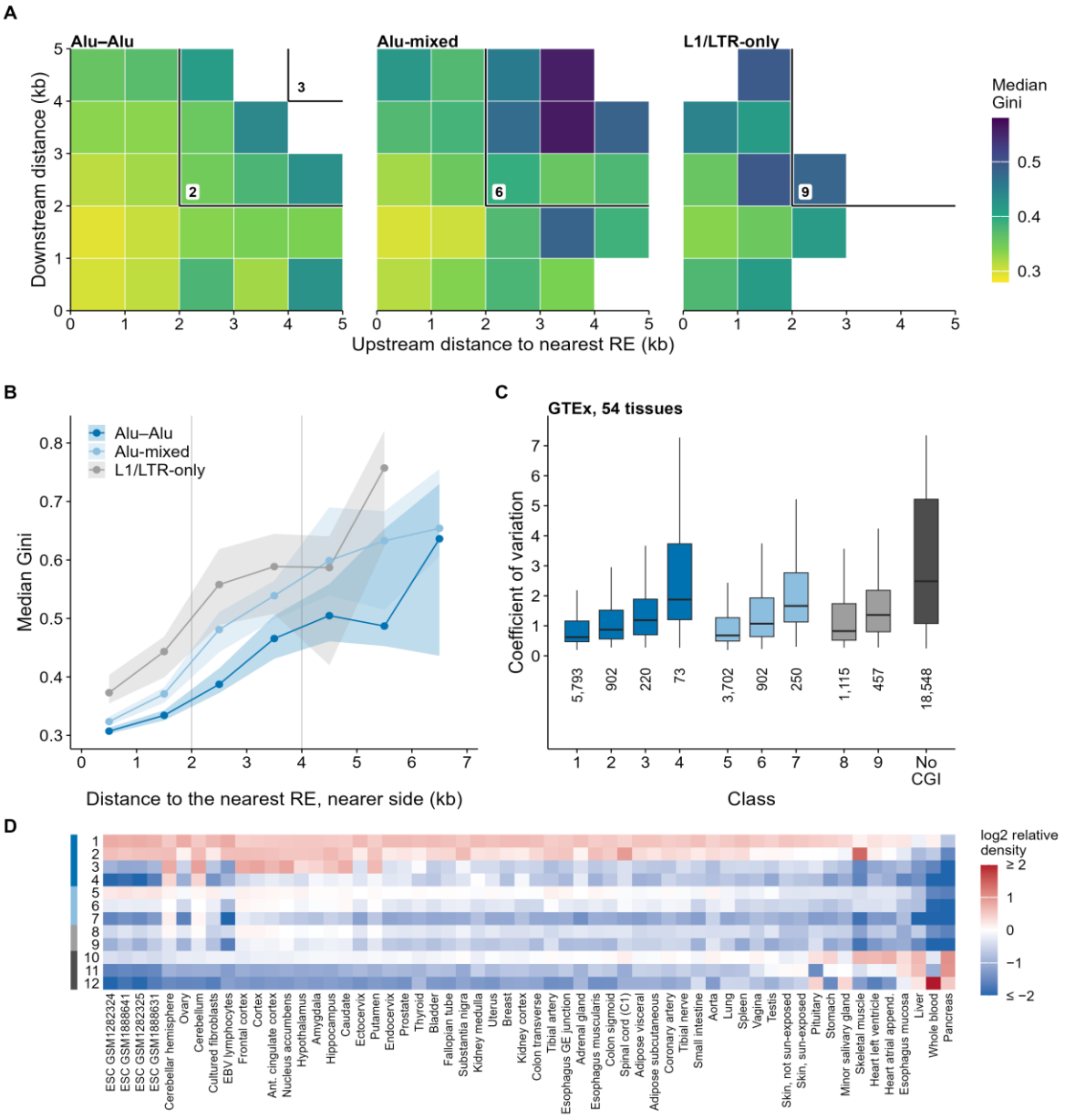


Supplemental Fig. S1. Distances from the TSS to the nearest retroelements. (A, B) Distance from the TSS to the nearest Alu, L1 and LTR element on each side, each type considered separately, for CpG-island genes (A; *n* = 13,414) and non-CpG-island genes (B; *n* = 18,548), as the percentage of genes per 100-bp bin. The first bin includes elements overlapping the TSS; genes without an element within 10 kb are not shown. Solid lines, upstream; dotted lines, downstream; gray lines, 2 and 4 kb. (C) Median Gini coefficient of CpG-island genes by distance to the nearest retroelement of any type on the upstream or the downstream side, each side considered separately. TSS, transcription start site.

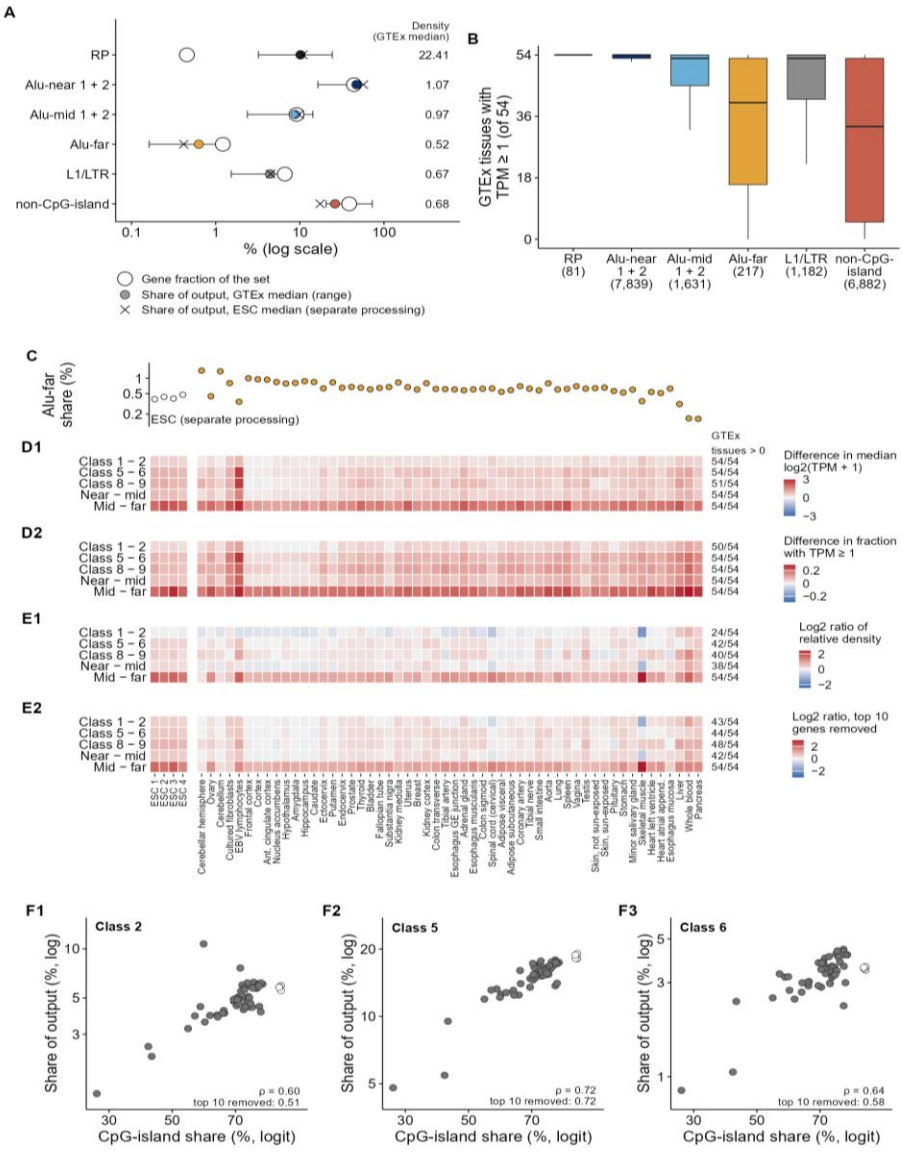


Supplemental Fig. S2. Per-type distances, Alu subfamilies, length-matched deviations and island-length quintiles. (A) Gini of CpG-island genes by distance to the nearest Alu, L1 or LTR element, each considered separately. (B) Median Gini of

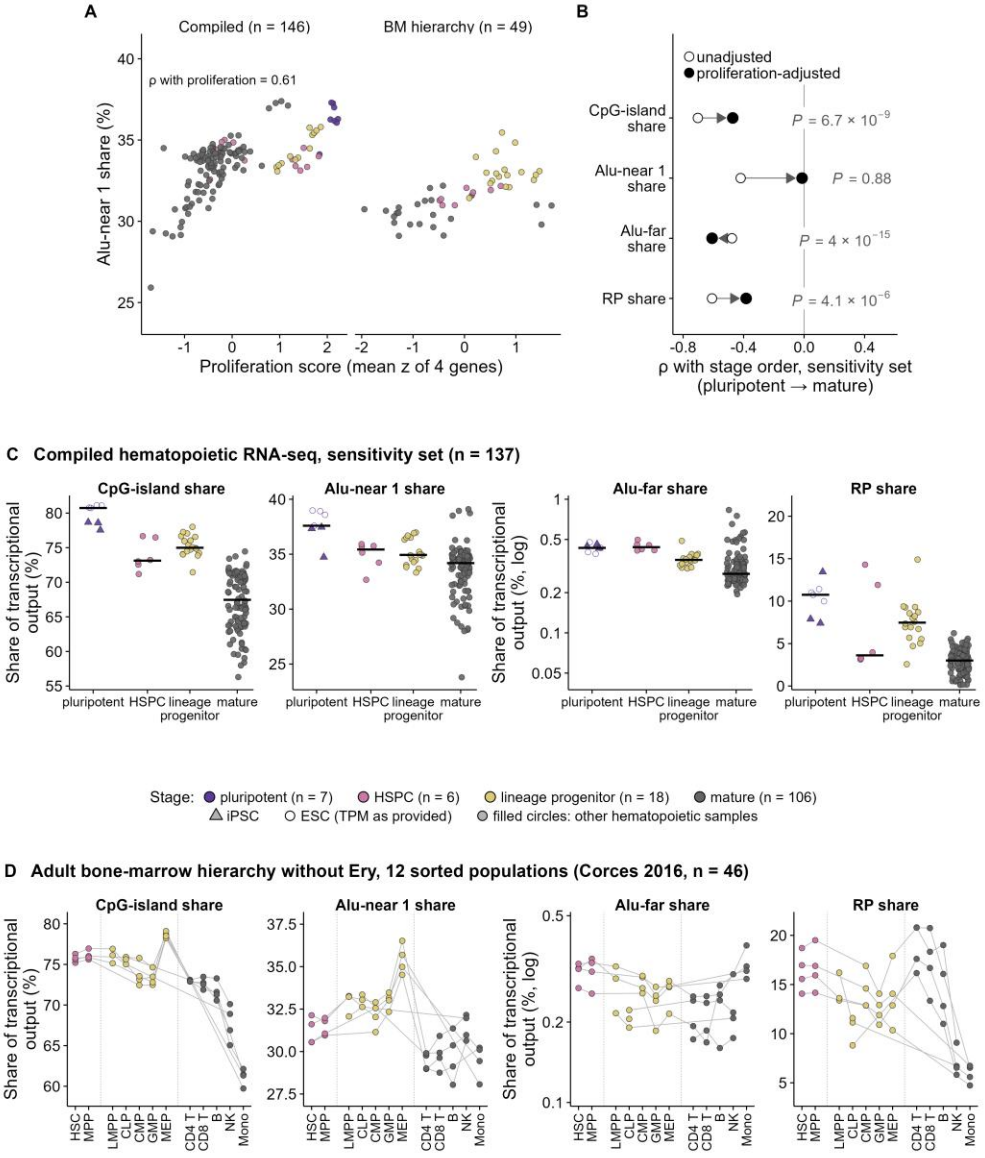
CpG-island genes by distance to the nearest Alu, by the subfamily of that Alu (post hoc). AluJ, n = 3,468. AluS, n = 7,677. AluY, n = 1,278. Other Alu (mainly FLAM and FRAM), n = 753. Lines are offset horizontally for clarity. Within island-length quintiles, the order held in all five quintiles for AluJ and AluS, whereas AluY and other Alu had too few genes at 2–4 or 4–10 kb in four quintiles. (C) Observed minus expected median coefficient of variation (CV) of CpG-island genes by distance to the nearest retroelement, as in Fig. 1D. (D, E) Median Gini (D) and CV (E) of CpG-island genes by distance to the nearest retroelement within each island-length quintile. (F) Gini of non-CpG-island genes by distance to and type of the nearest retroelement, as in Fig. 1E. RE, retroelement.



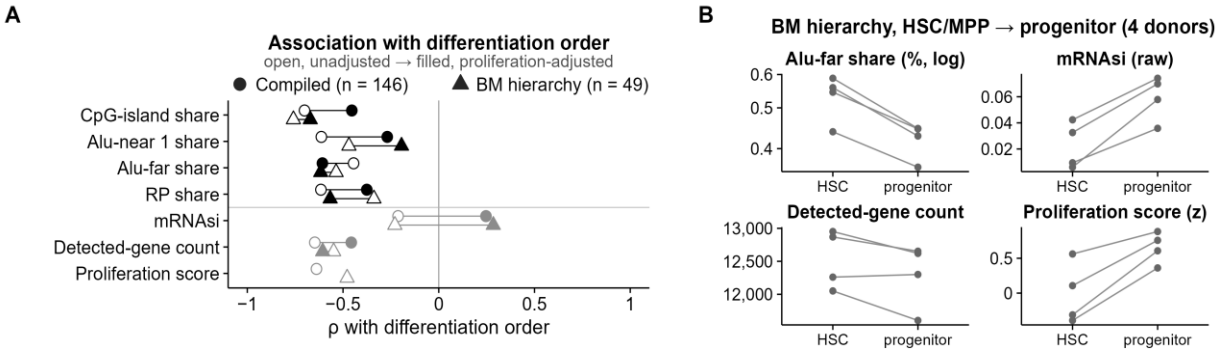
Supplemental Fig. S3. Evenness of expression by distance on each side and by promoter class. (A) Median Gini coefficient of CpG-island genes by upstream (x) and downstream (y) distance to the nearest retroelement in 1-kb bins, for Alu–Alu, Alu-mixed and L1/LTR-only configurations; black lines, lower bounds of classes 2, 3, 6 and 9. (B) Median Gini by distance to the nearest retroelement on the nearer side; bands, 95% bootstrap intervals; gray lines, 2 and 4 kb. (C) Coefficient of variation across 54 GTEx tissues by class; numbers, genes per class. (D) Relative transcript density (share of output divided by fraction of genes, log2) of classes 1–12 in four ESC samples and 54 GTEx tissues. ESC, embryonic stem cell.



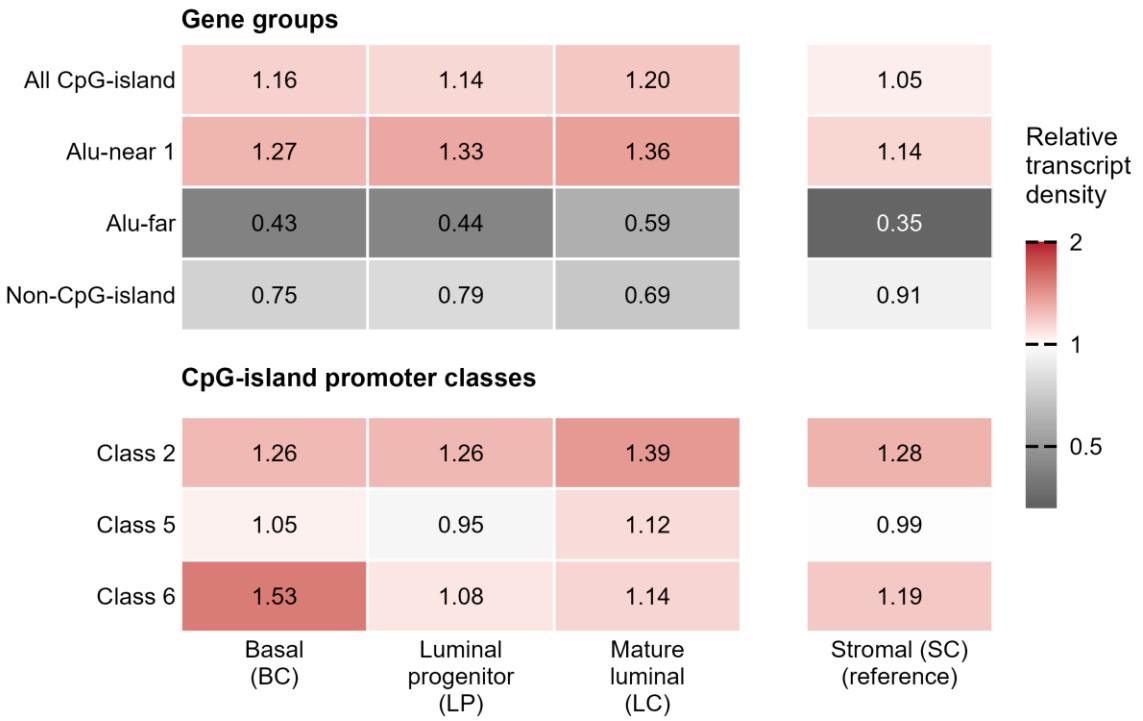
Supplemental Fig. S4. Promoter-class shares and gene-level differences across GTEx tissues. (A) Fraction of genes (open circles) and share of transcript output in GTEx tissues (filled circles, median and range) and in ESC samples (crosses) for promoter groups of the 17,832-gene set; numbers, relative transcript density (GTEx median). Alu-near 1 + 2, classes 1 and 5 (7,839 non-RP genes); Alu-mid 1 + 2, classes 2, 3 and 6 (1,631 genes). (B) Number of GTEx tissues with TPM ≥ 1 by group; numbers in parentheses, genes. (C) Alu-far share by tissue. (D) Differences between adjacent classes in median log2(TPM + 1) (D1) and in the fraction of genes with TPM ≥ 1 (D2) in each tissue; right, tissues with a positive difference. (E) Log2 ratio of relative density between adjacent classes, all genes (E1) and with each class's ten largest contributors removed (E2). (F) Shares of classes 2, 5 and 6 against the CpG-island share across tissues; ρ, Spearman correlation with all genes and with the ten largest contributors removed; open circles, ESC samples. ESC, embryonic stem cell; RP, ribosomal protein; TPM, transcripts per million.



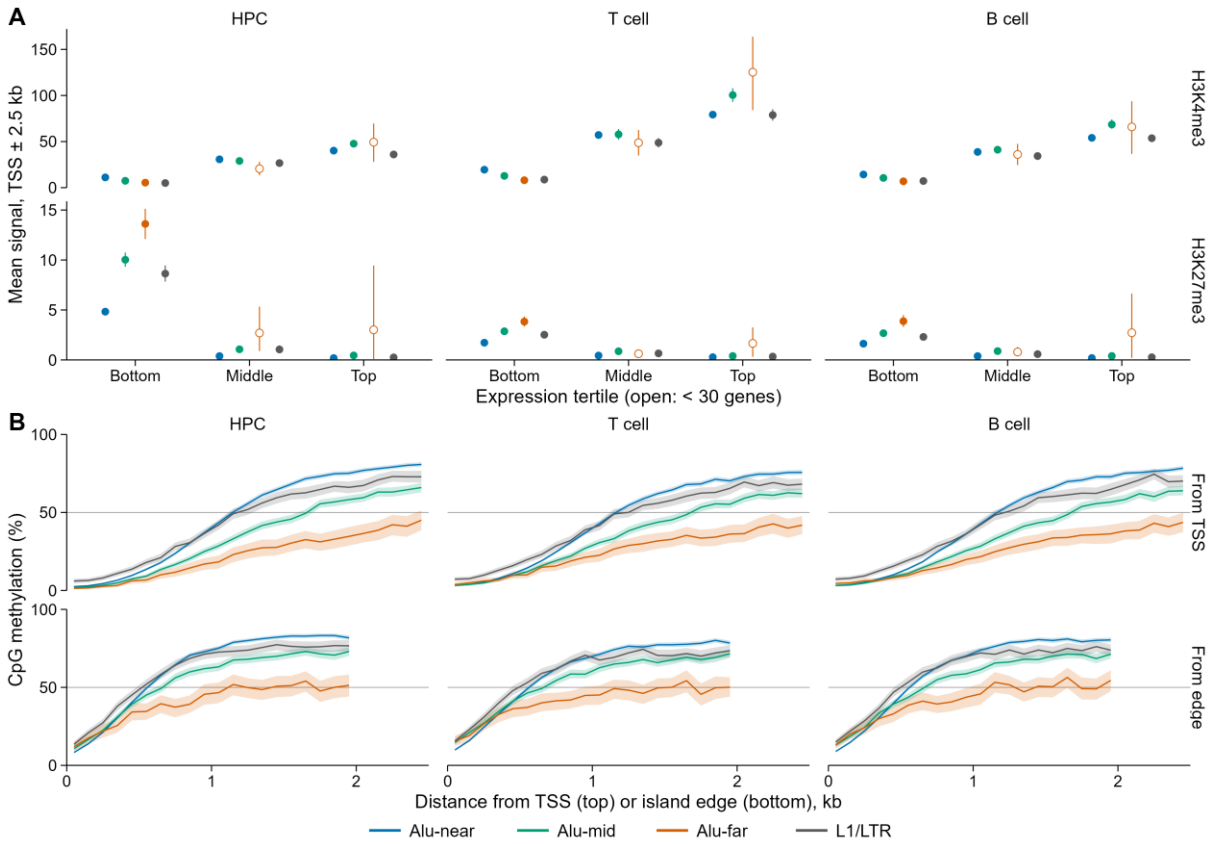
Supplemental Fig. S5. Proliferation and sample-set sensitivity. (A) Alu-near 1 share against the proliferation score (mean z-score of *PCNA*, *MKI67*, *TOP2A* and *CCNB1*) in the compiled dataset and the Corces hierarchy. (B) Spearman correlation with differentiation stage before (open) and after (filled) adjustment for proliferation in the sensitivity set. (C) Shares in the compiled sensitivity set (137 samples, after removing two dendritic-cell samples from GSE89322, six child bone-marrow CD34^{+} samples from GSE65996 and one leukemic iPSC, and moving four thymic single-positive samples from mature to lineage progenitor, giving pluripotent 7, HSPC 6, lineage progenitor 18 and mature 106). (D) Shares in the Corces hierarchy without erythroblasts (12 populations, 46 samples). BM, bone marrow; ESC, embryonic stem cell; HSPC, hematopoietic stem and progenitor cell; iPSC, induced pluripotent stem cell; RP, ribosomal protein; population abbreviations as in Fig. 4.



Supplemental Fig. S6. Associations with differentiation stage with counts-per-million normalization. (A) Spearman correlation of each share and of mRNAsi, the detected-gene count and the proliferation score with differentiation stage, before (open) and after (filled) adjustment for proliferation, in the compiled dataset (circles) and the Corces hierarchy (triangles); CPM version of Figs. 4D and 5C. (B) HSC/MPP and progenitor values of the four Corces donors for the Alu-far share, mRNAsi, the detected-gene count and the proliferation score. BM, bone marrow; CPM, counts per million; HSC/MPP, hematopoietic stem cells and multipotent progenitors.



Supplemental Fig. S7. Promoter classes in a mammary epithelial hierarchy. Relative transcript density (share of output divided by fraction of genes) of gene groups (top; all CpG-island, Alu-near 1, Alu-far and non-CpG-island genes) and of classes 2, 5 and 6 (bottom) in basal (BC), luminal progenitor (LP) and mature luminal (LC) cells, with stromal cells (SC) as reference (Pellacani et al. 2016; one pooled sample per population). Pre-specified predictions from the bone-marrow hierarchy were not supported (0 of 4; Table 1).



Supplemental Fig. S8. Histone marks by expression and methylation measured from the TSS and from the island edge. (A) Mean H3K4me3 and H3K27me3 signal over TSS ±2.5 kb by expression tertile within each cell type (Corces TPM) for Alu-near, Alu-mid, Alu-far and L1/LTR promoters, in HPC, T cells and B cells. Bars, 95% gene-level bootstrap intervals; open symbols, fewer than 30 genes. (B) CpG methylation by distance from the TSS (top, 0–2.5 kb) and outward from the CpG-island edge (bottom, 0–2 kb), with the upstream and downstream sides averaged per gene, in HPC, T cells and B cells. CpGs overlapping RepeatMasker elements excluded; bands, 95% gene-level bootstrap intervals; horizontal line, 50%. RP genes excluded, even-numbered autosomes. Groups as in Fig. 6. HPC, CD34⁺ hematopoietic progenitor cells (ENCODE); RP, ribosomal protein; TSS, transcription start site.

Supplemental Methods

Abbreviations are as defined in the main text.

Genes and TSS

The 31,962 genes of the classification set were derived from 63,512 Ensembl genes in the UCSC Genome Browser (GRCh38) by the steps in Supplemental Table S1. The TSS was taken from the transcript with the longest genomic span in a transcript table exported from the UCSC Genome Browser (47,339 Ensembl genes, 101,792 transcripts), whose transcript versions correspond to GENCODE release 45. For one gene (ENSG00000257262) the CpG island begins at the base after the annotated TSS, a difference between 0- and 1-based coordinates. This gene was kept as a CpG-island gene but omitted from analyses of island position (13,413 genes, Fig. 2). The unmasked island track was downloaded on October 23, 2026 (UCSC file date March 7, 2014) and the RepeatMasker track on January 19, 2026. The distances to the nearest Alu and the Alu counts used in the regression models were regenerated from this RepeatMasker table and the current TSS and were identical for all genes with a value.

Promoter classes

The three pre-specified alternative outer bounds were 7 kb upstream and 8 kb downstream for both configurations, per-side 95th percentiles, and separate bounds for the Alu and non-Alu sides. Classes were assigned by stepwise lower bounds, so that a gene falls in the most distant class whose bounds it exceeds on both sides, and ties between element types were resolved in the order Alu, L1, LTR.

Permutation and regression

Permutations used seed 20261015 and bootstrap intervals seed 20261022. Two-sided permutation P values used the +1 rule. The regression models of Supplemental Table S3 are described in its footnote. Regional features were computed in windows clipped to the chromosome, Alu base fraction from the same RepeatMasker table and GC fraction from the hg38 genome sequence, and gene counts from GENCODE v26 protein-coding genes with the gene itself excluded. Quintiles of regional Alu fraction and GC fraction were computed over the 13,320 CpG-island genes with the nearest retroelement within 10 kb, and a quintile was evaluable when each distance group had at least 30 genes.

Hematopoietic data and indices

Raw counts of the compiled dataset were obtained through Stemformatics (files dated August 2024) and converted to TPM with one GENCODE v26 gene-length table (summed exon length of the longest transcript). The four ESC samples (GSM1282324 and GSM1282325 from GSE53094, GSM1888631 and GSM1888641 from GSE73211) had no counts, so their TPM values were used as provided. Donor centering for the AUC subtracted each donor's mean over the samples in the comparison, using donors with both groups and counting ties as 0.5. Replacing the proliferation score with the S-phase and G2/M lists of Tirosh et al. (2016) did not change the sign of any adjusted correlation. The stress-response score used 20 dissociation-response genes (van den Brink et al. 2017), *FOS*, *FOSB*, *JUN*, *JUNB*, *JUND*, *EGR1*, *EGR2*, *EGR3*, *IER2*, *IER3*, *ATF3*, *DUSP1*, *ZFP36*, *KLF2*, *KLF4*, *KLF6*, *NR4A1*, *PPP1R15A*, *HSPA1A* and *HSPA1B*, none of which is an Alu-far gene. mRNAsi used the Progenitor Cell Biology Consortium signature of TCGAbiolinks 2.38.0. The detected-gene count was the mean of 10 downsamples of raw counts to 2,121,775 reads (compiled data) or 3,103,963 reads (Corces data), with seed 20261014. The

registered island-length match of Supplemental Table S7 used nearest-neighbor matching without replacement, from the longest Alu-far island down. Random matches sampled genes without replacement within 20 bins of masked island length over the 10,869 CpG-island genes of classes 1–9 in the share set (1,000 draws, seed 20261031).

Independent datasets

Registration was sequential (Granja et al. 2019, Ainciburu et al. 2023, then E-MTAB-13862). Single-cell data were aggregated per donor and author cell label (at least 30 cells) on the CPM scale. A prediction required all four donors in E-MTAB-13862, with TPM and CPM in agreement, a majority of donors in the Ainciburu data, and the registered AUC threshold in the Granja data (≥ 0.75 for a predicted increase and ≤ 0.25 for a predicted decrease). The registered share denominators excluded RP genes in E-MTAB-13862 and included them in the Ainciburu data, and the Granja source matrix contains no RP genes. Mammary RPKM values were renormalized within the 17,832 genes, with absent genes set to 0.

In the Human Cell Atlas bone-marrow data (HCAData, accessed October 22, 2026), cells with at least 500 detected genes and less than 20% mitochondrial UMIs were kept (278,577 of 378,000) and labeled with SingleR 2.12.0 and celldex 1.20.0. Cells labeled HSC_CD34+ or HSC_-G-CSF were taken as HSCs, and units with fewer than 30 cells were excluded.

Chromatin and DNA methylation

Relations with expression were analyzed for DNA methylation only, because it is measured as a fraction at single-CpG resolution and can be compared between positions and genes. The Cox models of the methylation boundary used expression percentile as the covariate, robust variance clustered by gene and a group × expression interaction, with and without adjustment for the TSS-to-edge distance. Hazard ratios are reported per 50 percentile points as summaries, because proportional hazards were not met in Alu-near and Alu-mid genes.

Software

Analyses used R 4.5.3 (Bioconductor 3.22) with data.table 1.18.6.1, ggplot2 4.0.3, patchwork 1.3.2, survival 3.8-6, GenomicRanges 1.62.1, rtracklayer 1.70.1, BSgenome.Hsapiens.UCSC.hg38 1.4.5, SummarizedExperiment 1.40.0, matrixStats 1.5.0, SingleR 2.12.0, celldex 1.20.0, CytoTRACE2 1.1.0, Seurat 5.6.0, TCGAbiolinks 2.38.0, clinfun 1.1.6 and lme4 2.0-6. Package versions for each analysis are recorded in the Supplemental Code.

Supplemental References

References cited here and in the main text are listed in the main reference list, except the following.

Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G et al. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352: 189–196.