Transposable Elements Shape Stemness in Normal and Leukemic Hematopoiesis

Giacomo Grillo[1,&], Bettina Nadorp[1,&], Aditi Qamra[1], Amanda Mitchell[1], Christopher Arlidge[1], Ankita Nand[1], Naoya Takayama[1], Alex Murison[1], Seyed Ali Madani Tonekaboni[1,2], Andrea Arruda[3], Jean C. Y. Wang[1,3,4], Mark D. Minden[1,2,3,4], Özgen Deniz[5,6], Héléna Boutzen[1], John E. Dick[1,7, *], Mathieu Lupien [1,2,8, *]

1. Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada, M5G 1L7
2. Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada, M5G 1L7
3. Division of Medical Oncology and Hematology, Department of Medicine, University Health Network, Toronto, Ontario, Canada, M5G 2M9
4. Department of Medicine, University of Toronto, Toronto, Ontario, Canada, M5G 1A1
5. Centre for Haemato-Oncology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London, UK, EC1M 6BQ
6. Queen Mary University of London Centre for Epigenetics, London, UK, E1 4NS
7. Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, M5S 1A8
8. Ontario Institute for Cancer Research, Toronto, Ontario, Canada, M5G 0A3

&. Equal contribution

* Corresponding author: John E. Dick: john.dick@uhn.ca and Mathieu Lupien: mathieu.lupien@uhn.ca

Keywords: chromatin accessibility, ATAC, epigenetics, cancer stem cell, tumor initiating cell, transposable elements, DNA repeats, repeat sequence, chromatin, hematopoiesis, leukemia, LSC, HSC, primitive, mature, acute myeloid leukemia, progenitor.

Abstract

Despite most acute myeloid leukemia (AML) patients achieving complete remission after induction chemotherapy, two-thirds will relapse with fatal disease within five years. AML is organized as a cellular hierarchy sustained by leukemia stem cells (LSC) at the apex, with LSC properties directly linked to tumor progression, therapy failure, and disease relapse [1–5]. Despite the central role of LSC in poor patient outcomes, little is known about the genetic determinants driving their stemness properties. As LSCs share many functional and molecular properties with normal hematopoietic stem cells (HSC) [6], we investigated accessible chromatin unique across normal hematopoietic and cancer cell states and identified transposable elements (TEs) as genetic determinants of both primitive populations in comparison with their downstream mature progeny. A clinically-relevant TE chromatin accessibility-based LSCTE121 signature was developed that enabled patient classification based on survival outcomes. Through functional assays, primitive cell specific-TE subfamilies were found to serve as docking sites for stem cell-associated regulators of genome topology or lineage-specific transcription factors, including LYL1 in LSCs. Finally, using chromatin editing tools, we establish that chromatin accessibility at LTR12C elements in LSCs are necessary to maintain stemness properties. Our work identifies TEs as genetic drivers of primitive versus mature cell states, where distinct TE subfamilies account for stemness properties in normal versus leukemic hematopoietic stem cells.

Main

Lifelong blood production originates from self-renewing, multipotent hematopoietic stem cells (HSCs) that give rise to the full hierarchy of progenitor and then mature hematopoietic populations [7]. The phenotypic heterogeneity across the hematopoietic hierarchy relies on population-specific genetic architectures, where distinct populations arise from the contribution of different sections of a shared genome [8–11]. For example, comparing chromatin accessibility across hematopoietic populations identified chromatin variants, defined as genetic elements varying in chromatin accessibility between populations along the differentiation axis [12], and revealed a role for CTCF in mediating changes to the three-dimensional genome organization to govern the transition of HSC from dormant, low primed to activated, primed states [13]. By contrast, chromatin variants specific to individual mature populations serve as cis-regulatory elements by acting as binding sites for lineage-specific transcription factors and promoting lineage-specific expression profiles [13,14]. Hence, identifying chromatin variants across hematopoietic populations reveals the DNA sequences undergoing chromatin reprogramming to drive phenotypic heterogeneity.

Akin to normal hematopoiesis, acute myeloid leukemia (AML) is hierarchically organized [6], with leukemia stem cells (LSCs) at the apex [1,2]. Because LSCs possess self-renewal capacity, they sustain the long-term generation of malignant populations and represent the source of evolutionary diversity, driving leukemic progression, therapy failure and relapse [3–5]. The clinical relevance of LSCs has been established in two ways. First, a LSC-specific gene expression signature score (LSC17) was developed and found to accurately predict both overall patient survival and response to standard induction therapy [15,16]. Second, mutational tracing studies in paired diagnosis/relapse AML samples showed the existence of genetically diverse, LSC-driven subclones present in AML diagnosis samples, where some LSC are already fated to survive induction therapy and fuel relapse [4]. Finally, the link between HSCs and LSCs during human leukemogenesis has also been made. Mutations that give rise to clonal hematopoiesis are amongst the first changes to the genetic architecture of HSCs that create a pre-leukemic HSC reservoir that further evolves leading to LSCs and fulminant AML [4,17,18]. However, we know little of the identity of additional non-genetic events such as chromatin variants that drive the transformation of HSCs into LSCs. Although there are examples where non-genetic events have been broadly linked to cancer stem cell properties [3,19,20], little is known of the specific features of such determinants. Thus, a clear understanding of the non-genetic events found in LSCs that are distinct from HSCs is required to better understand how cancer stemness is governed and to devise antagonistic strategies to improve therapies by ensuring that LSC are eradicated.

To date, the search for chromatin variants that represent determinants of phenotypic heterogeneity across normal hematopoietic populations and AML has mainly focused on the non-repetitive genome, but this excludes transposable elements (TEs), which compose ~50% of the human genome [21,22]. However, TEs account for diverse genetic functions relevant to development and pathogenesis [23]. TEs are classified in subfamilies composed of tens to thousands of individual elements traced back to a single ancestral unit [24]. In AML, as in many other cancer types, the expression of specific subfamilies is repressed to prevent activation of the viral mimicry response otherwise activated by the double-stranded RNAs they encode [25–27]. Other TE subfamilies lie in cancer-associated chromatin variants, where they function as *cis*-regulatory elements rather than encoding transcripts [28,29]. Such TE subfamilies potentiate the expression of neighboring genes and/or impact the three-dimensional genome organization by recruiting sequence-specific transcription factors, including CTCF [30–32]. This results in TE mediated oncogene expression in AML, as in various other cancer states [29,31,33]. Despite their reported contribution to cancer, the role of TEs as determinants of stemness properties, such as in LSCs, remains unknown. In the present study, we show how different TE subfamilies are enriched in chromatin variants identified across the continuum of both normal hematopoietic and LSC-driven leukemic hierarchies to modulate stemness properties.

Results

Chromatin variants from primitive and mature populations contain distinct TE subfamilies

To systematically study TE subfamily enrichment within accessible chromatin across hematopoietic populations, we used the Assay for Transposase-Accessible Chromatin Sequencing (ATAC-Seq; see Methods) data generated across six highly purified hematopoietic stem and progenitor populations (HSPC), including long-term (LT) and short-term (ST) HSCs, common myeloid progenitors (CMP), megakaryocyte-erythroid progenitors (MEP), multi-lymphoid progenitors (MLP) and granulocyte-monocyte progenitors (GMP), as well as seven mature cell populations exhibiting myeloid and lymphoid fates isolated from human cord blood donors (Extended Data Fig. 1A) [13]. These populations were selected according to the sorting scheme we developed (see Methods), which achieves the highest purity compared to other studies [14] and allows the isolation of HSCs with distinct repopulation and self-renewal potential (~30%) for improved resolution at the apex of the normal hematopoietic hierarchy (see Methods). We selected chromatin accessibility because it enables the identification of chromatin variants that define cell state identity [12–14,34,35].

The enrichment of all TE subfamilies (n=971, hg38, see Methods) within the accessible chromatin landscape of hematopoiesis was calculated using ChromVAR (see Methods). We used the genomic coordinates of TE subfamily elements instead of transcription factor DNA recognition sequence positions (see details in Methods) while still controlling for GC content. Unsupervised clustering of accessibility Z-scores for all TE subfamilies in different hematopoietic populations revealed that LT-, ST-HSCs and later progenitor populations, collectively referred to as HSPC, share similar accessible TE subfamily enrichment profiles and collectively, these are distinct from mature populations (Fig. 1A). The megakaryocyte-erythroid lineage is the exception, with MEPs clustering with erythroid precursors that are more similar to other mature rather than HSPC populations (Fig. 1A) and could be caused by a strong lineage-specific enrichment of TE subfamilies. To show that the megakaryocyte-erythroid lineage-specific TE subfamilies are driving the separation in the unsupervised clustering, we identified the strongest megakaryocyte-erythroid-specific TE subfamilies (Benjamini-Hochberg corrected p-value (q-value)<0.0001, Extended Data Fig. 1B) and subsequently removed them from the list of TE subfamilies used in unsupervised clustering. This resulted in a clear-cut separation of HPSC versus mature populations (Extended Data Fig. 1C). Thus, our data demonstrate that the chromatin variants reported from chromatin accessibility data across specific TE subfamilies distinguish HSPC from mature hematopoietic cells.

To identify which differentially accessible TE subfamilies are distinct between HSPC versus mature populations, we compared the accessibility Z-scores of all TE subfamilies using a two-sided Wilcoxon signed rank test. We found 81 and 128 TE subfamilies belonging to 12 TE families were enriched in the HSPC and mature populations, respectively ($q < 0.01$, Fig. 1B, Extended Data Fig. 1D, Extended Data Table 1). HSPC populations were mainly enriched in subfamilies belonging to ERV1 and ERV3 TE families (Fig. 1B). Enriched ERV1 subfamilies included MER61A (q=1.10e-05) and LTR39 (q=1.87e-06), while enriched ERV3 subfamilies included LTR16E1 (q=4.61e-07) and LTR33 (q=9.05e-07) (Fig 1C, Extended Data Fig. 1E), exemplified at unique elements (Fig. 1D). In contrast, mature populations were mainly enriched in SINE1, SINE2 and Transposon subfamily members (Fig. 1B), such as AluJB (q=1.26e-07) and AluSx (q=1.08e-07) of the SINE1 family, as well as MIR (q=2.93e-07) and MIR3 (q=2.93e-07) of the SINE2 family (Fig. 1C, Extended Data Fig. 1E). Despite finding a similar number of accessible DNA elements (aka: peaks) in each mature population (Extended Data Fig. 1F), the enrichment of SINE1, SINE2 and Transposons subfamily members relied on population-specific accessible elements. This contrasts with HSPC, where the same elements from

ERV1, ERV3 subfamily members are accessible across all stem and progenitor populations. In conclusion, our study demonstrates distinct chromatin accessibility patterns of TE subfamilies in HSPC versus mature populations, highlighting their potential role in cell state identity within hematopoiesis.

Accessible TE subfamilies that serve as docking sites for stem factors define primitive hematopoietic populations

We next assessed if chromatin variants over TE subfamilies could distinguish LT-HSCs (the most primitive HSC subset) from any other hematopoietic population (Extended Data Fig. 1A). Comparison of the accessibility Z-scores of all TE subfamilies did not identify differentially-enriched TE subfamilies across all one-to-one comparisons (Extended Data Table 2). We next compared stem versus progenitors and mature populations. Comparison of stem populations against the collection of mature populations revealed 61 TE subfamilies enriched in stem populations versus 119 in mature populations ($q < 0.01$, Extended Data Fig. 2A, Extended Data Table 3), while comparison of progenitor to mature populations identified 56 TE subfamilies enriched in progenitor versus 89 in mature populations ($q < 0.01$, Extended Data Table 4). Of all accessible TE subfamilies found to be distinct between HSPC compared to mature populations (Fig. 1A), twenty-three TE subfamilies were specific to the stem populations (Extended Data Fig. 2B-C). Interestingly, LTR85a, MLT1E2, and HERV4_I-int were the only two also enriched in stem populations compared to progenitor populations ($q < 0.05$, Extended Data Fig. 2D, Extended Data Table 5). Notably, MER4B-int, MER57B1, MER57B2, MER57-int, LTR16A2, LTR67B, and LTR85a TE subfamilies showed the highest Z-scores in the LT/HSPC and/or Act/HSPC signatures of chromatin accessibility that we previously showed could discriminate between LT-HSC, ST-HSC and other progenitor populations [13] (Extended Data Table 6). Collectively, our results reveal that chromatin variants over specific TE subfamilies discriminate HSPC from mature populations, with a limited set of twenty-three that are unique to the stem populations.

Each TE subfamily harbors DNA recognition sequences allowing them to recruit transcription factors and guide cell differentiation [29,30,36–39]. We used the ReMap atlas of 485 transcription factor cistromes (see Methods) to examine the enrichment of these sequences over the twenty-three TE subfamilies that were specific to stem populations. We found enrichment of HOXA9, a well-known HSC factor [40], at elements of the MER57B1 (q=2.88e-13, logOR=27.69) and MER57B2 (q=0.0007, logOR=11.95) (Fig. 1E, Extended Data Fig. 2E) subfamilies. The cistrome for RUNX1, another well-known HSCs transcription factor [41], was found to be enriched over elements of the MLT1E2 (q=0.03, logOR=2.12) (Extended Data Fig. 2F). While none of the 485 cistromes in ReMap were enriched over HERV4_I-int, HERVK9-int, LTR80A, LTR85a, LTR9A1, MamGypLTR1b, MER127, MER4B-int, MER57-int, MER83B-int, MLT2E and PRIMA41-int elements, the cistromes of CTCF and/or cohesin complex factors (SMC1A2, SMC3 and RAD21), known regulators of genome topology [42], were enriched over MamGypsy2-LTR and LTR16A2 elements (Fig. 1E, Extended Data Fig. 2G, Extended Data Table 7). We validated the enrichment of these stem transcription factors by computing an independent score, the GIGGLE enrichment score [43]. Our analysis revealed enrichment of HOXA9 cistrome over MER57B1 and MER57B2 elements (Fig. 1F, Extended Data Fig. 2H), enrichment of RUNX1 cistrome over MLT1E2 elements (Extended Data Fig. 2I) and enrichment of cistromes from CTCF and cohesin complex factors over MamGypsy2-LTR and LTR16A2 elements (Fig. 1F, Extended Data Fig. 2J). In parallel, we performed sequence analysis and identified the enrichment of the HOXA9 or CTCF DNA binding motifs within MER57B1 or MamGypsy2-LTR elements, respectively (Fig. 1G). Collectively, these results suggest that TE subfamilies uniquely accessible in stem populations provide docking sites for known “stem” transcription factors, including HOXA9, RUNX1, and components of the three-dimensional genome organization (CTCF, cohesin).

LSCs exhibit HSPC-like TE subfamily enrichment within accessible chromatin

Considering the role of LSCs in AML etiology, we next examined whether chromatin variants over TE subfamilies could be used to discriminate populations enriched for self-renewing LSCs from leukemic cell fractions lacking LSCs as well as from normal HSPC and mature hematopoietic populations. Uncultured AML patient samples (n=15) were sorted into fractions based on CD34 and CD38 expression, and fractions were functionally classified as LSC+ (n=11) or LSC- (n=24) based on their ability to generate a leukemic graft upon xenotransplantation into immunodeficient mice, as described previously [15,44]. In parallel, we performed ATAC-Seq and measured TE subfamily enrichment within the accessible chromatin profiles of LSC- and LSC+ fractions. Unsupervised clustering of Z-scores of all TE subfamilies in LSC+ and LSC- fractions, as well as normal hematopoietic populations, showed clustering of all LSC+ fractions with the HSPC populations (Fig. 2A). In contrast, most LSC- fractions (17/24) clustered not with LSC+ fractions but rather with mature hematopoietic populations (Fig. 2A). The discordance between the TE subfamily enrichment profile classification (as HSPC-like) and the functional classification (as LSC-) for the 7 LSC- fractions that clustered with the LSC+ fractions (2 CD34+/CD38-, 4 CD34+/CD38+, 1 CD34-/CD38+; Fig. 2A-B) may draw from low LSC frequency within these populations. In conclusion, our analysis reveals an association between chromatin variants over TE subfamilies and the discrimination of LSC populations, providing a nuanced classification for AML biology.

Direct comparison of accessibility Z-scores of all TE subfamilies between LSC+ and LSC- fractions revealed 44 significantly enriched accessible TE subfamilies in LSC+ and 77 in LSC- fractions (Fig. 2B-C, Extended Data Fig. 3A-B, Extended Data Table 8). LSC+ fractions were mostly enriched for ERV1 and ERV3 families, whereas LSC- fractions showed enrichment of SINE1, SINE2, and Transposon families (Fig. 2C). Next, we compared accessible TE families and subfamilies across normal hematopoietic populations and LSC fractions. Overall, the proportions of TE families discriminating HSPC vs mature and LSC+ vs LSC- fractions were skewed compared to the proportions of TEs distributed across the human genome (Fig. 2D). Specifically, accessible TE families enriched in HSPC populations and LSC+ fractions consisted preferentially of ERV1 and ERV3 families (Fig. 2D). In contrast, accessible TE families enriched in mature populations and LSC- fractions had higher representation of SINE1 and Transposons families (Fig. 2D). The majority of TE subfamilies enriched in the accessible chromatin of LSC+ or LSC- fractions were shared with HSPC (21/44) or mature populations (53/77), respectively (Fig. 2E). In line with the similarities observed in the overall TE family distribution, the majority of common subfamilies between LSC+ fractions and HSPC populations (14/21 subfamilies) belonged to the ERV1 and ERV3 families (Fig. 2E, Extended Data Fig. 3C-D). In contrast, the majority of common subfamilies between LSC- fractions and mature populations belonged to the SINE1 and Transposons families (Fig. 2E, Extended Data Fig. 3C-D). Z-score comparisons of the enrichment in the common subfamilies between HSPC/LSC+ or mature/LSC- showed largely concordant enrichment (Extended Data Fig. 3D). Two TE subfamilies enriched in LSC+, but not HSPC accessible chromatin, namely LTR12C (ERV1 family) and L1PA7 (LINE1 family), showed the largest discordance between HSPC and LSC+ populations (Extended Data Fig. 3E). Although not discriminating LSC+ from LSC- fractions, all LINE1 families enriched in HSPC versus mature hematopoietic populations were more enriched in LSC+ compared to HSPC populations (Extended Data Fig. 3E); in line with the DNA hypomethylation over LINE1 elements reported across many cancers [45]. Collectively, our data suggest that LSCs can be discriminated from committed leukemic cells based on the enrichment of specific TE subfamilies within their accessible chromatin that are largely shared with normal HSPC populations.

An accessible TE subfamily-based signature shows clinical value as a marker of relapse in AML

Stemness is related to survival and relapse in AML and other cancer types [46,47]. We hypothesized that TE subfamilies enriched in LSC+ and LSC- fractions (44 and 77 families enriched in LSC+ and LSC- respectively, as shown in Fig 2B-C, hereafter referred to as LSCTE121) could serve as a signature for stemness in AML and accordingly that this would be associated with clinical outcomes. To this end, we performed ATAC-seq on peripheral blood samples from two independent cohorts of AML patients (cohort1, n=29; cohort2, n=60) (Extended Data Table 9) and assessed each sample for their LSCTE121 signature score. Unsupervised clustering based on the LSCTE121 score grouped 9 of the cohort1 AML samples with the LSC- fractions whereas the other 20 AML samples clustered in the same main branch as the LSC+ fractions (Extended Data Fig. 4A). Similarly, 20 of the cohort2 AML samples clustered with the LSC- fractions whereas the other 40 clustered in the same main branch as the LSC+ fractions (Extended Data Fig. 4B). This classification of bulk AML samples into LSC-like signature score and non-LSC-like signature score is reminiscent of recent gene expression-based patient stratification, lending credence to our approach [48,49]. To simplify the interpretation of TE subfamily enrichment, we created a combined TE Z-score restricted to the LSCTE121. This score uses Stouffer's method to combine the enrichment of LSC+ and the depletion of LSC- TE subfamilies (see Methods). Comparison of the combined LSCTE121 Z-scores of samples in the major branches of the heatmaps (Extended Data Fig. 4A-B) showed the highest scores in the branch containing the LSC+ fractions (red branch). As expected, LSC+ fractions had higher LSCTE121 Z-scores than LSC- fractions (Extended Data Fig. 4C). Further, there was no correlation between the blast count of a given sample and the LSCTE121 Z-score in both cohorts (cohort1: R=0.031, p=0.88; cohort2: R=-0.17, p=0.21; Extended Data Fig. 4D-E), suggesting that the combined LSCTE121 Z-score stratification is a reflection of stemness and not of blast count. Kaplan-Meier estimates comparing the top 25th percentile with the bottom 25th percentile of LSCTE121 Z-score AML patients from cohort1 showed significant differences in disease-free and overall survival (Fig. 3A-C, Extended Data Fig. 4F), with in high and low LSCTE121 subgroups. Additionally, when compared with the LSC17 score, a gene expression-based score shown to predict overall patient survival and response to standard induction therapy [15], there was variation in how these two scoring methods classified samples [15] (Fig. 3A and 3D). We validated the prediction of overall survival in our patient cohort by using the LSC17 score (Extended Data Fig. 4G, log-rank p-value=2e-04) and we then sought to determine the similarity between the LSCTE121 Z-score and the LSC17 score. We observed low correlation between the LSC17 score and the LSCTE121 Z-score (R=0.27, p=0.16; Fig. 3D). However, the LSCTE121 score was significantly higher in fractions with the highest LSC frequency, assessed with limiting dilution assays, supporting the capture of stemness properties by the LSCTE121 Z-score (Fig. 3E). Furthermore, significant differences in disease-free survival were also observed with cohort2 AML patients (Fig. 3F-H, Extended Data Fig. 4H). Collectively, our results suggest that the LSCTE121 Z-score captures a different underlying biology than the LSC17 score and that each score provides risk stratification based on complementary discriminators of stemness.

TE subfamilies accessible in LSCs are docking sites of AML essential factors

Using the ReMap cistrome data, we next assessed if accessible TE subfamilies that were shared or unique across primitive hematopoietic populations and LSC+ fractions (Fig. 2E; Extended

Data Tables 10-12) could provide docking sites for different transcription factors. Cistromes for four transcription factors were found to be enriched across accessible TE subfamilies in LSC+ and HSPC populations (Fig. 4A-B and Extended Data Fig. 5A-B). These included ERG, RUNX1, LMO2 and TRIM28. Each of these transcription factors is documented to contribute to normal hematopoiesis and/or leukemia development [41,50–52]. Beyond its reported role in erythroblast differentiation [51], TRIM28 is also known to suppress TE function [53], suggesting that it may play a role in the transition from HSPC and LSC+ to mature fates. In agreement with previous studies reporting an interplay between ERG and RUNX1 in hematopoiesis and AML [54], the ERG and RUNX1 cistromes were enriched over more than 70% of the same accessible TE subfamilies (10/15 for HSPC-specific, 8/10 for LSC+ specific and 6/8 for TEs shared between HSPC and LSC+ Extended Data Fig. 5C), including LTR78 (ERG: q=0.005, logOR=1.74; RUNX1: q=0.0004, logOR=1.9), LTR67B (ERG: q=1.12e-08, logOR=2.73; RUNX1: q=5.77e-17, logOR=4.09) and MLT1E2 (ERG: q=0.0003, logOR=2.02; RUNX1: q=0.009, logOR=1.73 and Extended Data Fig. 5D, Extended Data Table 10-12). Over half of the TE subfamilies enriched for ERG and/or RUNX1 cistromes also preferentially harbored their DNA recognition sequences (8/25 for ERG motif and 11/32 for RUNX1 motif; Extended Data Fig. 5E-F). Collectively, these results reflect the shared biology across HSPC hematopoietic populations and LSC, relying on four core transcription factors, namely ERG, RUNX1, LMO2, and TRIM28, using accessible elements from TE subfamilies as direct docking sites on the genome.

Cistrome enrichment analysis over accessible TEs identified eleven transcription factors specific to HSPC populations, including HIF1A, SETDB1, GATA1, ZNF143, TEAD4, PBX2, EZH2, CBX3, SMC3, MAX, HDAC2 (Fig. 4B, Extended Data Fig. 5A). While no reports implicate CBX3 and MAX in hematopoiesis, a previous characterization of PBX2 knock-out mice showcased that PBX2 is not essential for normal hematopoiesis or immune response [55]. Similarly, TEAD4 favors embryonic to blood cell state commitment [56], and GATA1, HIF1A, SETB1, EZH2, and HDAC2 are known regulators of HSCs self-renewal and/or differentiation [57–61]. ZNF143 and SMC3 define a unique class of transcription factors that regulate the three-dimensional genome organization by binding to chromatin interaction anchors in collaboration with CTCF [62–64]. In HSPC populations, ZNF143, SMC3, CTCF cistromes, and those of their known partners SMC1A, RAD21 and CHD8 [62,65,66], enrich over LTR41B elements ($q \leq 3.22e-34$, $OR \leq 9.36$) (Extended Data Fig. 5D, Extended Data Table 10), in line with the role in 3D chromatin organization of LTR41B elements [67]. The accessibility of TE subfamilies serving as docking sites for known HSC factors and/or chromatin interaction factors parallels the enrichment of the GATA1 and CTCF DNA recognition sequences at some of these TEs (Extended Data Fig. 5G) and is in keeping with the role of CTCF as a gatekeeper of stemness in hematopoietic populations [13]. Together, these results suggest that specific TE subfamilies are accessible in HSPC populations, serving as DNA docking sites for HSC factors and/or regulators of genome topology.

Finally, we found four transcription factors whose cistromes were uniquely enriched over accessible TE subfamilies in LSC, namely LYL1, NFYA, NFYB, and POU5F1 (Fig. 4A-B, Extended Data Table 11). From the GIGGLE enrichment score analysis, we observed a bias for this enrichment for cistromes generated in blood or bone marrow samples for LYL1, NFYA and NFYB (Fig. 4C). TE subfamilies associated with NFYA and NFYB binding were also enriched for the DNA recognition sequence for these transcription factors (Fig. 4D). We next assessed the necessity of these transcription factors to growth in AML cell lines. Analyzing the CRISPR-based knock-down essentiality screen results from the Cancer Dependency Map (DepMap) project (see Methods) revealed negative essentiality scores for all four transcription factors across the majority (12/23) of AML cell lines (Extended Data Fig. 6A). While the median essentiality score across all 23 AML cell lines was not significantly lower for POU5F1 compared to all other genes, LYL1, NFYA and NFYB were consistently reported as essential (p-value<0.05; Extended Data Fig. 6A). No significant essentiality for these transcription factors were detected from the DepMap RNAi-based knock-down as opposed to knock-out screen (Extended Data Fig. 6B). When comparing essentiality scores across cells classified by cancer types, LYL1 was significantly more essential in AML cell lines than cells from other cancer types (CRISPR-based screen: 97%, 34/35; RNAi-based screen: 75%, 17/21) (Fig. 4E, Extended Data Fig. 6C, Extended Data Tables 13,14), in agreement with its reported LYL1 pro-oncogenic role in AML [68]. NFYA and NFYB factors were found essential across all cancer types (Fig. 4E, Extended Data Fig. 6C). Collectively, our results suggest that the accessibility of TE subfamilies specific to LSC defines docking sites for transcription factors, including LYL1, NFYA, and NFYB essential in AML.

TE subfamilies accessible in LSC are essential for stemness properties

We undertook functional studies to assess the contribution of TEs to stemness properties using the OCI-AML22 model of AML, known to harbor functionally assessed LSC+ and LSC- populations that can be enriched based on cell surface marker expression (CD34+CD38- fraction, very high LSC frequency; CD34+CD38+, very low LSC frequency; and CD34- (either CD38+ or CD38-) fractions, LSC-) [44]. From ATAC-seq profiles generated across each OCI-AML22 fraction, we measured TE subfamily enrichment within accessible chromatin. Using the LSCTE121 signature for unsupervised clustering, we observed that the LSC+ fraction of OCI-AML22 cells (CD34+/CD38-) clustered with LSC+ AML fractions (Fig. 5A). In contrast, the OCI-AML22 fractions functionally classified as LSC- (CD34-) [44] clustered with the majority of LSC- AML fractions (Fig. 5A). This demonstrates that the OCI-AML22 model faithfully recapitulates the heterogeneity in chromatin accessibility found across LSC+ and LSC- populations in primary AML samples, where LSC+ cells drive the production of LSC- populations. In agreement with increased LSC+ content in the CD34+/CD38- versus CD34+/CD38+ fractions in OCI-AML22 [44], the LSCTE121 score was higher in the CD34+CD38- compared to CD34+/CD38+ OCI-AML22 fraction (Extended Data Fig. 7A-B). This observation reflects the value of the LSCTE121 score to quantify stemness potential in AML.

To directly test whether TE subfamilies are functionally required to drive stemness properties we turned to novel chromatin editing technologies that allow altering the chromatin state across elements from an individual TE subfamily. We focused on the LTR12C TE subfamily, which exhibited one of the most differential chromatin accessibility between LSC+ and HSPC populations (Extended Data Fig. 3C). As expected, the LTR12C subfamily was more enriched in accessible chromatin in LSC+ compared to LSC- and LTR12C elements were overall more accessible in LSC+ compared to LSC- fractions (Fig. 5B-C). Furthermore, the LTR12C subfamily was more enriched in chromatin accessibility in CD34+/CD38- LSC+ compared to other fractions from OCI-AML22, and the enrichment decreased concomitantly as cells lost stemness and became more mature across the cellular hierarchy (Fig. 5D). To functionally assess the role of the LTR12C subfamily towards LSC+ properties, we stably expressed the CRISPR/dCas9-KRAB (CRISPRi) construct [12,69,70] in the OCI-AML22 model and evaluated if altering the chromatin state of LTR12C elements could inhibit LSC+ stemness and its ability to regenerate a cellular hierarchy (Fig. 5E, Extended Data Fig. 7C-D, see Methods). By using the Repguide tool [71], we generated a combo of 5 guide RNAs targeting specifically the LTR12C subfamily (n=2765 genomic copies), predicted to recognize 2400 copies (87%). As a negative control, we used a combo of 5 scramble guide RNAs (Extended Data Table 15). To test the specificity of our CRISPRi targeting strategy, we profiled chromatin accessibility by ATAC-seq in CRISPRi expressing OCI-AML22 cells co-expressing either scramble or LTR12C guide RNAs. All three biological replicates revealed a decreased chromatin accessibility at LTR12C elements ($p \leq 8e-04$) with limited off-target effect when assessed on phylogenetically proximal TE subfamilies, such as LTR12D ($p > 0.5$) (Fig. 5F-

G, Extended Data Fig. 7E-G), reflecting the specificity of the chromatin editing approach towards LTR12C elements. Functional flow based analysis showed that when, chromatin was repressed over LTR12C elements there was a reduced percentage of the LSC+ cellular fraction (CD34+/CD38- fraction) as compared to control guides (CD34+/CD38- percentages: 62.5% in control and 44.3% in LTR12C, Fig. 5H). In parallel there was an increase of CD34+/CD38+ cells that represent more committed phenotype (CD34+/CD38+ percentages: 27.2% in control and 45.4% in LTR12C, Fig. 5H). Collectively, these results establish that accessibility of LTR12C elements is a stemness determinant in LSC.

Discussion

Our study establishes that enrichment of specific TE subfamilies within chromatin variants serves as a hallmark of stemness in both normal hematopoiesis and AML, with direct clinical implications for AML patients. Many of the TE subfamilies enriched in chromatin variants of both LSC and HSPC populations are shared with those found enriched in chromatin variants identified in embryonic stem versus mature cells [29,72]. This includes members of the endogenous retrovirus (ERV) families known to act as docking sites for stem transcription factors, such as OCT4 and NANOG [29,36,37]. Expanding on the role of TEs as a source of *cis*-regulatory elements in stem and mature cells [28,29,36–39,72], our results show that HSPC-accessible TEs, such as LTR41, LTR41B and LTR55 are permissive for binding by regulators of the three-dimensional genome organization, including ZNF143, CTCF, and cohesin elements. CTCF-driven changes to the genome topology in primitive hematopoietic populations were identified as a regulator of quiescence exit and long-term self-renewal [13]. Here, additional transcription factors, namely ERG, RUNX1, LMO2, and TRIM28, were found based on the presence of their occupancy across elements from the LTR78, LTR67B and MLT1E2 TE subfamilies enriched in chromatin variants from LSC and HSPC populations. ERG, RUNX1, LMO2, and TRIM28 correspond to known hematopoietic transcription factors and reinforce their direct role in AML [41,50–53]. Collectively, our results suggest a contribution of distinct TE subfamilies to serve as determinants of stem properties, shaping both genome topology and transcriptional regulation and defining the cell state hierarchy in both normal and disease settings.

By revealing how differentially enriched TE subfamilies distinguish LSCs from mature leukemic cells, we provide a chromatin-based approach to interrogate repeat DNA sequences and discriminate AML samples and/or isolated populations according to their leukemia-initiating potential. When applied to AML patient samples, our LSCTE121 scoring scheme can stratify patients to identify those with high rates of relapse and shorter overall survival. The lack of correlation between our LSCTE121 and the LSC17 [15] scoring scheme argues for their complementarity in assessing the LSC content in primary patient samples. Although there are large AML datasets of RNA, there are few that interrogate chromatin variants at scale; our work points to the need for a deeper study of the repetitive genome and its contribution towards stemness properties in leukemia, but also cancer in general.

In cancer, chromatin variants are predominantly found at TEs [73] and can lead to oncogene activation [31]. However, studies on the role of TEs in cancer have primarily focused on their expression. For instance, high levels of ERV1 expression are linked to worse survival in patients with kidney cancer [74], and a 14 TE-based signature predicts survival in AML [75]. Our work extends beyond this TE expression-based perspective to the role of TEs as regulatory elements. Among the TE subfamilies accessible in LSCs, for example, we observed the LTR12C, whose elements work as enhancers in AML [33] and were previously shown to act as alternative promoters in hepatocellular carcinoma [76]. We further show that accessible TEs provide docking sites for oncogenic transcription factors in LSCs.

Using this approach, we identified LYL1 as an oncogenic driver of AML, in agreement with previous reports [68]. We further showed that, unlike NFYA/B factors that were essential across all cancer states, LYL1 is preferentially essential across AML cell lines as compared to cell lines from other cancer states. Altogether, our work suggests a direct contribution of the repetitive genome towards the governance of oncogenic transcription factors in AML.

Previous studies reported how altering the chromatin state at TEs with CRISPRi approaches could impact tumor growth in AML and prostate cancer [29,33]. We used the same technology to showcase the essential role of LSC+-specific TE subfamilies in maintaining stemness. We were specifically able to target hundreds of LTR12C elements, promoting their repression at the chromatin level without altering the underlying sequence. Such chromatin repression caused a drop in the percentage of LSC+ cells and an increase in more mature cell types in OCI-AML22 cells. These results reveal the essential role of LSC+-specific TEs on stemness properties and uncover LSCs' genetic vulnerabilities that, when targeted, should not affect HSPC populations. Altogether, the work presented here provides a framework to capture the primitive nature of cancers by revealing the role of TE subfamilies as genetic determinants of stemness properties in normal and leukemic stem populations.

Methods

Patient samples

All biological samples were collected with informed consent according to procedures approved by the Research Ethics Board of the University Health Network (UHN; REB# 01-0573-C) and viably frozen in the Leukemia Tissue Bank at Princess Margaret Cancer Centre/ University Health Network. Patient details can be obtained from Extended Data Table 9.

Cell sorting and Xenotransplantation assay

Hematopoietic populations were selected using the highest purity sorting scheme available [77,78]. Peripheral blood AML samples were sorted into fractions based on the expression of CD34 and CD38. LSC activity in each fraction was assessed by xenotransplantation into $NOD.Prkdc^{scid}.Il2rg^{null}$(NSG) mice as previously described [15]. Antibodies used are shown in Extended Data Table 16.

Bulk ATAC-seq

ATAC-seq was used to profile the accessible chromatin landscape of 11 LSC+, 24 LSC- fractions, and two independent cohorts of peripheral blood AML bulk samples (cohort1: n=29 12 LSC17 high, 17 LSC17 low; cohort2: n=60, unknown LSC17 score). 50,000 cells from each sample were processed as previously described [79]. Libraries were sequenced with 50bp single-end (LSC fractions) or 50bp paired-end (AML bulk samples) reads and mapped to hg38. Reads were filtered to remove duplicates, unmapped or poor quality (Q < 30) reads, mitochondrial reads, ChrY reads, and those overlapping the ENCODE blacklist. Following alignment, accessible chromatin regions/peaks were called using MACS2. Default parameters were used except for the following: --keep-dup all -B --nomodel --SPMR -q 0.05. ATAC-seq profiles of hematopoietic cell populations were downloaded from GSE125064 and processed similarly. We only included LSC fractions from patients without NPM1 mutation and set a threshold of at least 10 million reads and more than 15k called peaks. ATAC-seq data of LSC fractions and bulk AML samples can be accessed at EGAS00001004893, EGAS00001004896, and EGAS00001007191, respectively. Narrowpeak files are submitted to GEO under ID GSE234237.

TE families and families

Bedfiles of repeats classified as TE families (excluding hence simple repeats, satellites, sn/rRNA, and repeats of unknown families) and with more than 100 individual genomic elements (n=971) in the hg38 human genome build were downloaded from the UCSC Genome Browser. Subfamilies were classified into families according to Repbase (https://www.girinst.org/repbase/) [80,81].

TE enrichment in accessible chromatin analysis

TE enrichment analysis was performed using ChromVAR [82]. The binary matrix for ChromVAR was generated by identifying the presence/absence of accessible chromatin peaks in each sample of interest. For the same set of peaks, a binary annotation matrix was generated with rows as peaks and columns as each TE subfamily. TE subfamily-to-peak assignment was generated by checking the overlap of TEs for each subfamily in the accessible chromatin peak set. ChromVAR was run with default parameters using the two input matrices, generating a bias-corrected deviation Z-score. Deviation Z-scores were then compared between samples of interest using a non-parametric two-sided Wilcoxon signed rank test. For plotting, heatmap colors for Z-scores were truncated to minimum and maximum values of -10 and 10, respectively, and clustering was based on Euclidean distance unless otherwise stated. For the LSC TE Z-score, we combined all LSC+/LSC- enriched TEs in separate .bed files and calculated the Z-score over all LSC+/LSC- enriched TEs separately. To combine both Z-scores, we used Stouffer's method with equal weights of the LSC+ Z-score and the inverse LSC- Z-score:

$$Z_{repeats} = \frac{0.5 \cdot Z_{LSC+} + 0.5 \cdot -(Z_{LSC-})}{\sqrt{(0.5)^2 + (0.5)^2}}$$

ATAC signal was quantified over LTR12C elements accessible in LSC+ and/or LSC- using deeptools [83]. Given the different number of samples for LSC+ and LSC- fractions (11 and 24 respectively), we reported ATAC signal of top and bottom 25% of samples based on LSCTE121 score. The top 25% corresponds to LSC+, while the bottom 25% is composed uniquely of LSC- and it was therefore named LSC-.

Transcription factor cistromes and motifs enrichment analysis

The enrichment of cistromes and motifs was assessed at the TE subfamilies enriched specifically in LT-HSCs, HSPC populations, LSC+ and commonly enriched in primitive populations and LSC+(5, 70, 21 and 23 TE families respectively; Extended Data Fig. 2C, Fig. 2E). Each TE subfamily was analyzed individually, restricting our analysis to the elements overlapping with a given consensus set of ATAC peaks (e.g. for a subfamily enriched in primitive populations, we inquired transcription factor cistromes and motifs enriched at the elements overlapping with the ATAC peaks catalog of primitive populations). Enrichment of transcription factor cistromes was performed using the Bioconductor package LOLA (https://bioconductor.org/packages/release/bioc/html/LOLA.html) [84], version 1.20.0. 485 transcription factor cistromes were obtained from ReMap (http://remap.univ-amu.fr/) [85]. Enriched transcription factor cistromes at individual TE subfamilies were called against all elements from the 971 TE subfamilies overlapping any of the primitive/mature/LSC+/LSC- ATAC peaks ($q < 0.05$). The top 5% most frequently enriched are shown in Fig. 4A, Extended Data Fig. 5A-B. HOMER motif discovery tool (findMotifsGenome.pl) version 4.7 was used to compute the enrichment of motifs. Enriched motifs within individual TE subfamilies (as explained above) were called against all elements from the 971 TE subfamilies overlapping any of the primitive/mature/LSC+/LSC- ATAC peaks (Benjamini $q < 0.05$). GIGGLE enrichment scores were calculated as described in [29] with the following parameters: Species - Human hg38, Data type in Cistrome - Transcription factor, chromatin regulator, Peak number of Cistrome sample to use - All peaks in each sample.

Essentiality scores

Essentiality scores were obtained from the Broad Institute project Achilles (DepMap) [86]. CRISPR (Avana 21Q1) and combined RNAi genetic dependencies data were downloaded from the "Download" page of the website https://depmap.org/portal/achilles/. For CRISPR data, we classified cell lines into cancer types following the sample information available on the DepMap website. In particular, the following 23 cell lines were classified as AML: HEL, HEL 92.1.7, MV411, OCI-AML2, THP1, NOMO-1, EOL-1, KASUMI-1, N-4, OCIAML3, MOLM13, TF1, U937, F36P, KO52, AML-193, M-07e, P31/FUJ, MONO-MAC-1, MOLM14, MUTZ-8, OCI-M2 and SHI-1. For RNAi data, our analysis was restricted to the following AML cell lines analyzed by the project Achilles: HL60, MOLM13, MOLM16, MV411, NB4, OCIAML2, OCIAML3, OCIAML5, THP1 and U937.

LSC17 scores

The LSC17 scores for all AML bulk samples were obtained using the clinical LSC17 assay. In general, the LSC17 score can be calculated as previously reported [15]. In short, log2 normalized values of each of the 17 genes were multiplied by its coefficient and then summed. For genes with more than one probe, probes with the highest overall signal were selected.

Survival analysis

The disease-free interval was defined as the time from remission until the date of the first new tumor progression event (locoregional recurrence, distant metastasis, development of a new primary tumor, or death). Survival analysis was performed using the Kaplan–Meier estimate method. P values comparing Kaplan–Meier survival curves were calculated using the log-rank (Mantel–Cox) test.

OCI-AML22 culturing

Cells were expanded in culture in the following medium, with final concentrations as indicated: X-VIVO 10 (Lonza, BE04-380Q) supplemented with 20% BIT 9500 Serum Substitute (StemCell technologies, 09500), 1x Glutamax Supplement (Thermo Fisher Scientific, 35050061), Primocin 0.1 mg/ml (invivogen), SCF (200 ng/ml; Miltenyi Biotech, 130-096-696), IL3 (20 ng/ml; Miltenyi Biotech, 130-095-069), TPO (20 ng/ml; Peprotech, 300-18), FLT3L (40 ng/ml; Peprotech, 300-19 B), IL6 (10 ng/ml; Miltenyi Biotech, 130-093-934), G-CSF (10 ng/ml; Miltenyi Biotech, 130-093-861). Cells were maintained at a density of $0.8x10^6$ cells/ml and passaged every 3-5 days in a 96-well flat bottom plate. The CD34+CD38- fraction or CD34+ fraction was regularly sorted to serially expand the cells.

Lentiviral production

VSV-G pseudotyped lentiviral vector particles were produced by polyethyleneimine (PEI)- based co-transfection[87] of 10.5 µg of pVSVg, 20.5 µg of pCMVR8.74 (both from Addgene) and 38 µg of transfer vector (pHR-SFFV-KRAB-dCas9-P2A-mCherry, from addgene, Plasmid #60954) per HEK293FT cells T175 flask. Viral particles harvested after 44 and 70 hrs were pooled and filtered (0.22-µm filter), concentrated 100x□using Lenti-X concentrator protocol (Takaka, ref 631232) optimized with Lenti-X concentrator beads at a ratio 1/8 (beads/virus supernatant). Concentrated virus were resuspended in X-VIVO 10 (Lonza) and stored at –80□°C until use.

Lentivirus titration

Virus titration was assessed on MOLM13 cells. For titration, MOLM13 cells were plated at 25,000 cells/well of 96 well round bottom plate, in 100 µl/well. Viral supernatant was added in 3 serial dilutions in duplicate. After mixing by pipetting cultures were incubated O/N. On the next day, 60-80 µl of supernatant was carefully removed, and 150 µl fresh medium was added. After 3 additional days transduction efficiency (%BFP+) was determined with a flow cytometer (Symphony, BD). Titers were calculated by using the mean of all 6 transductions of duplicate serial dilutions considering poisson-distribution: TU/mL= ln(%untransduced cells/100)* 25,000 cells/(vector input eg. 0.5 µl/1000).

Generation of dCas9-KRAB (CRISPRi) OCI-AML22 and transduction with LTR12C or control vectors

After expansion, the OCI-AML22 CD34+CD38- fraction was sorted (Moflo, Astrio, BD), washed, and resuspended in its own media. Cells were mixed with concentrated virus produced as previously described, at a 1/10 ratio supplemented with LentiBOOST research grade, Mayflower Biosciences, ref SB-P-LV-101-03 (1/100). Cells were then spun at 1200 rpm for 10 min before leaving them overnight in the incubator. The following day, the medium was replaced with OCI-AML22 culturing medium. To obtain enough cells, OCI-AML22 transduced cells were expanded over 1.5 months before being sorted for the mcherry+CD34+ population. After 2 additional months of cell expansion, mcherry+CD34+ were sorted again (Moflo, BD) to obtain 82 million cells. 1.79 million OCI-AML22 CD34+mcherry+ cells were collected and used for the next steps and are referred to as the dcas9-KRAB (CRISPRi) OCI-AML22. The mcherry+CD34+ sorted cells were transduced the day following sort, with 3 independently produced virus (4 wells for each condition) in a 96-well plate. Each well received 160 000 OCI-AML22 mcherry+CD34+ cells, the virus of interest concentrated as described in the virus section (LTR12C or control) at a dilution 1/10 and LentiBOOST research grade, Mayflower Biosciences, ref SB-P-LV-101-03 (1/100). Cells were then spun at 1200 rpm for 10 min before leaving them overnight in the incubator. The following day, the medium containing virus was washed and replaced with OCI-AML22 culturing medium. 5 days after transduction with LTR12C or control vector guides lentivirus, cell surface marker, mcherry, viability, and GFP expression were assessed on the Symphony, BD using CD34-APC (BD, clone 8G12, ref 340441, 1/100), CD38-PeCy7 (BD, clone Hb7, ref:335790 1/100). 7 days post-transduction, the mcherry+GFP+ population was sorted (MoFlow, BD) for each of the replicate conditions. Enough cells were collected to perform on each replicate ATAC-Seq.

Design of guide RNAs targeting TEs

Guide RNAs targeting LTR12C elements were designed using Repguide (https://tanaylab.github.io/repguide/; [88], using default parameters. We selected the combination of 5 guide RNAs targeting the highest number of LTR12C elements with limited targeting of other LTR12 families (LTR12, LTR12_, LTR12B, LTR12D, LTR12E and LTR12F). Five negative control guide RNAs [71] were used as negative control guide RNAs combination. To perform our experiments guide RNAs were cloned in a lentiviral vector (VectorBuilder.com), carrying also the gene coding for the GFP. Sequences of guide RNAs used in this study can be found in Extended Data Table 15.

Protein extraction and western blot analysis

OCI-AML22 cells stably expressing dCas9-KRAB (CRISPRi) were lysed in modified RIPA (10 mmol/L Tris-HCl, pH 8.0; 1 mmol/L EDTA; 140 mmol/L NaCl; 1% Triton X-100; 0.1% SDS; 0.1% sodium deoxycholate) containing protease inhibitor (cOmplete, EDTA-free Protease Inhibitor combination, Roche). Lysates were sonicated for 5 minutes (30s ON, 30s OFF) using a Diagenode Bioruptor 300. Cell debris was removed by centrifugation (10min, 15,000 rpm at 4^{o}C) followed by protein quantification using the BCA Protein Assay Kit (Thermo Fisher Scientific). Western blot was performed by using the Protein Simple Automated system (Jess, Bio-Techne) following manufacturer instructions. We used an anti-Cas9 antibody (Diagenode, C15200203 - 1:5000) and anti-GAPDH (Cell Signaling, GAPDH (14C10) Rabbit mAb #2118 - 1:300).

Research reproducibility and code availability

Code for data processing, analysis, and plotting can be found on CodeOcean (https://codeocean.com/capsule/4918318/tree). The analytical pipeline is also made reusable on the CoBE platform (www.pmcobe.ca).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors Contributions

G.G., B.N.., A.Q. and M.L. designed the studies. G.G. and B.N. performed the computational analysis with help from A.Q. A.Mi. performed the functional xenograft transplantation while C.A. and N.T. prepared ATAC-seq libraries. A.Mu. processed part of ATAC-seq data and S.A.M.T gave computational input. Ö.D. and H.B. performed experimental validations. A.A. provided clinical annotation for AML samples. M.D.M provided AML samples. G.G, B.N., A.Q., Ö.D., H.B., J.C.Y.W., M.D.M., J.E.D., M.L. interpreted the data; M.L. and J.E.D. supervised the study. G.G., B.N. and M.L. wrote the manuscript.

Acknowledgements

We acknowledge the Princess Margaret Genomics and Bioinformatics group for providing the infrastructure required to conduct analyses included in this work. This work was funded by grants to M.L. and J.D. from the Princess Margaret Cancer Foundation, Ontario Institute for Cancer Research with funding from the Province of Ontario, Canadian Institutes for Health Research (CIHR), CEEHRC team grant from the CIHR, Medicine by Design, the Canadian Cancer Society Research Institute; to J.D. from the Terry Fox Research Institute Program Project and a Canada Research Chair; to N.T. from JSPS KAKENHI Grant Number JP17K09899. M.L. holds an Investigator Award from the Ontario Institute for Cancer Research and a Bernard and Francine Dorval Award from the Canadian Cancer Society.

Supplementary information

Extended Data Table 1. Differential accessible TEs in HSPC vs mature hematopoietic populations
Extended Data Table 2. Differential accessible TEs in LT-HSCs vs ST-HSCs hematopoietic populations
Extended Data Table 3. Differential accessible TEs in stem vs mature hematopoietic populations
Extended Data Table 4. Differential accessible TEs in progenitor vs mature hematopoietic populations
Extended Data Table 5. Differential accessible TEs in stem vs progenitor hematopoietic populations
Extended Data Table 6. Differential accessible TE Zscores in hematopoietic signatures
Extended Data Table 7. Transcription regulator cistromes enriched at stem TEs
Extended Data Table 8. Differential accessible TEs in LSC+ vs LSC- populations
Extended Data Table 9. Clinical information AML patients
Extended Data Table 10. Top 5% transcription regulator cistromes enriched at HSPC populations TEs
Extended Data Table 11. Top 5% transcription regulator cistromes enriched at LSC+ TEs
Extended Data Table 12. Top 5% transcription regulator cistromes enriched at HSPC and LSC+ TEs
Extended Data Table 13. CRISPR-based essentiality scores of LSC+ specific transcription regulators
Extended Data Table 14. RNAi-based essentiality scores of LSC+ specific transcription regulators
Extended Data Table 15. List of guide RNA sequences used in this study
Extended Data Table 16. Antibodies used for FACS sorting of CD34/CD38 fractions

Figure legends

Fig. 1: Distinct TE subfamilies populate the accessible chromatin of primitive and mature hematopoietic cells

A) Heatmap displaying differential accessibility Z-scores of all TE subfamilies across bulk ATAC-seq profiles of HSC, progenitor and mature hematopoietic populations; note that HSC and progenitors show similar TE subfamily enrichment profiles.

B) Association of TE families with 'stemness'. Direct comparison of HSC and progenitor (“primitive”) versus mature cells. Y-axis shows all TE families with enriched TE subfamilies (each dot denotes a different TE subfamily). LINE1, SINE1, SINE2, and Transposons subfamilies are enriched in mature, whereas ERV1 and ERV3 families are enriched in HSPC cells (left). Number of enriched TE subfamilies in HSPC or mature cells, divided by TE family (right). TE families are ordered from most highly enriched in HSPC to mature cells. n.s., not significant.

C) Examples of HSPC and mature enriched TE subfamilies of the ERV1, ERV3, SINE1 and SINE2 TE families. Boxplots show differential accessibility Z-scores in hematopoietic stem, progenitor and mature populations.

D) Genome browser view of ATAC-seq signal from hematopoietic populations for one element of each primitive enriched TE subfamily shown in C).

E) Transcription factor cistromes enrichment at 2 of the 23 TE subfamilies enriched in stem populations (HSCs). Every dot corresponds to a transcription factor cistrome. Red dashed lines correspond to -log10(q)=1.3 (q=0.05) and logOddsRatio(logOR)=1 thresholds.

F) Individual HOXA9 (left) or three-dimensional genome organization factor (right) cistromes enriched over HSC-specific TE subfamilies (according to Fig. 1E). Box plots showing enrichment GIGGLE scores of individual cistromes profiled in cell lines derived from a variety of tissue states.

G) Enrichment of HOXA9 and CTCF DNA recognition sequences within TE subfamilies enriched in HSCs. Bars represent -log10(q) for each TE subfamily for transcription factor DNA recognition sequences. The red dashed line corresponds to -log10(q)=1.3 (q=0.05) threshold.

Abbreviations: Long-term/Short-Term hematopoietic stem cells (LT-/ST-HSC), common myeloid progenitors (CMP), megakaryocyte-erythroid progenitors (MEP), myeloid-lymphoid progenitors (MLP), granulocyte-monocyte progenitor (GMP).

Fig. 2: Leukemia stem cells containing fractions cluster with HSPCs and show distinct and common subfamily enrichment.

A) Heatmap of accessibility Z-scores of all TE subfamilies (rows) across bulk ATAC-seq profiles of HSPC, mature hematopoietic populations, LSC+, and LSC- fractions clustered based on correlation. LSC+ fractions clustered with HSPC while LSC- fractions mostly clustered with mature populations.

B) Heatmap of TE subfamilies with differential accessibility Z-scores (rows, $q < 0.01$) between ATAC-seq profiles of LSC+ (n=11) and LSC- fractions (n=23).

C) Enrichment of TE families across LSC+ versus LSC- samples. The y-axis shows all TE families; points denote TE subfamilies. SINE1, SINE2, and Transposons families are enriched in LSC- samples, whereas ERV1 and ERV3 families are enriched in LSC+ samples. Number of enriched TE subfamilies in LSC+ or LSC- samples, divided by TE family (right). TE families are ordered from most highly enriched in LSC+ to LSC-. n.s., not significant.

D) Comparison of the proportion of TE families enriched in HSPC and hematopoietic mature populations in LSC- and LSC+ fractions and the overall distribution of TE families within the hg38 build of the human genome.

E) UpSet plot showing the intersection of TE subfamilies enriched in LSC+/HSPC (21 families) and LSC-/mature hematopoietic populations (53 families). Colors within bar graphs denote TE families as annotated in panel D.

Fig. 3: The LSCTE121 signature stratifies AML patients with distinct disease-free intervals

A) Heatmap showing differential accessibility Z-scores for LSCTE121 (q < 0.01 LSC+ vs LSC-) in bulk cohort1 AML tumors. Note that bulk tumors did not cluster according to their LSC17 scores with LSC+ or LSC- fractions.

B) Box plot displaying the LSCTE121 Z-score in cohort1 patients. Patients were divided into LSCTE121 high or low based on the heatmaps in A and B. p-values results of the Wilcoxon test are showcased on the box plots.

C) Kaplan–Meier estimates using Disease Free Interval as a clinical endpoint for LSCTE121 Z-score top vs bottom 25 percentile in cohort1 (n=8/group).

D) Pearson correlation between LSC17 score and LSCTE121 Z-score in cohort1 bulk AML patients.

E) Boxplot of LSCTE121 Z-score in LSC fractions. Note that engrafting fractions show higher LSCTE121 Z-score than non-engrafting fractions.

F) Same as panel A but for cohort2 patients.

G) Same as panel C but for cohort2 patients.

H) Same as panel D but for cohort2 patients (LSCTE121high: n=15; LSCTE121low: n=45).

Fig. 4: TE subfamilies enriched in accessible chromatin in LSC provide binding sites to transcription factors essential for AML

A) Frequency plot of transcription factor cistromes enriched at TE subfamilies in accessible chromatin in LSC+ fractions. The top 5% most frequently enriched transcription factor cistromes are shown.

B) UpSet plot showing the intersection of transcription factor cistromes enriched in LSC+/HSPC populations.

C) GIGGLE score for individual LYL1, NFYA, and NFYB cistromes profiled in cell lines derived from various tissue states when compared to TE subfamilies specifically accessible in LSC+.

D) Enrichment of NFY DNA recognition sequences within TE subfamilies enriched in the accessible chromatin of LSC+. The red dashed line corresponds to -log10(q)=1.3 (q=0.05) threshold.

E) Overview of essentiality scores of LSC+ specific transcription factors across all cancer types available in DepMap with more than five cell lines, based on CRISPR data. The distribution of the essentiality scores of each transcription factor calculated in AML cell lines was compared to the distribution of the same transcriptional factor calculated in cell lines from every other cancer type. Rectangles inner color corresponds to the median essentiality score, while border color corresponds to the pairwise t-test, Benjamini-Hochberg corrected p-values (adjusted p-value).

Fig. 5: Accessibility at LTR12C elements is essential for LSC stemness properties

- *A) Heatmap showing differential accessibility Z-scores for LSC TEs ($q < 0.01$ LSC+ vs LSC-) between ATAC-seq profiles of LSC+, LSC-, OCI-AML22 fractions, OCI-AML3 and MOLM13 cell lines, clustered based on correlation.*
- *B) Box plot displaying the LTR12C accessibility Z-scores in LSC+ and LSC-.*
- *C) ATAC-seq profile across all LTR12C elements accessible in LSC+ and/or LSC- fractions.*
- *D) Same as B but for OCI-AML22 fractions.*
- *E) Schematic overview of the CRISPRi chromatin editing approach.*
- *F) Violin plot for the chromatin accessibility signal distribution in OCI-AML22-LSC+ stably expressing both CRISPRi and scramble (control in gray) or LTR12C (purple) guide RNA combos over LTR12C or LTR12D elements. A representative biological replicate of three is shown. Every dot represents an individual TE element. p-values results of paired samples Wilcoxon test are showcased on the violin plot.*
- *G) Genome browser views of ATAC-seq signal reduction at individual LTR12C elements in OCI-AML22-LSC+ expressing both CRISPRi and LTR12C guide RNA combo (purple) compared to scramble gRNA combo (control in gray).*
- *H) Percentage of CD34+/CD38-, CD34+/CD38+, CD34-/CD38+ and CD34-/CD38- cells upon CRISPRi-mediated chromatin editing at LTR12C elements in OCI-AML22-LSC+. The results are shown as means ± SD from three independent transductions with the vector promoting the expression of either 5 scramble guide RNAs or 5 guide RNAs targeting LTR12C elements. p-value was generated by a two-sided t-test.*

Extended Data Fig. 1:

- *A) Schematic representation of hematopoiesis.*
- *B) Heatmap of differential accessibility Z-scores of TE subfamilies with $q < 0.01$ across bulk ATAC-seq profiles comparing megakaryocyte-erythroid lineage and all other hematopoietic samples.*
- *C) Heatmap displaying differential accessibility Z-scores of all TE subfamilies across bulk ATAC-seq profiles of HSCs, progenitors and mature hematopoietic populations excluding megakaryocyte-erythroid lineage TEs.*
- *D) Volcano plot showing the median difference in accessibility Z-scores for each TE subfamily between the stem/progenitor and differentiated hematopoietic cell populations versus the −log10 q-value for that difference.*
- *E) Examples of non-stem and stem-enriched families of the ERV1, ERV3, SINE1 and SINE2 TE families. Boxplots are showing differential accessibility Z-scores in the different hematopoietic populations.*
- *F) Number of called peaks in the different hematopoietic cell populations, note that stem and progenitor cells have a higher number of called peaks than mature hematopoietic cells.*

Extended Data Fig. 2:

A) Heatmap of differential accessibility Z-scores of TE subfamilies with $q < 0.01$ across bulk ATAC-seq profiles comparing stem and all mature hematopoietic samples.

B) UpSet plot showing the intersection of TE subfamilies enriched in stem or progenitor hematopoietic populations, note that 23 TE subfamilies are enriched exclusively in stem populations

C) 23 TE subfamilies enriched exclusively in stem samples. Bar plots show differential accessibility Z-scores in the stem, progenitor and mature cell state groups.

D) Heatmap of differential accessibility Z-scores of TE subfamilies with q < 0.05 across bulk ATAC-seq profiles comparing stem and progenitor hematopoietic samples.

E), F) and G) Transcription factor cistromes enriched at 3 of the 23 TE subfamilies enriched in stem populations (HSCs). Every dot corresponds to a transcription factor cistrome taken into account in the analysis. Red dashed lines correspond to -log10(q)=1.3 (q=0.05) and logOddsRatio(logOR)=1 thresholds.

H), I) and J) Individual HOXA9 (H), RUNX1 (I) or three-dimensional genome organization factors (J) cistromes enriched over HSCs specific TE subfamilies and with enrichment of the HOXA9, RUNX1 or three-dimensional genome organization factors cistrome (according to E, F and G). Box plots showing enrichment GIGGLE scores of individual cistromes profiled in cell lines derived from a variety of tissue states.

Extended Data Fig. 3:

A) Volcano plot showing the median difference in bias-corrected accessibility deviation Z-scores for each TE subfamily between the LSC+ and LSC- fractions versus the −log10(q-value) for that difference.

B) Bar plots showing the number of total or accessible TE elements belonging to the 44 TE subfamilies enriched in LSC+ fractions (left) or in the 77 TE subfamilies enriched in LSC- fractions (right).

C) Genome browser views of ATAC-seq signal reduction at individual TE elements in HSPC, LSC+, Mature and LSC- samples.

D) TE subfamilies are differentially enriched only in LSC+, LSC-, Stem/Progenitor and Differentiated populations of the hematopoietic system, as shown in Fig. 2e. The Cleveland dot plot shows the median z-scores for the relevant comparisons.

E) Shared TE subfamilies between LSC+, LSC-, Stem/Progenitor and Differentiated populations of the hematopoietic system as shown in Fig. 2E, Cleveland dot plot shows the median Z-scores for the relevant comparisons.

Extended Data Fig. 4:

A) and B) Heatmap showing differential accessibility Z-scores for LSC TEs (q < 0.01 LSC+ vs LSC-) between ATAC-seq profiles of LSC+, LSC- fractions and bulk AML tumors (cohort1 in A and cohort2 in B*), clustered based on correlation. Note that 20 and 40 samples cluster with LSC+ fractions in cohort1 and cohort2 bulk AML patients respectively.*

C) Boxplot of LSCTE121 Z-score in LSC fractions, note that engrafting fractions show higher LSCTE121 Z-score than non-engrafting fractions. p values results of the Wilcoxon test are showcased on the box plots.

D) Pearson correlation between LSCTE121 Z-score and blast count in peripheral blood for cohort1 bulk AML patients.

E) Pearson correlation between LSCTE121 Z-score and blast count in peripheral blood for cohort2 bulk AML patients.

F) Kaplan–Meier estimates using Overall Survival as a clinical endpoint for LSCTE121 Z-score top vs bottom 25 percentile in cohort1 (n=8/group).

G) Kaplan-Meier estimates using Overall Survival as a clinical endpoint, patient division based on LSC17 high vs low scores in cohort1 bulk AML patients; LSC17 high: n = 12, LSC17 low: n = 17.

H) Same as I but for cohort2 patients (LSCTE121high: n=15; LSCTE121low: n=45).

Extended Data Fig. 5:

A) *Frequency plot of transcription factor cistromes enriched at TE subfamilies enriched in primitive populations. The top 5% most frequently enriched transcription factor cistromes are shown.*
B) *Frequency plot of transcription factor cistromes enriched at TE subfamilies enriched in primitive and LSC+ fractions. The top 5% most frequently enriched transcription factor cistromes are shown.*
C) *UpSet plot showing the intersection of TE subfamilies showing enrichment for ERG and/or RUNX1 cistromes in primitive and/or LSC+. Note that most of the TE subfamilies are shared between ERG and RUNX1.*
D) *Example of TE subfamilies enriched in primitive and/or LSC+ with enrichment of both ERG and RUNX1 cistromes and TE subfamily enriched in primitive populations with enrichment of ZNF143, CTCF and other looping factor cistromes. Dots correspond to each transcription factor cistrome taken into account in the analysis. Red dashed lines correspond to -log10(q)=1.3 (q=0.05) and logOddsRatio(logOR)=1 thresholds.*
E) *Enrichment of ERG DNA recognition sequences within TE subfamilies enriched in LSC+ and/or primitive populations. Bars represent -log10(q). The red dashed line corresponds to -log10(q)=1.3 (q=0.05) threshold.*
F) *Same as E but for the RUNX1 DNA recognition sequences.*
G) *Same as E but for the GATA1 and CTCF DNA recognition sequences within TE subfamilies enriched in primitive populations.*

Extended Data Fig. 6:

A) *Violin plots for essentiality scores of LSC+ specific transcription factors, based on CRISPR data. For each transcription factor the distribution of the essentiality scores calculated in 23 AML cell lines was compared to the distribution of all the other genes tested. p-values based on two-sided t-tests are reported. The red line corresponds to the essentiality score = 0 threshold.*
B) *Violin plots for essentiality scores of LSC+ specific transcription factors, based on RNAi data. For each transcription factor the distribution of the essentiality scores calculated in 10 AML cell lines was compared to the distribution of all the other genes tested. p-values based on two-sided t-tests are reported. The red line corresponds to the essentiality score = 0 threshold.*
C) *Overview of essentiality scores of LSC+ specific transcription factors across all cancer types available in DepMap with more than five cell lines, based on RNAi data. The distribution of the essentiality scores of each transcription factor calculated in AML cell lines was compared to the distribution of the same transcriptional factor calculated in cell lines from each other cancer type. Rectangles inner color corresponds to the median essentiality score, while border color corresponds to Pairwise t-test Benjamini-Hochberg corrected p-values (adjusted p-value).*

Extended Data Fig. 7:

A) *Box plot displaying the LSCTE121 Z-score in LSCs, OCI-AML22 fractions, OCI-AML3 and MOLM13 cell lines. p-values results of the Wilcoxon test are showcased on the box plots.*
B) *Box plot displaying the LTR12C accessibility Z-scores in OCI-AML3 and MOLM13 cell lines.*
C) *dCas9-KRAB expression levels in OCI-AML22 cells. Protein Simple (western blot) was performed on whole extracts, showcasing dCas9-KRAB expression in positive control [29]. The star indicates the dCas9-KRAB band in CRISPRi OCI-AML22.*

D) Quantification of the chemiluminescence signal presented in C. Signal peaks corresponding to GAPDH (loading control) and dCas9-KRAB (CRISPRi) are indicated in the Fig..

E) Violin plots showcasing the chromatin accessibility signal distribution in replicate #1 over all LTR12 TE subfamilies, except LTR12C and LTR12D (in Fig. 5E). Every dot represents an individual TE element. p-values results of paired samples Wilcoxon test are showcased on the violin plot.

F) Violin plots showcasing the chromatin accessibility signal distribution in replicate #2 OCI-AML22-LSC+ stably expressing both CRISPRi and scramble (control in gray) or LTR12C (purple) guide RNA combos, over LTR12C or LTR12D (left) and all other LTR12 TE subfamily elements (right). Every dot represents an individual TE element. Note that the loss of chromatin accessibility is specific to LTR12C elements only. p-values results of paired samples Wilcoxon test are showcased on the violin plot.

G) Same as D but for replicate #3 OCI-AML22-LSC+ stably expressing both CRISPRi and scramble or LTR12C guide RNA combos.

References:

1. Warner, J. K., Wang, J. C. Y., Hope, K. J., Jin, L. & Dick, J. E. Concepts of human leukemic development. *Oncogene* **23**, 7164–7177 (2004).
2. Dick, J. E. Acute Myeloid Leukemia Stem Cells. *Annals of the New York Academy of Sciences* vol. 1044 1–5 Preprint at https://doi.org/10.1196/annals.1349.001 (2005).
3. Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).
4. Shlush, L. I. *et al.* Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547**, 104–108 (2017).
5. Eppert, K. *et al.* Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.* **17**, 1086–1093 (2011).
6. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3**, 730–737 (1997).
7. Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333**, 218–221 (2011).
8. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–970 (2008).
9. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
10. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
11. Deblois, G., Tonekaboni, S. A. M., Grillo, G. & Martinez, C. Epigenetic switch–induced viral mimicry evasion in chemotherapy-resistant breast cancer. *Cancer Discov.* (2020).
12. Grillo, G. & Lupien, M. Cancer-associated chromatin variants uncover the oncogenic role of transposable elements. *Curr. Opin. Genet. Dev.* **74**, 101911 (2022).
13. Takayama, N. *et al.* The Transition from Quiescent to Activated States in Human Hematopoietic Stem Cells Is Governed by Dynamic 3D Genome Reorganization. *Cell Stem Cell* (2020)

doi:10.1016/j.stem.2020.11.001.

14. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
15. Ng, S. W. K. *et al.* A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* **540**, 433–437 (2016).
16. Zeng, A. G. X. *et al.* A cellular hierarchy framework for understanding heterogeneity and predicting drug response in acute myeloid leukemia. *Nat. Med.* **28**, 1212–1223 (2022).
17. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature Publishing Group* **506**, 328–333 (2014).
18. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
19. Fennell, K. A. *et al.* Non-genetic determinants of malignant clonal fitness at single-cell resolution. *Nature* **601**, 125–131 (2022).
20. Kreso, A. *et al.* Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* **339**, 543–548 (2013).
21. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).
22. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
23. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).
24. Britten, R. J. & Kohne, D. E. Repeated Sequences in DNA. *Science* **161**, 529–540 (1968).
25. Roulois, D. *et al.* DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* **162**, 961–973 (2015).
26. Chiappinelli, K. B. *et al.* Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* **164**, 1073 (2016).
27. Colombo, A. R. *et al.* Suppression of Transposable Elements in Leukemic Stem Cells. *Sci. Rep.*

7, 7029 (2017).

28. Barakat, T. S. *et al.* Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**, 276–288.e8 (2018).
29. Grillo, G. *et al.* Transposable elements are co-opted as oncogenic regulatory elements by lineage-specific transcription factors in prostate cancer. *Cancer Discov.* (2023) doi:10.1158/2159-8290.CD-23-0331.
30. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
31. Jang, H. S. *et al.* Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 611–617 (2019).
32. Kruse, K. *et al.* Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. *Cold Spring Harbor Laboratory* 523712 (2019) doi:10.1101/523712.
33. Deniz, Ö. *et al.* Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nat. Commun.* **11**, 3506 (2020).
34. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
35. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535–1537.e16 (2018).
36. Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
37. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
38. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).
39. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
40. Lawrence, H. J. *et al.* Loss of expression of the Hoxa-9 homeobox gene impairs the proliferation

and repopulating ability of hematopoietic stem cells. *Blood* **106**, 3988–3994 (2005).

41. Ichikawa, M. *et al.* A role for RUNX1 in hematopoiesis and myeloid leukemia. *Int. J. Hematol.* **97**, 726–734 (2013).

42. Cheng, H., Zhang, N. & Pati, D. Cohesin subunit RAD21: From biology to disease. *Gene* **758**, 144966 (2020).

43. Layer, R. M. *et al.* GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods* **15**, 123–126 (2018).

44. Boutzen, H. *et al.* A primary hierarchically organized patient-derived model enables in depth interrogation of stemness driven by the coding and non-coding genome. *Leukemia* **36**, 2690–2704 (2022).

45. Xiao-Jie, L., Hui-Ying, X., Qi, X., Jiang, X. & Shi-Jie, M. LINE-1 in cancer: multifaceted functions and potential clinical implications. *Genet. Med.* **18**, 431–439 (2016).

46. Merlos-Suárez, A. *et al.* The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8**, 511–524 (2011).

47. Li, Y. *et al.* Suppression of cancer relapse and metastasis by inhibiting cancer stemness. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 1839–1844 (2015).

48. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265–1281.e24 (2019).

49. Xie, S. Z. *et al.* Sphingosine-1-phosphate receptor 3 potentiates inflammatory programs in normal and leukemia stem cells to promote differentiation. *Blood Cancer Discov* **2**, 32–53 (2021).

50. Knudsen, K. J. *et al.* ERG promotes the maintenance of hematopoietic stem cells by restricting their differentiation. *Genes Dev.* **29**, 1915–1929 (2015).

51. Hosoya, T., Clifford, M., Losson, R., Tanabe, O. & Engel, J. D. TRIM28 is essential for erythroblast differentiation in the mouse. *Blood* **122**, 3798–3807 (2013).

52. Morishima, T. *et al.* LMO2 activation by deacetylation is indispensable for hematopoiesis and T-ALL leukemogenesis. *Blood* **134**, 1159–1175 (2019).

53. Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**,

237–240 (2010).

54. Mandoli, A. *et al.* The Hematopoietic Transcription Factors RUNX1 and ERG Prevent AML1-ETO Oncogene Overexpression and Onset of the Apoptosis Program in t(8;21) AMLs. *Cell Rep.* **17**, 2087–2100 (2016).

55. Selleri, L. *et al.* The TALE homeodomain protein Pbx2 is not essential for development and long-term survival. *Mol. Cell. Biol.* **24**, 5324–5331 (2004).

56. Goode, D. K. *et al.* Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Dev. Cell* **36**, 572–587 (2016).

57. Imanirad, P. *et al.* HIF1α is a regulator of hematopoietic progenitor and stem cell development in hypoxic sites of the mouse embryo. *Stem Cell Res.* **12**, 24–35 (2014).

58. Koide, S. *et al.* Setdb1 maintains hematopoietic stem and progenitor cells by restricting the ectopic activation of nonhematopoietic genes. *Blood* **128**, 638–649 (2016).

59. Herviou, L., Cavalli, G., Cartron, G., Klein, B. & Moreaux, J. EZH2 in normal hematopoiesis and hematological malignancies. *Oncotarget* **7**, 2284–2296 (2016).

60. Heideman, M. R. *et al.* Sin3a-associated Hdac1 and Hdac2 are essential for hematopoietic stem cell homeostasis and contribute differentially to hematopoiesis. *Haematologica* **99**, 1292–1303 (2014).

61. Ferreira, R., Ohneda, K., Yamamoto, M. & Philipsen, S. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol. Cell. Biol.* **25**, 1215–1227 (2005).

62. Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **2**, 6186 (2015).

63. Zhou, Q. *et al.* ZNF143 mediates CTCF-bound promoter-enhancer loops required for murine hematopoietic stem and progenitor cell function. *Nat. Commun.* **12**, 43 (2021).

64. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110–1121 (2016).

65. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).

66. Heidari, N. *et al.* Genome-wide map of regulatory interactions in the human genome. *Genome Res.* **24**, 1905–1917 (2014).

67. Hyle, J. *et al.* Auxin-inducible degron 2 system deciphers functions of CTCF domains in transcriptional regulation. *Genome Biol.* **24**, 14 (2023).

68. Meng, Y.-S., Khoury, H., Dick, J. E. & Minden, M. D. Oncogenic potential of the transcription factor LYL1 in acute myeloblastic leukemia. *Leukemia* **19**, 1941–1947 (2005).

69. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

70. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).

71. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).

72. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* **9**, e1003504 (2013).

73. Burns, K. H. Transposable elements in cancer. *Nature Publishing Group* **17**, 415–424 (2017).

74. Zapatka, M. *et al.* The landscape of viral associations in human cancers. *Nat. Genet.* **52**, 320–330 (2020).

75. Colombo, A. R., Triche, T., Jr & Ramsingh, G. Transposable Element Expression in Acute Myeloid Leukemia Transcriptome and Prognosis. *Sci. Rep.* **8**, 16449 (2018).

76. Hashimoto, K. *et al.* CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors. *Genome Res.* **25**, 1812–1824 (2015).

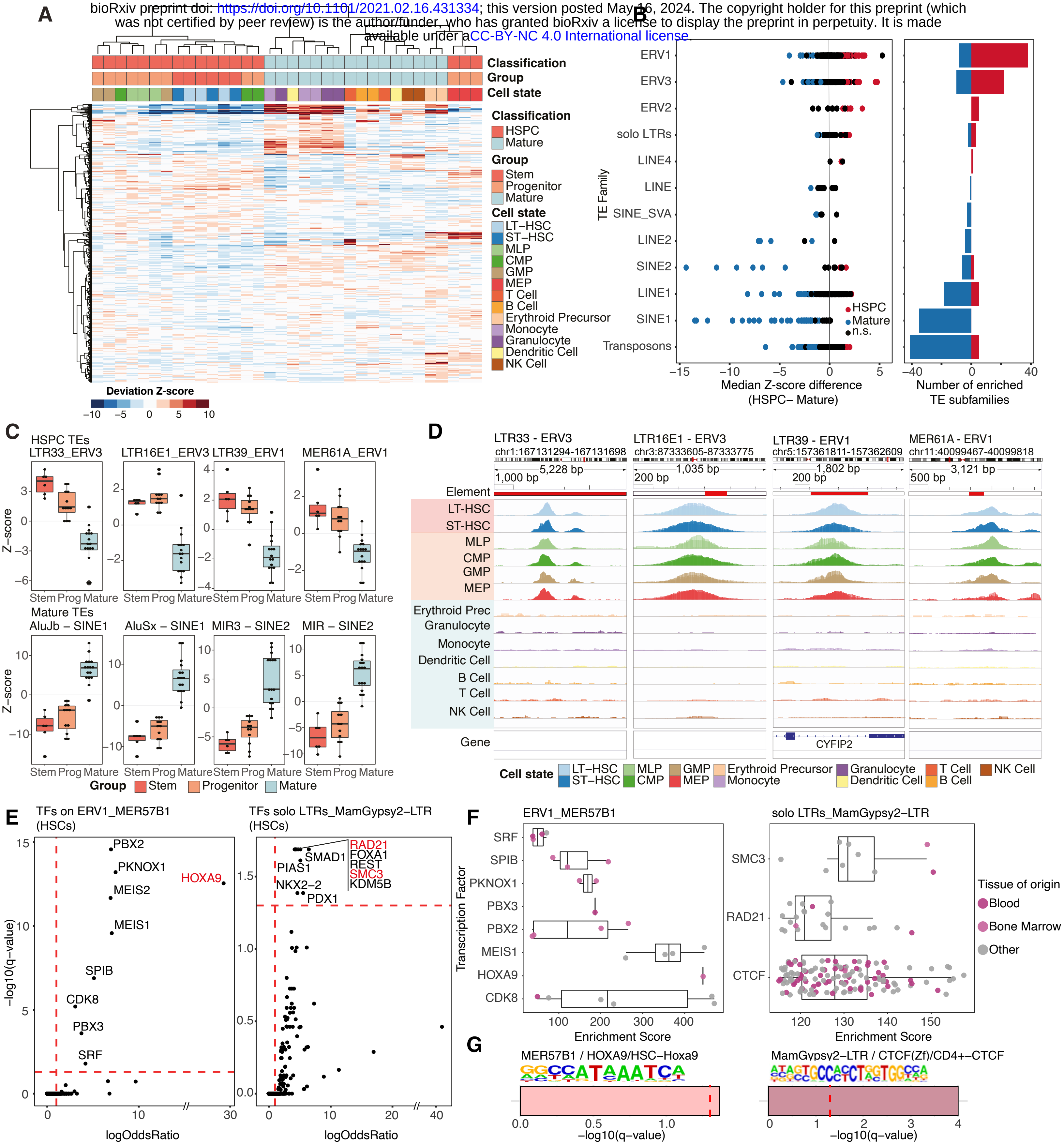
77. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).

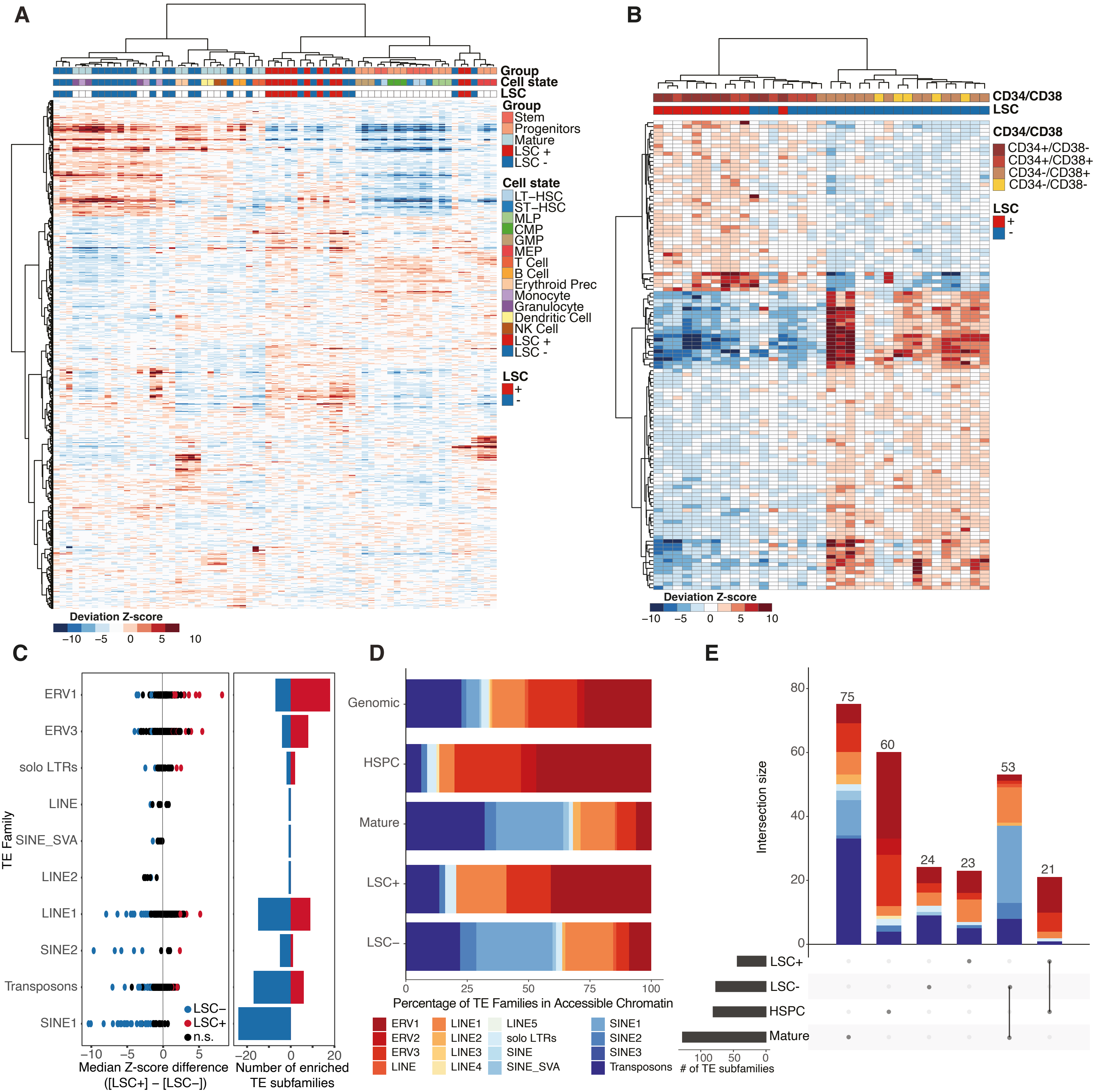
78. Laurenti, E. *et al.* CDK6 levels regulate quiescence exit in human hematopoietic stem cells. *Cell Stem Cell* **16**, 302–313 (2015).

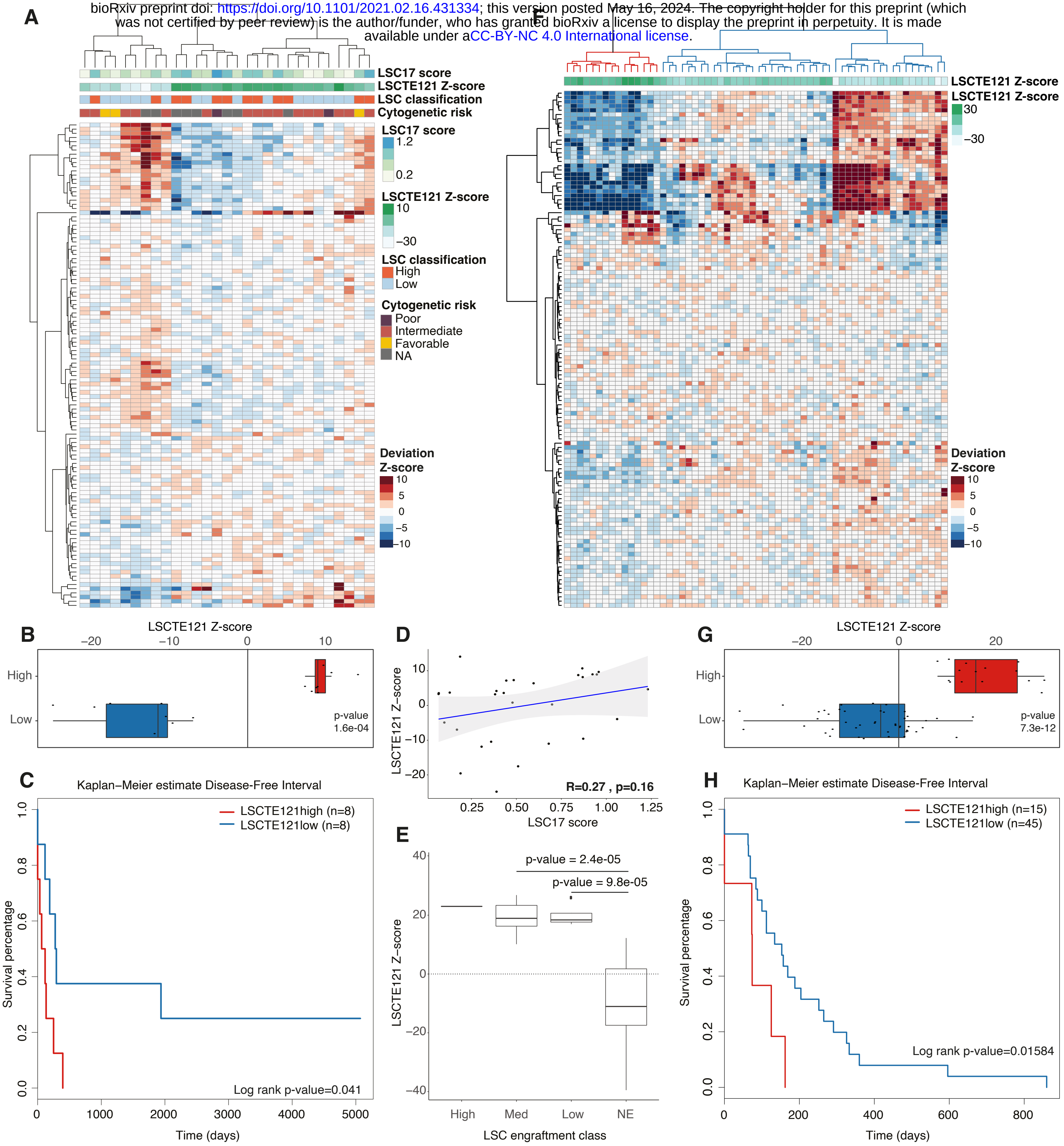
79. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of

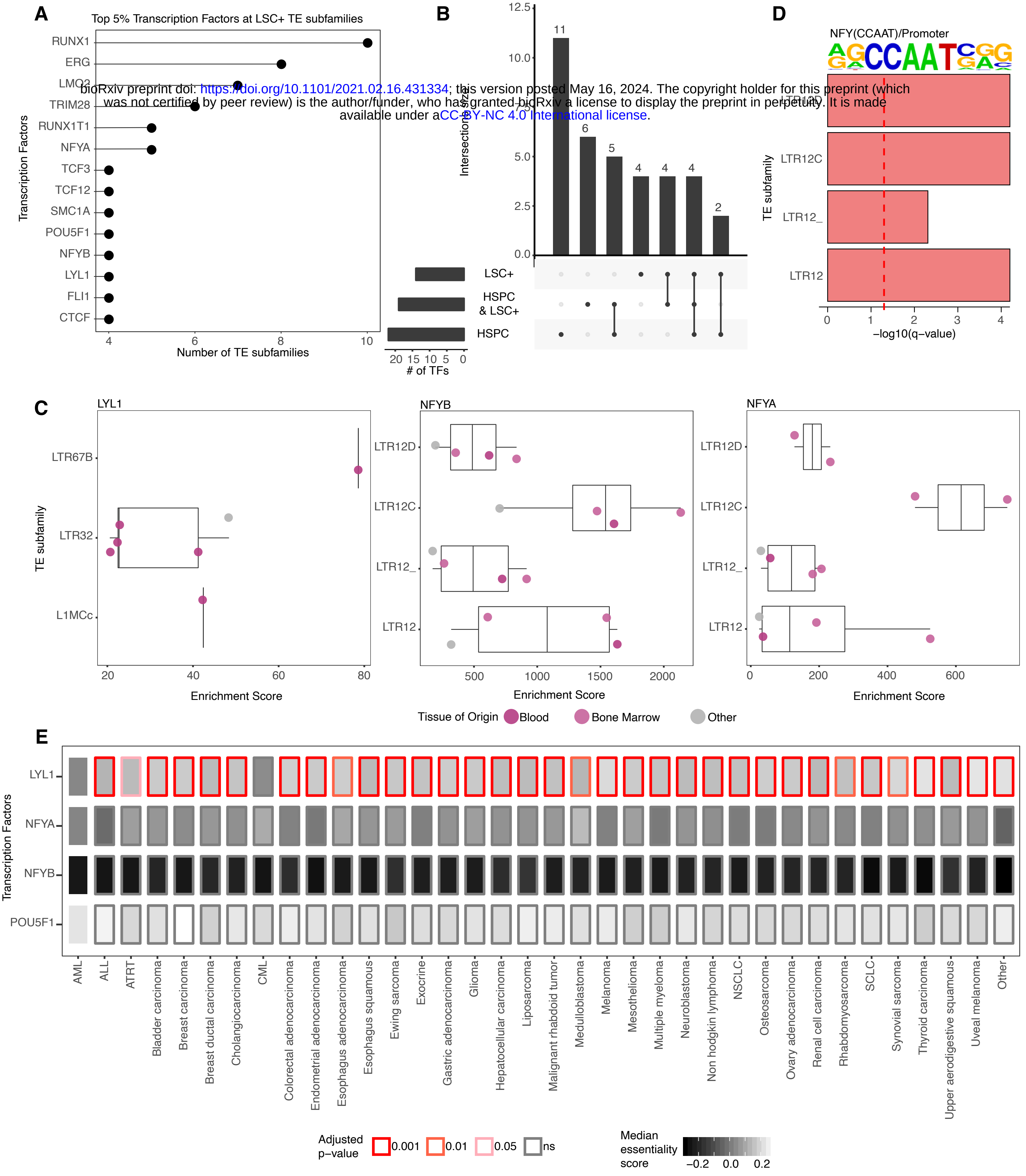
native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

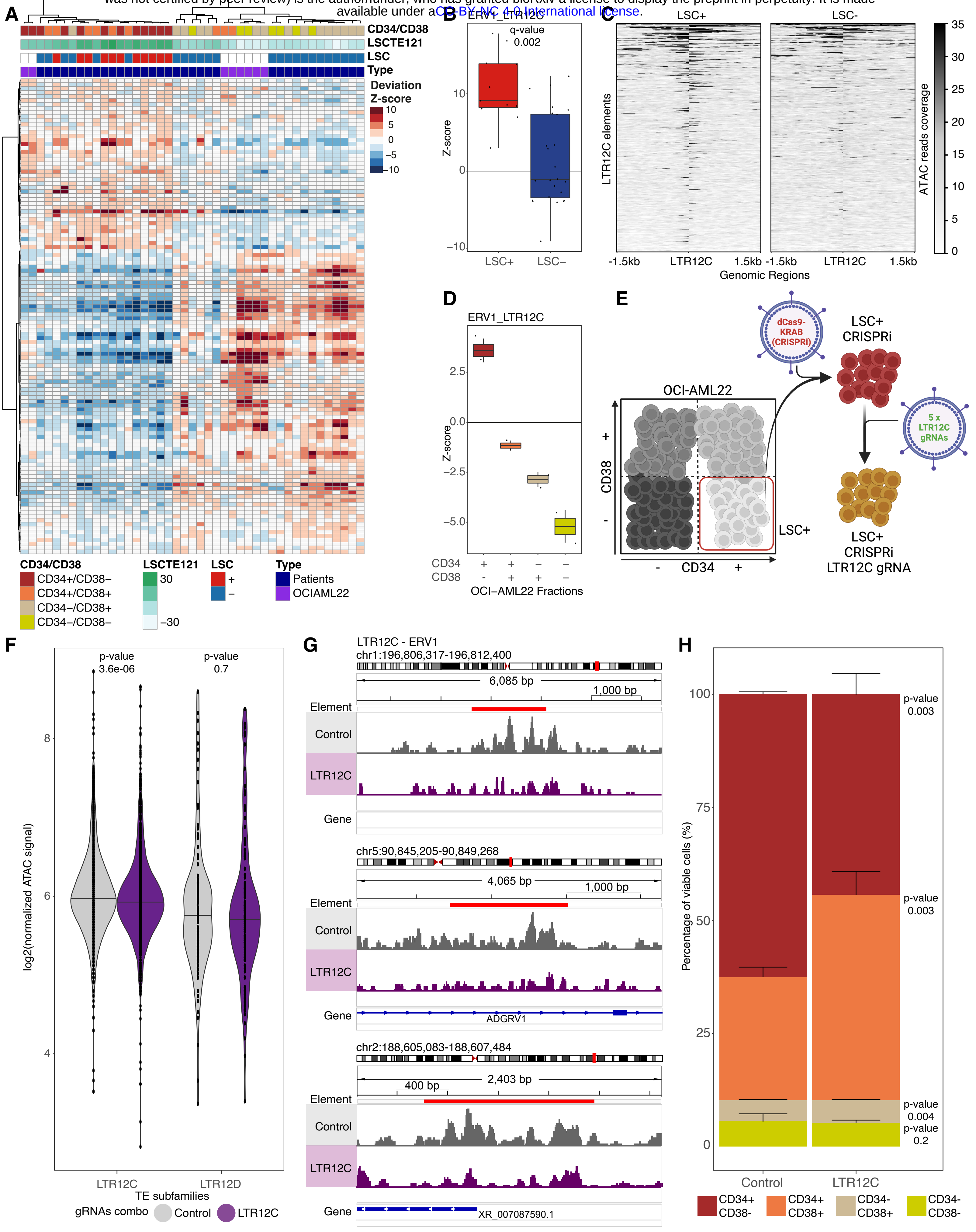
80. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
81. Jurka, J. Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.* **8**, 333–337 (1998).
82. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
83. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).
84. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).
85. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018).
86. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–570.e16 (2017).
87. Kneissl, S. *et al.* Measles virus glycoprotein-based lentiviral targeting vectors that avoid neutralizing antibodies. *PLoS One* **7**, e46667 (2012).
88. Fuentes, D. R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife* **7**, (2018).

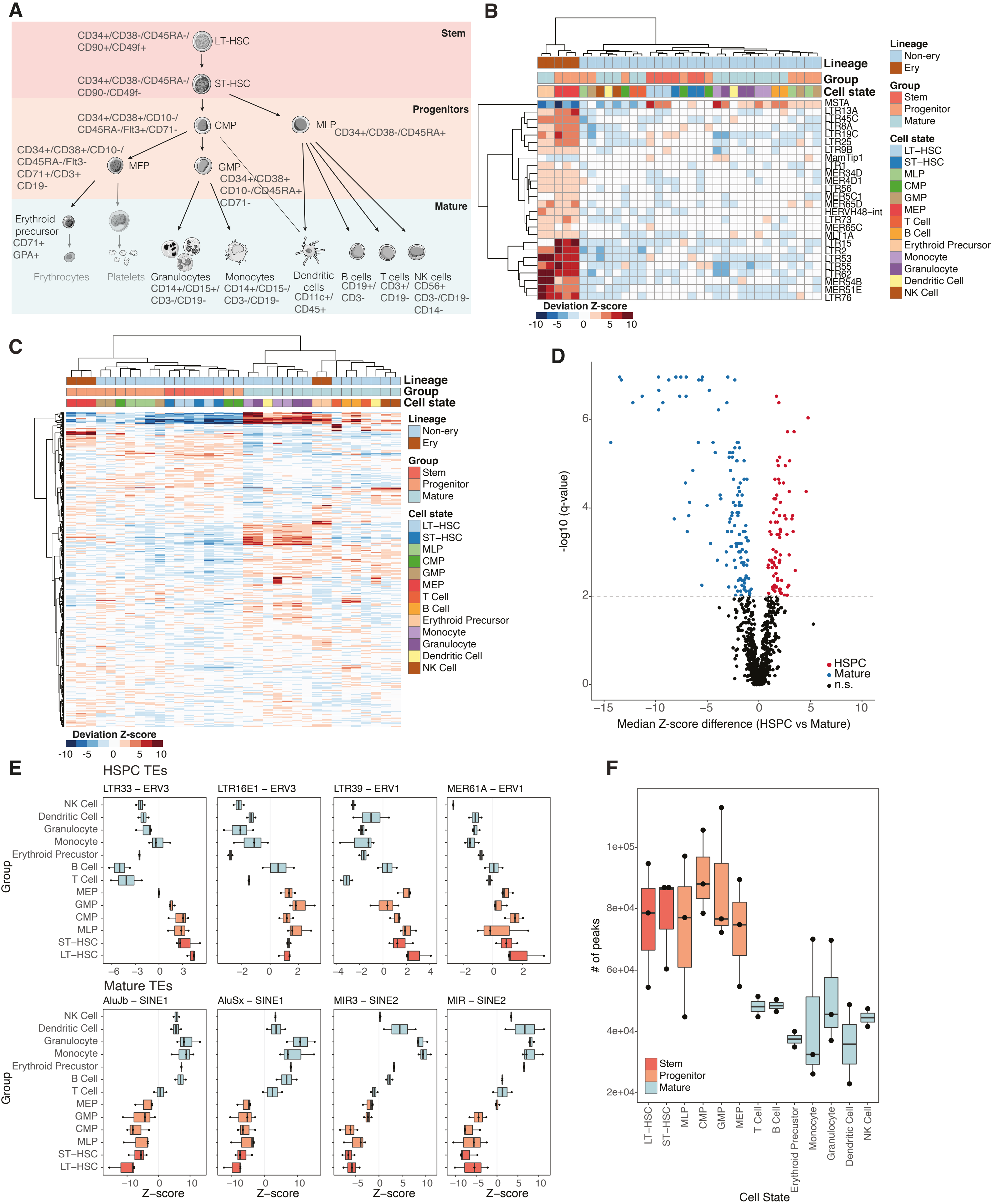

A
Classification
Group
Cell state
Classification
HSPC
Mature
Group
Stem
Progenitor
Mature
Cell state
LT−HSC
ST−HSC
MLP
CMP
GMP
MEP
T Cell
B Cell
Erythroid Precursor
Monocyte
Granulocyte
Dendritic Cell
NK Cell
Deviation Z-score
−10 −5 0 5 10
B
ERV1
ERV3
ERV2
solo LTRs
LINE4
LINE
SINE_SVA
LINE2
SINE2
LINE1
SINE1
Transposons
TE Family
HSPC
Mature
n.s.
−15 −10 −5 0 5
Median Z-score difference (HSPC− Mature)
−40 0 40
Number of enriched TE subfamilies
C
HSPC TEs
LTR33_ERV3
LTR16E1_ERV3
LTR39_ERV1
MER61A_ERV1
Z−score
Stem Prog Mature
Mature TEs
AluJb − SINE1
AluSx − SINE1
MIR3 − SINE2
MIR − SINE2
Group
Stem
Progenitor
Mature
D
LTR33 - ERV3
chr1:167131294-167131698
5,228 bp
1,000 bp
LTR16E1 - ERV3
chr3:87333605-87333775
1,035 bp
200 bp
LTR39 - ERV1
chr5:157361811-157362609
1,802 bp
200 bp
MER61A - ERV1
chr11:40099467-40099818
3,121 bp
500 bp
Element
LT-HSC
ST-HSC
MLP
CMP
GMP
MEP
Erythroid Prec
Granulocyte
Monocyte
Dendritic Cell
B Cell
T Cell
NK Cell
Gene
CYFIP2
Cell state
LT−HSC
ST−HSC
MLP
CMP
GMP
MEP
Erythroid Precursor
Monocyte
Granulocyte
Dendritic Cell
T Cell
B Cell
NK Cell
E
TFs on ERV1_MER57B1 (HSCs)
PBX2
PKNOX1
HOXA9
MEIS2
MEIS1
SPIB
CDK8
PBX3
SRF
−log10(q−value)
logOddsRatio
0 10 30
TFs solo LTRs_MamGypsy2−LTR (HSCs)
RAD21
SMAD1
FOXA1
PIAS1
REST
SMC3
KDM5B
NKX2−2
PDX1
0 10 20 40
logOddsRatio
F
ERV1_MER57B1
SRF
SPIB
PKNOX1
PBX3
PBX2
MEIS1
HOXA9
CDK8
Transcription Factor
100 200 300 400
Enrichment Score
solo LTRs_MamGypsy2−LTR
SMC3
RAD21
CTCF
120 130 140 150
Enrichment Score
Tissue of origin
Blood
Bone Marrow
Other
G
MER57B1 / HOXA9/HSC−Hoxa9
GGCCATAAATCA
0.0 0.5 1.0
−log10(q−value)
MamGypsy2−LTR / CTCF(Zf)/CD4+−CTCF
0 1 2 3 4
−log10(q−value)

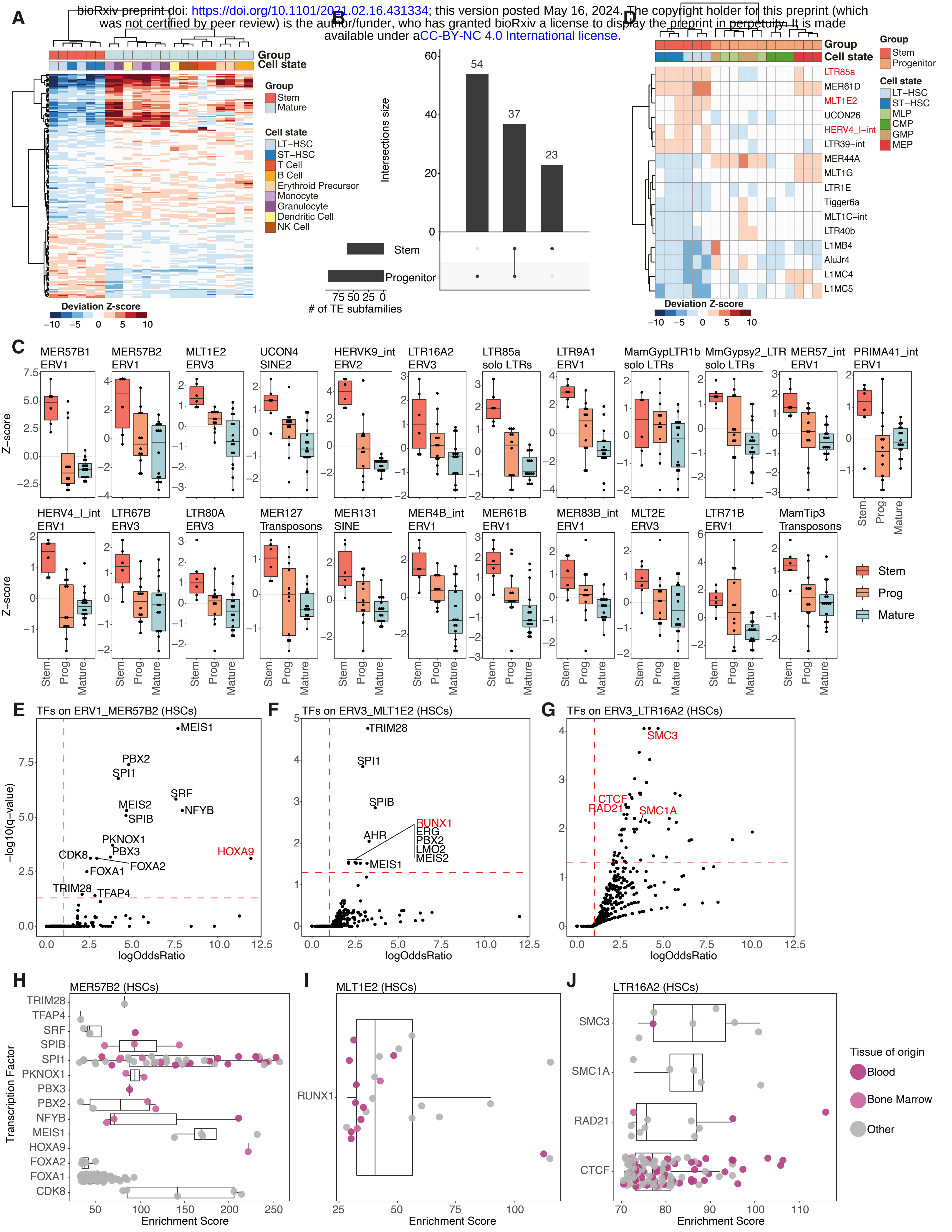

A
Group
Cell state
LSC
Group
Stem
Progenitors
Mature
LSC +
LSC -
Cell state
LT−HSC
ST−HSC
MLP
CMP
GMP
MEP
T Cell
B Cell
Erythroid Prec
Monocyte
Granulocyte
Dendritic Cell
NK Cell
LSC +
LSC -
LSC
+
-
Deviation Z-score
−10
−5
0
5
10
B
CD34/CD38
LSC
CD34/CD38
CD34+/CD38-
CD34+/CD38+
CD34-/CD38+
CD34-/CD38-
LSC
+
-
Deviation Z-score
−10
−5
0
5
10
C
TE Family
ERV1
ERV3
solo LTRs
LINE
SINE_SVA
LINE2
LINE1
SINE2
Transposons
SINE1
LSC−
LSC+
n.s.
−10
−5
0
5
Median Z-score difference
([LSC+] − [LSC−])
−20
0
20
Number of enriched
TE subfamilies
D
Genomic
HSPC
Mature
LSC+
LSC−
0
25
50
75
100
Percentage of TE Families in Accessible Chromatin
ERV1
ERV2
ERV3
LINE
LINE1
LINE2
LINE3
LINE4
LINE5
solo LTRs
SINE
SINE_SVA
SINE1
SINE2
SINE3
Transposons
E
Intersection size
0
20
40
60
80
75
60
24
23
53
21
LSC+
LSC-
HSPC
Mature
100
50
0
# of TE subfamilies

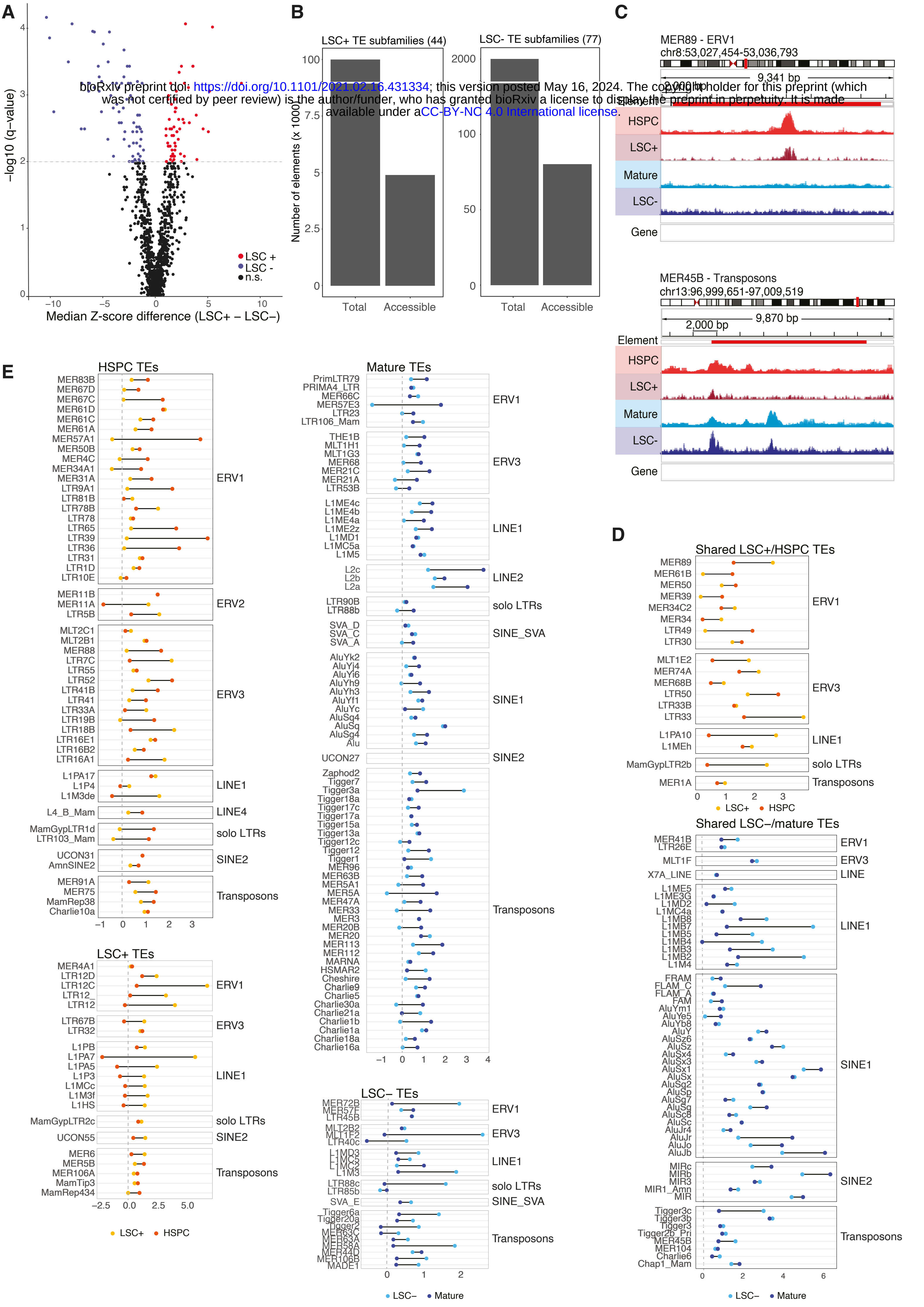

A
LSC17 score
LSCTE121 Z-score
LSC classification
Cytogenetic risk
LSC17 score
1.2
0.2
LSCTE121 Z-score
10
−30
LSC classification
High
Low
Cytogenetic risk
Poor
Intermediate
Favorable
NA
Deviation
Z-score
10
5
0
−5
−10
F
LSCTE121 Z-score
LSCTE121 Z-score
30
−30
Deviation
Z-score
10
5
0
−5
−10
B
LSCTE121 Z-score
−20
−10
0
10
High
Low
p-value
1.6e-04
C
Kaplan−Meier estimate Disease-Free Interval
LSCTE121high (n=8)
LSCTE121low (n=8)
Survival percentage
0.0
0.2
0.4
0.6
0.8
1.0
Log rank p-value=0.041
0
1000
2000
3000
4000
5000
Time (days)
D
LSCTE121 Z-score
10
0
−10
−20
R=0.27 , p=0.16
0.25
0.50
0.75
1.00
1.25
LSC17 score
E
LSCTE121 Z-score
40
20
0
−20
−40
p-value = 2.4e-05
p-value = 9.8e-05
High
Med
Low
NE
LSC engraftment class
G
LSCTE121 Z-score
−20
0
20
High
Low
p-value
7.3e-12
H
Kaplan−Meier estimate Disease-Free Interval
LSCTE121high (n=15)
LSCTE121low (n=45)
Survival percentage
0.0
0.2
0.4
0.6
0.8
1.0
Log rank p-value=0.01584
0
200
400
600
800
Time (days)

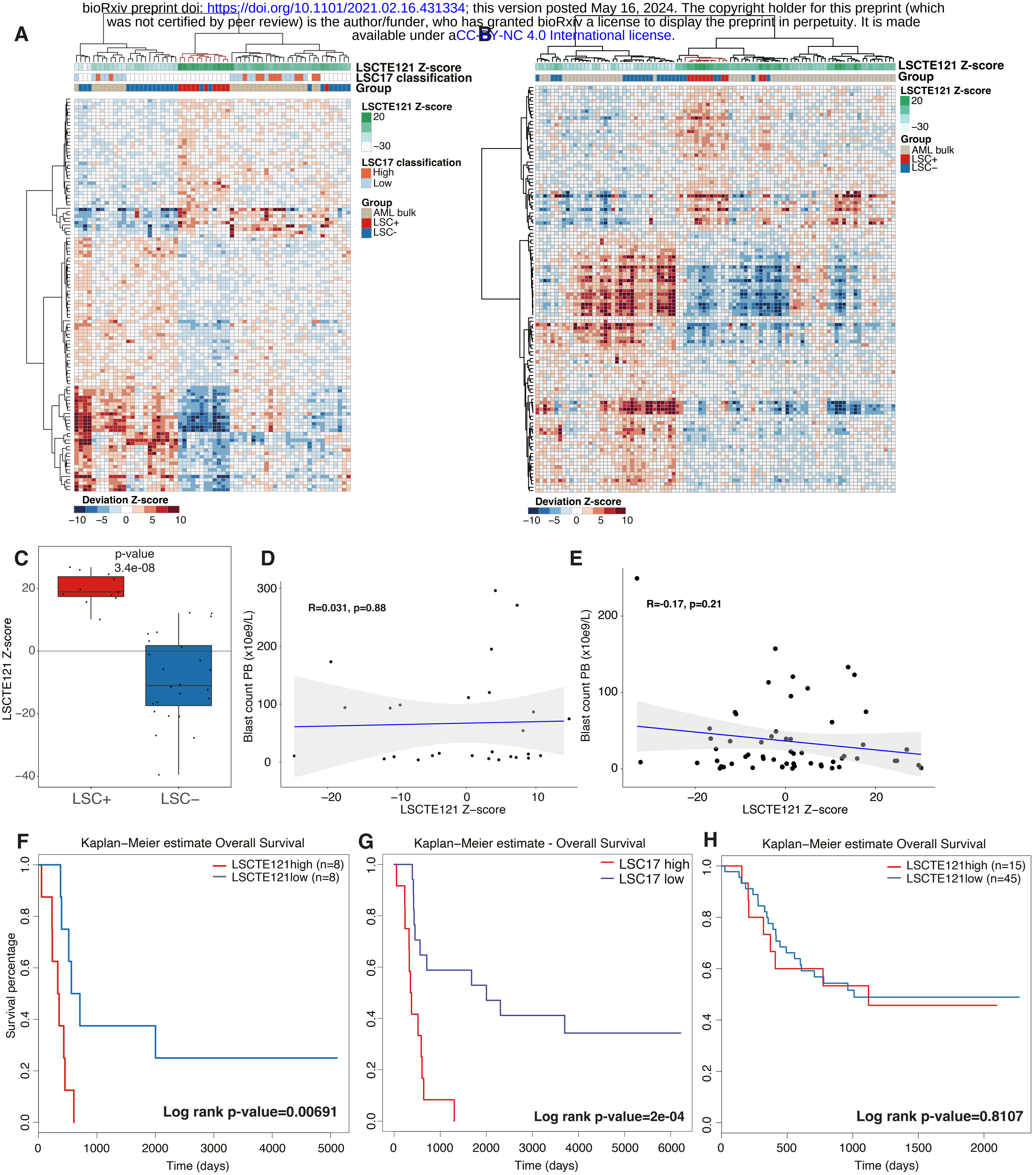

A
Top 5% Transcription Factors at LSC+ TE subfamilies
Transcription Factors
RUNX1
ERG
LMO2
TRIM28
RUNX1T1
NFYA
TCF3
TCF12
SMC1A
POU5F1
NFYB
LYL1
FLI1
CTCF
4
6
8
10
Number of TE subfamilies
B
Intersection size
12.5
10.0
7.5
5.0
2.5
0.0
11
6
5
4
4
4
2
LSC+
HSPC & LSC+
HSPC
20 15 10 5 0
# of TFs
D
NFY(CCAAT)/Promoter
TE subfamily
LTR12D
LTR12C
LTR12_
LTR12
0
1
2
3
4
−log10(q−value)
C
LYL1
NFYB
NFYA
TE subfamily
LTR67B
LTR32
L1MCc
20
40
60
80
Enrichment Score
LTR12D
LTR12C
LTR12_
LTR12
500
1000
1500
2000
Enrichment Score
LTR12D
LTR12C
LTR12_
LTR12
0
200
400
600
Enrichment Score
Tissue of Origin
Blood
Bone Marrow
Other
E
Transcription Factors
LYL1
NFYA
NFYB
POU5F1
AML
ALL
ATRT
Bladder carcinoma
Breast carcinoma
Breast ductal carcinoma
Cholangiocarcinoma
CML
Colorectal adenocarcinoma
Endometrial adenocarcinoma
Esophagus adenocarcinoma
Esophagus squamous
Ewing sarcoma
Exocrine
Gastric adenocarcinoma
Glioma
Hepatocellular carcinoma
Liposarcoma
Malignant rhabdoid tumor
Medulloblastoma
Melanoma
Mesothelioma
Multiple myeloma
Neuroblastoma
Non hodgkin lymphoma
NSCLC
Osteosarcoma
Ovary adenocarcinoma
Renal cell carcinoma
Rhabdomyosarcoma
SCLC
Synovial sarcoma
Thyroid carcinoma
Upper aerodigestive squamous
Uveal melanoma
Other
Adjusted p−value
0.001
0.01
0.05
ns
Median essentiality score
−0.2
0.0
0.2

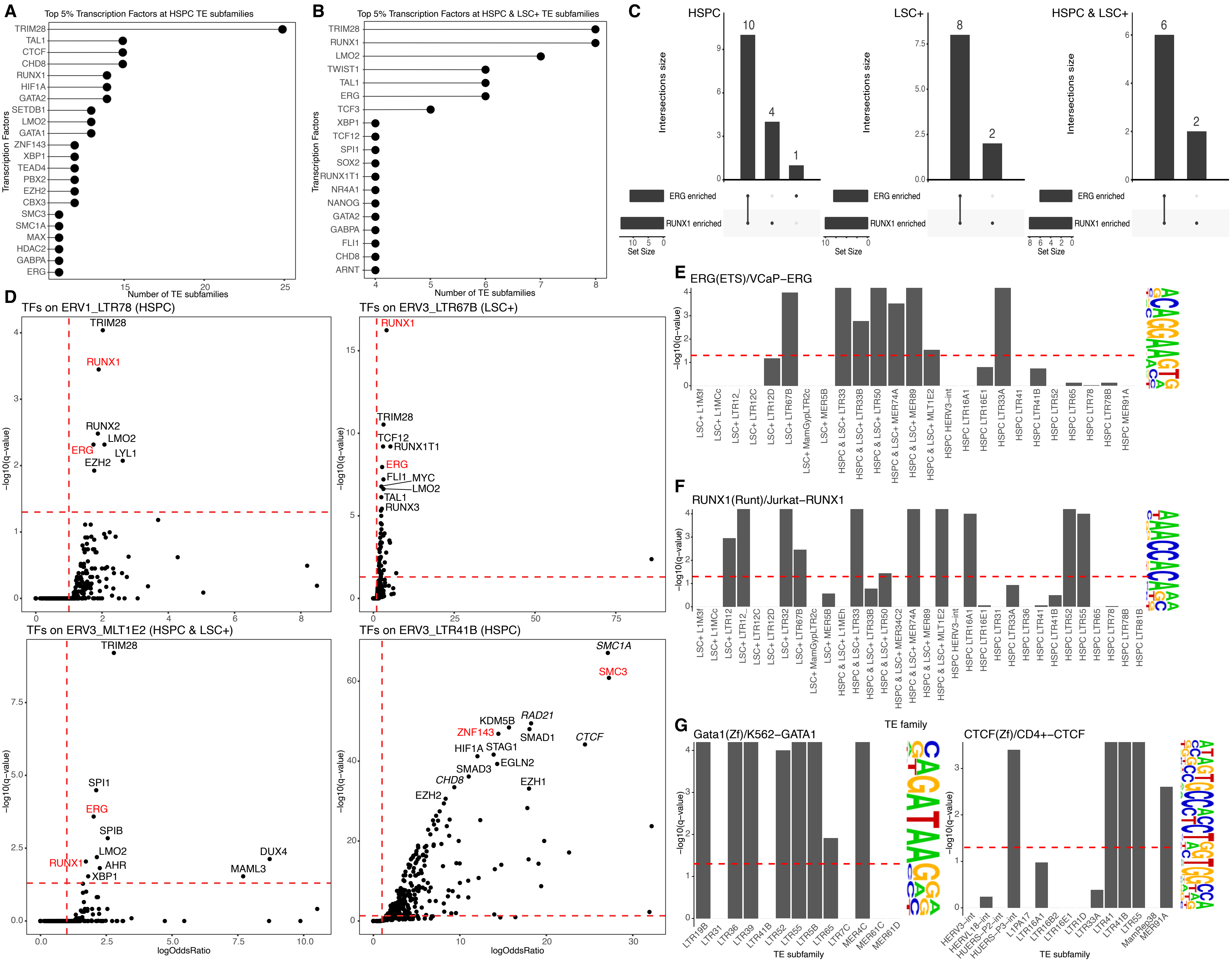

A
CD34/CD38
LSCTE121
LSC
Type
Deviation Z-score
10
5
0
−5
−10
CD34/CD38
CD34+/CD38−
CD34+/CD38+
CD34−/CD38+
CD34−/CD38−
LSCTE121
30
−30
LSC
+
−
Type
Patients
OCIAML22
B
ERV1_LTR12C
q-value 0.002
Z-score
10
0
−10
LSC+
LSC−
C
LSC+
LSC-
LTR12C elements
-1.5kb
LTR12C
1.5kb
Genomic Regions
ATAC reads coverage
35
30
25
20
15
10
5
0
D
ERV1_LTR12C
Z-score
2.5
0.0
−2.5
−5.0
CD34 + + − −
CD38 - + + −
OCI−AML22 Fractions
E
OCI-AML22
CD38
CD34
+
-
LSC+
dCas9-KRAB (CRISPRi)
LSC+ CRISPRi
5 x LTR12C gRNAs
LSC+ CRISPRi LTR12C gRNA
F
p-value 3.6e-06
p-value 0.7
log2(normalized ATAC signal)
8
6
4
LTR12C
LTR12D
TE subfamilies
gRNAs combo
Control
LTR12C
G
LTR12C - ERV1
chr1:196,806,317-196,812,400
6,085 bp
1,000 bp
Element
Control
LTR12C
Gene
chr5:90,845,205-90,849,268
4,065 bp
1,000 bp
Element
Control
LTR12C
Gene
ADGRV1
chr2:188,605,083-188,607,484
2,403 bp
400 bp
Element
Control
LTR12C
Gene
XR_007087590.1
H
Percentage of viable cells (%)
100
75
50
25
0
p-value 0.003
p-value 0.003
p-value 0.004
p-value 0.2
Control
LTR12C
CD34+ CD38-
CD34+ CD38+
CD34- CD38+
CD34- CD38-

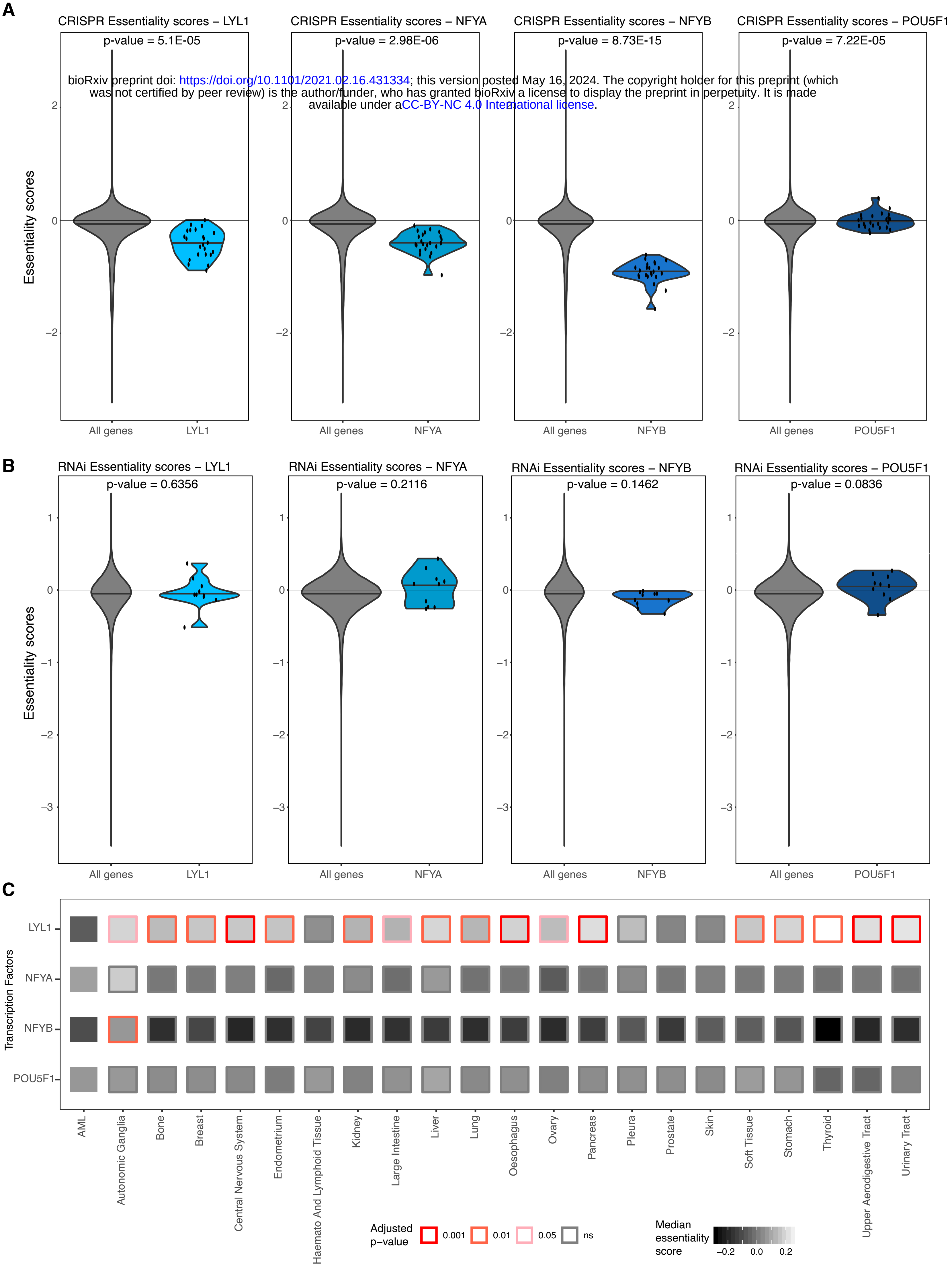

A
Stem
CD34+/CD38-/CD45RA-/
CD90+/CD49f+
LT-HSC
CD34+/CD38-/CD45RA-/
CD90-/CD49f-
ST-HSC
Progenitors
CD34+/CD38+/CD10-/
CD45RA-/Flt3+/CD71-
CMP
MLP
CD34+/CD38-/CD45RA+
CD34+/CD38+/CD10-/
CD45RA-/Flt3-/
CD71+/CD3+
CD19-
MEP
GMP
CD34+/CD38+
CD10-/CD45RA+
CD71-
Mature
Erythroid
precursor
CD71+
GPA+
Erythrocytes
Platelets
Granulocytes
CD14+/CD15+/
CD3-/CD19-
Monocytes
CD14+/CD15-/
CD3-/CD19-
Dendritic
cells
CD11c+/
CD45+
B cells
CD19+/
CD3-
T cells
CD3+/
CD19-
NK cells
CD56+
CD3-/CD19-
CD14-
B
Lineage
Group
Cell state
MSTA
LTR13A
LTR45C
LTR8A
LTR19C
LTR25
LTR9B
MamTip1
LTR1
MER34D
MER4D1
LTR56
MER5C1
MER65D
HERVH48–int
LTR73
MER65C
MLT1A
LTR15
LTR2
LTR53
LTR55
LTR62
MER54B
MER51E
LTR76
Deviation Z-score
-10 -5 0 5 10
Lineage
Non-ery
Ery
Group
Stem
Progenitor
Mature
Cell state
LT–HSC
ST–HSC
MLP
CMP
GMP
MEP
T Cell
B Cell
Erythroid Precursor
Monocyte
Granulocyte
Dendritic Cell
NK Cell
C
Lineage
Group
Cell state
Lineage
Non-ery
Ery
Group
Stem
Progenitor
Mature
Cell state
LT–HSC
ST–HSC
MLP
CMP
GMP
MEP
T Cell
B Cell
Erythroid Precursor
Monocyte
Granulocyte
Dendritic Cell
NK Cell
Deviation Z-score
-10 -5 0 5 10
D
-log10 (q-value)
0
2
4
6
−15 −10 −5 0 5 10
Median Z-score difference (HSPC vs Mature)
HSPC
Mature
n.s.
E
HSPC TEs
LTR33 – ERV3
LTR16E1 – ERV3
LTR39 – ERV1
MER61A – ERV1
NK Cell
Dendritic Cell
Granulocyte
Monocyte
Erythroid Precustor
B Cell
T Cell
MEP
GMP
CMP
MLP
ST–HSC
LT–HSC
Group
-6 -3 0 3
-2 0 2
-4 -2 0 2 4
-2 0 2
Mature TEs
AluJb – SINE1
AluSx – SINE1
MIR3 – SINE2
MIR – SINE2
NK Cell
Dendritic Cell
Granulocyte
Monocyte
Erythroid Precustor
B Cell
T Cell
MEP
GMP
CMP
MLP
ST–HSC
LT–HSC
Group
-10 0 10
-10 0 10
-5 0 5 10
-10 -5 0 5 10
Z–score
Z–score
Z–score
Z–score
F
# of peaks
2e+04
4e+04
6e+04
8e+04
1e+05
Stem
Progenitor
Mature
LT–HSC
ST–HSC
MLP
CMP
GMP
MEP
T Cell
B Cell
Erythroid Precustor
Monocyte
Granulocyte
Dendritic Cell
NK Cell
Cell State

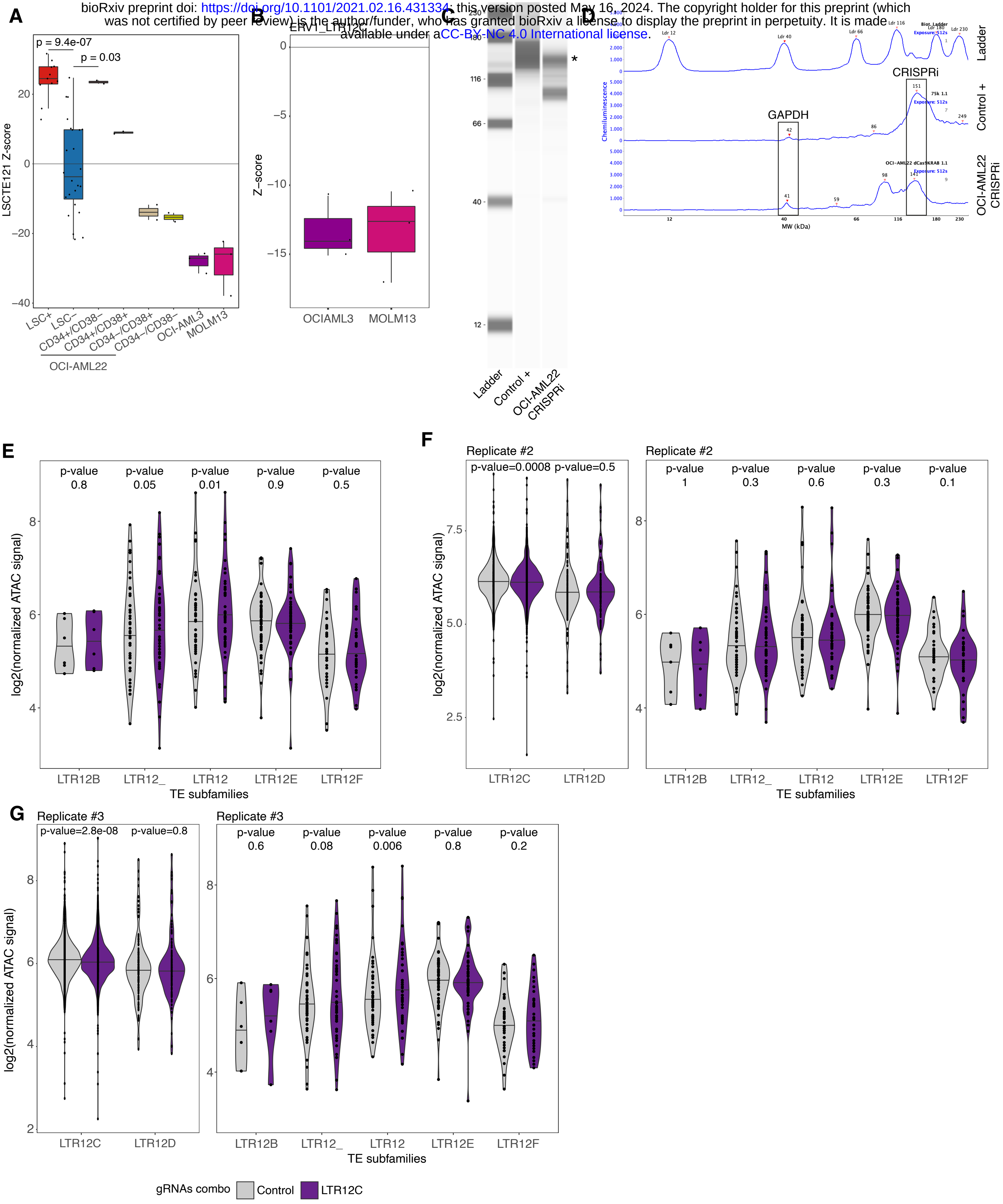

A
Group
Cell state
Group
Stem
Mature
Cell state
LT–HSC
ST–HSC
T Cell
B Cell
Erythroid Precursor
Monocyte
Granulocyte
Dendritic Cell
NK Cell
Deviation Z-score
−10 −5 0 5 10
B
Intersections size
60 40 20 0
54
37
23
Stem
Progenitor
75 50 25 0
# of TE subfamilies
D
Group
Cell state
LTR85a
MER61D
MLT1E2
UCON26
HERV4_I–int
LTR39–int
MER44A
MLT1G
LTR1E
Tigger6a
MLT1C–int
LTR40b
L1MB4
AluJr4
L1MC4
L1MC5
Group
Stem
Progenitor
Cell state
LT–HSC
ST–HSC
MLP
CMP
GMP
MEP
Deviation Z-score
−10 −5 0 5 10
C
Z–score
MER57B1 ERV1
MER57B2 ERV1
MLT1E2 ERV3
UCON4 SINE2
HERVK9_int ERV2
LTR16A2 ERV3
LTR85a solo LTRs
LTR9A1 ERV1
MamGypLTR1b solo LTRs
MmGypsy2_LTR solo LTRs
MER57_int ERV1
PRIMA41_int ERV1
HERV4_I_int ERV1
LTR67B ERV3
LTR80A ERV3
MER127 Transposons
MER131 SINE
MER4B_int ERV1
MER61B ERV1
MER83B_int ERV1
MLT2E ERV3
LTR71B ERV1
MamTip3 Transposons
Stem
Prog
Mature
Stem
Prog
Mature
E
TFs on ERV1_MER57B2 (HSCs)
−log10(q−value)
logOddsRatio
MEIS1
PBX2
SPI1
SRF
NFYB
MEIS2
SPIB
PKNOX1
PBX3
CDK8
FOXA2
FOXA1
HOXA9
TRIM28
TFAP4
F
TFs on ERV3_MLT1E2 (HSCs)
logOddsRatio
TRIM28
SPI1
SPIB
AHR
RUNX1
ERG
PBX2
LMO2
MEIS2
MEIS1
G
TFs on ERV3_LTR16A2 (HSCs)
logOddsRatio
SMC3
CTCF
RAD21
SMC1A
H
MER57B2 (HSCs)
Transcription Factor
TRIM28
TFAP4
SRF
SPIB
SPI1
PKNOX1
PBX3
PBX2
NFYB
MEIS1
HOXA9
FOXA2
FOXA1
CDK8
Enrichment Score
50 100 150 200 250
I
MLT1E2 (HSCs)
RUNX1
Enrichment Score
25 50 75 100
J
LTR16A2 (HSCs)
SMC3
SMC1A
RAD21
CTCF
Enrichment Score
70 80 90 100 110
Tissue of origin
Blood
Bone Marrow
Other


A
−log10 (q−value)
Median Z-score difference (LSC+ − LSC−)
LSC +
LSC -
n.s.
B
LSC+ TE subfamilies (44)
LSC- TE subfamilies (77)
Number of elements (x 1000)
Total
Accessible
C
MER89 - ERV1
chr8:53,027,454-53,036,793
9,341 bp
2,000 bp
Element
HSPC
LSC+
Mature
LSC-
Gene
MER45B - Transposons
chr13:96,999,651-97,009,519
9,870 bp
2,000 bp
D
Shared LSC+/HSPC TEs
MER89
MER61B
MER50
MER39
MER34C2
MER34
LTR49
LTR30
ERV1
MLT1E2
MER74A
MER68B
LTR50
LTR33B
LTR33
ERV3
L1PA10
L1MEh
LINE1
MamGypLTR2b
solo LTRs
MER1A
Transposons
LSC+
HSPC
Shared LSC−/mature TEs
MER41B
LTR26E
ERV1
MLT1F
ERV3
X7A_LINE
LINE
L1ME5
L1ME3G
L1MD2
L1MC4a
L1MB8
L1MB7
L1MB5
L1MB4
L1MB3
L1MB2
L1M4
LINE1
FRAM
FLAM_C
FLAM_A
FAM
AluYm1
AluYe5
AluYb8
AluY
AluSz6
AluSz
AluSx4
AluSx3
AluSx1
AluSx
AluSq2
AluSp
AluSg7
AluSg
AluSc8
AluSc
AluJr4
AluJr
AluJo
AluJb
SINE1
MIRc
MIRb
MIR3
MIR1_Amn
MIR
SINE2
Tigger3c
Tigger3b
Tigger3
Tigger2b_Pri
MER45B
MER104
Charlie6
Chap1_Mam
Transposons
LSC−
Mature
E
HSPC TEs
MER83B
MER67D
MER67C
MER61D
MER61C
MER61A
MER57A1
MER50B
MER4C
MER34A1
MER31A
LTR9A1
LTR81B
LTR78B
LTR78
LTR65
LTR39
LTR36
LTR31
LTR1D
LTR10E
ERV1
MER11B
MER11A
LTR5B
ERV2
MLT2C1
MLT2B1
MER88
LTR7C
LTR55
LTR52
LTR41B
LTR41
LTR33A
LTR19B
LTR18B
LTR16E1
LTR16B2
LTR16A1
ERV3
L1PA17
L1P4
L1M3de
LINE1
L4_B_Mam
LINE4
MamGypLTR1d
LTR103_Mam
solo LTRs
UCON31
AmnSINE2
SINE2
MER91A
MER75
MamRep38
Charlie10a
Transposons
LSC+ TEs
MER4A1
LTR12D
LTR12C
LTR12_
LTR12
ERV1
LTR67B
LTR32
ERV3
L1PB
L1PA7
L1PA5
L1P3
L1MCc
L1M3f
L1HS
LINE1
MamGypLTR2c
solo LTRs
UCON55
SINE2
MER6
MER5B
MER106A
MamTip3
MamRep434
Transposons
LSC+
HSPC
Mature TEs
PrimLTR79
PRIMA4_LTR
MER66C
MER57E3
LTR23
LTR106_Mam
ERV1
THE1B
MLT1H1
MLT1G3
MER68
MER21C
MER21A
LTR53B
ERV3
L1ME4c
L1ME4b
L1ME4a
L1ME2z
L1MD1
L1MC5a
L1M5
LINE1
L2c
L2b
L2a
LINE2
LTR90B
LTR88b
solo LTRs
SVA_D
SVA_C
SVA_A
SINE_SVA
AluYk2
AluYj4
AluYi6
AluYh9
AluYh3
AluYf1
AluYc
AluSq4
AluSq
AluSg4
Alu
SINE1
UCON27
SINE2
Zaphod2
Tigger7
Tigger3a
Tigger18a
Tigger17c
Tigger17a
Tigger15a
Tigger13a
Tigger12c
Tigger12
Tigger1
MER96
MER63B
MER5A1
MER5A
MER47A
MER33
MER3
MER20B
MER20
MER113
MER112
MARNA
HSMAR2
Cheshire
Charlie9
Charlie5
Charlie30a
Charlie21a
Charlie1b
Charlie1a
Charlie18a
Charlie16a
Transposons
LSC− TEs
MER72B
MER57F
LTR45B
ERV1
MLT2B2
MLT1F2
LTR40c
ERV3
L1MD3
L1MC5
L1MC2
L1M3
LINE1
LTR88c
LTR85b
solo LTRs
SVA_E
SINE_SVA
Tigger6a
Tigger20a
Tigger2
MER63C
MER63A
MER58A
MER44D
MER106B
MADE1
Transposons
LSC−
Mature


A
LSCTE121 Z-score
LSC17 classification
Group
LSCTE121 Z-score
20
−30
LSC17 classification
High
Low
Group
AML bulk
LSC+
LSC-
Deviation Z-score
−10 −5 0 5 10
B
LSCTE121 Z-score
Group
LSCTE121 Z-score
20
−30
Group
AML bulk
LSC+
LSC−
Deviation Z-score
−10 −5 0 5 10
C
p-value
3.4e-08
LSCTE121 Z-score
20
0
−20
−40
LSC+
LSC−
D
R=0.031, p=0.88
Blast count PB (x10e9/L)
300
200
100
0
−20 −10 0 10
LSCTE121 Z−score
E
R=-0.17, p=0.21
Blast count PB (x10e9/L)
200
100
0
−20 0 20
LSCTE121 Z−score
F
Kaplan−Meier estimate Overall Survival
LSCTE121high (n=8)
LSCTE121low (n=8)
Survival percentage
1.0
0.8
0.6
0.4
0.2
0.0
Log rank p-value=0.00691
0 1000 2000 3000 4000 5000
Time (days)
G
Kaplan−Meier estimate - Overall Survival
LSC17 high
LSC17 low
1.0
0.8
0.6
0.4
0.2
0.0
Log rank p-value=2e-04
0 1000 2000 3000 4000 5000 6000
Time (days)
H
Kaplan−Meier estimate Overall Survival
LSCTE121high (n=15)
LSCTE121low (n=45)
1.0
0.8
0.6
0.4
0.2
0.0
Log rank p-value=0.8107
0 500 1000 1500 2000
Time (days)


A
Top 5% Transcription Factors at HSPC TE subfamilies
Transcription Factors
TRIM28
TAL1
CTCF
CHD8
RUNX1
HIF1A
GATA2
SETDB1
LMO2
GATA1
ZNF143
XBP1
TEAD4
PBX2
EZH2
CBX3
SMC3
SMC1A
MAX
HDAC2
GABPA
ERG
15
20
25
Number of TE subfamilies
B
Top 5% Transcription Factors at HSPC & LSC+ TE subfamilies
Transcription Factors
TRIM28
RUNX1
LMO2
TWIST1
TAL1
ERG
TCF3
XBP1
TCF12
SPI1
SOX2
RUNX1T1
NR4A1
NANOG
GATA2
GABPA
FLI1
CHD8
ARNT
4
5
6
7
8
Number of TE subfamilies
C
HSPC
Intersections size
10
4
1
ERG enriched
RUNX1 enriched
10 5 0
Set Size
LSC+
8
2
10 0
HSPC & LSC+
6
2
8 6 4 2 0
D
TFs on ERV1_LTR78 (HSPC)
TRIM28
RUNX1
RUNX2
LMO2
ERG
LYL1
EZH2
-log10(q-value)
TFs on ERV3_LTR67B (LSC+)
RUNX1
TRIM28
TCF12
RUNX1T1
ERG
FLI1
MYC
LMO2
TAL1
RUNX3
TFs on ERV3_MLT1E2 (HSPC & LSC+)
TRIM28
SPI1
ERG
SPIB
LMO2
RUNX1
AHR
XBP1
DUX4
MAML3
logOddsRatio
TFs on ERV3_LTR41B (HSPC)
SMC1A
SMC3
KDM5B
RAD21
ZNF143
SMAD1
CTCF
HIF1A
STAG1
EGLN2
SMAD3
CHD8
EZH1
EZH2
logOddsRatio
E
ERG(ETS)/VCaP-ERG
-log10(q-value)
LSC+ L1M3f
LSC+ L1MCc
LSC+ LTR12_
LSC+ LTR12C
LSC+ LTR12D
LSC+ LTR67B
LSC+ MamGypLTR2c
LSC+ MER5B
HSPC & LSC+ LTR33
HSPC & LSC+ LTR33B
HSPC & LSC+ LTR50
HSPC & LSC+ MER74A
HSPC & LSC+ MER89
HSPC & LSC+ MLT1E2
HSPC HERV3-int
HSPC LTR16A1
HSPC LTR16E1
HSPC LTR33A
HSPC LTR41
HSPC LTR41B
HSPC LTR52
HSPC LTR65
HSPC LTR78
HSPC LTR78B
HSPC MER91A
F
RUNX1(Runt)/Jurkat-RUNX1
-log10(q-value)
LSC+ L1M3f
LSC+ L1MCc
LSC+ LTR12
LSC+ LTR12_
LSC+ LTR12C
LSC+ LTR12D
LSC+ LTR32
LSC+ LTR67B
LSC+ MamGypLTR2c
LSC+ MER5B
HSPC & LSC+ L1MEh
HSPC & LSC+ LTR33
HSPC & LSC+ LTR33B
HSPC & LSC+ LTR50
HSPC & LSC+ MER34C2
HSPC & LSC+ MER74A
HSPC & LSC+ MER89
HSPC & LSC+ MLT1E2
HSPC HERV3-int
HSPC LTR16A1
HSPC LTR16E1
HSPC LTR31
HSPC LTR33A
HSPC LTR36
HSPC LTR41
HSPC LTR41B
HSPC LTR52
HSPC LTR55
HSPC LTR65
HSPC LTR78
HSPC LTR78B
HSPC LTR81B
G
TE family
Gata1(Zf)/K562-GATA1
-log10(q-value)
LTR19B
LTR31
LTR36
LTR39
LTR41B
LTR52
LTR55
LTR5B
LTR65
LTR7C
MER4C
MER61C
MER61D
TE subfamily
CTCF(Zf)/CD4+-CTCF
HERV3-int
HERVL18-int
HUERS-P2-int
HUERS-P3-int
L1PA17
LTR16A1
LTR16B2
LTR16E1
LTR1D
LTR33A
LTR41
LTR41B
LTR55
MamRep38
MER91A
TE subfamily


A
CRISPR Essentiality scores – LYL1
p-value = 5.1E-05
CRISPR Essentiality scores – NFYA
p-value = 2.98E-06
CRISPR Essentiality scores – NFYB
p-value = 8.73E-15
CRISPR Essentiality scores – POU5F1
p-value = 7.22E-05
Essentiality scores
2
0
−2
All genes
LYL1
NFYA
NFYB
POU5F1
B
RNAi Essentiality scores – LYL1
p-value = 0.6356
RNAi Essentiality scores – NFYA
p-value = 0.2116
RNAi Essentiality scores – NFYB
p-value = 0.1462
RNAi Essentiality scores – POU5F1
p-value = 0.0836
Essentiality scores
1
0
−1
−2
−3
All genes
LYL1
NFYA
NFYB
POU5F1
C
Transcription Factors
LYL1
NFYA
NFYB
POU5F1
AML
Autonomic Ganglia
Bone
Breast
Central Nervous System
Endometrium
Haemato And Lymphoid Tissue
Kidney
Large Intestine
Liver
Lung
Oesophagus
Ovary
Pancreas
Pleura
Prostate
Skin
Soft Tissue
Stomach
Thyroid
Upper Aerodigestive Tract
Urinary Tract
Adjusted p−value
0.001
0.01
0.05
ns
Median essentiality score
−0.2
0.0
0.2


A
p = 9.4e-07
p = 0.03
LSCTE121 Z-score
20
0
−20
−40
LSC+
LSC−
CD34+/CD38−
CD34+/CD38+
CD34−/CD38+
CD34−/CD38−
OCI-AML3
MOLM13
OCI-AML22
B
ERV1_LTR12C
Z-score
0
−5
−10
−15
OCIAML3
MOLM13
C
230
180
116
66
40
12
*
Ladder
Control +
OCI-AML22 CRISPRi
D
Ldr 12
Ldr 40
Ldr 66
Ldr 116
Biot. Ladder
Ldr 180
Exposure: 512s
Ldr 230
Ladder
CRISPRi
151
75k 1.1
Exposure: 512s
249
86
GAPDH
42
Control +
Chemiluminescence
5.000
4.000
3.000
2.000
1.000
0
OCI-AML22 dCas9KRAB 1.1
Exposure: 512s
98
141
41
59
OCI-AML22 CRISPRi
12
40
66
116
180
230
MW (kDa)
E
log2(normalized ATAC signal)
p-value 0.8
p-value 0.05
p-value 0.01
p-value 0.9
p-value 0.5
8
6
4
LTR12B
LTR12_
LTR12
LTR12E
LTR12F
TE subfamilies
F
Replicate #2
p-value=0.0008
p-value=0.5
log2(normalized ATAC signal)
7.5
5.0
2.5
LTR12C
LTR12D
Replicate #2
p-value 1
p-value 0.3
p-value 0.6
p-value 0.3
p-value 0.1
8
6
4
LTR12B
LTR12_
LTR12
LTR12E
LTR12F
TE subfamilies
G
Replicate #3
p-value=2.8e-08
p-value=0.8
log2(normalized ATAC signal)
8
6
4
2
LTR12C
LTR12D
Replicate #3
p-value 0.6
p-value 0.08
p-value 0.006
p-value 0.8
p-value 0.2
8
6
4
LTR12B
LTR12_
LTR12
LTR12E
LTR12F
TE subfamilies
gRNAs combo
Control
LTR12C