Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Hyeon-Chun Park
The Catholic University of Korea
Haeyoun Choi
The Catholic University of Korea
Sukil Kim
The Catholic University of Korea
Jinyong Choi
The Catholic University of Korea
Jung-Hwan Oh
The Catholic University of Korea
Mun-Gan Rhyu
The Catholic University of Korea
Seung-Jin Hong
hongsjin@catholic.ac.kr
The Catholic University of Korea

Research Article

Keywords: CpG islands, Alu retroelements, promoter architecture, transcriptome composition, hematopoietic differentiation, evenness of expression

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11682879/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

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

Hyeon-Chun Park[1,5], Haeyoun Choi[1,2,5], Sukil Kim[3], Jinyong Choi[1], Jung-Hwan Oh[4], Mun-Gan Rhyu[1], Seung-Jin Hong[1]

[1]Department of Microbiology, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea.

[2]Catholic Hematopoietic Stem Cell Bank, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea.

[3]Department of Preventive Medicine, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea.

[4]Department of Internal Medicine, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea.

[5]These authors contributed equally to this work.

Corresponding author: Seung-Jin Hong, Department of Microbiology, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea. hongsjin@catholic.ac.kr

Running Title: Alu distance and hematopoietic transcript output

Abstract

Differentiation shifts transcript output among thousands of genes, but how output is partitioned among large gene groups remains unclear. To ask whether promoter architecture marks such groups, we classify 31,962 human genes by promoter CpG island and nearest retroelement. CpG-island genes with a retroelement, mostly an Alu, within 2 kb are expressed more evenly across tissues than those with more distant ones at matched island length. By distance to the nearest Alu, CpG-island promoters form an axis from Alu-near (within 2 kb) through Alu-mid to Alu-far promoters (none within 4–8 kb), with longer islands and less even expression. In blood progenitors and lymphocytes, Alu-near promoters carry more H3K4me3 and Alu-far promoters more H3K27me3. The methylation boundary lies about 1 kb from the transcription start site at Alu-near promoters, nearer in less expressed genes, and farther at Alu-mid and Alu-far promoters. In a bone-marrow hierarchy, the share of Alu-near genes flanked by Alu on both sides, about one third of output, is larger in progenitors than in hematopoietic stem cells and follows proliferation, whereas the Alu-far share, below 1%, is larger in stem cells, an excess led by *MEIS1* and *HOX* regulators. Pre-specified tests support both directions in an independent bulk dataset and each in one of two single-cell datasets. Alu distance thus defines promoter classes whose shares differ in opposite directions between these cells, a difference hidden in the total CpG-island share.

Keywords. CpG islands, Alu retroelements, promoter architecture, transcriptome composition, hematopoietic differentiation, evenness of expression.

Introduction

As multicellular organisms develop, cells change the expression of thousands of genes while continuing to express the genes that all cells need (Orkin and Zon 2008; Graf and Enver 2009). Stem and progenitor cells have long been viewed as states from which lineage programs are selected, through asymmetric divisions and fluctuations in gene expression (Morrison and Kimble 2006; Chang et al. 2008; Losick and Desplan 2008). Lineage-specifying transcription factors and enhancers switch individual genes on or off (Hnisz et al. 2013; Laurenti and Göttgens 2018), and global regulators raise or lower the total output of a cell (Lin et al. 2012; Nie et al. 2012). How output is divided among large groups of genes between these two scales, and which features of the genome mark such groups, remain unclear.

The total mRNA content of a cell scales with cell size (Padovan-Merhar et al. 2015), so groups of genes are naturally compared by the share of this total that each holds. The usual summaries of transcriptomes do not report this share. Marker genes and housekeeping gene lists differ between studies, and the expression of listed housekeeping genes differs between differentiation stages (Eisenberg and Levanon 2013; Ong et al. 2022). Indices of differentiation potential, such as the number of detected genes, on which CytoTRACE is based (Gulati et al. 2020), or a pluripotency-derived signature, mRNAsi (Malta et al. 2018), are learned from expression data and do not show which classes of genes account for the differences between cells. A classification of genes fixed by genome sequence, independent of the expression data being analyzed, would allow such shares to be compared between cell populations and predictions to be specified in advance. Two abundant sequence features of promoters are candidates for such a classification.

One is the CpG island, a characteristic feature of vertebrate genomes that overlaps the promoters of most protein-coding genes (Gardiner-Garden and Frommer 1987; Deaton and Bird 2011). CpG islands serve both broadly expressed genes and developmental regulator genes, and the islands of the latter tend to be longer and are often marked by H3K27me3 (Vinogradov 2005; Mikkelsen et al. 2007; Deaton and Bird 2011). CpG-island status alone therefore does not separate these two groups of genes.

The other is the retroelement. Retroelements make up nearly half of the human genome (International Human Genome Sequencing Consortium 2001), and Alu elements, a primate-specific family of more than one million copies (Batzer and Deininger 2002), are enriched in gene-dense regions near broadly expressed genes, whereas L1 and LTR elements are depleted there (Medstrand et al. 2002; Versteeg et al. 2003; Eller et al. 2007; Urrutia et al. 2008). The distribution of Alu and L1 elements is associated with CpG-island length, and methylation at the island flanks, the CpG-island shores, varies between tissues and with the nearest retroelement (Kang et al. 2006; Kim et al. 2006; Irizarry et al. 2009). Alu elements can also affect transcription and chromatin at nearby promoters (Ferrari et al. 2020; Payer et al. 2021; Costallat et al. 2022). Because retroelements are numerous near most promoters, their type and distance relative to the CpG island could distinguish groups of promoters across the genome at once, whether or not the elements themselves act on expression.

Here we classify 31,962 human genes from sequence alone by promoter CpG-island status and the type and distance of the nearest Alu, L1 or LTR element. We ask whether the distance to the nearest element orders promoters in island length, evenness of expression, chromatin marks and DNA methylation, and whether the shares of output from the resulting classes differ between hematopoietic populations, from pluripotent cells to mature blood cells, in ways that the total CpG-island share does not. Hematopoiesis offers sorted populations with public expression and epigenome data (Corces et al. 2016; Notta et al. 2016), and we test predictions specified in advance in independent bulk and single-cell datasets. We find that distance to the nearest Alu orders CpG-island promoters along one axis of island length, evenness of expression, chromatin marks and methylation boundary that persists after adjustment for island length and regional Alu density, and that the shares of the two ends of this axis differ in opposite directions between hematopoietic stem cells and progenitors in a bone-marrow hierarchy, with partial support in independent datasets.

Results

CpG-island status and distance to the nearest retroelement stratify evenness of expression across tissues

We classified 31,962 human genes (Supplemental Table S1) by whether a CpG island (UCSC cpgIslandExt) overlapped the transcription start site (TSS) and by the type and distance of the nearest Alu, L1 or LTR element on each side of the TSS (Fig. 1A; Supplemental Fig. S1), using the TSS of each gene's longest transcript. Evenness of expression across 54 GTEx tissues was measured with the Gini coefficient, higher for less even expression (GTEx Consortium 2020). Genes with a TSS-overlapping CpG island (13,414; 42.0%) were expressed more evenly than genes without one (18,548; median Gini 0.355 vs 0.794, Mann–Whitney $P < 0.001$; Fig. 1B), and the coefficient of variation (CV) and genes with a single annotated transcript gave the same result (Supplemental Table S2).

Among CpG-island genes, the Gini coefficient rose with distance to the nearest Alu (median 0.322 within 2 kb vs 0.729 beyond 10 kb) but changed little with distance to the nearest L1 or LTR element (0.332 vs 0.383 and 0.331 vs 0.358) (Fig. 1C; Supplemental Fig. S2A). The same rise was seen within each Alu subfamily (Supplemental Fig. S2B). Distance to the nearest retroelement of any type, most often an Alu, gave a similar gradient (0.330 vs 0.763). We thus classified promoters by both the type and the distance of the nearest retroelement.

Because genes with a distant nearest retroelement also tend to have longer CpG islands, we compared each distance group with genes from the same island-length quintile (10,000 permutations; Fig. 1D). Genes with the nearest element within 2 kb were slightly below this expectation (observed minus expected median Gini −0.017), whereas those at 2–4 kb and 4–10 kb were less evenly expressed (+0.080 and +0.214; adjusted $P \leq 1 \times 10^{-4}$; CV in Supplemental Fig. S2C), and this order held within every island-length quintile (Supplemental Fig. S2D, E). In linear models of the Gini coefficient, distance to the nearest Alu explained more variance than island length (R^2 0.064 vs 0.009, both on a log scale) and still explained 5.6% of the remaining variance after adjustment for island length. Its contribution fell to 1.1% after adjustment for the number of Alu elements within 10 kb and

to 0.8% when GC content, CpG observed/expected ratio, gene length and mean expression were also included (partial R^2). Adding regional Alu density, GC content and gene density to the island-length model reduced the contribution of Alu distance from 5.6% to 1.3%. The gradient in evenness was also present among genes with similar regional Alu density (pre-specified) or similar GC content (Supplemental Table S3). Within each type of nearest element, a closer element went with a lower Gini (Fig. 1E; Supplemental Fig. S2F for non-CpG-island genes).

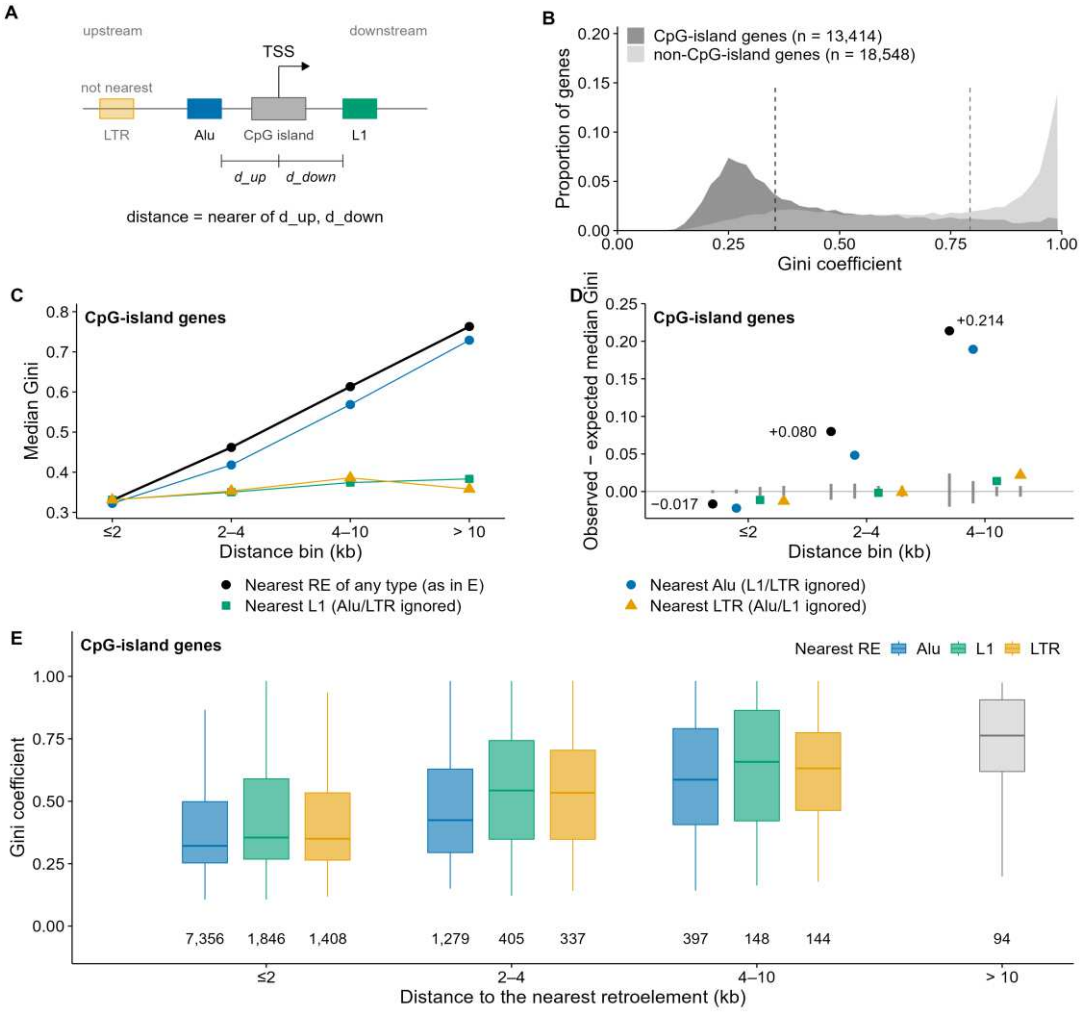


Fig. 1. CpG-island status, nearest retroelement and evenness of expression across tissues. (A) Distance and type of the nearest Alu, L1 or LTR element, taken on the nearer side of the TSS. (B) Gini coefficient across 54 GTEx tissues for CpG-island (n = 13,414) and non-CpG-island genes (n = 18,548); dashed lines, medians. (C) Median Gini of CpG-island genes by distance to the nearest retroelement of any type (black) or of each type. (D) Observed minus expected median Gini by distance to the nearest retroelement of any type,

expected from 10,000 permutations within island-length quintiles; gray bars, 95% permutation range. (E) Gini by distance to and type of the nearest retroelement; > 10 kb, none within 10 kb; whiskers, 1.5 × interquartile range. Alu, blue; L1, green; LTR, orange. RE, retroelement.

Alu elements flank CpG islands on both sides

Around the 11,283 CpG islands at the TSSs of CpG-island genes, Alu coverage rose steeply within 2 kb of the island edge and stayed near 0.20 up to 10 kb, about twice its genome-wide coverage (0.105), whereas L1 and LTR coverage stayed below their genome-wide levels (Fig. 2A). The nearest retroelement was an Alu on 68% of upstream and 73% of downstream sides of CpG-island genes, against 51% and 53% for non-CpG-island genes (Fig. 2B). Of CpG-island genes, 52.1% had an Alu as the nearest element on both sides (Alu–Alu), 36.2% on one side (Alu-mixed) and 11.7% on neither (L1/LTR-only) (Fig. 2C). Within every island-length quintile, Alu–Alu genes had the lowest and L1/LTR-only genes the highest median Gini (0.307–0.425 vs 0.384–0.514), with non-overlapping 95% bootstrap intervals, as pre-specified (Fig. 2D).

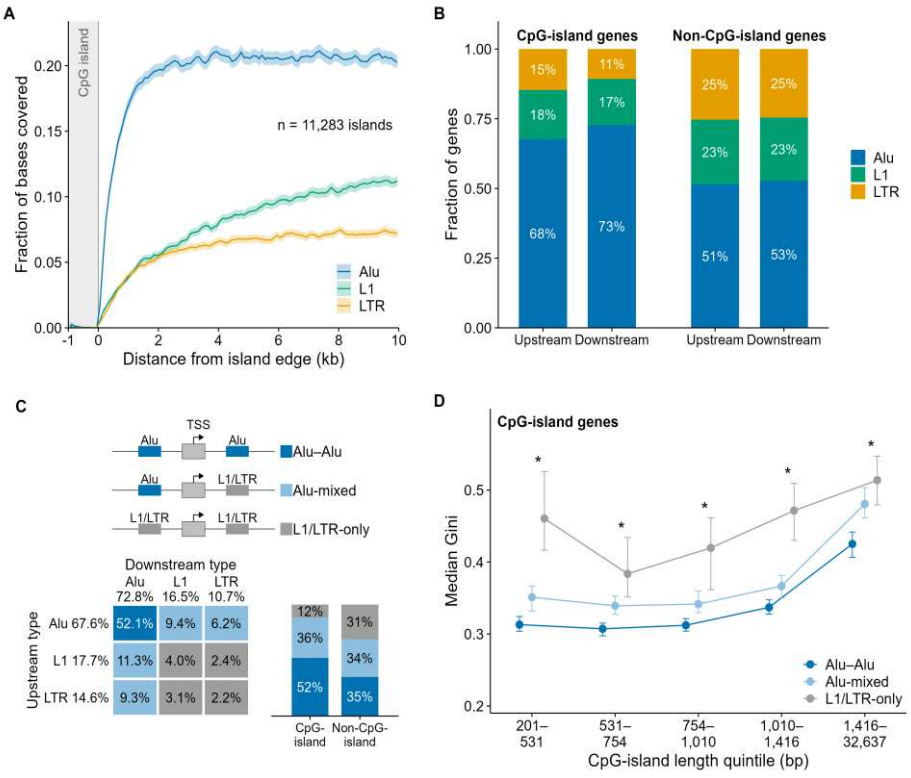


Fig. 2. Retroelements around CpG islands. (A) Fraction of bases covered by Alu, L1 and LTR elements outward from the edges of 11,283 CpG islands at the TSSs of 13,413 CpG-island genes (2,086 islands lie at the TSSs of two to six genes and were counted once); gray, inside the island; bands, 95% bootstrap intervals. (B) Type of the nearest retroelement on each side of CpG-island (n = 13,413, one gene omitted, Supplemental Methods) and non-CpG-island genes (n = 18,548). (C) Upstream and downstream types combined as Alu–Alu, Alu-mixed and L1/LTR-only. (D) Median Gini of CpG-island genes by configuration within island-length quintiles; bars, 95% bootstrap intervals; asterisks, non-overlapping Alu–Alu and L1/LTR-only intervals (pre-specified).

Promoter features change with distance to the nearest Alu

Combining configuration and distance gave nine CpG-island and three non-CpG-island classes (Fig. 3A). The nearest Alu was most often 0.6–1.4 kb from the TSS of CpG-island genes and became less frequent beyond 2 kb (Supplemental Fig. S1A), so we divided CpG-island genes at 2 kb, and Alu–Alu genes also at 4 kb (Supplemental Fig. S3A). Outer thresholds, which enclose most genes of each configuration, separated the most distant genes (Methods). Evenness of expression changed continuously with distance (Fig. 3B; Supplemental Fig. S3B), so the classes divide a gradient rather than mark breakpoints. We refer to class 1 (an Alu as the nearest element on both sides, one within 2 kb; 5,793 genes) as Alu-near 1, class 5 (the Alu-mixed counterpart; 3,702) as Alu-near 2, classes 2 and 3 as Alu-mid 1 (1,122), class 6 as Alu-mid 2 (902), and classes 4 and 7 (nearest element beyond the outer thresholds on both sides; 323) as Alu-far**.**

From Alu-near 1 to Alu-mid 1 and Alu-far promoters, median island length increased (820, 1,340 and 1,923 bp) and evenness of expression decreased (median Gini 0.32, 0.43 and 0.69). Of the housekeeping genes in two lists (Eisenberg and Levanon 2013; Hounkpe et al. 2021), 44.9% and 47.0% had class 1 promoters, although class 1 holds 27.0% of protein-coding genes. Transcription factors made up 29.4% of Alu-far and 6.6% of class 1 genes (odds ratio 5.88, and 4.42 after adjustment for island length; Supplemental Table S4).

Promoter-class shares across tissues

To compare shares of total transcript output between GTEx and the hematopoietic datasets, we used a fixed set of 17,832 protein-coding genes present in all expression tables, with the 81 ribosomal-protein (RP) genes as a separate group (Supplemental Table S1). Classes in this set were smaller (Alu-near 1, 4,834 genes; Alu-far, 217). Across 54 GTEx tissues, Alu-near 1 genes supplied a median of 31.5% of output, and 38.8% in embryonic stem cells (ESCs), where the CpG-island share was 80.9% against a median of 72.4% across tissues (Fig. 3C). Across tissues, the Alu-near 1, Alu-near 2 and Alu-far shares each correlated with the CpG-island share (Spearman ρ 0.73, 0.72 and 0.64; Fig. 3D; Supplemental Figs. S3, S4).

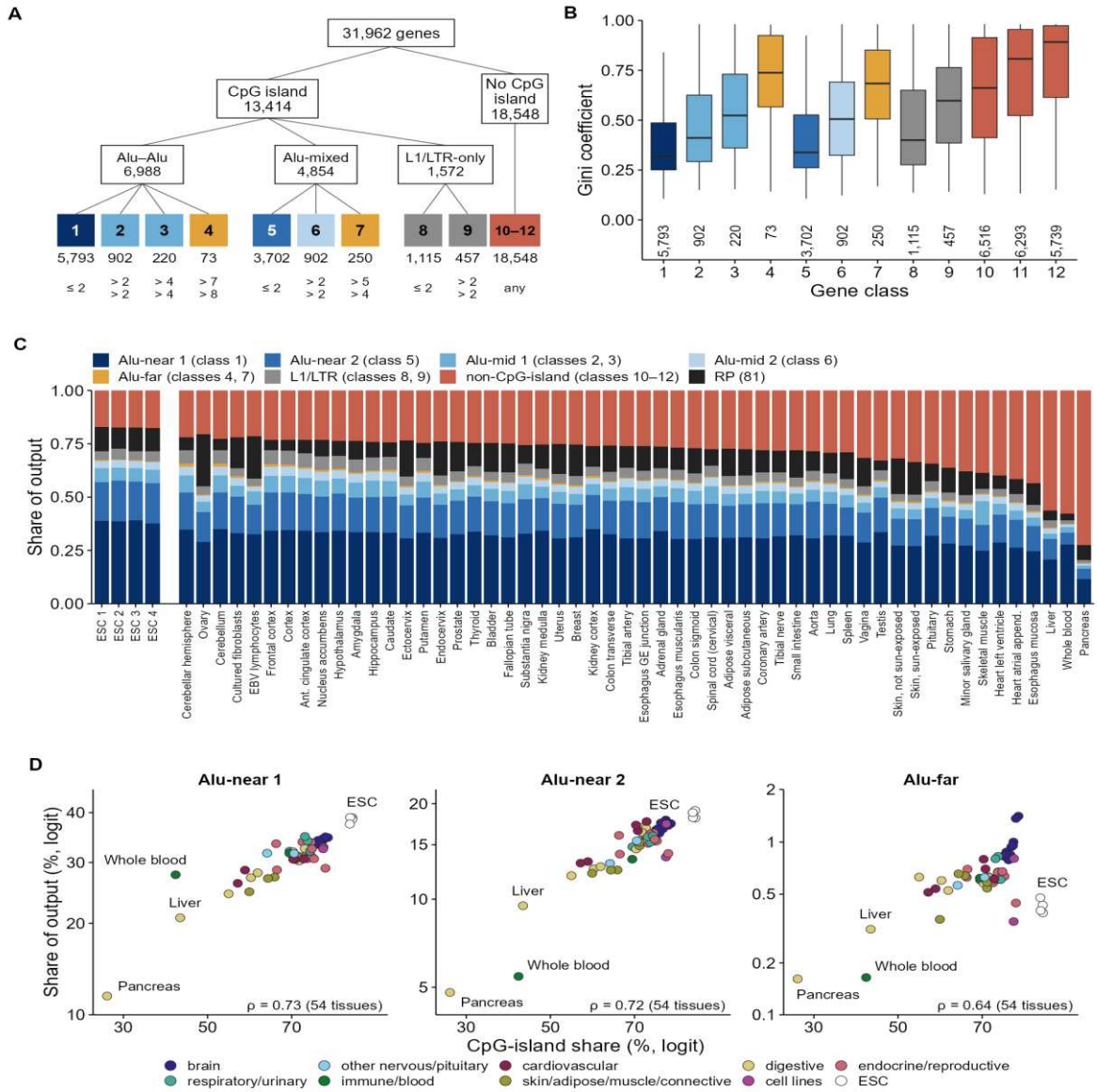


Fig. 3. Promoter classes defined by CpG islands and the distance to the nearest retroelement. (A) Classification of 31,962 genes by promoter CpG island and distance (kb, upstream above downstream) to the nearest Alu, L1 or LTR element; ≤ 2, within 2 kb on at least one side; numbers, genes per class. (B) Gini coefficient across 54 GTEx tissues by class. (C) Share of transcript output by group in 54 GTEx tissues, ordered by CpG-island share, and in four embryonic stem cell (ESC) samples (17,832-gene set, RP genes separate). (D) Alu-near 1, Alu-near 2 and Alu-far shares against the CpG-island share (logit scale); ρ, Spearman correlation across tissues (ESC excluded); colors, organ systems. RP, ribosomal protein.

Promoter-class shares of the hematopoietic transcriptome

We next examined two bulk hematopoietic datasets (Fig. 4A, B; Supplemental Table S5). The first was a compiled set of 146 samples from seven GEO series, comprising pluripotent cells, hematopoietic stem and progenitor cells (HSPCs), lineage progenitors and mature cells. The second was the bone-marrow hierarchy of Corces et al., in which 13 sorted populations from 4 donors (49 samples) separate hematopoietic stem cells and multipotent progenitors (HSC/MPP) from lineage progenitors and mature cells. Because the same donors were sorted into each population, the Corces data served to define differences and the compiled data as a second set. The CpG-island share was lowest in mature cells in both datasets (median 68.6% in the compiled data and 70.6% in the Corces data, against 73.1–79.7% and 75.4–75.8% in the other groups). Shares were computed with transcripts per million (TPM) and, to check that they did not depend on gene-length correction, with counts per million (CPM), and values below are for TPM unless stated. The Alu-near 1 share lay within 29.4–35.3% across the Corces populations, and the median Alu-far share was 0.317% in HSC/MPP and 0.262% in progenitors (Fig. 4B).

We summarized each difference between stem and progenitor groups (HSPCs in the compiled data, HSC/MPP in the Corces data) as an area under the receiver operating characteristic curve (AUC), the probability that a sample of the first group has the higher share. In the compiled

data, the Alu-far share was higher in HSPCs than in lineage progenitors (AUC 0.935), but the Alu-near 1 share differed little (0.357) and the total CpG-island share was lower in HSPCs (0.250) (Fig. 4A). In the Corces hierarchy, the Alu-near 1 share was lower in HSC/MPP than in common myeloid, granulocyte–macrophage and megakaryocyte–erythroid progenitors (CMP, GMP and MEP) (AUC 0.010) and the Alu-far share higher (0.969), whereas the total CpG-island share differed in direction among progenitor types (Fig. 4B, C; Supplemental Fig. S5). The opposite differences of the two shares were therefore partial in the compiled data and clear in the Corces hierarchy.

The Alu-near 1 share correlated with a proliferation score (mean z-score of *PCNA*, *MKI67*, *TOP2A* and *CCNB1*; Spearman ρ 0.61–0.74 across datasets and normalizations), and its association with differentiation stage, the rank order of the sample groups, weakened after adjustment for proliferation (compiled ρ −0.353 to +0.077 with TPM and −0.613 to −0.269 with CPM; Fig. 4D; Methods). Its lower share in mature cells than in HSCs of the Corces hierarchy was spread over many genes, the largest contributor, *HSP90AB1*, accounting for 6% of the difference (Supplemental Data S1). The Alu-far share showed little correlation with proliferation ($|\rho| \leq 0.10$).

The higher Alu-far share in HSC/MPP held under the Alu–Alu outer thresholds applied to Alu-mixed genes, without *CCND2*, which supplied 28–30% of the share, and for the 171 of its 217 genes that were transcription factors or carried H3K27me3 within 1 kb of the TSS in $CD34^+$ hematopoietic progenitor cells (HPC, as designated by ENCODE; Supplemental Table S6) (Corces AUC 0.917–1.000; Supplemental Table S7). Under the alternative thresholds its correlation with proliferation rose in the compiled data (ρ 0.43, above the pre-specified limit of 0.20), and in the Corces data the Alu-far share under the thresholds in use correlated with a dissociation-stress score (van den Brink et al. 2017) that was highest in HSC/MPP (ρ 0.73). After adjustment for stress and proliferation, its association with stage remained in the compiled data (ρ −0.57) but not in the Corces data once *CCND2* was removed (ρ −0.04).

In a registered comparison, 217 other CpG-island genes matched to Alu-far genes for island length showed the same difference (Corces AUC 1.000 with TPM and CPM), so the prediction that the difference was specific to Alu distance was not supported. In this single deterministic match the

difference came mainly from *HSP90AB1*, *NPR3* and *DUSP1*, and in 1,000 random matches the median AUC was 0.46 (TPM) and 0.50 (CPM), with stress-response and housekeeping genes rather than regulators leading (post hoc). Genes matched to Alu-near 1 genes for island length differed in the opposite direction to Alu-near 1 (AUC 0.813 and 0.938; random-match medians 0.75 and 0.91; Supplemental Table S7).

Ten highly expressed genes, such as *GAPDH* and *HSPA8*, supplied 17–31% of the Alu-mid 1 share, which differed in direction between the two datasets even without them (AUC 0.958 in the Corces data and 0.065 in the compiled data).

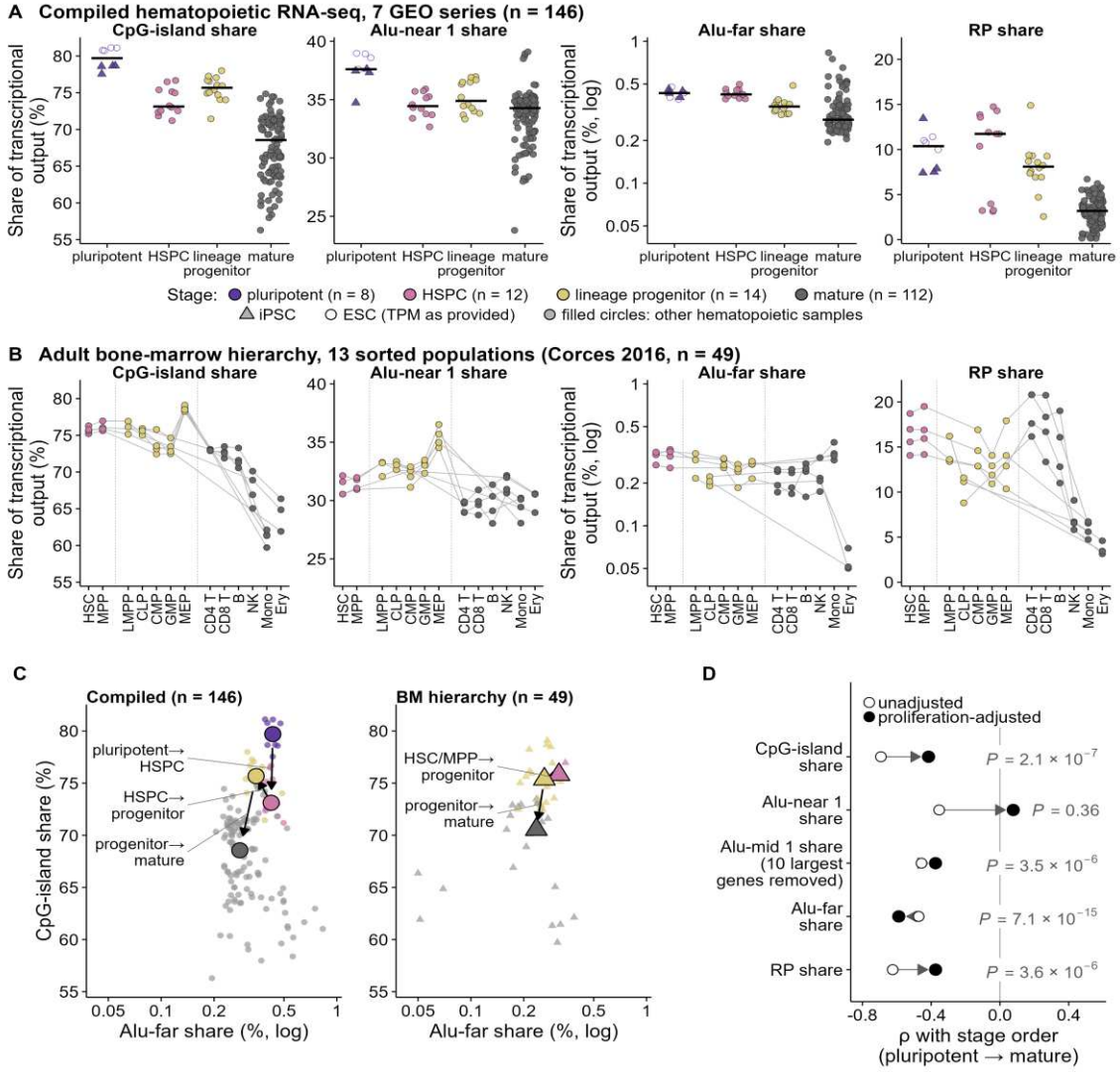


Fig. 4. Promoter-class shares in hematopoietic cells. (A) CpG-island, Alu-near 1, Alu-far and RP shares in the compiled dataset (146 samples); bars, medians; triangles, iPSC; open circles, ESC. (B) The same shares in 13 sorted bone-marrow populations (Corces et al.

2016; 49 samples); lines connect samples of the same donor. (C) CpG-island against Alu-far share; large symbols, group medians; arrows connect compared groups. (D) Spearman correlation with differentiation stage (pluripotent, HSPC, lineage progenitor, mature) in the compiled dataset before (open) and after (filled) adjustment for proliferation; P, P value of the adjusted correlation; Alu-mid 1 shown without its ten largest contributors. TPM shown; CPM in Supplemental Fig. S6. BM, bone marrow; CLP, common lymphoid progenitor; Ery, erythroblast; iPSC, induced pluripotent stem cell; Mono, monocyte; NK, natural killer cell; other abbreviations as in the text.

Pre-specified tests in independent datasets

We tested the Corces directions in three independent datasets, bulk bone-marrow RNA-seq from four donors (E-MTAB-13862, Lalioti et al. 2025) and single-cell data analyzed as pseudo-bulk from three donors (Granja et al. 2019) and eight donors (Ainciburu et al. 2023), with predictions and decision rules registered before each dataset was read (Fig. 5A; Table 1; Methods). The higher Alu-near 1 share in progenitors was supported in E-MTAB-13862 and Ainciburu but not in Granja, and the lower Alu-far share in progenitors in E-MTAB-13862 and Granja but not in Ainciburu. The RP prediction was supported in Ainciburu only, and predictions for the total CpG-island share were not supported (Table 1).

A post hoc analysis of gene contributions showed what carried the Alu-far difference (Supplemental Data S1). In the Corces hierarchy, 69–77% of the HSC/MPP excess came from higher expression of detected Alu-far genes rather than from more genes being detected. *MEIS1* and the *HOX* genes *HOXA5*, *HOXA6* and *HOXB5* together supplied 40% of the difference with TPM, and *PCDH9* alone about half with CPM. Nine or ten of the ten leading genes had the same sign in each of the three independent datasets. In the Ainciburu data they were also higher in hematopoietic stem cells (HSCs), but other Alu-far genes, led by *ATP5F1E*, *TIMM13* and *MRPS34*, which encode mitochondrial proteins, were lower, and the total share was lower in HSCs. A registered test of the 11 leading regulators in the Human Cell Atlas bone-marrow data (Hay et al. 2018), in which we had already

examined the Alu-far share, was not supported ((2/4 donors for HSC-like cells vs CMP, GMP and MEP, and 0/8 for HSPCs vs mature cells; Supplemental Data S2).

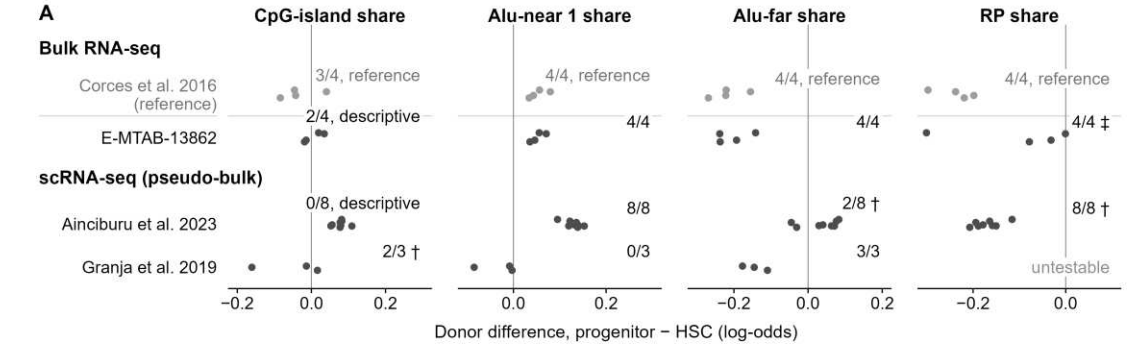
As a descriptive comparison, we set the shares beside two expression-based indices, a pluripotency-derived signature (mRNAsi) and the number of detected genes, a proxy for CytoTRACE (Fig. 5B–D). mRNAsi was lower in HSCs than in progenitors and followed proliferation (ρ 0.58–0.81). The detected-gene count was higher in HSC/MPP in the Corces hierarchy (AUC 0.802) but not consistently in the other datasets, and CytoTRACE 2 (Kang et al. 2025) scored HSCs higher in Granja (3/3, descriptive) and above lineage-restricted progenitors but not lymphoid-primed multipotent progenitors (LMPP) in Ainciburu. In a pre-specified comparison, the Alu-far share added little beyond the detected-gene count, with which it was correlated (ρ 0.61). With the count regressed out, its Corces AUC fell from 0.969 to 0.708, below the registered 0.75, although its association with stage after adjustment for the count and proliferation remained (ρ −0.33 in Corces and −0.50 in the compiled data). Recomputed within the most highly expressed genes of each class, the Alu-far and total CpG-island shares kept the sign of their adjusted association with stage in both datasets (Fig. 5D).

In a mammary epithelial hierarchy with one pooled sample per population (Pellacani et al. 2016), none of four predictions derived from the bone-marrow results was supported (0/4; Table 1; Supplemental Fig. S7). There the CpG-island classes separated epithelial from stromal cells (relative transcript density of CpG-island genes 1.14–1.20 vs 1.05), and the largest difference among epithelial populations was in Alu-mid 2, higher in basal cells (1.53 vs 1.08–1.14).

Table 1. Pre-specified predictions in independent datasets.

Dataset	Alu-near 1	Alu-far	RP	CpG island
E-MTAB-13862 (bulk)	**4/4**	**4/4**	NS	—
Granja et al. 2019 (scRNA)	0/3	**3/3**	—	2/3, NS
Ainciburu et al. 2023 (scRNA)	**8/8**	2/8†	**8/8†**	NS

Predictions, Alu-near 1 share higher, and Alu-far and RP shares lower, in progenitors than in HSCs. Values, donors with the predicted direction; bold, supported; NS, not supported; —, not registered or not testable. Comparisons, HSC vs $CD34^{+}CD38^{+}$ cells (E-MTAB-13862), HSC vs progenitors (Granja), HSC vs LMPP (Ainciburu; CpG island, progenitors vs mature myeloid cells). Support required all four donors with both TPM and CPM (E-MTAB-13862), the registered AUC threshold (Granja) or a majority of donors (Ainciburu). RP in E-MTAB-13862, 4/4 with TPM but 1/4 with CPM. †, the prediction also required HSC vs lineage-restricted progenitors (Alu-far 0/8, RP 8/8), and Alu-far gave 2/8 with CPM and 0/8 with TPM.



B

Index	Input	Reference · re-fitting	Direction along BM hierarchy	Correlation with proliferation
Promoter-class shares	promoter-class shares of TPM	none · none		
CpG-island			– ↓	◐
Alu-near 1			↑ ↓	○
Alu-far			↓ –	●
RP			↓ ↓	◐
mRNAsi	similarity to pluripotent cells	PCBC pluripotent samples · none	↑ ↓	○
Detected-gene count (CytoTRACE basis)	detected genes (raw counts)	none · depth matching	↓ ↓	◐

Direction (left, HSC → progenitor; right, HSC → mature):
↓ earlier stage higher (AUC ≥ 0.75) ↑ later stage higher (AUC ≤ 0.25) – neither
● |ρ| with proliferation < 0.2 in both datasets ○ > 0.5 in both ◐ otherwise

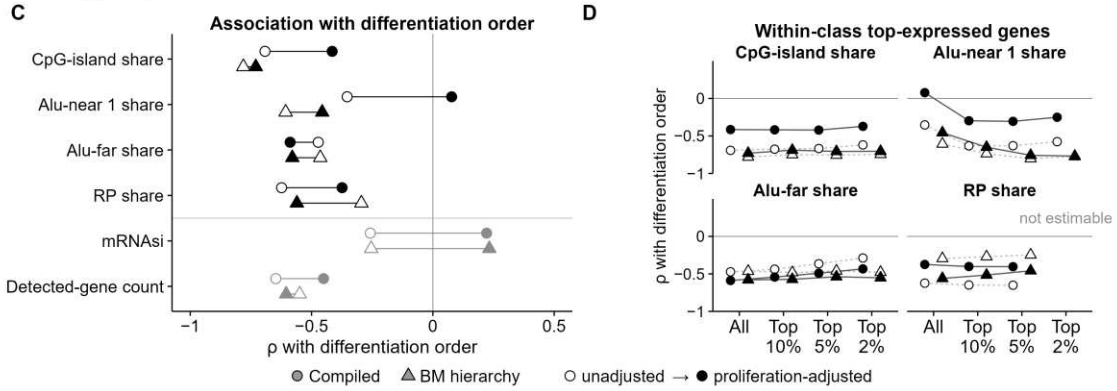


Fig. 5. Pre-specified tests in independent datasets and comparison with expression-based indices. (A) Donor differences in share between progenitors and HSCs (donor-

centered log-odds) in the Corces hierarchy (reference, HSC/MPP) and three independent datasets; numbers, donors with the predicted direction; "descriptive", not pre-specified; †, the prediction included a second comparison; ‡, TPM shown, CPM gave 1/4; verdicts in Table 1. (B) Direction along the bone-marrow hierarchy and correlation with proliferation of the shares and two expression-based indices, with symbols as keyed in the panel. (C) Spearman correlation with stage before (open) and after (filled) adjustment for proliferation in the compiled dataset (circles; detected-gene count, n = 142) and the Corces hierarchy (triangles). (D) As in C, within the 10%, 5% or 2% most highly expressed genes of each class (TPM). BM, bone marrow; PCBC, Progenitor Cell Biology Consortium.

Histone marks and the methylation boundary at Alu-near, Alu-mid and Alu-far promoters

In ENCODE data from HPC, T cells and B cells (ENCODE Project Consortium 2020; Supplemental Table S6), CpG-island promoters were compared in four groups, Alu-near (Alu-near 1 and 2), Alu-mid (Alu-mid 1 and 2), Alu-far and L1/LTR. In all three cell types, H3K4me3 around the TSS decreased from Alu-near to Alu-mid and Alu-far promoters (Alu-far 0.33–0.40 times the Alu-near signal), and H3K27me3 increased in the same order (Alu-far 4.7–7.5 times) (Fig. 6A). Promoters with an H3K27me3 peak were less frequent among Alu-near genes (odds ratio 0.33–0.36) and more frequent among Alu-far genes (5.99–6.99) (Supplemental Table S8). Within expression tertiles, H3K4me3 differed little between groups, but in the bottom tertile H3K27me3 was lowest at Alu-near and highest at Alu-far promoters (Supplemental Fig. S8A).

Methylation of CpGs outside repeats was low at the TSS in all groups (1.5–7.4%). In HPC it reached 50% about 1.0 kb upstream and 1.2 kb downstream of the TSS at Alu-near promoters and 1.6 and 1.7 kb at Alu-mid promoters but not within 2.5 kb at Alu-far promoters (34–39% at 2 kb), and positions in T and B cells were within 100–200 bp of these (Fig. 6A). Measured from the island edge, methylation in the first 500 bp outside it was similar in all groups (21–30%), so the earlier rise at Alu-near promoters largely followed the shorter distance from the TSS to the island edge (Supplemental Fig. S8B). L1/LTR promoters, with TSS-to-edge distances close to those of Alu-near promoters,

showed a similar profile, although their histone marks were closer to those of Alu-mid promoters.

We took as the methylation boundary the first position outward from the TSS at which methylation reached 50% on each side (Methods). In HPC, the median boundary of Alu-near genes lay 1.25, 1.45 and 1.75 kb from the TSS in the bottom, middle and top expression tertiles (Fig. 6B), and across cell types the top-tertile boundary lay 500–700 bp farther from the TSS than the bottom-tertile boundary in Alu-near genes and 400–500 bp farther in Alu-mid genes (Supplemental Table S8). In Alu-far genes the boundary showed no monotone relation with expression. In bottom-tertile Alu-far genes it lay at a median of 3.45 kb in HPC, and across cell types 34–41% of their sides had not reached 50% by 5 kb, against 6–7% of bottom-tertile Alu-near sides. In Cox models, the hazard ratio of the boundary per 50 expression-percentile points was 0.65–0.73 in Alu-near, 0.85–0.88 in Alu-mid and 1.40–1.61 in Alu-far genes (group × expression interaction $P < 0.001$ in each cell type; Supplemental Table S8).

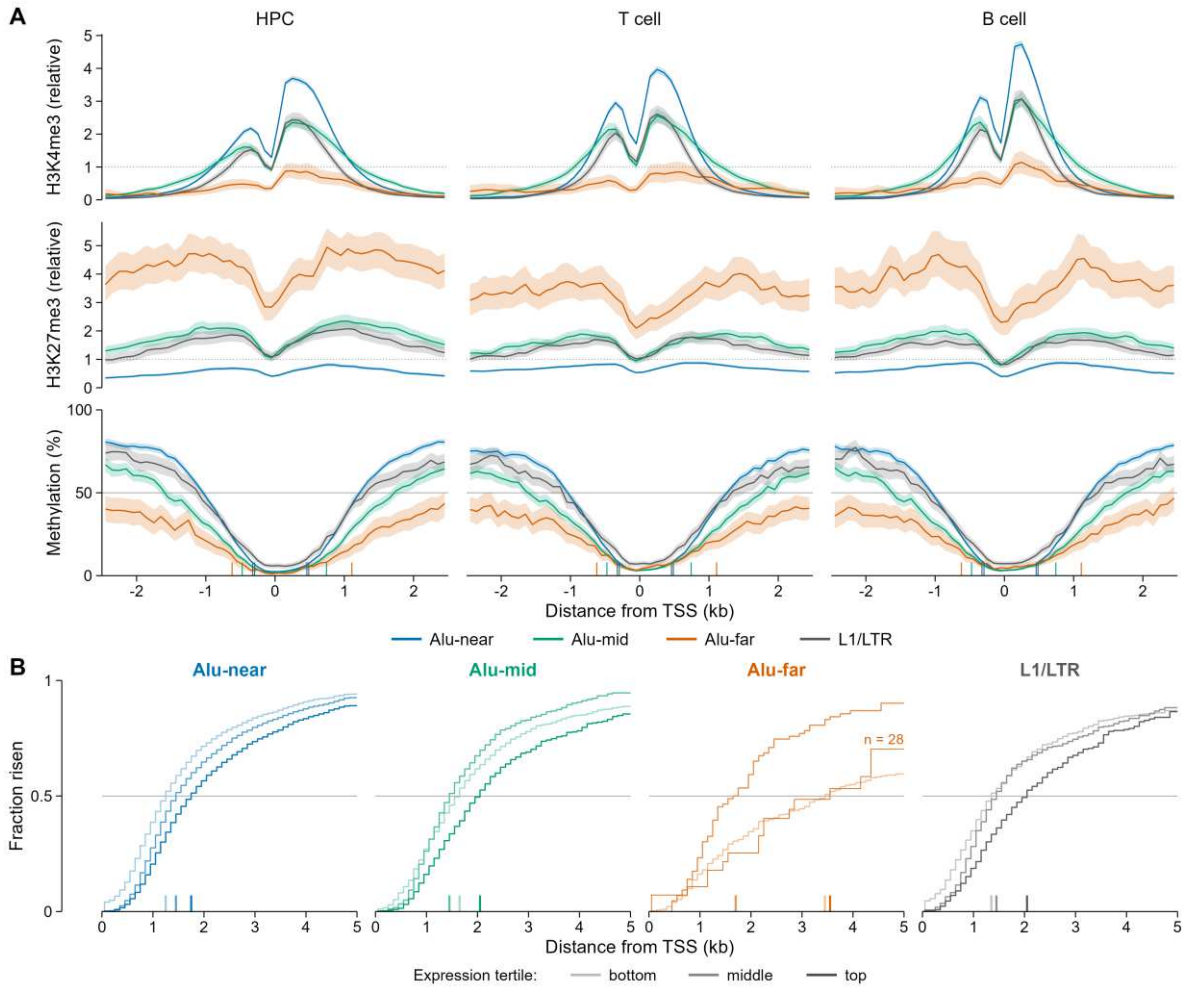


Fig. 6. Histone marks and the methylation boundary at promoter groups. (A) H3K4me3 and H3K27me3 signal relative to the mean of the four groups, and methylation of CpGs outside repeats, within 2.5 kb of the TSS in HPC, T cells and B cells, for Alu-near (n = 3,986), Alu-mid (n = 920), Alu-far (n = 169) and L1/LTR promoters (n = 708; even-numbered autosomes, RP genes excluded); bands, 95% bootstrap intervals; ticks, median TSS-to-edge distances. (B) Fraction of promoter sides on which methylation has reached 50% (the methylation boundary) by distance from the TSS in HPC, by expression tertile (Kaplan–Meier estimates); ticks, tertile medians; thin line, Alu-far top tertile (n = 28 sides). Ranges over HPC, T cells and B cells in Supplemental Table S8.

Discussion

Differentiation changes the output of thousands of genes. Most known mechanisms act on individual genes, through enhancers and repressors that activate or silence them, or on the whole cell, by raising or lowering total transcription, and less is known about how output is shifted between large groups of genes. We asked whether promoter classes defined from sequence alone behave as such groups. Alu-near 1 promoters formed a large class of evenly expressed genes holding about one third of output in hematopoietic cells, and Alu-far promoters a small class of 323 genes rich in transcription factors. Their shares differed in opposite directions between HSCs and progenitors, unlike the total CpG-island share, with pre-specified support for each direction in two of three independent datasets, and genes of other classes matched for island length showed the reverse of the Alu-near difference. Retroelements near CpG-island promoters may therefore provide a sequence-encoded frame for quantitative regulation of groups of genes, complementing the switch-like control of individual genes.

CpG-island promoters are known to be heterogeneous, with short islands at broadly expressed genes and long islands, often marked by Polycomb-associated H3K27me3, at developmental regulators (Vinogradov 2005; Mikkelsen et al. 2007; Ku et al. 2008; Deaton and Bird 2011). At the same island length, genes with a retroelement, most often an Alu, within 2 kb were more evenly expressed than genes whose nearest element lay farther away. Because Alu proximity and short

islands largely coincide, the classification treats them as one sequence feature, which orders CpG-island promoters from the evenly expressed Alu-near class to the Alu-far class, rich in transcription factors and consistent with Polycomb targets that are active in some lineages and repressed in others (Mikkelsen et al. 2007; Ku et al. 2008). The intermediate Alu-mid class showed no consistent direction, so the informative contrast lies between the two ends.

Shares measure how the steady-state mRNA pool is divided between gene classes. Total transcription scales with cell volume through a limiting factor shared across genes (Padovan-Merhar et al. 2015), and this transcriptional dose can be raised globally, as under elevated MYC activity (Lin et al. 2012; Nie et al. 2012) and in stem and progenitor cells (Percharde et al. 2017; Kim et al. 2023). Within any given total, output directed to lineage products leaves less for other genes. TPM and CPM remove differences in dose (Lovén et al. 2012) while keeping the proportion of output that each predefined group of genes receives. The lower CpG-island share of mature cells therefore does not indicate silencing, and is consistent with CpG-island genes keeping a reduced output for basic functions while lineage products take a larger share. The Alu-near 1 share followed proliferation more closely than stage, in line with lower protein synthesis in HSCs (Signer et al. 2014) and higher total transcription in proliferating progenitors (Bulut-Karslioglu et al. 2018), and its lower share in mature cells was spread across many genes rather than concentrated in a few.

Alu proximity was associated with an active promoter state that island length and margin methylation do not fully explain. L1/LTR promoters have islands as short as those of Alu-near promoters and a similar methylation profile, yet their histone marks resembled those of Alu-mid promoters. Within island-length quintiles, genes flanked by Alu on both sides were also more evenly expressed than genes flanked only by L1 or LTR elements. Alu elements carry transcription-factor binding sites (Polak and Domany 2006; Laperriere et al. 2007), resemble enhancers (Su et al. 2014), promote histone acetylation through TFIIIC (Ferrari et al. 2020) and are enriched near genes activated at embryonic genome activation (Töhönen et al. 2015; Ge 2017). A nearby Alu is therefore associated with a state of even expression across tissues, most evident in pluripotent cells, where Alu-near 1 genes supplied 38.8% of output against a median of 31.5% in adult tissues.

Retroelements can also contribute to inactivation. An Alu insertion raised methylation and repressive marks around a neighboring gene and lowered its expression (Byun et al. 2012), and upstream SINE elements led to promoter methylation *in vitro* (Estécio et al. 2012). Methylation just outside the island was similar across classes, as observed earlier for the CpGs between an island and its nearest retroelement (Kang et al. 2006; Kim et al. 2006), and such margin methylation accumulates with age, earlier in the more rapidly renewing gastric antrum (Hong et al. 2013; Oh et al. 2014). What differed was the distance from the margin to the TSS, shortest at Alu-near promoters and longest at Alu-far promoters, which kept a wide unmethylated region marked by H3K27me3, as described for Polycomb targets (Xie et al. 2013; Jeong et al. 2014). The position of the boundary was also most closely related to expression at Alu-near promoters, where it lay nearer the TSS in less expressed genes. These observations are consistent with a model we proposed from gastric mucosa, where Alu-flanked CpG-island genes became overmethylated earlier than those flanked by LTR or distant elements (Rhyu et al. 2012). If margin methylation accrues with division, it would reach the TSS first at Alu-near promoters and lower many broadly expressed genes gradually and together, which we call tidal downregulation. In the bone-marrow hierarchy, the Alu-near 1 share was lower in mature cells than in HSC/MPP, as this model would predict, but higher in progenitors, where it followed proliferation. Because DNA methylation is copied at each division, such lowering could serve as an epigenetic memory that keeps differentiation directional and frees output for lineage genes.

Shares of sequence-defined classes differ from expression-derived indices in how their genes are chosen. The gene groups are fixed before any data are read, so predictions can be registered in advance, and each group has properties of its own, such as island length and transcription-factor content, that make a difference in share interpretable. Among expression-derived indices, CytoTRACE 2 (Kang et al. 2025), trained on annotated developmental potency, ranked HSCs above lineage-restricted progenitors in the single-cell data, whereas mRNAsi (Malta et al. 2018), derived from pluripotent cells, ranked them below progenitors and followed proliferation. The detected-gene count, on which CytoTRACE is based (Gulati et al. 2020), was correlated with the Alu-far share, and once the count was taken into account, the share added little to the separation of HSCs from

progenitors. Unlike these indices, however, the share can be divided among the genes that make it up, and its excess in HSCs came mainly from regulators such as *MEIS1* and the *HOX* genes, whereas in gene sets matched for island length it came from stress-response and housekeeping genes.

Support in independent datasets was partial. Both failures of the two main predictions occurred in single-cell data, where populations are defined by clustering rather than by surface markers, and in the Ainciburu data the leading Alu-far regulators were higher in HSCs while other genes of the same class, led by genes encoding mitochondrial proteins, were lower and the total share was lower in HSCs, as expected for a small class. Whether the Alu-far difference depends on Alu distance or on island length is also unresolved, because a registered set of genes matched for island length reproduced it but most random matched sets did not. The Human Cell Atlas test rested on automated labels whose HSC reference includes mobilized peripheral-blood $CD34^{+}$ cells. In a mammary hierarchy, none of four predictions held. The CpG-island classes instead separated epithelial from stromal cells, and among epithelial populations Alu-mid 2 rather than Alu-near 1 or Alu-far differed most, so the classes whose shares differ between populations may differ between tissues, although each population was represented by one pooled sample.

This study has further limitations. The TSS was taken from the longest transcript, although single-transcript genes gave similar results. Predictions for the independent datasets were derived after the Corces results had been examined and were tested with share denominators that differed between datasets. Pluripotent cells were not included in the methylation comparisons, because their high non-CpG methylation depends on *DNMT3A* and *DNMT3B* (Lister et al. 2011; Liao et al. 2015). The chromatin and methylation data cover three cell types with one methylation replicate each, come in part from one donor or from short-term cultured cord-blood cells, and share no sample with the expression data, and the reported profiles rest on genes of even-numbered autosomes. Shares are relative quantities of steady-state mRNA, so they reflect mRNA stability as well as transcription, are not comparable in absolute terms between datasets processed differently and do not show whether the output of a class fell or that of other classes rose, which per-cell transcript counts would resolve (Kim et al. 2023). All associations here are cross-sectional, and because Alu elements number more than

one million, their effect cannot be tested by manipulating single loci.

Promoter architecture defined from genome sequence provides a fixed frame for comparing transcriptome composition between cell populations, with gene groups set before any data are read. In human hematopoiesis, the shares of the large Alu-near class and the small Alu-far class differed in opposite directions between HSCs and progenitors, differences that are not visible in the total CpG-island share. Whether these differences reflect the tidal downregulation proposed here can now be tested by following methylation and expression of the same promoters across divisions and by measuring total output per cell. Comparing species whose SINE families arose independently would show whether the frame depends on Alu elements themselves.

Methods

Gene sets

We analyzed 31,962 Ensembl genes (UCSC Genome Browser, GRCh38) with GTEx v8 expression profiles (filtering steps in Supplemental Table S1). This set, which includes RP genes, was used for classification, evenness of expression, island length and gene annotations. Shares of transcript output were computed in a fixed set of 17,832 protein-coding genes (GENCODE v26) present in all hematopoietic expression tables, with its 81 RP genes as a separate group. For each gene, the TSS was taken from the transcript with the longest genomic span, independently of expression data, and analyses of evenness were repeated for the 15,896 genes with a single annotated transcript (Supplemental Table S2).

CpG islands, retroelements and promoter classes

Genes whose TSS overlapped an island of the UCSC cpgIslandExt track (hg38) were CpG-island genes (13,414), and the longest overlapping island was used. Island-length analyses were repeated with the unmasked track (cpgIslandExtUnmasked). Alu, L1 and LTR elements were taken from the UCSC RepeatMasker track for hg38 (rmsk table, RepeatMasker 3.0.1, RepBase library of March 2, 2010). On each side of the TSS we recorded the nearest element of each type and of any type,

measuring the distance from the TSS to the nearer end of the element regardless of strand (0 for elements overlapping the TSS).

CpG-island genes were grouped by the type of the nearest element on the two sides as Alu–Alu, Alu-mixed (Alu on one side) or L1/LTR-only, and each configuration was divided by distance, giving nine classes (Fig. 3A). Alu–Alu genes were class 1 when the nearest element was within 2 kb on at least one side, class 2 beyond 2 kb on both sides, class 3 beyond 4 kb and class 4 beyond 7 kb upstream and 8 kb downstream. Alu-mixed genes were class 5 (within 2 kb), class 6 (beyond 2 kb) and class 7 (beyond 5 kb upstream and 4 kb downstream), and L1/LTR-only genes were class 8 (within 2 kb) and class 9 (beyond 2 kb). Non-CpG-island genes formed classes 10–12 by configuration. The 2- and 4-kb thresholds follow the distribution of distances to the nearest Alu (Supplemental Fig. S1A), and the outer thresholds enclose 95–96% of Alu–Alu genes and 77–91% of Alu-mixed genes on each side. Three pre-specified alternative outer bounds gave the same order of evenness of expression (Supplemental Methods; Supplemental Data S2), and applying the Alu–Alu bounds to Alu-mixed genes is examined in Supplemental Table S7. Class 1 is called Alu-near 1, class 5 Alu-near 2, classes 2 and 3 Alu-mid 1, class 6 Alu-mid 2, classes 4 and 7 Alu-far, and classes 8 and 9 L1/LTR.

Evenness of expression across tissues

Evenness across the 54 GTEx tissues was measured with the Gini coefficient and the CV of median TPM (Kryuchkova-Mostacci and Robinson-Rechavi 2017; O'Hagan et al. 2018). The Gini coefficient was $\Sigma_i \Sigma_j |x_i - x_j| / (2n^2\bar{x})$ and the CV was $\sigma/\bar{x}$, where x is the TPM in each of n = 54 tissues. Tissues without detected expression were kept as 0, and no gene of the set was zero in all tissues. Because island length co-varies with retroelement distance, distance groups of the 13,320 CpG-island genes with an element within 10 kb were compared with an expectation from 10,000 permutations of distance labels within island-length quintiles, and P values were adjusted over the three distance groups (Benjamini–Hochberg). Retroelement coverage was computed in 100-bp bins up to 10 kb from each edge of the 11,283 islands at TSSs of CpG-island genes, each island counted once (Fig. 2A). The

comparison of configurations within island-length quintiles (Fig. 2D) was pre-specified as non-overlap of the 95% bootstrap intervals of the Alu–Alu and L1/LTR-only medians.

Housekeeping genes and transcription factors

Housekeeping genes were taken from Eisenberg and Levanon (2013) and the HRT Atlas (Hounkpe et al. 2021), and transcription factors from Lambert et al. (2018; version 1.01, 1,637 genes). Housekeeping fractions were computed among the 18,730 protein-coding genes, because both lists consist of protein-coding genes, and transcription-factor enrichment among the 13,414 CpG-island genes, with Fisher's exact test and Mantel–Haenszel odds ratios over island-length quintiles.

Hematopoietic expression data and shares

The compiled dataset comprised 146 RNA-seq samples from seven GEO series (Supplemental Table S5), grouped as pluripotent (8), HSPC (12), lineage progenitor (14) and mature cells (112) and ranked 1–4 in this order. The bone-marrow hierarchy of Corces et al. (2016) comprised 13 sorted populations (49 samples from 4 donors), grouped as HSC/MPP, progenitors (LMPP, CMP, GMP, MEP and common lymphoid progenitors, CLP) and mature cells and ranked 1–3 in this order. GTEx v8 median TPM was obtained from the GTEx Portal, hematopoietic RNA-seq counts through Stemformatics (GSE107011, GSE69239, GSE65996, GSE89322, GSE89778, GSE73211 and GSE53094) (Choi et al. 2019) and from GEO (Corces et al., GSE74246; Granja et al., GSE139369; Ainciburu et al., GSE180298; Pellacani et al., supplemental tables of the article, raw data EGAS00001000552), E-MTAB-13862 from ArrayExpress, the Human Cell Atlas data through HCAData, and ChIP-seq and bisulfite sequencing data from the ENCODE portal (Supplemental Table S6). Raw counts were converted to TPM with one GENCODE v26 gene-length table and to CPM. The share of a promoter group was the summed expression of its genes divided by the summed expression of the 17,832 genes in the same sample. CPM served as a check that conclusions did not depend on length correction and was the primary scale for 3′ single-cell data.

Differences between HSC/MPP and CMP, GMP and MEP in the Corces data, and between

HSPCs and lineage progenitors in the compiled data, were summarized as the AUC, the probability that a sample of the earlier stage exceeds a sample of the comparison group, after centering each donor's values in the Corces data. LMPP and CLP were not included in this comparison because they were sorted in part or entirely from other donors (LMPP from two of the four donors and CLP from none), whereas HSC, MPP, CMP, GMP and MEP came from the same four donors. Associations with stage were Spearman correlations, with partial correlations adjusted for a proliferation score (mean z-score of log2(TPM + 1) of *PCNA*, *MKI67*, *TOP2A* and *CCNB1*) and, where stated, a dissociation-stress score (van den Brink et al. 2017). Gene contributions to a share were the gene's expression divided by the same denominator, and the gene-level analysis of the Alu-far difference was defined after the share results had been seen. For comparison, we computed the mRNAsi index (Malta et al. 2018), the number of detected genes after downsampling to a common depth, a proxy for CytoTRACE, and CytoTRACE 2 scores in single-cell data. Further details are given in Supplemental Methods.

Pre-specified tests

Predictions and decision rules were written into a version-controlled analysis record before the corresponding values were computed, and analyses so recorded are labeled pre-specified. Supplemental Data S2 lists the 42 of 120 recorded predictions that bear on results reported here. Of these, 17 were registered before any related result had been seen and 25 after partial knowledge of related results, and 24 were supported, 16 not supported and 2 partially supported. Directional predictions from the Corces results were tested in bulk bone-marrow RNA-seq from four donors (Lalioti et al. 2025, E-MTAB-13862), in single-cell data aggregated per donor and population from three donors (Granja et al. 2019) and eight donors (Ainciburu et al. 2023), and in a mammary epithelial hierarchy (Pellacani et al. 2016). A test of the leading Alu-far regulators was also registered in the Human Cell Atlas bone-marrow data of eight donors (Hay et al. 2018), with cells labeled by SingleR (Aran et al. 2019) against the Human Primary Cell Atlas reference (Mabbott et al. 2013). Registration was sequential, so later predictions were written with knowledge of earlier verdicts, and

the Alu-far share had been examined in the Human Cell Atlas data before its gene-list test was registered. A prediction was supported when the registered fraction of donors showed the registered direction and any registered AUC threshold was met. Before submission, two further pre-specified tests examined whether the distance gradient of evenness held within quintiles of regional Alu density and whether genes matched to Alu-far genes for island length reproduced the Alu-far difference (Supplemental Methods). Other comparisons were descriptive.

Chromatin and DNA methylation

ENCODE H3K4me3 and H3K27me3 ChIP-seq signal p-value tracks and whole-genome bisulfite sequencing data for $CD34^+$ HPC, T cells and B cells were used (Supplemental Table S6). Each cell type had one methylation dataset, and histone and methylation data came partly from different donors, none shared with the expression datasets. Promoters were analyzed in four groups without RP genes, Alu-near (classes 1 and 5), Alu-mid (classes 2, 3 and 6), Alu-far (classes 4 and 7) and L1/LTR (classes 8 and 9). Profiles were explored on odd-numbered autosomes and are reported for even-numbered autosomes. Signals were averaged in 100-bp bins over TSS ±2.5 kb (strand-oriented), within genes and then over genes, with 95% intervals from 1,000 gene-level bootstrap draws. Histone signal is shown relative to the mean of all four groups in the same cell type (divisors in Supplemental Table S8). Methylation was computed from CpGs outside RepeatMasker elements, because short reads map poorly within repeats. Promoters were H3K27me3-marked when an ENCODE H3K27me3 peak lay within 1 kb of the TSS, and enrichment was summarized as Mantel–Haenszel odds ratios over island-length quintiles.

On each side of each promoter, the methylation boundary was the first 100-bp bin outward from the TSS, up to 5 kb, that had at least two CpGs outside repeats, a mean methylation of at least 50% and a next informative bin also at or above 50%. Sides without such a bin were censored at their last informative bin. Expression tertiles were defined on all autosomes from Corces TPM of the matching population (HSC and MPP for HPC, $CD4^+$ T cells for T cells and B cells for B cells). Boundary distributions were estimated with Kaplan–Meier curves (both sides of a gene in one curve, intervals

clustered by gene), and median differences between the top and bottom tertiles were given with 2,000-draw gene-level bootstrap intervals. Cox models with a group × expression interaction summarized the relation between expression and the boundary (Supplemental Methods).

Statistics

Two-sided tests were used unless stated otherwise. Group differences used the Mann–Whitney U test, associations Spearman correlation and enrichment Fisher's exact test or Mantel–Haenszel odds ratios. Uncertainty of medians was given as 95% bootstrap intervals from 2,000 resamples of genes. Analyses used R v4.5.3 with the packages listed in Supplemental Methods. All comparisons are cross-sectional.

Software availability

The code that reproduces the figures and tables from the stored result tables, together with the scripts used to derive these tables, is provided as Supplemental Code. Inputs that are not public, the processed input tables and the version-controlled analysis repository, including the registration commits listed in Supplemental Data S2, are available from the authors on request.

Data access

No new sequencing data were generated in this study. The promoter classification of all genes, per-sample shares and gene contributions are provided as Supplemental Data S1, and the registered predictions bearing on the reported results as Supplemental Data S2. All analyzed datasets are public, with accessions listed in Methods and Supplemental Tables S5 and S6.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Sang-Uk Seo and Sun Shin (The Catholic University of Korea) for their assistance in interpreting the genomic data, and Hye-Jung Lee and Ji-Yon Jang (The Catholic University of Korea) for their assistance in data collection and organization. This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (Grant Nos. 2022R1F1A1073487 and RS-2023-NR076336).

Author contributions: HCP performed data analysis and interpretation of genome-wide datasets related to hematopoietic differentiation. HC analyzed transcriptomic data and generated data visualizations. SK contributed to statistical analysis and methodological evaluation. JC participated in study investigation and interpretation. JHO conducted data curation and dataset organization. MGR supervised the study and provided overall guidance. SJH performed statistical analyses and was a major contributor to manuscript writing. All authors read and approved the final manuscript.

References

Ainciburu M, Ezponda T, Berastegui N, Alfonso-Pierola A, Vilas-Zornoza A, San Martin-Uriz P, Alignani D, Lamo-Espinosa J, San-Julian M, Jimenez-Solas T et al. 2023. Uncovering perturbations in human hematopoiesis associated with healthy aging and myeloid malignancies at single-cell resolution. *eLife* **12**: e79363.

Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR et al. 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**: 163–172.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370–379.

Bulut-Karslioglu A, Macrae TA, Oses-Prieto JA, Covarrubias S, Percharde M, Ku G, Diaz A, McManus MT, Burlingame AL, Ramalho-Santos M. 2018. The Transcriptionally Permissive Chromatin State of Embryonic Stem Cells Is Acutely Tuned to Translational Output. *Cell Stem Cell* **22**: 369–383.

Byun HM, Heo K, Mitchell KJ, Yang AS. 2012. Mono-allelic retrotransposon insertion addresses epigenetic transcriptional repression in human genome. *J Biomed Sci* **19**: 13.

Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. 2008. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* **453**: 544–547.

Choi J, Pacheco CM, Mosbergen R, Korn O, Chen T, Nagpal I, Englart S, Angel PW, Wells CA. 2019. Stemformatics: visualize and download curated stem cell data. *Nucleic Acids Res* **47**: D841–D846.

Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ et al. 2016. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**: 1193–1203.

Costallat M, Batsche E, Rachez C, Muchardt C. 2022. The 'Alu-ome' shapes the epigenetic environment of regulatory elements controlling cellular defense. *Nucleic Acids Res* **50**: 5095–5110.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010–1022.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29**: 569–574.

Eller CD, Regelson M, Merriman B, Nelson S, Horvath S, Marahrens Y. 2007. Repetitive sequence environment distinguishes housekeeping genes. *Gene* **390**: 153–165.

ENCODE Project Consortium. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: 699–710.

Estécio MR, Gallegos J, Dekmezian M, Lu Y, Liang S, Issa JP. 2012. SINE retrotransposons cause epigenetic reprogramming of adjacent gene promoters. *Mol Cancer Res* **10**: 1332–1342.

Ferrari R, de Llobet Cucalon LI, Di Vona C, Le Dilly F, Vidal E, Lioutas A, Oliete JQ, Jochem L, Cutts E, Dieci G et al. 2020. TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping and Histone Acetylation. *Mol Cell* **77**: 475–487.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *J Mol Biol* **196**: 261–282.

Ge SX. 2017. Exploratory bioinformatics investigation reveals importance of "junk" DNA in early embryo development. *BMC Genomics* **18**: 200.

Graf T, Enver T. 2009. Forcing cells to change lineages. *Nature* **462**: 587–594.

Granja JM, Klemm S, McGinnis LM, Kathiria AS, Mezger A, Corces MR, Parks B, Gars E, Liedtke M, Zheng GXY et al. 2019. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology* **37**: 1458–1465.

GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**: 1318–1330.

Gulati GS, Sikandar SS, Wesche DJ, Manjunath A, Bharadwaj A, Berger MJ, Ilagan F, Kuo AH, Hsieh RW, Cai S et al. 2020. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**: 405–411.

Hay SB, Ferchen K, Chetal K, Grimes HL, Salomonis N. 2018. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp Hematol* **68**: 51–61.

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA. 2013. Super-

enhancers in the control of cell identity and disease. *Cell* **155**: 934–947.

Hong SJ, Lee HJ, Oh JH, Jung SH, Min KO, Choi SW, Rhyu MG. 2013. Age-related methylation patterning of housekeeping genes and tissue-specific genes is distinct between the stomach antrum and body. *Epigenomics* **5**: 283–299.

Hounkpe BW, Chenou F, de Lima F, De Paula EV. 2021. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res* **49**: D947–D955.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M et al. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41**: 178–186.

Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, Zhang X, Chavez L, Wang H, Hannah R et al. 2014. Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat Genet* **46**: 17–23.

Kang M, Gulati GS, Brown EL, Qi Z, Avagyan S, Armenteros JJA, Gleyzer R, Zhang W, Steen CB, D'Silva JP et al. 2025. Improved reconstruction of single-cell developmental potential with CytoTRACE 2. *Nat Methods* **22**: 2258–2263.

Kang MI, Rhyu MG, Kim YH, Jung YC, Hong SJ, Cho CS, Kim HS. 2006. The length of CpG islands is associated with the distribution of Alu and L1 retroelements. *Genomics* **87**: 580–590.

Kim YH, Hong SJ, Jung YC, Kim SJ, Seo EJ, Choi SW, Rhyu MG. 2006. The 5'-end transitional CpGs between the CpG islands and retroelements are hypomethylated in association with loss of heterozygosity in gastric cancers. *BMC Cancer* **6**: 180.

Kim YK, Cho B, Cook DP, Trcka D, Wrana JL, Ramalho-Santos M. 2023. Absolute scaling of single-cell transcriptomes identifies pervasive hypertranscription in adult stem and progenitor cells. *Cell Rep* **42**: 111978.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-

specificity metrics. *Brief Bioinform* **18**: 205–214.

Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS et al. 2008. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4**: e1000242.

Lalioti ME, Romero-Mulero MC, Karabacz N, Mess J, Demollin H, Rettkowski J, Schuldes K, Mitterer M, Wadle C, Shoumariyeh K et al. 2025. Differentiation, ageing and leukaemia alter the metabolic profile of human bone marrow haematopoietic stem and progenitor cells. *Nat Cell Biol* **27**: 1367–1380.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **172**: 650–665.

Laperriere D, Wang TT, White JH, Mader S. 2007. Widespread Alu repeat-driven expansion of consensus DR2 retinoic acid response elements during primate evolution. *BMC Genomics* **8**: 23.

Laurenti E, Göttgens B. 2018. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**: 418–426.

Liao J, Karnik R, Gu H, Ziller MJ, Clement K, Tsankov AM, Akopian V, Gifford CA, Donaghey J, Galonska C et al. 2015. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat Genet* **47**: 469–478.

Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, Lee TI, Young RA. 2012. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**: 56–67.

Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S et al. 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**: 68–73.

Losick R, Desplan C. 2008. Stochasticity and cell fate. *Science* **320**: 65–68.

Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. 2012. Revisiting global gene expression analysis. *Cell* **151**: 476–482.

Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. 2013. An expression atlas of human

primary cells: inference of gene function from coexpression networks. *BMC Genomics* **14**: 632.

Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kaminska B, Huelsken J, Omberg L, Gevaert O et al. 2018. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173**: 338–354.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12**: 1483–1495.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**: 553–560.

Morrison SJ, Kimble J. 2006. Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature* **441**: 1068–1074.

Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green DR, Tessarollo L, Casellas R et al. 2012. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**: 68–79.

Notta F, Zandi S, Takayama N, Dobson S, Gan OI, Wilson G, Kaufmann KB, McLeod J, Laurenti E, Dunant CF et al. 2016. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**: aab2116.

O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB. 2018. GeneGini: Assessment via the Gini Coefficient of Reference "Housekeeping" Genes and Diverse Human Transporter Expression Profiles. *Cell Syst* **6**: 230–244.

Oh JH, Rhyu MG, Jung SH, Choi SW, Kim SI, Hong SJ. 2014. Slow overmethylation of housekeeping genes in the body mucosa is associated with the risk for gastric cancer. *Cancer Prev Res (Phila)* **7**: 585–595.

Ong SLM, Baelde HJ, van IJzendoorn DGP, Bovée JVMG, Szuhai K. 2022. Identification of stable housekeeping genes for induced pluripotent stem cells and -derived endothelial cells for drug testing. *Sci Rep* **12**: 16160.

Orkin SH, Zon LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**: 631–

644.

Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, Wu AR, Churchman LS, Singh A, Raj A. 2015. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol Cell* **58**: 339–352.

Payer LM, Steranka JP, Kryatova MS, Grillo G, Lupien M, Rocha PP, Burns KH. 2021. Alu insertion variants alter gene transcript levels. *Genome Res* **31**: 2236–2248.

Pellacani D, Bilenky M, Kannan N, Heravi-Moussavi A, Knapp D, Gakkhar S, Moksa M, Carles A, Moore R, Mungall AJ et al. 2016. Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *Cell Rep* **17**: 2060–2074.

Percharde M, Bulut-Karslioglu A, Ramalho-Santos M. 2017. Hypertranscription in Development, Stem Cells, and Regeneration. *Dev Cell* **40**: 9–21.

Polak P, Domany E. 2006. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7**: 133.

Rhyu MG, Oh JH, Hong SJ. 2012. Epigenetic implication of gene-adjacent retroelements in Helicobacter pylori-infected adults. *Epigenomics* **4**: 527–535.

Signer RA, Magee JA, Salic A, Morrison SJ. 2014. Haematopoietic stem cells require a highly regulated protein synthesis rate. *Nature* **509**: 49–54.

Su M, Han D, Boyd-Kirkup J, Yu X, Han JJ. 2014. Evolution of Alu elements toward enhancers. *Cell Rep* **7**: 376–385.

Töhönen V, Katayama S, Vesterlund L, Jouhilahti EM, Sheikhi M, Madissoon E, Filippini-Cattaneo G, Jaconi M, Johnsson A, Burglin TR et al. 2015. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nature Communications* **6**: 8207.

Urrutia AO, Ocana LB, Hurst LD. 2008. Do Alu repeats drive the evolution of the primate transcriptome? *Genome Biol* **9**: R25.

van den Brink SC, Sage F, Vertesy A, Spanjaard B, Peterson-Maduro J, Baron CS, Robin C, van Oudenaarden A. 2017. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods* **14**: 935–936.

Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13**: 1998–2004.

Vinogradov AE. 2005. Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends Genet* **21**: 639–643.

Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D et al. 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**: 1134–1148.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementinformationRS20261110.pdf
- SupplementalDataS1.xlsx
- SupplementalDataS2.xlsx
- SupplementalCode.zip