Accessibility Over Transposable Elements Reveals Genetic Determinants of Stemness Properties in Normal and Leukemic Hematopoiesis

Bettina Nadorp[1,7], Giacomo Grillo[1,7], Aditi Qamra[1], Amanda Mitchell[1], Christopher Arlidge[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], John E. Dick[1,5,*], Mathieu Lupien[1,2,6,*]

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. Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, M5S 1A8
6. Ontario Institute for Cancer Research, Toronto, Ontario, Canada, M5G 0A3
7. Equal contribution

* Corresponding author: John E. Dick: jdick@uhnresearch.ca and Mathieu Lupien: mlupien@uhnresearch.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 of patients will relapse with fatal disease within 5 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 of the genetic determinants of their stemness properties [6–8]. Although much AML research focuses on mutational processes and their impact on gene expression programs, the genetic determinants of cell state properties including stemness expand beyond mutations, relying on the genetic architecture captured in the chromatin of each cell [9–11]. As LSCs share many functional and molecular properties with normal hematopoietic stem cells (HSC), we identified genetic determinants of primitive populations enriched for LSCs and HSCs in comparison with their downstream mature progeny by investigating their chromatin accessibility. Our work reveals how distinct transposable element (TE) subfamilies are used in primitive versus mature populations, functioning as docking sites for stem cell-associated regulators of genome topology, including CTCF, or lineage-specific transcription regulators in primitive and mature populations, respectively. We further show how TE subfamilies accessible in LSCs define docking sites for several oncogenic drivers in AML, namely FLI1, LYL1 and MEIS1. Using chromatin accessibility profiles from a cohort of AML patients, we further show the clinical utility of our TE accessibility-based LSCTE121 scoring scheme to identify patients with high rates of relapse. Collectively, our work reveals how different accessible TE subfamilies serve as genetic determinants of stemness properties in normal and leukemic hematopoietic stem cells.

Introduction

Lifelong blood production originates from self-renewing, multipotent hematopoietic stem cells (HSCs) that give rise to the full complement of progenitor and mature hematopoietic populations, including granulocytes, erythrocytes, natural killer, B and T cells [12]. This phenotypic heterogeneity relies on population-specific genetic architectures, where phenotypically distinct populations arise from the contribution of different sections of a shared genome [11,13–15]. For instance, chromatin accessibility across hematopoietic populations reveals how DNA recognition sequences for CTCF contribute to the transition of dormant, low primed to activated, primed HSC and mediate changes to the three-dimensional genome organization [16]. In contrast, DNA recognition sequences of lineage-specific transcription regulators are uniquely accessible to serve as cis-regulatory elements promoting lineage-specific expression profiles across each mature population [16,17]. Hence, characterizing the chromatin accessibility across hematopoietic populations identifies genetic determinants of phenotypic heterogeneity.

Akin to normal hematopoiesis, acute myeloid leukemia (AML) is hierarchically organized [18], with leukemia stem cells (LSC) standing at the apex [19,20]. As LSCs possess self-renewal capacity they serve to sustain long-term maintenance of the malignant population and represent the source of evolutionary diversity, driving leukemic progression, therapy failure and relapse [1–3]. The clinical relevance of LSCs is reflected in the predictive value of an LSC-derived gene expression signature (LSC17) on overall patient survival and response to standard induction therapy [6]. Mutational tracing studies of genetically diverse subclones, present at AML diagnosis, have demonstrated that LSCs survive induction therapy and cause relapse [2]. Mutations that give rise to clonal hematopoiesis are amongst the first changes to the genetic architecture of HSCs and promote the development of AML by creating a pre-leukemic precursor reservoir [2,21,22]. However, we know little of the additional genetic determinants that drive the transformation of HSCs into LSCs. Thus, a clear understanding of the genetic architecture in LSCs in comparison to HSCs is required to find the sum of genetic determinants of stemness and to devise strategies to block oncogenesis.

To date, the search for genetic determinants of phenotypic heterogeneity across normal hematopoietic populations and AML has mainly focused on the non-repetitive genome, excluding transposable elements (TEs) that compose close to 50% of the human genome [23,24]. However, TEs account for diverse genetic functions relevant to development and pathogenesis[25]. TEs are classified in subfamilies composed of tens to thousands of individual elements traced back to a single ancestral unit [26]. 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 [27–29]. Other TE subfamilies lie in accessible chromatin where they function as *cis*-regulatory elements rather than encoding transcripts [30]. Such TE subfamilies potentiate the expression of neighbouring genes and/or impact the three-dimensional genome organization by recruiting sequence-specific transcription regulators, including CTCF [31–33]. This results in TE onco-exaptation mediating oncogene expression in AML, as in various other cancer types [32,34,35]. Despite their reported contribution to cancer, the contribution of TEs as genetic determinants of stemness properties in LSCs remains unknown. In the present study, we show how different TE subfamilies are accessible across the continuum of both normal hematopoietic and LSC-driven leukemic hierarchies, acting as *cis*-regulatory elements to serve as genetic determinants of stemness properties.

Results

Accessible TEs distinguish primitive from mature cells

To systematically study TE enrichment within accessible chromatin across hematopoietic populations, we used Assay for Transposase-Accessible Chromatin Sequencing (ATAC-Seq; see Methods) data generated across six hematopoietic stem and progenitor populations (LT-HSCs, ST-HSCs, common myeloid progenitors-CMP, megakaryocyte-erythroid progenitors-MEP, multi-lymphoid progenitors-MLP and granulocyte-monocyte progenitors-GMP) and seven mature cell populations exhibiting myeloid and lymphoid fates isolated from human cord blood (Suppl. Fig. 1a) [16]. These populations were selected according to the sorting scheme we developed [36,37], which achieves the

highest purity compared to other studies [17] and allows isolation of HSCs with intermediate potential (~30%) for improved resolution at the apex of the normal hematopoietic hierarchy [36] (see Methods).

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 and sequencing depth. 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 primitive, share similar TE enrichment profiles that are distinct from mature populations (Fig. 1a). The megakaryocyte-erythroid lineage is the exception, where strong lineage-specific TE subfamily enrichment results in the clustering of MEP with erythroid precursors that are more similar to other mature rather than primitive hematopoietic populations (Fig. 1a). To show that the megakaryocyte-erythroid lineage specific TEs are driving the separation in the unsupervised clustering, we removed the strongest lineage-specific TEs (Benjamini-Hochberg corrected p-value (q-value)<0.0001, Suppl. Fig. 1b) from the list of TEs used in unsupervised clustering. This resulted in a clear cut separation of HSCs and progenitors (primitive) versus mature populations (Suppl. Fig. 1c). Thus, our data suggest that the differential TE accessibility across specific subfamilies distinguishes primitive from mature hematopoietic cells.

To identify differentially-accessible TE subfamilies between primitive versus mature populations, we compared the accessibility Z-scores of all TE subfamilies, using two-sided wilcoxon signed rank test. We found a total of 93 and 138 TE subfamilies belonging to 13 TE families enriched in the primitive and mature populations, respectively (q < 0.01, Fig. 1b, Suppl. Fig. 1d, Suppl. Table 1). Primitive populations were mainly enriched in subfamilies belonging to HERV1 and HERV3 TE families (Fig. 1b). Enriched HERV1 subfamilies included MER61A (q=2.57e-04) and LTR39 (q=1.65e-04), while enriched HERV3 subfamilies included LTR16E1 (q=2.94e-04) and LTR33 (q=8.10e-05)(Fig 1c). Examples of specific elements of each subfamily uniquely enriched in primitive populations can be seen in Fig. 1d. In contrast, mature populations were mainly enriched in SINE1, SINE2 and Transposons family members (Fig. 1b),

such as AluJB (q=7.59e-05) and AluSx (q=2.9e-04) of the SINE1 subfamily, as well as MIR (q=8.29e-05) and MIR3 (q=7.59e-05) of the SINE2 subfamily (Fig. 1c, Suppl. Fig. 1e). Despite the thousands of accessible sites found across all populations (Suppl. Fig. 1f), no individual TE element of these subfamilies was accessible across all mature populations and absent from primitive populations.

Accessible TEs as docking sites for stem factors

We next assessed if chromatin accessibility at TE subfamilies could distinguish LT-HSCs (the most primitive HSC subset) from any other hematopoietic population (Suppl. Fig. 1a). Comparison of the accessibility Z-scores of all TE subfamilies failed to identify differentially-enriched TE subfamilies across all one-to-one comparisons. No differentially-accessible TE subfamilies could discriminate LT-HSCs or ST-HSCs from the collection of progenitor populations (Suppl. Table 2 and Suppl. Table 3). Comparison of LT-HSCs against the collection of mature populations revealed 25 TE subfamilies enriched in LT-HSCs versus 70 in mature populations ($q < 0.01$, Suppl. Fig. 2a, Suppl. Table 4). In contrast, we could not identify TE subfamilies differentially-enriched between ST-HSCs versus mature populations ($q < 0.01$). Comparison of progenitors to mature populations identified 59 TE subfamilies enriched in progenitor versus 83 in mature populations ($q < 0.01$, Suppl. Fig. 2b, Suppl. Table 5). Of all accessible TE subfamilies distinct between LT-HSC, progenitors or primitive populations compared to mature populations, five TE subfamilies were found to be specific to LT-HSCs (Suppl. Fig. 2c), namely MER57B1, MER57B2, LTR67B, MamGypLTR1b and MamGypsy2-LTR (Suppl. Fig 2d). In agreement with these results, MER57B1, MER57B2 and LTR67B subfamilies showed the highest Z-scores in the LT/HSPC signature of chromatin accessibility previously reported to discriminate long-term from short-term self-renewing and other progenitor populations [16] compared to all other hematopoietic chromatin accessibility signatures (Suppl. Table 6). Collectively, our results reveal that specific TE subfamilies discriminate primitive from mature populations, with a limited set of five that are uniquely accessible in the self-renewing LT-HSC population.

TEs harbor DNA recognition sequences to recruit transcription regulators and guide cell differentiation [31,38–41]. We used the ReMap atlas of 485 transcription regulator cistromes (see Methods)

to examine enrichment of these sequences over the five TE subfamilies that were specific to LT-HSCs. We found enrichment of HOXA9, a well-known HSC factor [42], at elements of the MER57B1 (q=1.49e-13, OR=29.21) and MER57B2 (q=0.007 , OR=10.09) (Fig. 1e) subfamilies. We further detected enrichment of the HOXA9 DNA recognition sequence across MER57B1 and MER57B2 elements (Fig. 1f). The cistromes for ERG and RUNX1, two known HSCs transcription regulators [43], were found to be enriched over elements of the LTR67B subfamily (ERG: q=9.31e-12, OR=4.06; RUNX1: q=6.71e-12, OR=4.17) (Fig. 1e). In parallel, ERG and RUNX1 DNA recognition sequences were enriched over LTR67 elements (Fig. 1f). None of the 485 cistromes in ReMap were enriched over MamGypLTR1b elements. However, the cistromes of cohesin complex factors (SMC3 and RAD21), known regulators of genome topology [44], were enriched over MamGypsy2-LTR elements (Fig. 1e). We also observed enrichment for the DNA recognition sequence of CTCF, known to colocalize with cohesin [45] and to serve as a gatekeeper to self-renewal in LT-HSCs [16], over MamGypsy2-LTR elements (Fig. 1f). Collectively, these results suggest that unique TE accessibility in LT-HSCs provides docking sites for known “stem” transcription regulators, including HOXA9, ERG, RUNX1 and components of the three-dimensional genome organization (CTCF, cohesin).

LSCs exhibit primitive-like accessible TE enrichment

Considering the role of LSC in AML etiology, we next examined whether accessible TE could be used to discriminate populations enriched for self-renewing LSCs from leukemic cell fractions that lacked LSC as well as from normal primitive versus mature hematopoietic populations. Peripheral blood samples from 15 AML patients were sorted into fractions based on CD34 and CD38 expression, and fractions were classified as LSC+ (n=11) or LSC- (n=24) based on their ability to generate a leukemic graft upon xenotransplantation into immunodeficient mice. In parallel, we performed ATAC-Seq and measured TE enrichment within the accessible chromatin profiles of each LSC fraction. 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 primitive hematopoietic 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 enrichment profile classification (as primitive) 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 be due to low LSC frequency within these populations (measured by limiting dilution assays) leading to functional misclassification.

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, Suppl. Fig. 3a, Suppl. Table 7). LSC+ fractions were mostly enriched for HERV1, HERV3 and LINE1 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 primitive vs mature and LSC+ vs LSC- fractions were skewed compared to the proportions of TEs distributed across the human genome (Fig. 2d). Specifically, accessible TEs enriched in primitive populations and LSC+ fractions consisted preferentially of HERV1 and HERV3 families (Fig. 2d). In contrast, accessible TEs enriched in mature populations and LSC- fractions had higher representation of SINE1 and Transposons families (Fig. 2d). Extending our analysis to subfamilies showed that the majority of TE subfamilies enriched in LSC+ or LSC- accessible chromatin were shared with primitive/LSC+ (23/44) or mature/LSC- populations (55/77), respectively (Fig. 2e). In line with the similarities observed in the overall TE family distribution, the majority of common subfamilies between LSC+ fractions and primitive populations (16/23 subfamilies) belonged to the HERV1 and HERV3 families (Fig. 2e, Suppl. Fig. 3b). In contrast, the majority of common subfamilies between LSC- fractions and mature populations belonged to the SINE1 and Transposons families (Fig. 2e, Suppl. Fig. 3b). Z-score comparisons of the enrichment in the common subfamilies between primitive/LSC+ or mature/LSC- showed largely concordant enrichment (Suppl. Fig. 3b). Two TE subfamilies enriched in LSC+, but not primitive accessible chromatin, namely LTR12C (HERV1 family) and L1PA7 (LINE1 family), showed the largest discordance between primitive and LSC+ populations (Suppl. Fig. 3c). Although not discriminating LSC+ from LSC- fractions, all LINE1 subfamilies enriched in primitive versus

mature hematopoietic populations were more strongly enriched in LSC+ compared to primitive hematopoietic populations (Suppl. Fig. 3c); this is in line with the DNA hypomethylation over LINE1 elements reported across many cancers [46]. Collectively, our data suggest that LSCs can be discriminated from committed leukemic cells based on the enrichment of specific TEs within their accessible chromatin that are largely shared with normal primitive hematopoietic populations.

Accessible TEs are docking sites of LSC 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 (23, 70 and 21 respectively) could provide docking sites for different transcription regulators (shown in Fig. 2e; Suppl. Tables 8-10; see Methods). Cistromes for five transcription regulators were found to be enriched across accessible TE subfamilies in LSC and/or primitive populations (Fig. 3a,b and Suppl. Fig. 4a,b). These included ERG, RUNX1, LMO2, TAL1 and TRIM28. Each of these transcription regulators are documented to contribute to normal hematopoiesis and/or leukemia development [43,47–50]. Beyond its reported role in erythroblast differentiation [48], TRIM28 is also known to suppress TEs [51], suggesting that it plays a role in the transition from primitive and LSC+ to mature fates. In agreement with previous studies reporting an interplay between ERG and RUNX1 in hematopoiesis and AML [52], the ERG and RUNX1 cistromes were enriched over more than 70% of the same accessible TE subfamilies (13/19 for primitive-specific, 8/10 for LSC+ specific and 6/7 for TEs shared between primitive and LSC+ Suppl. Fig. 4c), including LTR78 (ERG: q=0.005, OR=1.74; RUNX1: q=0.0003, OR=1.9), LTR67B (ERG: q=1.09e-08 , OR=2.74 ; RUNX1: q=5.38e-17 , OR=4.09) and MLT1E2 (ERG: q=0.0003 , OR=2.03 ; RUNX1: q=0.009 , OR=1.73 and Suppl. Fig. 4d, Suppl. Table 8-10). Over a third of the TE subfamilies enriched for ERG and/or RUNX1 cistromes also preferentially encoded their DNA recognition sequences (11/29 for ERG motif and 11/34 for RUNX1 motif; Suppl. Fig. 4e,f). Collectively, these results reflect the shared biology across primitive hematopoietic populations and LSC, relying on five core transcription regulators, namely ERG, RUNX1, LMO2, TAL1 and TRIM28, using accessible TEs as direct docking sites on the genome.

Cistrome enrichment analysis over accessible TEs identified four transcription regulators specific to primitive HSC populations: CHD8, CTCF, SMC1A and XBP1 (Suppl Fig. 4a). While XBP1 is known to bind enhancers and promoters to regulate gene expression [53], CTCF and SMC1A instead bind to anchors of chromatin interactions to regulate the three-dimensional genome organization [45]. CHD8 is a known interactor of CTCF [54], whose cistrome enriches over LTR41 elements with CTCF and SMC1A cistromes (q≤4.81e-29, OR≥6.32) (Suppl. Fig. 4d, Suppl. Table 8). At LTR41 elements, we further identified enrichment of the cistrome of ZNF143, another regulator of chromatin interactions [55,56] (q=1.02e-55, OR=12.97) (Suppl. Fig. 4d, Suppl. Table 8). The accessibility of TE subfamilies serving as docking sites for CTCF and other elements regulating genome topology in primitive populations parallels the enrichment of the CTCF DNA recognition sequences at some of these TEs (Suppl. Fig. 4g), and is in keeping with the role of CTCF as a gatekeeper of stemness in hematopoietic populations [16]. Together, these results suggest that specific TE subfamilies are accessible in primitive populations, serving as genomic docking sites for regulators of genome topology.

Finally, we found seven transcription regulators whose cistromes were commonly enriched over accessible TEs unique to LSC, namely FLI1, GABPA, LYL1, MEIS1, NFYA, RUNX1T1 and TCF3 (Fig. 3a,b, Suppl. Table 9). In parallel, the DNA recognition sequences for FLI1 and NFYA were also enriched in accessible TE subfamilies (Fig. 3c). Among these factors, FLI1 is aberrantly expressed in AML and operates in concert with ERG to demarcate targets for aberrant epigenetic regulation [57,58], LYL1 has oncogenic potential in AML[59], MEIS1 assists in AML maintenance [60] and RUNX1T1 fusion with RUNX1 drives one of the most common subtypes of AML [61]. Our results suggest that additional transcription regulators, namely GABPA, NFYA and TCF3, contribute to the stemness properties typical of LSCs in AML. Hence, we next assessed the contribution of these transcription regulators to growth in AML cell lines. Analyzing the CRISPR-based essentiality screen results from the Cancer Dependency Map (DepMap) project [62] revealed negative essentiality scores for all seven transcription regulators across the majority (14/23) of AML cell lines (Suppl. Fig. 5a). FLI1, GABPA, LYL1 and NFYA median essentiality score across all 23 AML cell lines was significantly different from the median essentiality scores

distribution for all other genes (p-value<0.05; Suppl. Fig. 5a). The essentiality for FLI1 in AML cell lines was further substantiated by the DepMap RNAi-based screen (p-value<0.05; Suppl. Fig. 5b). In comparison to other cancer types, FLI1 and LYL1 were significantly more essential in AML cell lines than cells from other cancer types (CRISPR-based screen: 94%, 33/35 for FLI1 and 97%, 34/35 for LYL1); RNAi-based screen: 75%, 16/21 and 17/21 for FLI1 and LYL1, respectively)(Fig. 3d,e, Suppl. Tables 11,12). Collectively, our results suggest that the accessibility of TE subfamilies specific to LSC define docking sites for transcription regulators essential in AML.

Accessible TEs as a marker of AML relapse

Stemness is related to survival and relapse in AML and other cancer types [63,64]. We hypothesized that TEs enriched in LSC+ and LSC- fractions (44 and 77 subfamilies 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 that may be associated with clinical outcomes. To this end, we performed ATAC-seq on 29 peripheral blood samples from AML patients (Suppl. Table 13) and determined the enrichment over all TE subfamilies in accessible chromatin in comparison to the LSC+ and LSC- fractions. Using the LSCTE121 for unsupervised clustering, we observed that 9 of the bulk samples clustered with the LSC- fractions whereas the other 20 samples clustered in the same main branch as the LSC+ fractions (Fig. 4a). This classification of bulk AML samples into LSC-like and non-LSC-like groups is reminiscent of recent gene expression-based patient stratification, lending credence to our approach [65,66].

To simplify the interpretation of TE 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- TEs (see Methods). Comparison of the combined LSCTE121 Z-scores of samples in the three major branches of the heatmap (Fig. 4a) showed the highest scores in the branch containing all but one of the LSC+ fractions (red branch). LSC+ fractions had higher LSCTE121 Z-scores than LSC- fractions (Suppl. Fig. 6a). Further, there was no correlation between the blast count of a given sample and the LSCTE121 Z-score (R=0.031, p=0.88; Suppl. Fig. 6b), suggesting that the combined LSCTE121 Z-score stratification is a reflection of stemness and not the blast count.

The LSC17 score can predict overall patient survival and response to standard induction therapy [6], therefore we sought to determine the similarity between the LSCTE121 Z-score and the LSC17 score. We calculated LSC17 scores for the 29 AML patient samples (shown in Fig. 2) as previously reported [6] and validated the prediction of overall survival in our patient cohort (Suppl. Fig. 6c, log rank p-value=2e-04). Restriction of unsupervised clustering to the AML samples according to the LSCTE121 Z-scores did not reflect the LSC17 score classification (high vs low for scores above or below the median of a reference cohort, respectively, Fig. 4b). Furthermore, correlation analysis did not show a strong link between the LSC17 score and the LSCTE121 Z-score (R=0.27,p=0.16; Suppl. Fig. 6d). Kaplan-Meier estimates comparing the top 25th percentile with the bottom 25th percentile of LSCTE121 Z-scores showed significant differences in overall survival (log rank p-value=0.00691) (Fig. 4c) and disease-free survival (log rank p-value=0.041), with both groups containing samples with high and low LSC17 scores (Fig. 4d). Collectively, our results suggest that the LSCTE121 Z-score captures a different underlying biology than the LSC17 score, and that the 2 scores can provide risk stratification based on complementary discriminators of stemness.

Discussion

Our study establishes that enrichment of specific TE subfamilies within the accessible chromatin serves as a hallmark of stemness in both normal hematopoiesis and AML, with direct clinical implications for AML patients. Many of the accessible TEs specific to both LSC and normal primitive hematopoietic populations are shared with those found accessible in embryonic stem versus mature cells [67]. This includes members of the human endogenous retrovirus (HERV) subfamilies known to act as docking sites for stem transcription regulators, such as OCT4 and NANOG [38,39]. Expanding on the role of TEs are a source of *cis*-regulatory elements in stem and mature cells [30,38–41,67], our results show that accessible TEs in primitive populations are permissive for binding by regulators of the three-dimensional genome organization, including CTCF and cohesin elements. CTCF-driven changes to the genome topology in primitive hematopoietic populations was recently identified as a regulator of quiescence exit and long-term self-renewal [16]. Here, additional transcription regulators, namely ERG, RUNX1, LMO2, TAL1 and TRIM28, were found based on the presence of their DNA recognition sequences in the accessible TEs in LSC and primitive hematopoietic populations. These correspond to known hematopoietic transcription regulators and reinforce their direct role in AML [43,47–51]. Collectively, our results suggest a direct role for specific TE subfamilies as genetic determinants, shaping both genome topology and transcriptional regulation that occurs in tissue hierarchies when stem cells differentiate into mature populations in both normal and disease settings.

In cancer, chromatin accessibility changes at TEs [68] and can lead to oncogene activation [32]. However, the role of TEs in cancer has primarily focused on their expression. For instance, high levels of HERV1 expression are linked to worse survival in patients with kidney cancer [69] and a 14 TE-based signature predicts survival in AML [70]. Our work expands the TE expression-based perspective to a role of accessible TEs as cis-regulatory elements. Among the TE subfamilies accessible in LSCs, for example, we observed the LTR12C subfamily, whose elements work as enhancers in AML [34] and were previously shown to act as alternative promoters in hepatocellular carcinoma [71]. We further show that accessible TEs provide docking sites for oncogenic transcription regulators in LSCs. Using this

approach, we identified FLI1, LYL1 and MEIS1 as oncogenic drivers of AML, in agreement with previous reports [57,59,60]. We further showed that FLI1 and LYL1 are preferentially essential across AML cell lines as compared to cell lines from other cancer types. Altogether, our work suggests a direct contribution of the repetitive genome towards the governance of oncogenic transcription regulators in AML.

By revealing how differentially-accessible TE subfamilies distinguish LSCs from mature leukemic cells, we provide a chromatin-based approach to 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 [6] scoring scheme argues for their complementarity in assessing the LSC content in primary patient samples. It also supports a deeper study of the repetitive genome and its contribution towards stemness properties in cancer. 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 Suppl. Table 13.

Cell sorting and Xenotransplantation assay

Hematopoietic populations were selected using the highest purity sorting scheme available [36,37]. Peripheral blood AML samples were sorted into fractions based on 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 [6]. Antibodies used are shown in Suppl. Table 14.

Bulk ATAC-seq

ATAC-seq was used to profile the accessible chromatin landscape of 11 LSC+, 24 LSC- fractions and 29 (12 LSC17 high, 17 LSC17 low) peripheral blood AML bulk samples. 50,000 cells from each sample were processed as previously described [72]. Libraries were sequenced with 50bp single-end (LSC fractions) or 50 bp 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 used LSC fractions from patients without NPM1 mutation and set a threshold of 10 million reads as well as 15k called peaks. ATAC-seq data of LSC fractions and bulk AML samples can be accessed at EGAS00001004893 and EGAS00001004896, respectively.

TE subfamilies and families

Bedfiles of repeats classified as TE subfamilies (excluding hence simple repeats, satellites, sn/rRNA and repeats of unknown subfamilies) (n=971) in the hg38 human genome build were downloaded from UCSC Genome Browser. Subfamilies were classified into families according to Repbase (https://www.girinst.org/repbase/) [73,74].

TE enrichment analysis

TE enrichment analysis was performed using ChromVAR [75]. The binary matrix for ChromVAR was generated by identifying 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 overlap of TEs for each subfamily in the accessible chromatin peakset. 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:

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

Transcription regulator cistromes and motifs enrichment analysis

The enrichment of cistromes and motifs was assessed at the TE subfamilies enriched specifically in LT-HSCs, primitive populations, LSC positive and commonly enriched in primitive populations and LSC positive (5, 70, 21 and 23 TE subfamilies respectively; Suppl. 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 regulator

cistromes and motifs enriched at the elements overlapping with the ATAC peaks catalogue of primitive populations). Enrichment of transcription regulator cistromes was performed using the Bioconductor package LOLA (https://bioconductor.org/packages/release/bioc/html/LOLA.html) [76], version 1.20.0. 485 transcription regulator cistromes were obtained from ReMap (http://remap.univ-amu.fr/) [77]. Enriched transcription regulator 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 2.5% most frequently enriched are shown in Fig. 3a, Suppl. Fig. 4a and d. 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 $q < 0.05$).

Essentiality scores

Essentiality scores were obtained from the Broad Institute project Achilles (DepMap). 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 in cancer types following the sample information available on the DepMap website. For RNAi data, our analysis was restricted to the AML cell lines analysed by the project Achilles, namely HL60, MOLM13, MOLM16, MV411, NB4, OCIAML2, OCIAML3, OCIAML5, THP1 and U937.

LSC17 score calculation

The LSC17 score was calculated as previously reported [6]. In short, the 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. The LSC17 scores for AML bulk samples were obtained using the clinical LSC17 assay.

Survival analysis

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.

Authors Contributions

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

Acknowledgement

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.

References:

1. Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).
2. Shlush, L. I. *et al.* Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547**, 104–108 (2017).
3. Eppert, K. *et al.* Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.* **17**, 1086–1093 (2011).
4. Hermann, P. C. *et al.* Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. *Cell Stem Cell* **1**, 313–323 (2007).
5. Balic, M. *et al.* Most early disseminated cancer cells detected in bone marrow of breast cancer patients have a putative breast cancer stem cell phenotype. *Clin. Cancer Res.* **12**, 5615–5621 (2006).
6. Ng, S. W. K. *et al.* A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* **540**, 433–437 (2016).
7. Malta, T. M. *et al.* Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173**, 338–354.e15 (2018).
8. Yassin, M. *et al.* A novel method for detecting the cellular stemness state in normal and leukemic human hematopoietic cells can predict disease outcome and drug sensitivity. *Leukemia* (2019) doi:10.1038/s41375-019-0386-z.
9. Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888–903 (2013).
10. Takayama, N. *et al.* The Transition from Quiescent to Activated States in Human Hematopoietic Stem Cells is Governed by Dynamic 3D Genome Reorganization. *Available at SSRN: https://ssrn.com/abstract=3611286 or http://dx.doi.org/10.2139/ssrn.3611286*.
11. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–970 (2008).
12. Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term

multilineage engraftment. *Science* **333**, 218–221 (2011).

13. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
14. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
15. Deblois, G., Tonekaboni, S. A. M., Grillo, G. & Martinez, C. Epigenetic switch–induced viral mimicry evasion in chemotherapy-resistant breast cancer. *Cancer Discov.* (2020).
16. 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.
17. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
18. 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).
19. 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).
20. Dick, J. E. Acute Myeloid Leukemia Stem Cells. *Annals of the New York Academy of Sciences* vol. 1044 1–5 (2005).
21. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature Publishing Group* **506**, 328–333 (2014).
22. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
23. 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).
24. Sotero-Caio, C. G., Platt, R. N., 2nd, Suh, A. & Ray, D. A. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol.* **9**, 161–177 (2017).
25. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol.* **19**,

199 (2018).

26. Britten, R. J. & Kohne, D. E. Repeated Sequences in DNA. *Science* **161**, 529–540 (1968).
27. Roulois, D. *et al.* DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* **162**, 961–973 (2015).
28. Chiappinelli, K. B. *et al.* Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* **164**, 1073 (2016).
29. Colombo, A. R. *et al.* Suppression of Transposable Elements in Leukemic Stem Cells. *Sci. Rep.* **7**, 7029 (2017).
30. Barakat, T. S. *et al.* Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**, 276–288.e8 (2018).
31. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
32. Jang, H. S. *et al.* Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 611–617 (2019).
33. Kruse, K. *et al.* Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. *Cold Spring Harbor Laboratory* 523712 (2019) doi:10.1101/523712.
34. Deniz, Ö. *et al.* Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nat. Commun.* **11**, 3506 (2020).
35. Zeng, Y. *et al.* Characterization of functional transposable element enhancers in acute myeloid leukemia. *Sci. China Life Sci.* **63**, 675–687 (2020).
36. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).
37. Laurenti, E. *et al.* CDK6 levels regulate quiescence exit in human hematopoietic stem cells. *Cell Stem Cell* **16**, 302–313 (2015).
38. Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
39. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human

embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).

40. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).
41. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
42. 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).
43. Ichikawa, M. *et al.* A role for RUNX1 in hematopoiesis and myeloid leukemia. *Int. J. Hematol.* **97**, 726–734 (2013).
44. Cheng, H., Zhang, N. & Pati, D. Cohesin subunit RAD21: From biology to disease. *Gene* **758**, 144966 (2020).
45. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110–1121 (2016).
46. 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).
47. Knudsen, K. J. *et al.* ERG promotes the maintenance of hematopoietic stem cells by restricting their differentiation. *Genes Dev.* **29**, 1915–1929 (2015).
48. 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).
49. Morishima, T. *et al.* LMO2 activation by deacetylation is indispensable for hematopoiesis and T-ALL leukemogenesis. *Blood* **134**, 1159–1175 (2019).
50. Vagapova, E. R., Spirin, P. V., Lebedev, T. D. & Prassolov, V. S. The Role of TAL1 in Hematopoiesis and Leukemogenesis. *Acta Naturae* **10**, 15–23 (2018).
51. Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237–240 (2010).
52. 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).

53. Chen, X. *et al.* XBP1 promotes triple-negative breast cancer by controlling the HIF1α pathway. *Nature* **508**, 103–107 (2015).
54. Ishihara, K., Oshimura, M. & Nakao, M. CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol. Cell* **23**, 733–742 (2006).
55. Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **2**, 6186 (2015).
56. Heidari, N. *et al.* Genome-wide map of regulatory interactions in the human genome. *Genome Res.* **24**, 1905–1917 (2014).
57. Kornblau, S. M. *et al.* Abnormal expression of FLI1 protein is an adverse prognostic factor in acute myeloid leukemia. *Blood* **118**, 5604–5612 (2011).
58. Martens, J. H. A. *et al.* ERG and FLI1 binding sites demarcate targets for aberrant epigenetic regulation by AML1-ETO in acute myeloid leukemia. *Blood* **120**, 4038–4048 (2012).
59. 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).
60. Liu, J. *et al.* Meis1 is critical to the maintenance of human acute myeloid leukemia cells independent of MLL rearrangements. *Ann. Hematol.* **96**, 567–574 (2017).
61. Al-Harbi, S., Aljurf, M., Mohty, M., Almohareb, F. & Ahmed, S. O. A. An update on the molecular pathogenesis and potential therapeutic targeting of AML with t(8;21)(q22;q22.1);RUNX1-RUNX1T1. *Blood Adv* **4**, 229–238 (2020).
62. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–570.e16 (2017).
63. 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).
64. 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).
65. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265–1281.e24 (2019).

66. 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).

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

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

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

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

71. 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).

72. 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).

73. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).

74. Jurka, J. Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.* **8**, 333–337 (1998).

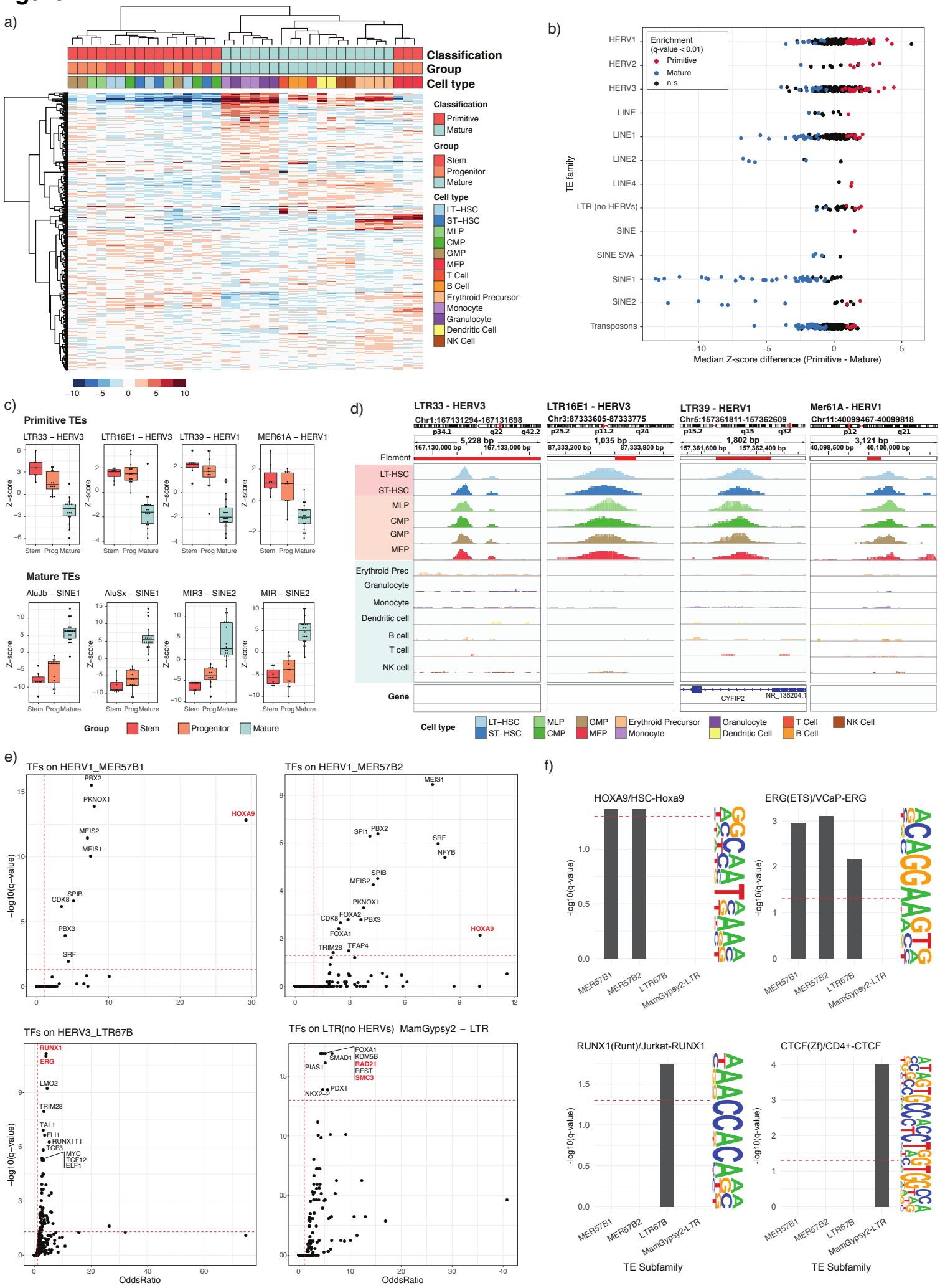
75. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

76. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).

77. 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).

Figures

Figure 1

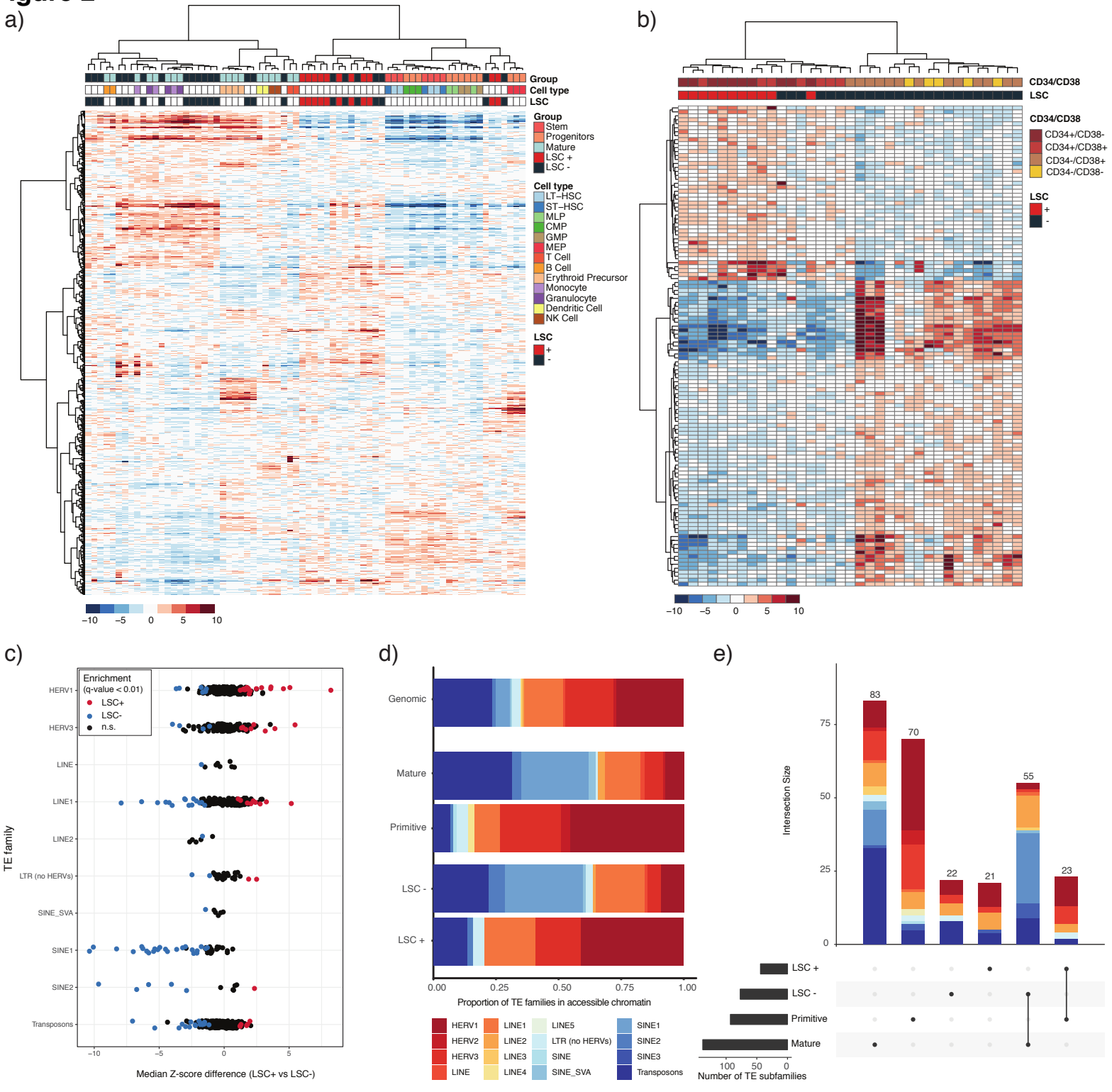


Distinct TE subfamilies populate the chromatin accessibility landscape of primitive and mature hematopoietic cells

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 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, points denote TE subfamilies. SINE1, SINE2 and Transposons families are enriched in mature, whereas HERV1 and HERV3 families are enriched in "primitive" cells.*
c) *Examples of primitive and mature enriched TE subfamilies of the SINE1, SINE2, HERV1 and HERV3 TE families. Bar plots 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 regulator cistromes enriched at 4 of the 5 TE subfamilies enriched in LT-HSCs. Every dot corresponds to a transcription regulator cistrome taken into account in the analysis. Red dashed lines correspond to -log10(q)=1.3 (q=0.05) and OddsRatio(OR)=1 thresholds.*
f) *Enrichment of HOXA9, ERG, RUNX1 and CTCF DNA recognition sequences within TE subfamilies enriched in LT-HSCs. Bars represent -log10(q) for each TE subfamily for transcription regulator 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).

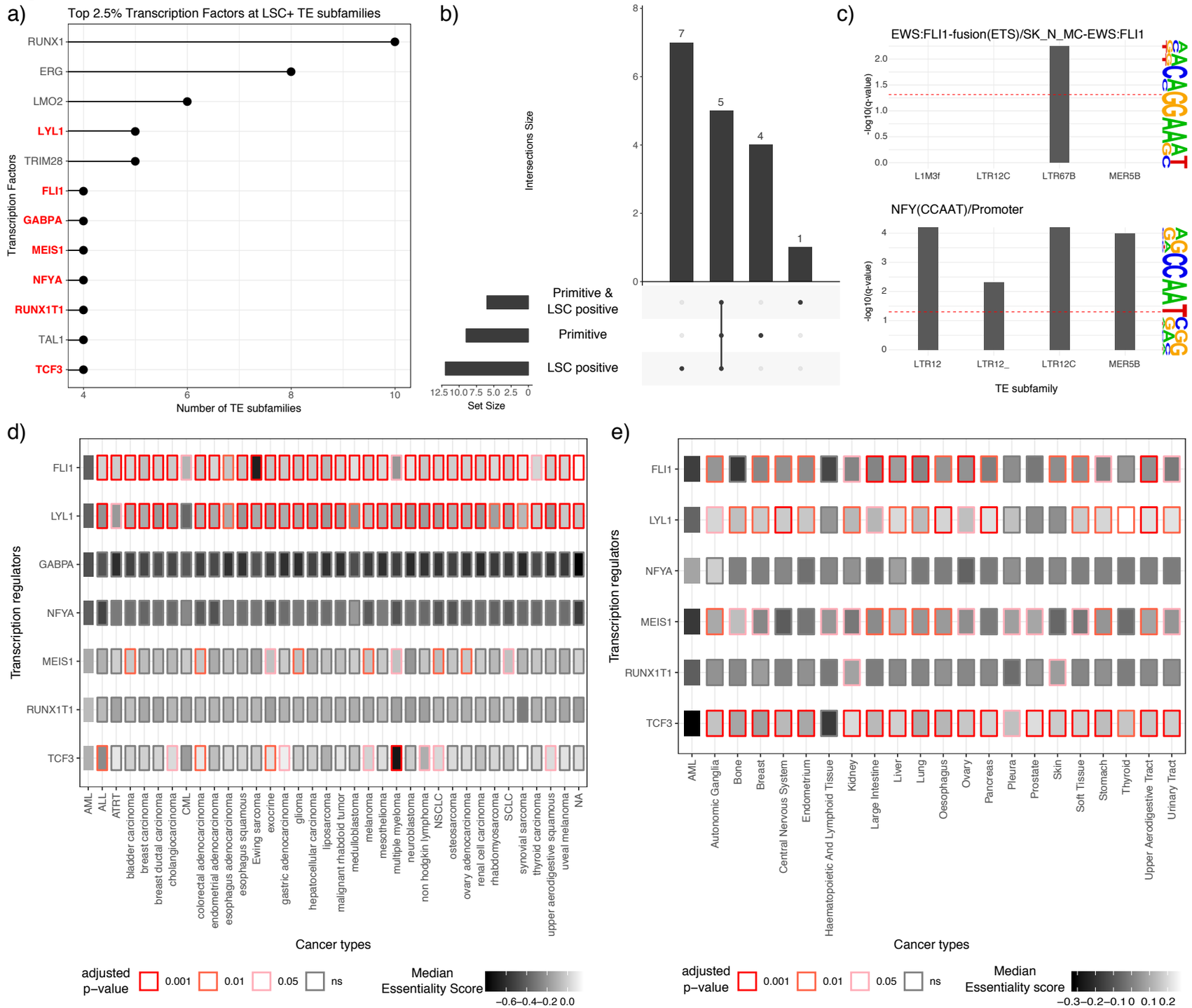
Figure 2


a)
Group
Cell type
LSC
Group
Stem
Progenitors
Mature
LSC +
LSC -
Cell type
LT–HSC
ST–HSC
MLP
CMP
GMP
MEP
T Cell
B Cell
Erythroid Precursor
Monocyte
Granulocyte
Dendritic Cell
NK Cell
LSC
+
-
−10 −5 0 5 10
b)
CD34/CD38
LSC
CD34/CD38
CD34+/CD38-
CD34+/CD38+
CD34-/CD38+
CD34-/CD38-
LSC
+
-
−10 −5 0 5 10
c)
Enrichment
(q-value < 0.01)
LSC+
LSC-
n.s.
HERV1
HERV3
LINE
LINE1
LINE2
LTR (no HERVs)
SINE_SVA
SINE1
SINE2
Transposons
TE family
−10 −5 0 5
Median Z-score difference (LSC+ vs LSC-)
d)
Genomic
Mature
Primitive
LSC -
LSC +
0.00 0.25 0.50 0.75 1.00
Proportion of TE families in accessible chromatin
HERV1
HERV2
HERV3
LINE
LINE1
LINE2
LINE3
LINE4
LINE5
LTR (no HERVs)
SINE
SINE_SVA
SINE1
SINE2
SINE3
Transposons
e)
Intersection Size
0 25 50 75
83
70
22
21
55
23
LSC +
LSC -
Primitive
Mature
100 50 0
Number of TE subfamilies

Leukemia stem cell containing fractions cluster with primitive hematopoietic cells, and show distinct and common subfamily enrichment.

a) *Heatmap of accessibility Z-scores of all TE subfamilies (rows) across bulk ATAC-seq profiles of HSCs, progenitors, mature hematopoietic populations, LSC+ and LSC- fractions clustered based on correlation. LSC+ fractions clustered with hematopoietic progenitors while LSC- fractions mostly clustered away 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. Y axis shows all TE families, points denote TE subfamilies. SINE1, SINE2 and Transposons families are enriched in LSC- samples, whereas HERV1 and HERV3 families are enriched in LSC+ samples.*
d) *Comparison of the proportion of TE families enriched in hematopoietic primitive and 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+/primitive hematopoietic populations (23 subfamilies) and LSC-/mature hematopoietic populations (55 subfamilies). Colors within bar graphs denote TE families as annotated in 2d.*

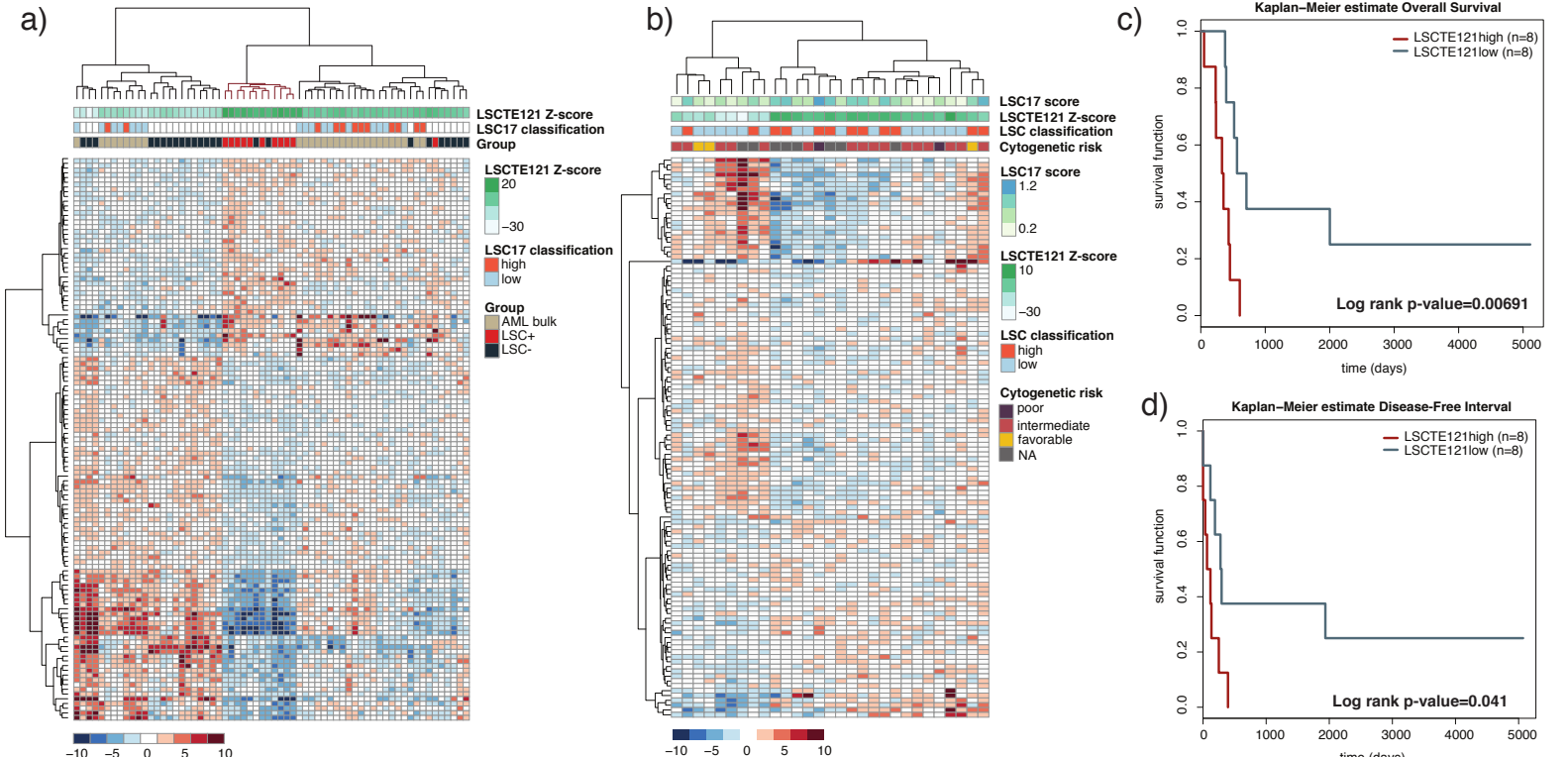
Figure 3


a)
Top 2.5% Transcription Factors at LSC+ TE subfamilies
Transcription Factors
RUNX1
ERG
LMO2
LYL1
TRIM28
FLI1
GABPA
MEIS1
NFYA
RUNX1T1
TAL1
TCF3
Number of TE subfamilies
b)
Intersections Size
Primitive & LSC positive
Primitive
LSC positive
Set Size
c)
EWS:FLI1-fusion(ETS)/SK_N_MC-EWS:FLI1
-log10(q-value)
L1M3f
LTR12C
LTR67B
MER5B
NFY(CCAAT)/Promoter
LTR12
LTR12_
TE subfamily
d)
Transcription regulators
FLI1
LYL1
GABPA
NFYA
MEIS1
RUNX1T1
TCF3
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
NA
Cancer types
adjusted p-value
0.001
0.01
0.05
ns
Median Essentiality Score
e)
Transcription regulators
FLI1
LYL1
NFYA
MEIS1
RUNX1T1
TCF3
AML
Autonomic Ganglia
Bone
Breast
Central Nervous System
Endometrium
Haematopoietic And Lymphoid Tissue
Kidney
Large Intestine
Liver
Lung
Oesophagus
Ovary
Pancreas
Pleura
Prostate
Skin
Soft Tissue
Stomach
Thyroid
Upper Aerodigestive Tract
Urinary Tract
Cancer types
adjusted p-value
0.001
0.01
0.05
ns
Median Essentiality score

TE subfamilies enriched in LSC+ provide binding sites to transcription regulators essential for AML

a) *Frequency plot of transcription regulator cistromes enriched at TE subfamilies enriched in LSC+ fractions. The top 2.5% most frequently enriched transcription regulator cistromes are shown. Note that transcription regulators specific to LSC+ fraction are in bold and red colored.*
b) *UpSet plot showing the intersection of transcription regulator cistromes enriched in LSC+/primitive hematopoietic populations.*
c) *Enrichment of FLI1 and NFY motifs within TE subfamilies enriched in LSC+. Bars represent -log10(q) for each TE subfamily for transcription regulator motifs. The red dashed line corresponds to -log10(q)=1.3 (q=0.05) threshold.*
d) *Overview of essentiality scores of LSC+ specific transcription regulators 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 regulator calculated in AML cell lines was compared to the distribution of the same transcriptional regulator calculated in each other cancer type cell lines. Rectangles inner color corresponds to median essentiality score, while border color corresponds to Pairwise t-test Benjamini-Hochberg corrected p-values (adjusted p-value).*
e) *As for d but based on the DepMap data from essentiality screen performed using RNAi.*

Figure 4

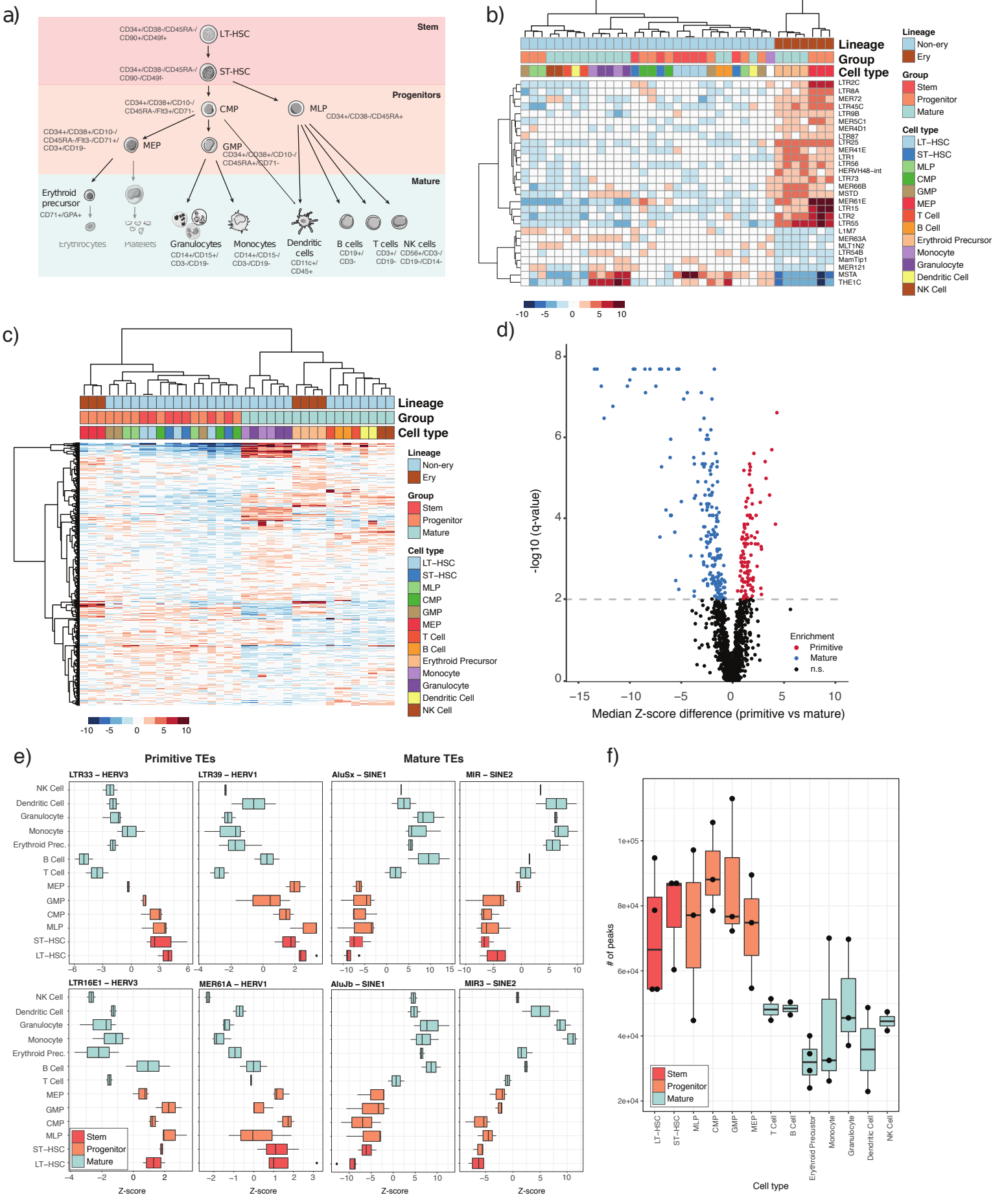

a)
LSCTE121 Z-score
LSC17 classification
Group
LSCTE121 Z-score
20
−30
LSC17 classification
high
low
Group
AML bulk
LSC+
LSC-
−10 −5 0 5 10
b)
LSC17 score
LSCTE121 Z-score
LSC classification
Cytogenetic risk
LSC17 score
1.2
0.2
LSCTE121 Z-score
10
−30
LSC classification
high
low
Cytogenetic risk
poor
intermediate
favorable
NA
−10 −5 0 5 10
c)
Kaplan–Meier estimate Overall Survival
LSCTE121high (n=8)
LSCTE121low (n=8)
survival function
time (days)
Log rank p-value=0.00691
d)
Kaplan–Meier estimate Disease-Free Interval
LSCTE121high (n=8)
LSCTE121low (n=8)
survival function
time (days)
Log rank p-value=0.041

TE enrichment over LSC TE subfamilies in accessible chromatin of AML bulk tumors stratifies patients overall survival and disease-free intervals

a) *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, note that bulk tumors did not cluster according to their LSC17 scores with LSC+ or LSC- fractions.*
b) *Heatmap showing differential accessibility Z-scores for LSC TEs in patients with low LSC17 scores, note that samples cluster into 3 distinct groups.*
c) *Kaplan–Meier estimates using Overall Survival as a clinical endpoint for LSCTE121 Z-score top vs bottom 25 percentile (n=8/group).*
d) *Kaplan–Meier estimates using Disease Free Interval as a clinical endpoint for for LSCTE121 Z-score top vs bottom 25 percentile (n=8/group).*

Supplementary Figures

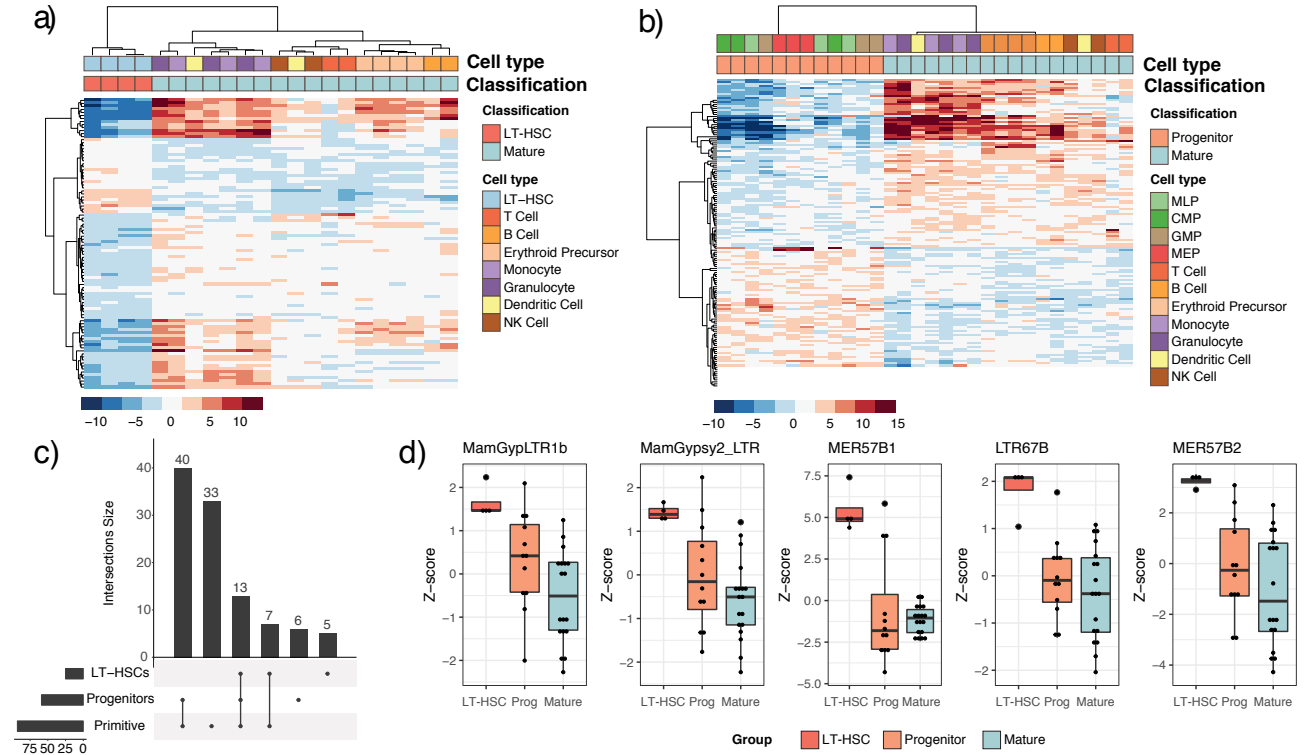
Supplementary Figure 1



Supplementary Figure 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 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 subfamilies of the SINE1, SINE2, HERV1 and HERV3 TE families. Barplots 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 numbers of called peaks than mature hematopoietic cells.*

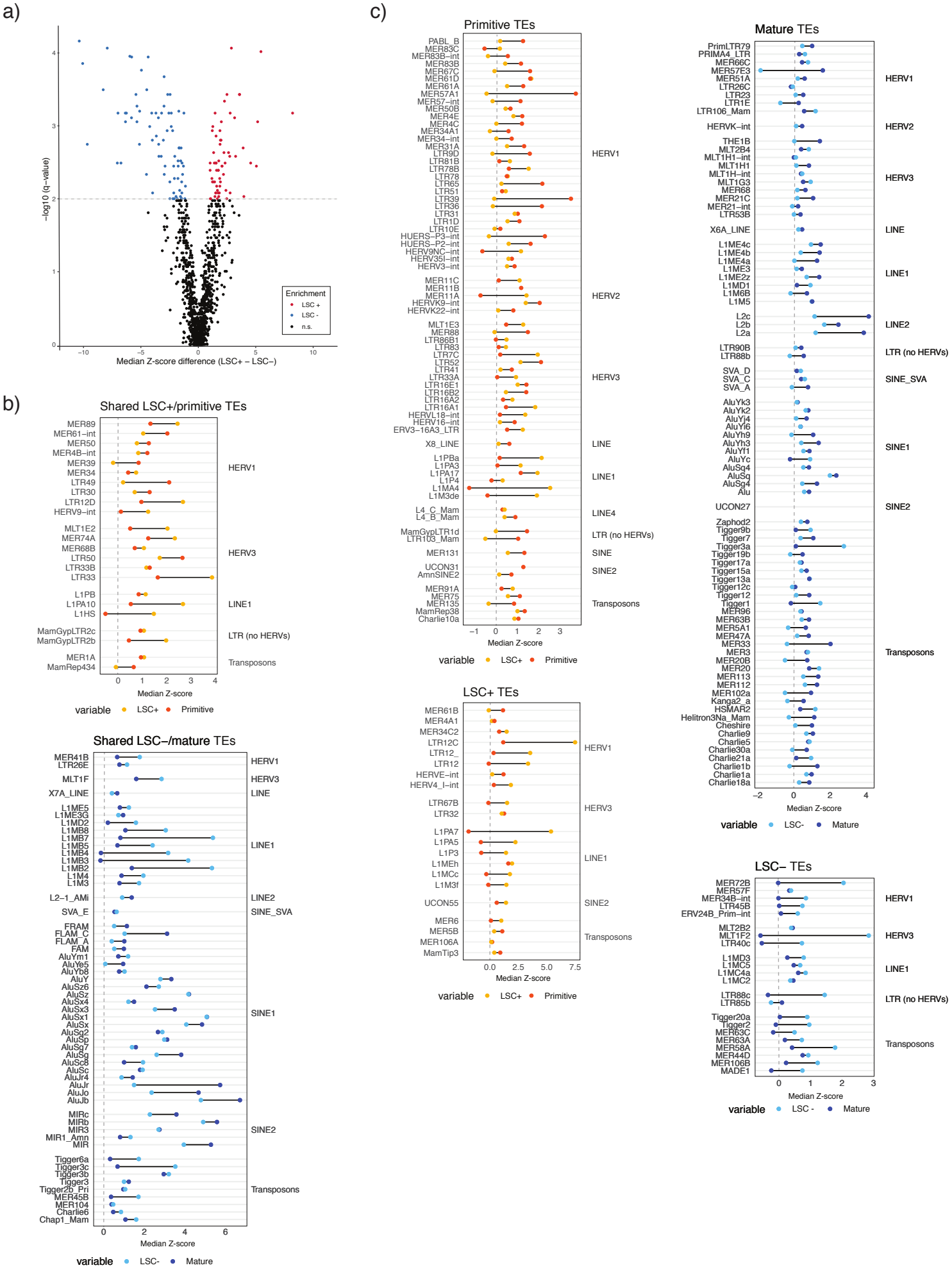
Supplementary Figure 2



Supplementary Figure 2:

a) *Heatmap of differential accessibility Z-scores of TE subfamilies with q < 0.01 across bulk ATAC-seq profiles comparing LT-HSC and all mature hematopoietic samples.*
b) *Heatmap of differential accessibility Z-scores of TE subfamilies with q < 0.01 across bulk ATAC-seq profiles comparing progenitor and all mature hematopoietic samples.*
c) *UpSet plot showing the intersection of TE subfamilies enriched in LT-HSC, progenitor and all primitive hematopoietic populations, note that 5 TE subfamilies are enriched exclusively in LT-HSCs.*
d) *5 TE subfamilies enriched exclusively in LT-HSCs. Bar plots show differential accessibility Z-scores in LT-HSCs, progenitor and mature cell type groups.*

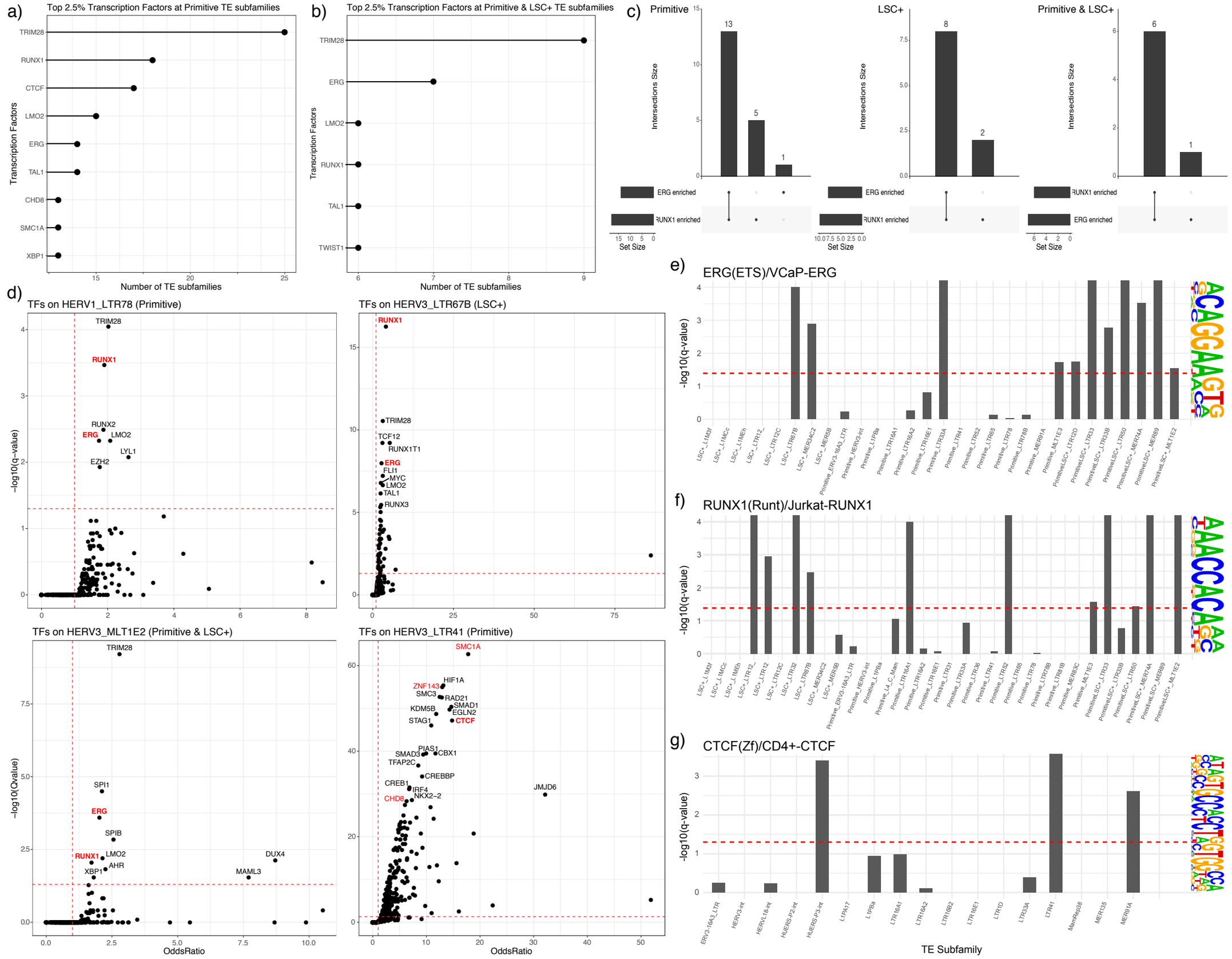
Supplementary Figure 3



Supplementary Figure 3:

a) *Volcano plot showing median difference in bias-corrected accessibility deviation Z-scores for each TE subfamily between the LSC+ and LSC- populations versus the −log10(q-value) for that difference.*
b) *TE subfamilies differentially enriched only in LSC+, LSC-, Stem/Progenitor and Differentiated populations of the hematopoietic system, as shown in Figure 2e. Cleveland dot plot shows the median z-scores for the relevant comparisons.*
c) *Shared TE subfamilies between LSC+, LSC-, Stem/Progenitor and Differentiated populations of the hematopoietic system as shown in Figure 5e, Cleveland dot plot shows the median z-scores for the relevant comparisons.*

Supplementary Figure 4

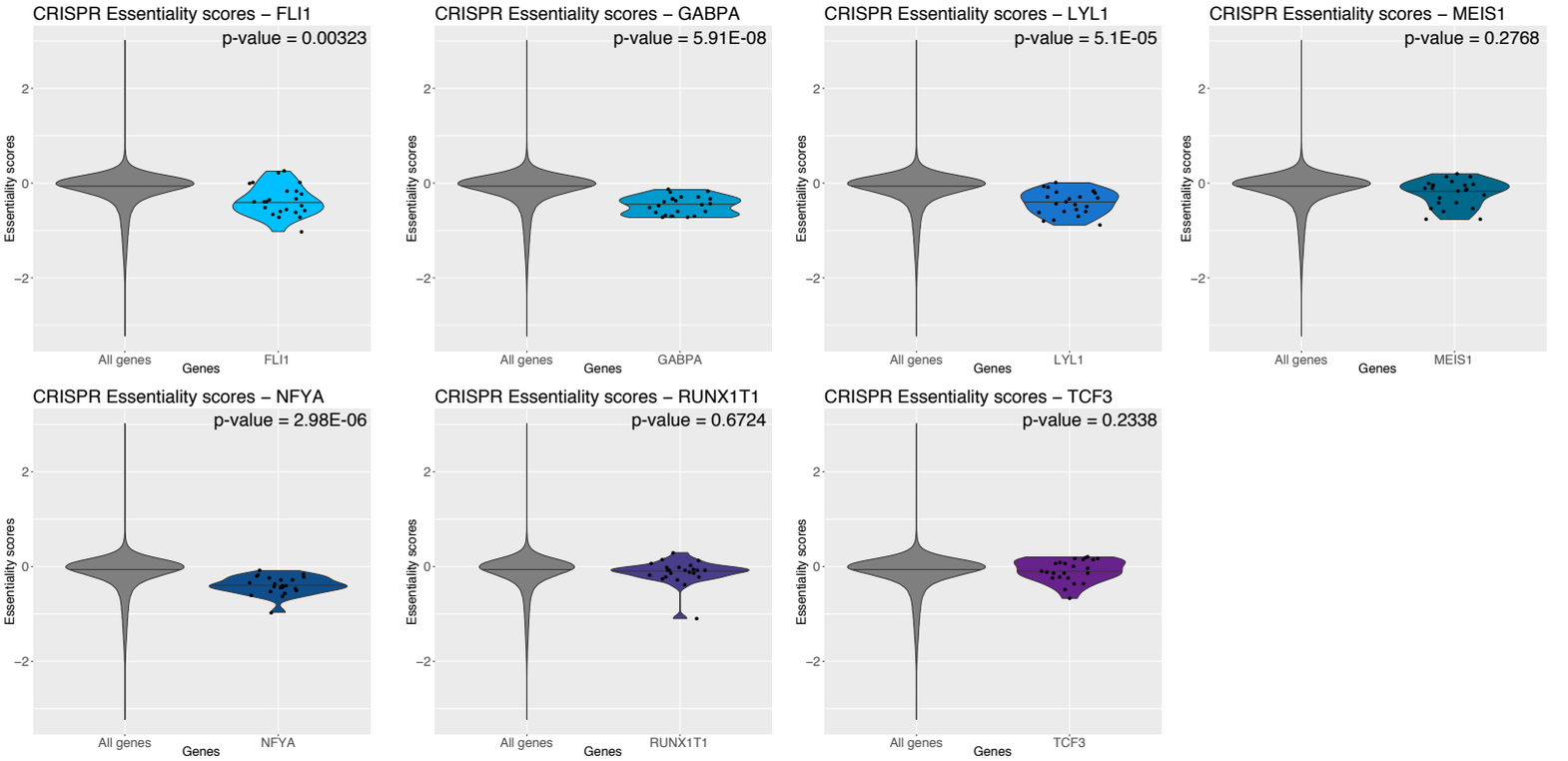


Supplementary Figure 4:

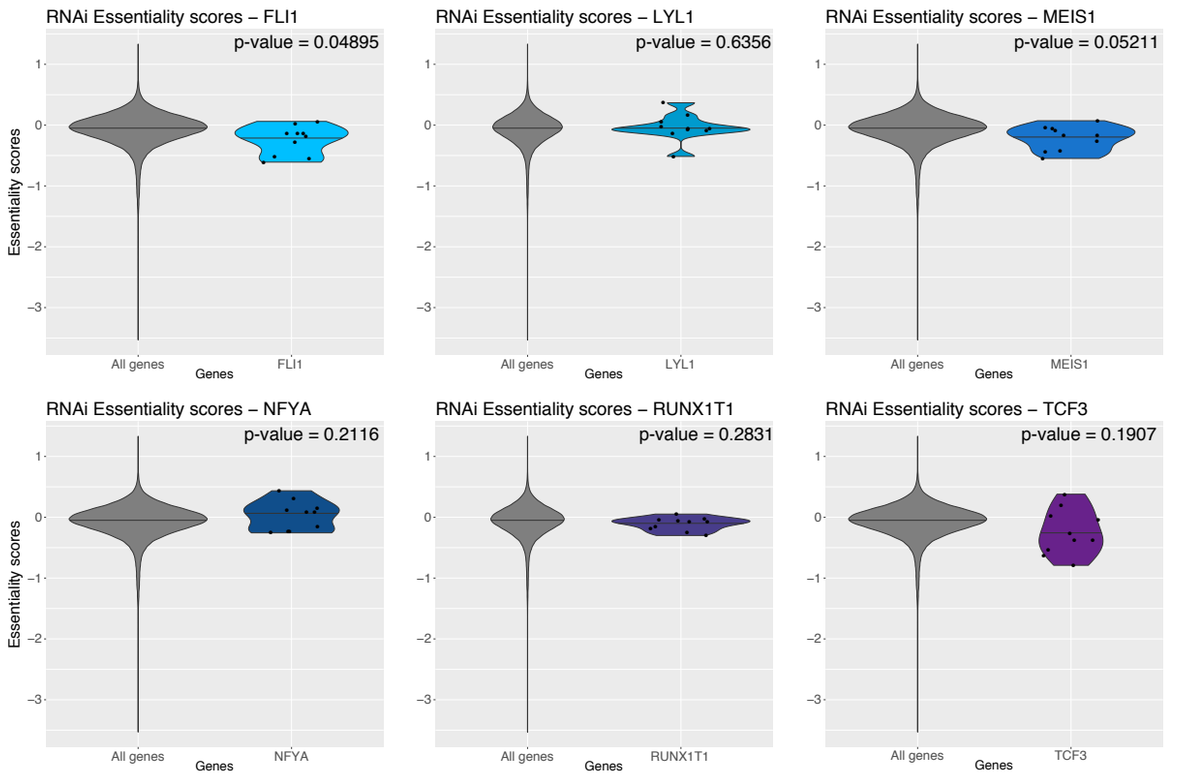
a) *Frequency plot of transcription regulator cistromes enriched at TE subfamilies enriched in primitive populations. The top 2.5% most frequently enriched transcription regulator cistromes are shown.*
b) *Frequency plot of transcription regulator cistromes enriched at TE subfamilies enriched in primitive and LSC+ fractions. The top 2.5% most frequently enriched transcription regulator cistromes are shown.*
c) *UpSet plot showing the intersection of TE subfamilies showing an enrichment for ERG and/or RUNX1 cistromes in primitive and/or LSC+. Note that most of 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 CTCF and other looping factor cistromes. Dots correspond to each transcription regulator cistrome taken into account in the analysis. Red dashed lines correspond to -log10(q)=1.3 (q=0.05) and OddsRatio(OR)=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) *As in e but for the RUNX1 DNA recognition sequences.*
g) *As in e but for the CTCF DNA recognition sequences.*

Supplementary Figure 5

a)



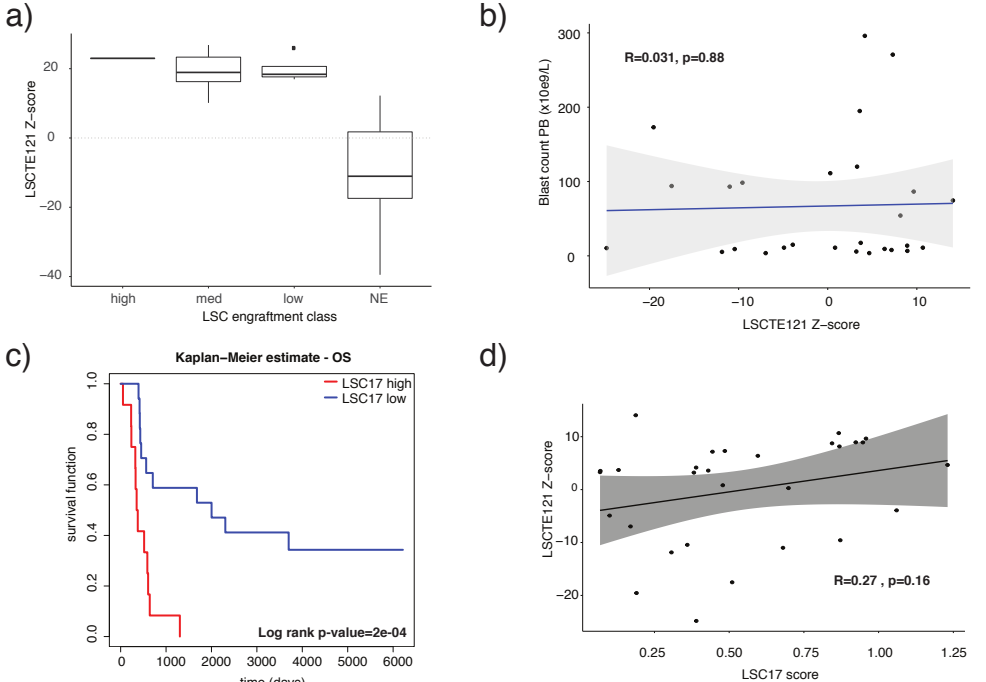
b)



Supplementary Figure 5:

a) *Violin plots for essentiality scores of LSC+ specific transcription regulators, based on CRISPR data. For each transcription regulator 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-test are reported.*
b) *Violin plots for essentiality scores of LSC+ specific transcription regulators, based on RNAi data. For each transcription regulator 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-test are reported.*

Supplementary Figure 6



Supplementary Figure 6:

a) *Boxplot of LSCTE121 Z-score in LSC fractions, note that engrafting fractions show higher LSCTE121 Z-score than non-engrafting fractions..*
b) *Pearson correlation between LSCTE121 Z-score and blast count in peripheral blood.*
c) *Kaplan-Meier estimates using Overall Survival as a clinical endpoint, patient division based on LSC17 high vs low scores; LSC17 high: n = 12, LSC17 low: n = 17.*
d) *Pearson correlation between LSC17 score and LSCTE121 Z-score.*

Supplementary Tables

Supplementary Table 1: Differential accessible TEs in primitive vs mature hematopoietic populations
Supplementary Table 2: Differential accessible TEs in LT-HSCs vs progenitor hematopoietic populations
Supplementary Table 3: Differential accessible TEs in ST-HSCs vs progenitor hematopoietic populations
Supplementary Table 4: Differential accessible TEs in LT-HSCs vs mature hematopoietic populations
Supplementary Table 5: Differential accessible TEs in progenitor vs mature hematopoietic populations
Supplementary Table 6: Differential accessible TE Zscores in hematopoietic signatures
Supplementary Table 7: Differential accessible TEs in LSC+ vs LSC- populations
Supplementary Table 8: Transcription regulator cistromes enriched at primitive populations TEs
Supplementary Table 9: Transcription regulator cistromes enriched at LSC+ TEs
Supplementary Table 10: Transcription regulator cistromes enriched at primitive and LSC+ TEs
Supplementary Table 11: CRISPR-based essentiality scores of LSC+ specific transcription regulators
Supplementary Table 12: RNAi-based essentiality scores of LSC+ specific transcription regulators
Supplementary Table 13: AML bulk patient annotation
Supplementary Table 14: Antibodies used for FACS sorting of CD34/CD38 fractions