Stem Cell Reports
Resource
ISSCR
OPEN ACCESS

High-Resolution Single-Cell DNA Methylation Measurements Reveal Epigenetically Distinct Hematopoietic Stem Cell Subpopulations

Tony Hui,[1,2] Qi Cao,[1] Joanna Wegrzyn-Woltosz,[2,3] Kieran O'Neill,[2,3,4] Colin A. Hammond,[4,5] David J.H.F. Knapp,[4,5] Emma Laks,[6] Michelle Moksa,[1] Samuel Aparicio,[3,6] Connie J. Eaves,[4,5,7] Aly Karsan,[2,3,4] and Martin Hirst[1,2,*]

[1]Department of Microbiology and Immunology and Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z3, Canada
[2]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 1L3, Canada
[3]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC V6T 2B5, Canada
[4]Terry Fox Laboratory, BC Cancer, Vancouver, BC V5Z 1L3, Canada
[5]Department of Medicine, University of British Columbia, Vancouver, BC V5Z 1M9, Canada
[6]Department of Molecular Oncology, BC Cancer, Vancouver, BC V5Z 1L3, Canada
[7]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6H 3N1, Canada
*Correspondence: mhirst@bcgsc.ca
https://doi.org/10.1016/j.stemcr.2018.07.003

SUMMARY

Increasing evidence of functional and transcriptional heterogeneity in phenotypically similar cells examined individually has prompted interest in obtaining parallel methylome data. We describe the development and application of such a protocol to index-sorted murine and human hematopoietic cells that are highly enriched in their content of functionally defined stem cells. Utilizing an optimized single-cell bisulfite sequencing protocol, we obtained quantitative DNA methylation measurements of up to 5.7 million CpGs in single hematopoietic cells. In parallel, we developed an analytical strategy (PDclust) to define single-cell DNA methylation states through pairwise comparisons of single-CpG methylation measurements. PDclust revealed that a single-cell epigenetic state can be described by a small (<1%) stochastically sampled fraction of CpGs and that these states are reflective of cell identity and state. Using relationships revealed by PDclust, we derive near complete methylomes for epigenetically distinct subpopulations of hematopoietic cells enriched for functional stem cell content.

INTRODUCTION

Hematopoietic stem cells (HSCs) are functionally defined cells that display evidence of extensive self-renewal of their ability to generate mature blood cells for the lifetime of the organism and following transplantation into myelosuppressed permissive hosts (Doulatov et al., 2012; Eaves, 2015). Clonal analyses of serially transplantable mouse HSCs have revealed that cells thus defined are composed of multiple distinct subpopulations that stably propagate specifically restricted abilities to produce different types of mature blood cell types (Benz et al., 2012; Dykstra et al., 2007; Kent et al., 2009; Sanjuan-Pla et al., 2013; Yamamoto et al., 2013).

Epigenetic modifications have been shown to be critical for the control of normal hematopoiesis as exemplified by the consequences of alterations incurred by disruption of *de novo* DNA methylation in primitive hematopoietic cells (Challen et al., 2012; Quivoron et al., 2011; Shlush et al., 2017). Moreover, in long-term HSC populations, lineage-specific enhancers appear to be epigenetically marked (Lara-Astiaso et al., 2014), and regulatory regions show gain or loss of DNA methylation during the differentiation of their progeny (Bock et al., 2012; Cabezas-Wallscheid et al., 2014). However, most of the epigenetic measurements underpinning these observations represent consensus values experimentally derived from thousands of cells partially enriched in HSCs or their progeny, thus failing to discern distinct epigenetic states within HSCs. Indeed, heterogeneity in methylation states of single CpGs is a common feature of cells assessed as bulk populations (Angermueller et al., 2016; Farlik et al., 2016; Hou et al., 2016; Hu et al., 2016; Qu et al., 2016). In addition, epigenetic heterogeneity has been observed across individual HSCs and clonally amplified HSC populations with preserved lineage potentialities (Farlik et al., 2016; Yu et al., 2016). Nevertheless, the degree to which heterogeneity in the methylome of HSCs is related to their defining properties remains poorly understood.

Assessment of the methylome of single cells is limited by measurement insensitivity and stochastic missing data. Current analytical strategies for single-cell DNA methylation measurements average DNA methylation in fixed genomic bins (Angermueller et al., 2016; Hou et al., 2016; Luo et al., 2017; Smallwood et al., 2014), or over defined genomic regions (Farlik et al., 2015, 2016; Hu et al., 2016). However, in many instances multiple regulatory regions are present within these genomic intervals and the relationship of their activity to average DNA methylation within an interval unknown. This is further complicated by the observations that the methylation state of a single CpG can affect transcription (Banet et al., 2000; Fürst

et al., 2012; Hashimoto et al., 2013; Jinno et al., 1995; Mamrut et al., 2013; Nile et al., 2008; Tsuboi et al., 2017; Zhou et al., 2017) by altering transcription factor binding affinity (Rishi et al., 2010; Yin et al., 2017). Imputation strategies leverage sequence context along with CpG methylation states across single cells to increase the resolution of genomic intervals (Angermueller et al., 2017). However, inference across cells (as well as sequence context) assumes homogeneity across cells, which is at cross-purposes with the generation of single-cell molecular measurements through the potential to mask rare subpopulations.

To address these limitations, we developed an automated plate-based high-resolution single-cell methylation protocol that we call Post-Bisulfite Adapter Ligation (PBAL), and analyzed the resulting sequence reads with an analytical pipeline (Pairwise Dissimilarity Clustering: PDclust) that leverages the methylation state of individual CpGs. We applied this single-cell methylation framework to profile primitive hematopoietic cells of mouse and human origin to identify epigenetically distinct subpopulations. Deep sampling of the CpG content of individual HSCs allowed for the near complete reconstitution of regulatory states from epigenetically defined subpopulations of HSCs and revealed a high level of redundancy of CpG methylation states within these phenotypically defined hematopoietic cell types.

RESULTS

Post-Bisulfite Adapter Ligation

PBAL is an adaption of the post-bisulfite adapter tagging (PBAT) strategy (Miura et al., 2012) optimized for library diversity. Previous single-cell PBAT-like strategies have used random primers extended with Illumina sequences to enable direct amplification (Angermueller et al., 2016; Smallwood et al., 2014). When comparing this approach with untagged random priming, we observed that extended randomers generated shorter double-stranded DNA fragments compared with randomers alone, suggesting inefficient priming (Figure S1). To circumvent this we used untagged random primers and ligated Illumina sequencing adapters to the resulting double-stranded DNA fragments. Pooling of single-cell PBAL libraries allowed the number of PCR cycles to be reduced and hence increased library diversity without compromising the minimum yield requirements for Illumina sequencing (see Experimental Procedures).

We applied this protocol to EPCR$^+$CD45$^+$CD48$^-$CD150$^+$ (ESLAM) cells isolated by fluorescence-activated cell sorting from adult mouse bone marrow (see Experimental Procedures). This extremely rare, phenotypically defined population is of interest because it is the most highly purified (~40% pure) source of functionally defined mouse hematopoietic cells with durable (>4–6 months) repopulating activity in transplanted hosts (Benz et al., 2012; Kent et al., 2009). As a comparator population, we sorted a related but less HSC-enriched lineage-negative Lin$^-$SCA1$^+$c-KIT$^+$ (LSK) phenotype (~3% HSCs) (Osawa et al., 1996) (Figure 1A). We also applied our PBAL protocol to the analogously human HSC-enriched Lin$^-$CD34$^+$CD38$^-$CD90$^+$CD45RA$^-$CD49f$^+$ population (hereafter referred to as "CD49f cells") isolated from two different cord blood donors and reported elsewhere (Knapp et al., 2017, 2018). Approximately 10% of these CD49f cells have long-term repopulating activity in transplanted immunodeficient mice, and contain all of those able to generate progeny with similar activity assessed in secondary hosts (Knapp et al., 2017, 2018; Notta et al., 2011).

Upon isolation, these mouse and human HSC-enriched cells were immediately lysed and fully unmethylated lambda and fully methylated T7 phage controls added, followed by bisulfite conversion and on-bead desulfonation (Domanico et al., 2013). The resulting single-stranded DNA was used as a template for random priming to generate double-stranded DNA for multiplexed library construction on an Agilent Bravo liquid handling platform (Figure 1B). Following library construction and low-cycle PCR amplification, failed wells were identified by qPCR targeted to repetitive sequences in the mouse or human sequences and removed (see Experimental Procedures) (Figures S2D and S2E). Libraries that passed this quality control (QC) threshold were pooled and sequenced together on an Illumina HiSeq platform.

An average of 38% of the resulting reads from all single cells sequenced aligned to the reference genome, compared with an average of 0.32% for no template controls (Figure S2A). Conversion of unmethylated cytosines was >99% efficient, while overconversion of methylated cytosines was <2% (Table S1). An average of 4 million sequence reads were generated per cell, enabling the measurement of the methylation states of 1.2 million CpGs per cell (Figure S2B). *In silico* merging of the resulting sequence reads confirmed that CpG recovery per cell is stochastic (Figure S2C). As an additional QC step, copy-number variation (CNV) in 5-Mb windows was assessed with Control-FREEC (Boeva et al., 2012) to identify and remove cells with uneven coverage that may represent biological or technical artifacts (Figures S2F and S2G). After removal of cells that failed any QC step (Supplemental Experimental Procedures), data for 64 LSK cells, 84 ESLAM cells, and 121 CD49f cells were available for analysis.

Methylation Adjacency in Individually Analyzed Cells

CpG methylation states derived from bulk cells are characterized by spatial correlation (Eckhardt et al., 2006; Zhang

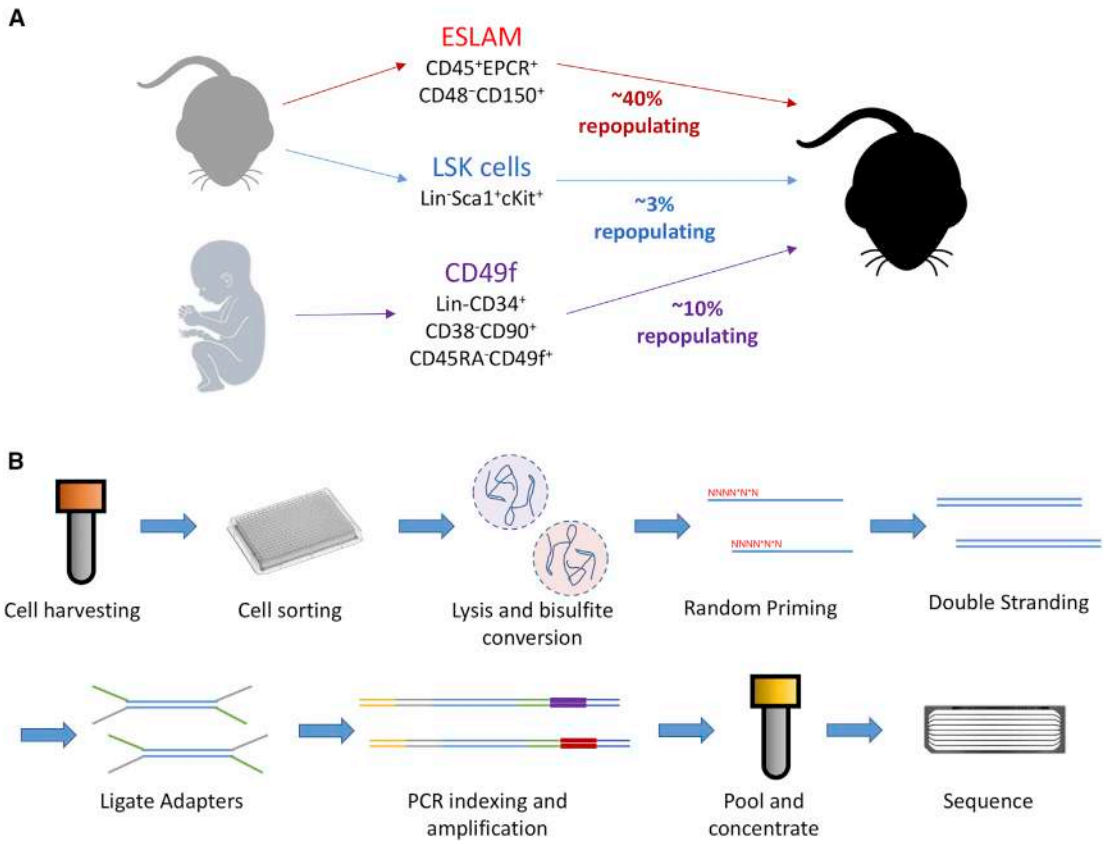

Figure 1. Overview of Experimental Methods

(A) Schematic of the phenotypes studied.

(B) Schematic of the PBAL method. Cells are lysed in 96-well plates to release genomic DNA (gDNA) and then subjected to bisulfite conversion that simultaneously converts and shears the gDNA into fragments. Random hexamers are then used to regenerate double-stranded DNA that is then end-repaired, a-tailed, and ligated with indexed adapters before low-cycle PCR amplification. Libraries are then pooled and sequenced on an Illumina sequencing platform.

See also Figures S1 and S2.

et al., 2015) and improvements in differentially methylated region (DMR) detection have been achieved through the development of algorithms that leverage spatial relationships (Hansen et al., 2012). Spatial correlation of CpG methylation state derived from bulk measurements has also provided a rationale for assigning the methylation state of single-CpG measurements to all CpGs within a genomic interval in single-cell methylation analyses (Farlik et al., 2015, 2016; Hou et al., 2016; Hu et al., 2016; Luo et al., 2017; Smallwood et al., 2014). However, the degree to which CpG methylation state is spatially correlated in individual cells has not been studied. To address this question, we calculated the probability that a CpG in single LSK cells was in the same methylation state with neighboring CpGs as a function of their genomic distance (Figure 2A) and, as a comparator, we generated adjacency measurements from PBAL datasets derived from 10,000 LSK cells using the calculated mean methylation difference between nearby CpGs to account for the continuous nature of bulk data.

This strategy revealed genome-wide concordance between CpGs up to 1 kb in *cis* in both the single cells and the bulk population, which then decreased to random chance at 2 kb (Figure 2B). Genome-wide CpG *cis* concordance was consistently higher in bulk compared with single-cell measurements and did not approach random chance at 5 kb, whereas single-cell concordance reached background levels at 4 kb (Figure 2B). Concordance between *cis* CpGs for single-cell and bulk-cell measurements was largely equivalent within CpG islands and promoters but was higher in bulk compared with single cells in all other genomic contexts (Figure 2C). In addition, we found that concordance for both single-cell and bulk samples decayed more rapidly within enhancers (Lara-Astiaso et al., 2014) compared with other genomic regions (Figure 2C).

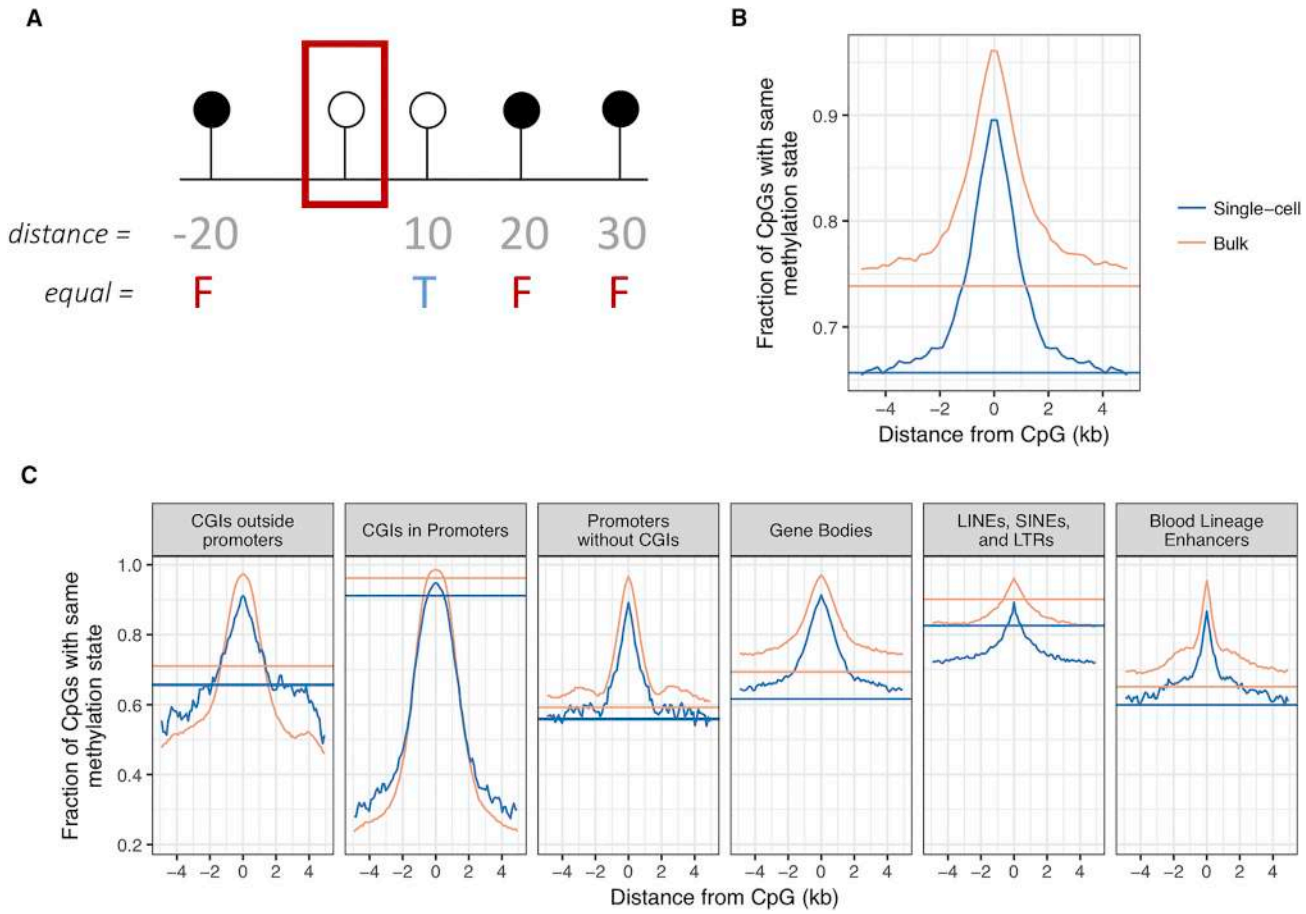

Figure 2. Concordance Analysis of Neighboring Methylated CpGs by Their Genomic Separation

(A) A schematic for how adjacency is calculated. For a randomly subsampled number of CpGs (CpG1), the concordance of methylation of 100 CpGs before and after CpG1 was recorded along with their separation.

(B) Analysis of bulk versus single LSK cells across CpG sites genome wide. Curves represent the weighted average in 100-bp bins. Horizontal lines indicate the probability that two randomly sampled CpG sites have the same methylation state based on their genome-wide methylation ("baseline concordance"). Baseline concordance was calculated as the probability of sampling two methylated or two unmethylated CpGs, which is equal to the square of the average fractional methylation rate plus 1 minus the square of the average fractional methylation rate.

(C) Analysis of bulk versus single cells across selected regions. For each panel, we considered only CpGs as CpG1 if they were found within each genomic feature.

Taken together, these findings suggest that analytical strategies that infer cytosine methylation within large bins (up to 100 kb, Luo et al., 2017) across the whole genome of single cells may lead to oversmoothening of CpG methylation, supporting a need for additional methodologies.

Identification of Epigenetic Subsets within HSC Populations at Single-CpG Resolution

Having established that spatial correlation of CpG methylation states decay rapidly toward random chance within 2 kb and are context specific within single cells, we sought to ask whether we could leverage the information content of single CpGs in single-cell methylation datasets. We (Gascard et al., 2015) and others (Bock et al., 2012; Hansen et al., 2012; Kundaje et al., 2015) have developed single-CpG resolution approaches for DMR detection from bulk DNA methylation measurements; however, these fail to address the sparse and stochastic methylation measurements characteristic of single-cell methylomes. To address this, we developed a measure of CpG methylation pairwise dissimilarity (PD) defined as the average of the absolute difference in methylation values at CpGs covered in each pairwise comparison (Figure 3A). We then calculated Euclidean

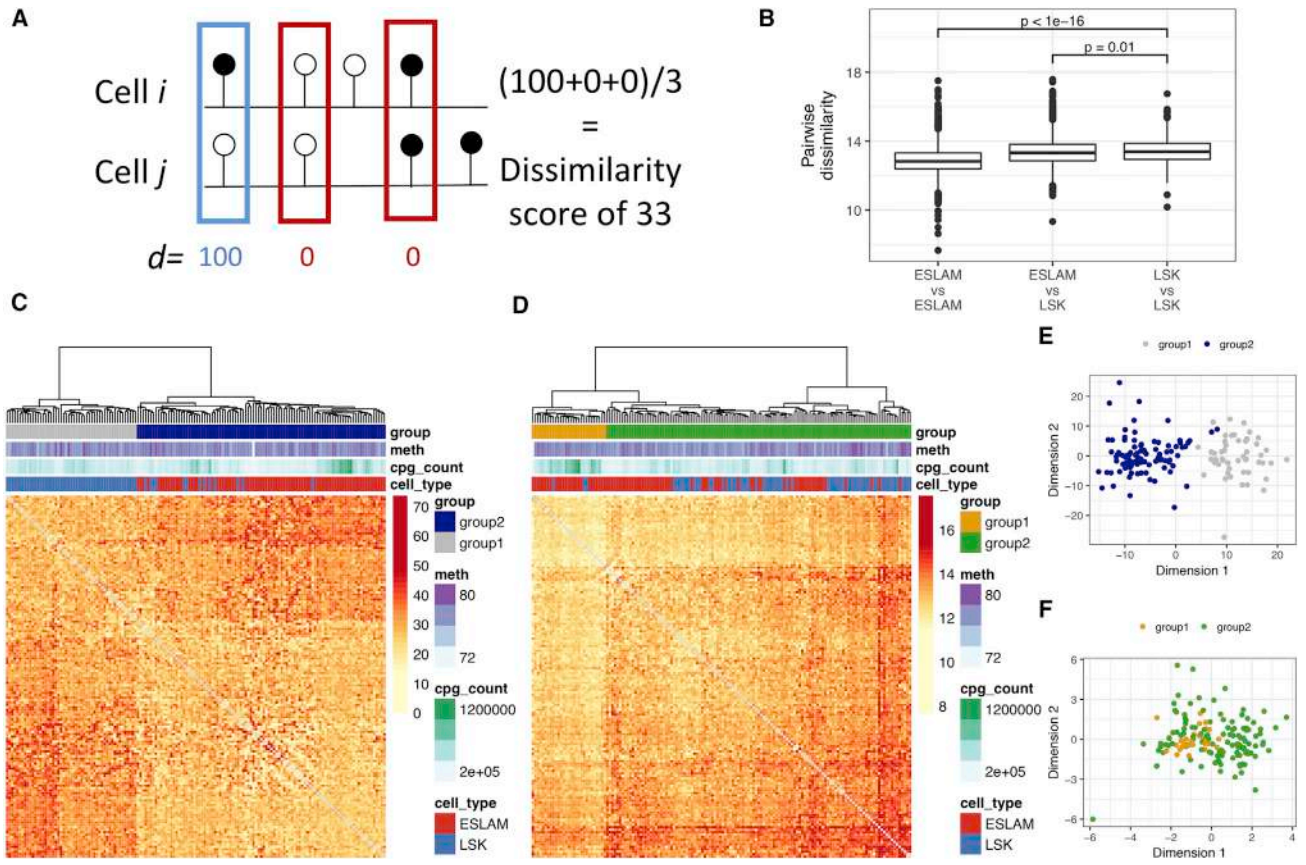

Figure 3. Pairwise Analysis of Single Cells Reveals Subsets within the Murine ESLAM Phenotype

(A) Schematic showing how the PD values are calculated between all paired comparisons of single cells.

(B) ESLAM cells have a lower overall PD compared with LSK cells and all cell types analyzed. Every pairwise comparison between cells denoted on the x axis is summarized as a box plot with the distribution of PD values shown on the y axis. p values were calculated using a two-sided t test.

(C) PDclust of CpGs associated with genes implicated in HSC function (Cabezas-Wallscheid et al., 2014) separate ESLAM and LSK cells, with some LSK cells exhibiting an ESLAM epigenetic signature. The rows and columns are symmetrical and represent single cells. The cells are shaded to represent the PD between each pair of cells, with red representing highly dissimilar and yellow representing highly similar. Meth, average genome-wide CpG methylation; cpg_count, number of distinct CpG sites recovered.

(D) Same as (C) but instead considering all CpGs regardless of their genomic position.

(E) Multidimensional scaling using PD calculated from (C) used directly as input.

(F) MDS analysis of (D) reveals group 1 at the epicenter of single ESLAM cells with group 2 surrounding the central cluster.

See also Figures S3–S5.

distances between each pair of cells using their PD values as features and performed hierarchical clustering (PDclust; see Experimental Procedures).

Application of this algorithm to the ELSAM and LSK single-cell datasets showed that ESLAM cells are epigenetically more similar to each other compared with LSK cells ($p < 0.01$), consistent with the knowledge that they are functionally less heterogeneous (Figure 3B). Unsupervised clustering on PDs derived from CpG sites within genomic regions previously implicated in the HSC-to-multipotent progenitor (MPP) transition (Cabezas-Wallscheid et al., 2014) generated two distinct subsets that were differentially enriched in single ESLAM and LSK cells (Figure 3C). Projection of PDs onto two-dimensional space with multidimensional scaling supported the existence of two epigenetic states defined by distinct DNA methylation signatures (Figure 3E)—one that was most enriched in the ESLAM population and the other in the LSK cells—with a proportion (13/64) of LSK cells showing an ESLAM-like epigenetic state.

Next, we enumerated PD values from CpGs genome wide to detect epigenetic subsets within the ESLAM and

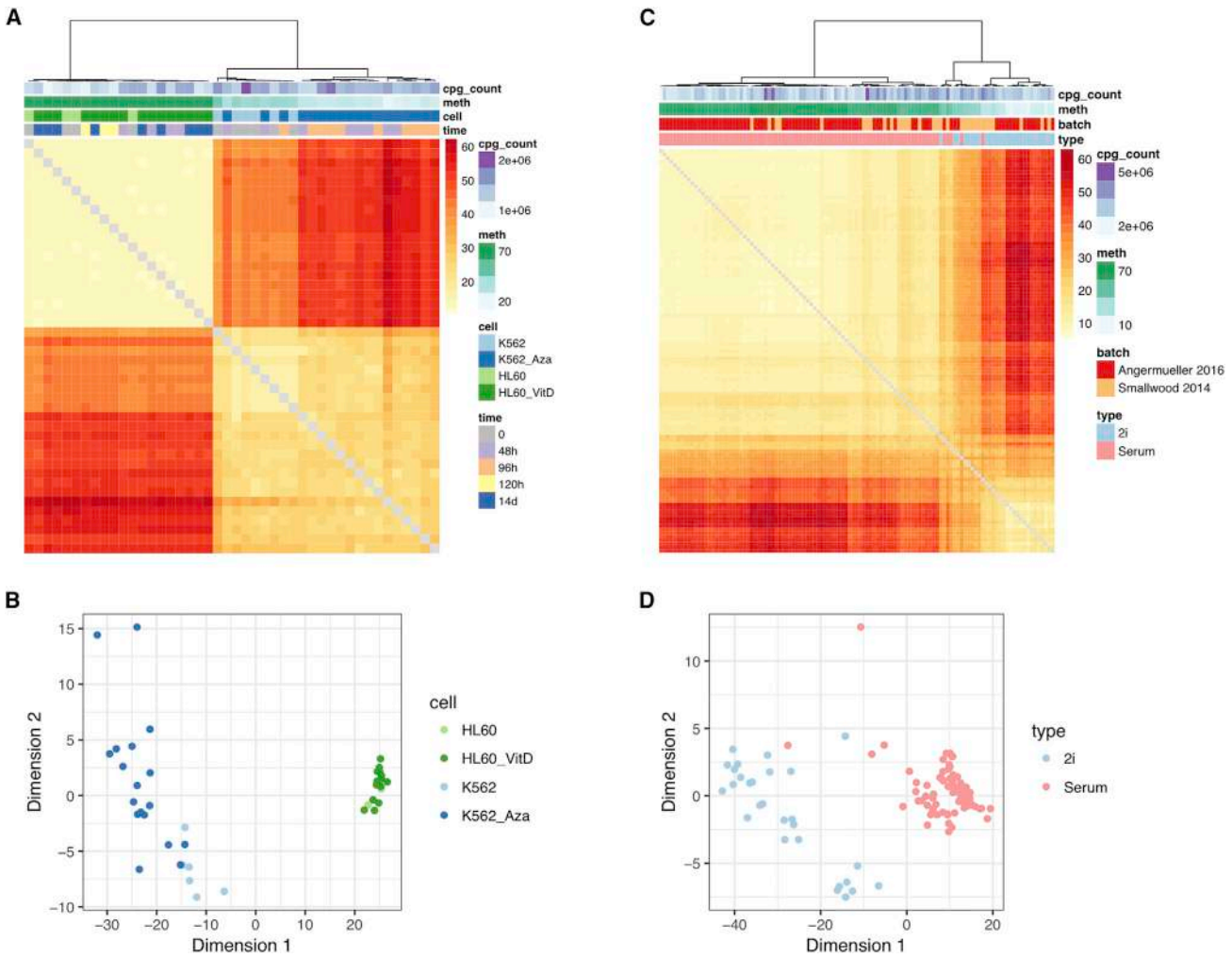

Figure 4. Pairwise Dissimilarity Applied to Existing Datasets Is Able to Distinguish Cells

Based on Cell Type and Treatment. (A and B) Clustering (A) and MDS scaling of PD values (B) calculated from HL60 and K562 cells (Farlik et al., 2015) separates cells by cell type, and reveals patterns of treatment-induced differentiation. Cells are shaded to represent the pairwise dissimilarity between each pair of cells, with red representing highly dissimilar and yellow representing highly similar. Meth, average genome-wide CpG methylation; cpg_count, number of distinct CpG sites recovered.

(C and D) Clustering (C) and MDS scaling (D) of embryonic stem cells (Smallwood et al., 2014; Angermueller et al., 2016) separates cells by the culture medium in which they were grown, and shows some cells in a transition state.

LSK populations. When all available CpG sites were used, PDclust revealed a subgroup of cells with the highest relative homogeneity (group 1) compared with all other cells (group 2) (Figure 3D). Multidimensional scaling (MDS) analysis confirmed these relationships by revealing a highly similar population (group 1) and a dispersed population (group 2) (Figure 3F). Group 1 included a higher proportion of ESLAM cells than LSK cells (26/84 versus 3/64 or 31% versus 5%). Interestingly, these proportions closely resemble the published biologically defined HSC content of both of these phenotypically defined populations. MDS plots further revealed a gradual and continuously increasing heterogeneity in the CpG profiles of LSK cells (Figure 3F). As a negative control, we considered only CpGs within a genomic region set that would not be expected to be relevant in HSCs (cortex enhancers; Hon et al., 2013) and showed no discernible relationships (Figure S3). As an additional validation, we applied PDclust to previously generated methylome data for single human leukemic cells (Farlik et al., 2015) and single murine embryonic stem cells (Angermueller et al., 2016; Smallwood et al., 2014) and separated cells accurately by type and treatment, with an exception for HL60 cells treated with vitamin D (Figure 4). Taken together, these results suggest

that the epigenetic state of a single cell can be accurately described by a small (<1%) stochastically sampled fraction of CpGs and that these states are reflective of cell identity and state.

Function of DMRs in Epigenetically Defined Subgroups of Mouse HSCs

To examine the functional significance of the epigenetically defined subsets of LSKs and ESLAM cells, we merged CpG calls *in silico* across all cells belonging to group 1 and separately to group 2 described above. We then inferred methylation values of all CpGs across the genome for these two groups using Bsmooth (Hansen et al., 2012) and identified DMRs between the two groups (Figure S4A and Experimental Procedures). This resulted in the identification of 9,922 DMRs that were hypomethylated in group 1 compared with group 2, and 10,047 DMRs that were the opposite state. For each DMR, we then used HOMER (Heinz et al., 2010) to annotate its overlapping genomic feature(s) as well as the nearest gene.

In general, CpG methylation is anti-correlated with expression when found in the promoter region or the first exon of a gene (Brenet et al., 2011), and active distal regulatory regions are typically hypomethylated (Stadler et al., 2011). We found that DMRs between groups 1 and 2 were enriched in promoters (±2 kb of transcription start site [TSS]), depleted in distal intergenic regions, and enriched in evolutionarily conserved genomic elements (Phastcon score [Siepel et al., 2005]; two-sided t test, $p \sim 0$) (Figure S4B). We next looked for subsets of DMRs that were either within the first exon, within the promoter, or intergenic and within 20 kb from an annotated TSS (Kundaje et al., 2015). We assigned each DMR of this class to the subgroup that reported the lower methylation value and associated each DMR group with the nearest genes using HOMER (Heinz et al., 2010). As a control, we performed the same analysis for data merged *in silico* from all LSK and ESLAM single cells. To identify genes uniquely associated with only group 1 or group 2, we removed those associated with DMRs from both populations and subjected the remaining group 1- or group 2-specific genes to gene set enrichment analysis (GSEA). Genes associated with hypomethylated DMRs in group 1 were significantly enriched (false discovery rate [FDR] adjusted q value < 0.1) in HSC proliferation terms as well as genes preferentially expressed in long-term HSCs and erythrocytes (Figure S4C). In contrast, genes associated with hypomethylated group 2 DMRs were enriched in genes specifically expressed in differentiated hematopoietic cells and genes that, when knocked out, lead to increased HSC numbers. Together these results suggest that DMRs specifically hypomethylated in group 1 compared with group 2 are associated with genes implicated in HSC function.

After identifying genes associated with hypomethylated DMRs, we sought to examine their expression in single ESLAM cells. To do this, we used previously generated single-cell RNA-sequencing data from $Lin^{-}c\text{-}KIT^{+}SCA1^{+}CD34^{-}FLT3^{-}CD48^{-}CD150^{+}$ HSCs (Wilson et al., 2015) and compared the expression of the DMR-associated genes thus identified with all genes. This showed that the DMR-associated genes were expressed at higher levels in these cells as compared with all other genes ($p < 0.01$, Mann-Whitney test) (Figure S5A). Further investigation of the levels of expression of these DMR-associated genes within the ESLAM population (Wilson et al., 2015) showed that a significant number showed heterogeneous expression profiles as might be expected if they were composed of one or more subpopulations (hypergeometric p value < 0.01) (Figure S5B). To identify potential surface markers that might allow their physical separation, we identified genes encoding plasma membrane proteins (as defined by gene ontology) and associated these with DMRs that were hypomethylated in group 1. The resulting list included *Cd82* (Figure S4A), a gene encoding a surface protein previously implicated in the maintenance of long-term HSCs *in vivo* (Hur et al., 2016).

Human HSCs and Their Derivatives Can Also Be Defined at Single-CpG Resolution

To determine whether single human hematopoietic cells could also be described using PDclust, we generated PBAL datasets for series of single human CD49f cells and obtained previously published datasets for all the major $CD34^{+}$ phenotypes in human cord blood (Farlik et al., 2016). Application of PDclust to the latter datasets showed that a majority of megakaryocytes and selected CD49f cells and other $CD34^{+}$ phenotypes had higher PD values and appeared to be outliers from the remaining cells (groups 2 and 3) (Figures S6A–S6C). These outlier cells also demonstrated significantly lower genome-wide average CpG methylation in comparison with all other cells, suggesting that they were either technical or biological outliers (Figure S6D). When they were removed, the single-cell methylome data clustered according to cell phenotype, with some overlap of the CD49f cells and MPPs (Figure 5A). Multidimensional analysis confirmed separation of CD49f cells and MPPs from other phenotypes and revealed a clear separation of progenitors of granulocytes and macrophages (GMPs) from all other phenotypes (Figure 5B). This analysis also showed that cells within the same phenotype had a lower dissimilarity as compared with other phenotypes ($p < 0.01$ for all comparisons) (Figure 5C). These results confirm that stochastic measurements of single-CpG states in single cells can accurately segregate multiple primitive phenotypically distinct human hematopoietic cell populations.

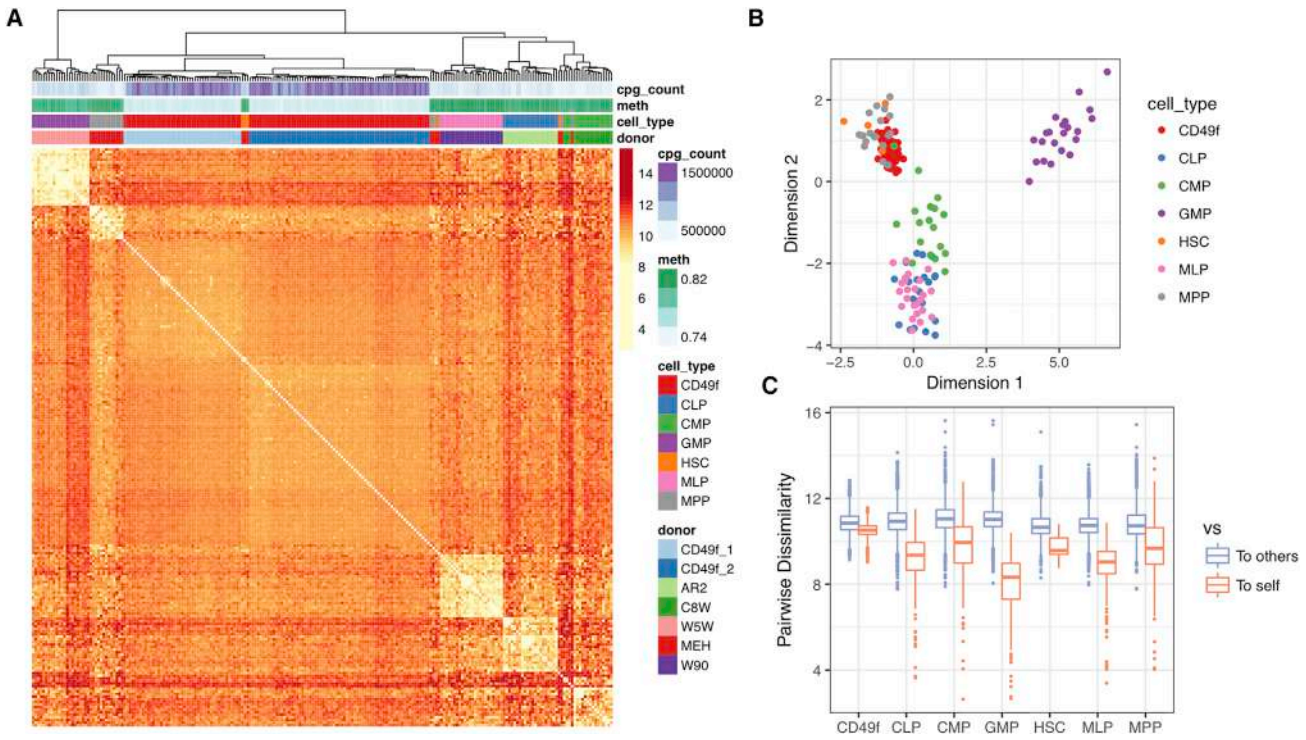

Figure 5. Pairwise Dissimilarities of Human Hematopoietic Cells with Different Phenotypes

(A) PD of CD49f and other CD34$^+$ subsets separate cells by phenotype with some overlaps. The rows and columns are symmetrical and represent single cells. Cells are shaded to represent the PD between each pair of cells, with red representing highly dissimilar and yellow representing highly similar.

(B) MDS of PD values shows that CD49f cells cluster in the middle, with GMPs, CLPs, and MLPs branching out in separate directions.

(C) Single cells have lower PD values compared with cells of the same type versus cells of a different phenotype. The distribution of PD values when a cell is compared with either a cell of the same or different phenotype is plotted as a box plot.

See also Figure S6.

Sources of Epigenetic Heterogeneity within the Human CD49f Compartment

The CD49f cells were sorted from cord blood derived from two individual donors, which enabled an examination of donor-specific contributions to methylation states at the single-cell level. Interestingly, this analysis revealed that that the data for single CD49f cells clustered separately by donor (Figure 5A). To investigate whether these donor-specific epigenetic differences were a result of genetic differences between the two donors, we identified genetic variants by applying MethylExtract (Barturen et al., 2014) to the donor *in silico* merged datasets. After combining non-identical homozygous variants for both donors (excluding homozygous variants involving a C or G), we masked CpGs within 200 bp of the donor-specific variants and recalculated the PD values. This decreased both inter- and intra-donor PD values, but donor-specific methylation states remained the major driver of variation within the CD49f population (Figures 6A and 6B). We also noted that the PD value increased when only CpGs within 200 bp of single-nucleotide variants (SNVs) were considered (Figure 6C). Taken together, this result suggests that genetic variation accounts only partly for the observed donor-associated epigenetic variation and that donor-specific epigenetic variation is a dominant feature across single and highly purified human CD49f cells.

Application of PDclust separately to the methylome data for the single CD49f cells from donors 1 and 2 then enabled the identification in each of a consistent subpopulation (group 1, comprising 9% and 11%, respectively, Figure 7A) that did not correlate with expression levels of the surface markers used to isolate them (CD3, 11b, 19, 34, 38, 90, 45RA, 49f) (Figure 7B). To test the sensitivity to the depth of sequence data obtained, we then sequenced the library from donor 1 to a 4-fold greater depth (an average of 7.8 million mapped reads per cell) (Figures S7A and S7B). Interestingly, this did not significantly alter the results (Figure S7C), suggesting that the cell-to-cell variation was already accurately captured by the original dataset.

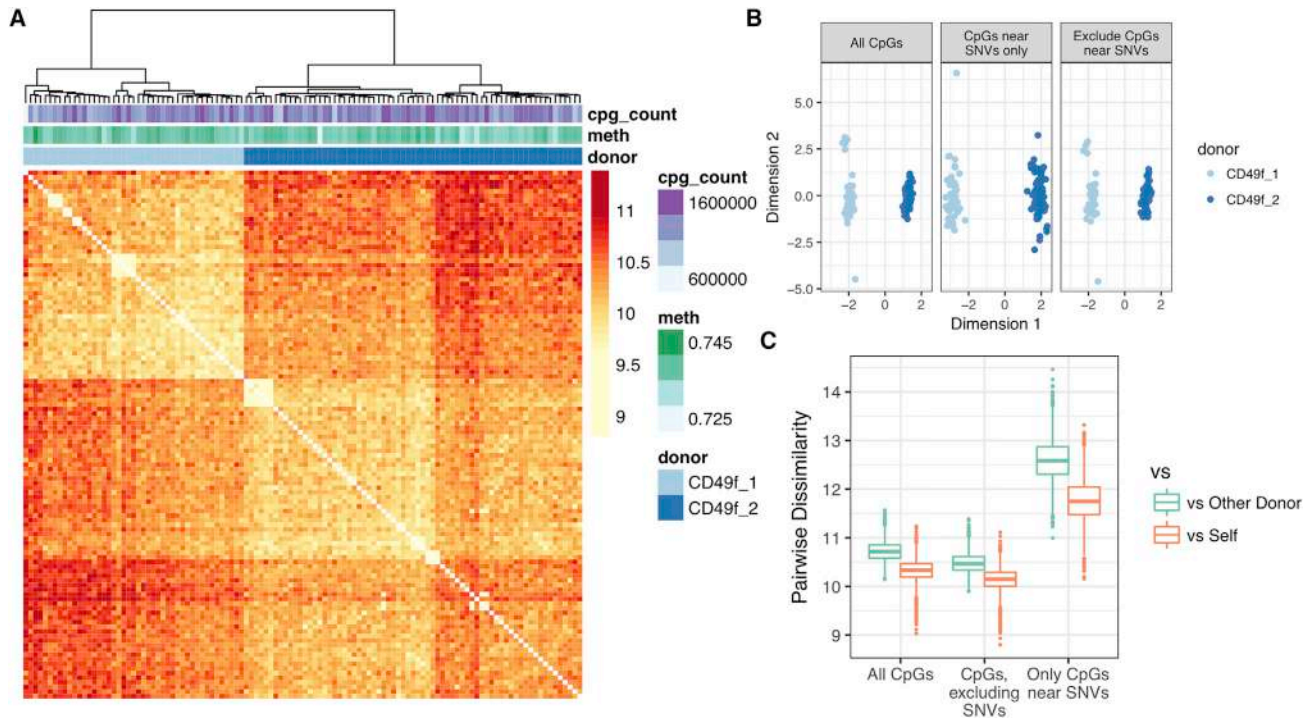

Figure 6. Donor Variation in CD49f Cell Methylomes

(A) Single CD49f cells still cluster by donor after removing CpGs within 200 bp of non-C and -G SNV locations.

(B) MDS projection of PD values onto 2D space remains largely unchanged despite taking into account SNVs.

(C) PD values as a function of comparisons versus cells either from the same or the other donor. Boxes are colored if only CpGs near SNVs, outside SNVs, or all CpGs were considered.

We then took advantage of the more deeply sequenced dataset to perform an *in silico* merging of the cells previously identified as group 1 (n = 6) and group 2 (n = 63) cells. *In silico* merging of the data from the first group 1 resulted in a nearly complete recapitulation of the human methylome (17.1 M CpGs in group 1 and 27.1 M CpGs in group 2) allowing for comprehensive epigenetic annotation of the two groups. As a comparator, we performed *in silico* merging of all available CD49f cell data as well as the published hematopoietic progenitor cell (HPC) data (common lymphoid progenitors [CLPs], common myeloid progenitors, GMPs, multilymphoid progenitors [MLPs], and MPPs) (Farlik et al., 2016). As before, we estimated the smoothened methylation values of all CpG sites in the genome for each group and called DMRs. This resulted in the identification of 18,035 DMRs that were hypomethylated in group 1 compared with group 2, and 15,283 DMRs that were in the opposite state. DMRs were enriched in promoters (i.e., sequences between 2 kb upstream and 500 bp downstream of coding gene TSSs), depleted in intergenic regions, and highly enriched in evolutionarily conserved elements (Phastcon score [Siepel et al., 2005]; two-sided t test, $p \sim 0$), suggesting that these DMRs have functional relevance (Figure 7C). GSEA of genes associated with DMRs that were hypomethylated in CD49f cells compared with HPCs showed these were enriched in pathways implicated in HSC differentiation (Figure 7D). Interestingly, DMRs hypomethylated in group 1 compared with group 2 were enriched in genes that are upregulated in leukemia (Casorelli et al., 2006) and genes whose expression is upregulated in hematopoietic progenitors (Ivanova et al., 2002) (Figure 7D). For example, *SERPING1*, a gene that is upregulated in acute promyeolocytic leukemia (Casorelli et al., 2006) and a prognostic marker for acute myeloid leukemia (Laverdière et al., 2016), was found to contain a DMR that was hypomethylated in group 1 cells (Figure 7E).

DISCUSSION

We describe an automated methodology for single-cell DNA methylation profiling of single cells optimized for genomic coverage. The increased coverage afforded by our methodology allowed epigenetically distinct subsets within highly purified phenotypes to be identified and distinguished from developmentally both closely and

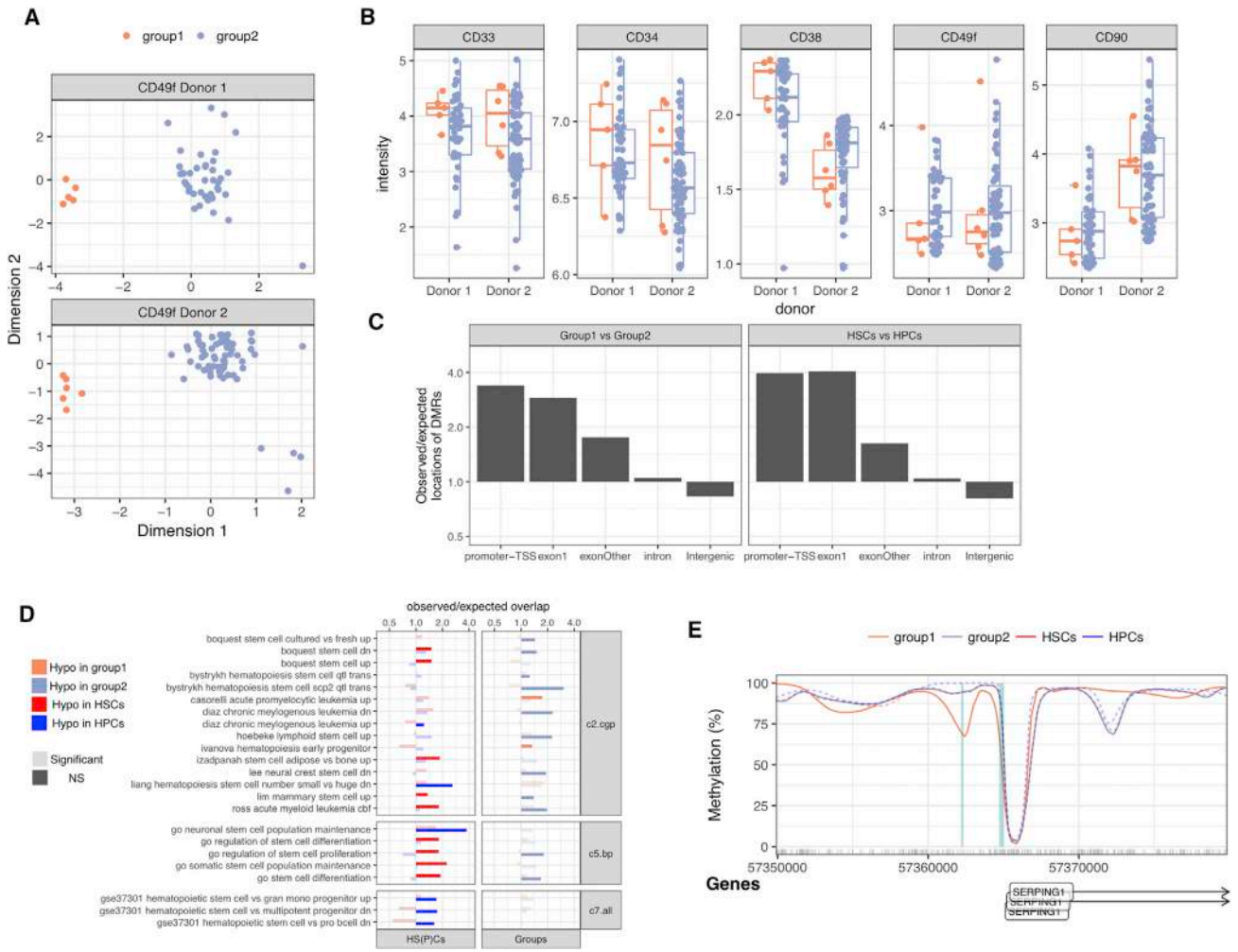

Figure 7. Heterogeneity within the Human CD49f Cell Compartment

(A) A rare subset of CD49f cells (group 1, 11% and 9% of all CD49f cells for donors 1 and 2, respectively) cluster away from the rest of the cells after projection of PD values with MDS.

(B) Distributions of surface markers obtained during index sorting of single cells belonging to cluster 1 or cluster 2, split by donor.

(C) Observed over expected enrichment of DMRs. The ratios were calculated by dividing the fraction of DMRs that overlap each region set by the fraction of the genome each region set occupies.

(D) Gene enrichment of DMRs that are hypomethylated in each comparison. The comparisons between group 1 and group 2 were separate from the comparisons between CD49f cells and published data for other CD34^{+} phenotypes. Bars are gray if their FDR-corrected binomial q value is <0.1.

(E) Example of a DMR near the *SERPING1* gene. Methylation values were smoothened for each population of cells with Bsmooth and plotted. Tick marks on the x axis represent the location of CpG dinucleotides. Computationally defined DMRs are highlighted in blue. The genes track shows high confidence protein-coding transcripts obtained from Gencode v75.

See also Figure S7.

more distally related cell types. We therefore anticipate that it will have utility in dissecting biologically important, but potentially subtle, epigenetic changes in cells that drive as yet poorly understood developmental and disease processes.

Existing single-cell DNA methylation experimental protocols can be broadly categorized into those that include random priming-based genomic preamplification and those that do not. In general, single-cell experimental protocols that involve rounds of random genomic priming (e.g., five in single-cell bisulfite sequencing [scBS-seq], two in PBAL, one in single-nucleus methylcytosine sequencing [snmC-seq], versus none in single-cell whole-genome bisulfite sequencing [scWGBS]) result in libraries

with increased fragment diversity independent of the number of subsequent PCR cycles (Luo et al., 2017). However, challenges associated with single-cell random priming include a consistent decrease in aligned sequence fraction and increased per-cell cost. Rounds of random priming are inversely correlated with lower mapping efficiency (average 25% in scBS-seq, 38% in PBAL, 53% in snmC-seq, versus 55% in scWGBS), suggesting that external nucleic acid sources are amplified in addition to the target genome. Additional sequencing of the resulting library pools can circumvent this problem as in the case of PBAL. We have designed the PBAL approach with these limitations in mind to optimize fragment diversity through a combination of two rounds of random priming and low-cycle PCR. We have developed a qPCR-based assay that identifies failed wells from all causes prior to pooling and sequencing. Taken together, these optimizations allow for the generation of single-cell methylomes at lower cost with diversities meeting or exceeding existing published strategies.

To identify epigenetic states from stochastic and sparse measurements characteristic of single-cell methylation datasets, we have developed an analysis package (PDclust; https://github.com/hui-tony-zk/PDclust) that leverages the information content of single CpGs without the need to infer the methylated states of missing data. We demonstrated its application to CpGs within regulatory regions (Cabezas-Wallscheid et al., 2014) through the identification of an epigenetically distinct subpopulation of cells in two different compartments of adult mouse bone marrow (LSK and ESLAM cells) that match their known content of HSCs defined by long-term repopulating assays: ~3% in LSKs (Osawa et al., 1996) and 40% in ESLAM cells (Benz et al., 2012; Kent et al., 2009). Analysis of the resulting DMRs further revealed a statistically significant number of hypomethylated regions in this subgroup that were associated with genes implicated in mouse HSC function, a majority of which have also been found to be transcriptionally active in mouse HSCs (Wilson et al., 2015). This supports previous evidence of a connection between CpG methylation status and expression examined at a single-cell level (Angermueller et al., 2016; Hu et al., 2016). A significant proportion of DMR-associated plasma membrane genes were also heterogeneously expressed among individual ESLAM cells including *Cd82*, a previously annotated marker of HSCs (Hur et al., 2016), suggesting that this might be useful in conjunction with the ESLAM protocol to further enrich for HSCs from adult mouse bone marrow.

Application of PDclust to single human hematopoietic cell datasets separated them according to established phenotypes and identified rare epigenetically distinct subpopulations from which near complete reconstitution of the methylome was achieved. Interestingly, within the highly purified CD49f population we identified donor as a significant source of consistent epigenetic heterogeneity, which was reduced but not eliminated by correcting for personal genetic variants. This observation is consistent with previous reports that showed genetic diversity as related to but not accountable for all DNA methylation differences (Gertz et al., 2011; Xie et al., 2012) and suggests that *in utero* environmental differences may be encoded within the HSC compartment, as has been reported for bulk analysis (Bommarito et al., 2017; Provençal and Binder, 2015).

The ability to detect distinct epigenetic states that distinguish phenotypically defined hematopoietic cell types from a few thousand stochastically sampled CpGs in single cells is an intriguing observation with potentially broad implications. One possible explanation for this observation is that phenotype-specific methylation signatures are characterized by extensive redundancy such that distinct epigenetic states can be accurately described by only a small fraction of single-CpG methylation states. In support of such a notion, the unique components of a DNA methylation "age" signature are contained in ~353 CpGs sites, presumably representing a random sample of a total age signature that involves many more sites not detected using the reduced representation strategies from which these signatures have been derived (Horvath, 2013). Regardless of the mechanism, the ability to describe epigenetic states in single cells from single-CpG measurements without the need to infer adjacent methylation states has significant implications for the design and analysis of single-cell methylation studies.

EXPERIMENTAL PROCEDURES

This study involved animal research and is approved by the University of British Columbia under ACC certificate number A14-0091. Human cord blood cells were obtained with informed consent from mothers of normal babies according to UBC-approved protocols (UBC REB Certificate H07-01945 to C.J.E.).

Sample Preparation

Murine bone marrow of 8- to 10-week-old mice was stained using LSK or ESLAM staining cocktails (Table S1). HSCs were then sorted using a FACSAria Fusion (Becton Dickinson, Franklin Lakes), flash-frozen, and stored at −80°C until processing. Human cord blood was obtained from two normal full-term deliveries. Cells were stained and sorted with the CD49f cocktail as described elsewhere (Knapp et al., 2017, 2018). Full details can be found in Supplemental Experimental Procedures.

PBAL Library Construction

All reagents were DNaseI or UV treated to remove ambient contaminants, and all samples were processed with negative controls. All liquid handling steps were carried out on the Bravo Automated Liquid Handling Platform (Agilent Technologies, G5409A). Plates

containing single cells were lysed, spiked-in with bisulfite conversion controls, and subject to automated bead-based bisulfite conversion (Domanico et al., 2013). The resulting single-stranded DNA was subject to random priming with random hexamers (3′-phosphothioate, NNNN*N*N) and Klenow exo- (NEB, M0212M) twice. The double-stranded DNA was then ligated with forked Illumina adapters (Lorzadeh et al., 2016) and PCR amplified for eight cycles. Successful wells were identified by qPCR targeted to repetitive regions of the mouse or human genome, pooled, and sequenced on an Illumina HiSeq 2,500. Full details can be found in Supplemental Experimental Procedures.

Data Processing

The first six bases of read1 and read2 were trimmed using Trimgalore v0.4.0 and aligned using Novoalign V3.02.10 (www.novocraft.com) to the mouse assembly GRCm38 (mm10) or human assembly GRCh37 (hg19). Duplicate reads were discarded using Picard V1.31 (http://picard.sourceforge.net), and methylation states were called using Novomethyl V1.01 (www.novocraft.com). Fractional methylation was merged for each CpG dinucleotide by taking a weighted average of each cytosine using bedtools (Quinlan and Hall, 2010). In most cases, only one base within a CpG dinucleotide had coverage. In these cases, the methylation information of the covered base was extrapolated to the other base. Processed CpG calls were imported into R (V3.3.2) for downstream analysis. We considered only autosomal CpG sites and CpG sites with a methylation value of 0% or 100%.

CNVs in 5-Mb windows were called using Control_FREEC V7.0 (Boeva et al., 2012) with default parameters. Single cells with conversion rates <96%, mappability <5%, <130,000 CpGs, and containing more than 50 windows with CNVs were removed. Full details can be found in Supplemental Experimental Procedures.

Methylation of Adjacent CpGs in Single Cells

For single cells, we randomly sampled up to 100,000 CpG sites either genome wide or within relevant genomic regions. For each randomly sampled CpG site (CpG1), we analyzed 100 CpG sites with coverage >0 before and after CpG1 and calculated the genomic distance to CpG1 in base pairs. We also recorded whether or not each CpG had the same methylation status as CpG1, resulting in a 2-column table containing the distance and equality status for each CpG for near CpG1. After combining all of the data in these tables for every CpG1, we binned distances into 100-bp bins and calculated the mean concordance as the fraction of CpGs in each bin that were equal. For single LSK cells, we did this for the ten single cells with the most CpG coverage. For the bulk LSK cell data that were continuous, we instead calculated the absolute difference in methylation between CpG1 and all nearby CpGs, using only CpGs with coverage ≥ 5 to avoid low coverage biasing potential CpG methylation values.

Pairwise Dissimilarity Clustering Using PDclust

For each pair of single cells, we calculated the average difference in DNA methylation of all pairwise-common CpG sites to obtain PD values. For genome-wide analysis, we took into consideration all CpGs, while for each genomic region sets we only considered CpG sites that lie within those respective regions. To group cells together with similar PD values, we calculated Euclidean distances between each pair of cells using the their PD values as features and performed hierarchical clustering with Ward's linkage (*ward.D2* in R). We used PD directly as input to multidimensional scaling (*cmdscale* in R) for visualization of cells in 2D space. Cell groups were annotated manually based on their visual distinctiveness and their resulting hierarchal clustering patterns. PDclust has been packaged into an R package, which can be downloaded from https://github.com/hui-tony-zk/PDclust.

Differentially Methylated Region Analysis

To group cells that belonged to the same cluster, we treated coverage of every CpG site as the number of cells with coverage at that site, and treated methylation fraction as the fraction of cells that had a methylated CpG at that site. We used Bsmooth (Hansen et al., 2012) to obtain estimated CpG methylation at all CpG sites in the genome. For each CpG comparison between the two groups, we calculated the Z score converted that into a two-tailed p value assuming a normal distribution (pnorm function in R) with the null hypothesis that the methylation is not different between the two CpGs. p values were multiple test-corrected using an FDR estimate, and CpGs with q value of <0.1 were called as dCpGs. We grouped dCpGs together into DMRs if they were within 500 bp of each other and only considered DMRs with ≥ 3 CpGs. We split these DMRs into two groups depending on which population had the lower methylation in each pair of comparisons, and associated DMRs to the nearest protein-coding transcript. After filtering for those DMRs in intergenic regions, exon1, or the promoter (+2 kb/−0.5 kb), we removed DMRs that associated with the same gene in both groups. We then performed GSEA for all DMR-associated genes for each group using a supervised set of ontologies. Full details can be found in Supplemental Experimental Procedures.

ACCESSION NUMBERS

Single-cell bisulfite sequencing (raw reads and CpG methylation calls) can be accessed from the Gene Expression Omnibus at GEO: GSE89545. Bulk LSK data can be accessed at GEO: GSE95697. Human CD49f CpG methylation data can be accessed at GEO: GSE106957. Raw reads for the human CD49f data can be conditionally accessed from the European Genome-Phenome Archive (EGAS00001002789).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and five tables and can be found with this article online at https://doi.org/10.1016/j.stemcr.2018.07.003.

AUTHOR CONTRIBUTIONS

T.H. and Q.C. developed the single-cell methylation protocol and generated the single-cell libraries. T.H. developed the analytical framework, analyzed and interpreted the data, and wrote the paper. K.O. helped with the CpG adjacency analysis. J.W.-W., C.A.H., and D.J.H.F.K. isolated the cells analyzed and assisted in

data interpretation. E.L. sorted test single cells for technology development. M.M. assisted with technology development, interpreted the data, and wrote the paper. S.A. and A.K. interpreted data. M.H. designed the study, interpreted data, and, with C.J.E., wrote the final draft of the manuscript.

ACKNOWLEDGMENTS

This work was supported by Terry Fox Research Institute Program Projects (grant no. 1021) awarded to M.H. and grant nos. 122869 and 1074 awarded to M.H., A.K., and C.J.E., Canadian Institutes of Health Research (CIHR), Genome Canada and Genome British Columbia (CIHR EP1-120589) awarded to M.H., a Canadian Cancer Society grant generously supported by the Lotte & John Hecht Memorial Foundation (grant no. 703489) to M.H., a CIHR and Canadian Cancer Society Research Institute grant to A.K., a CIHR-National Science and Engineering Research Council of Canada grant (CHRP 413633) to C.E., and a Terry Fox Research Institute New Investigator Award (grant no. 1039) to M.H. K.O. was supported by A Michael Smith Foundation for Health Research Trainee Award (no. 16127). T.H. was supported by a Canada Graduate Scholarship - Master's Award (CGS-M), D.J.H.F.K. by a Vanier Scholarship, and C.A.H. by a CIHR Frederick Banting and Charles Best Doctoral Scholarship. A.K. was supported by the John Auston BC Cancer Foundation Clinical Investigator Award. This research was enabled in part by support provided by WestGrid and Compute Canada (www.computecanada.ca) and Canada Foundation for Innovation (nos. 31343 and 31098). The authors wish to acknowledge Canada's Michael Smith Genome Sciences Center, Vancouver, Canada for computational resources and support and the Stem Cell Assay of the BC Cancer Agency for assistance in obtaining and isolating the cord blood cells used. A full list of other funders of infrastructure and research supporting the services accessed is available at www.bcgsc.ca/about/funding_support.

Received: March 14, 2018
Revised: July 8, 2018
Accepted: July 9, 2018
Published: August 2, 2018

REFERENCES

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S.A., Ponting, C.P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat. Methods *13*, 229–232.

Angermueller, C., Lee, H.J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. *18*, 67.

Banet, G., Bibi, O., Matouk, I., Ayesh, S., Laster, M., Molner Kimber, K., Tykocinski, M., de Groot, N., Hochberg, A., and Ohana, P. (2000). Characterization of human and mouse H19 regulatory sequences. Mol. Biol. Rep. *27*, 157–165.

Barturen, G., Rueda, A., Oliver, J.L., and Hackenberg, M. (2014). MethylExtract: high-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. Version 2. F1000Res. *2*, 217.

Benz, C., Copley, M.R., Kent, D.G., Wohrer, S., Cortes, A., Aghaeepour, N., Ma, E., Mader, H., Rowe, K., Day, C., et al. (2012). Hematopoietic stem cell subtypes expand differentially during development and display distinct lymphopoietic programs. Cell Stem Cell *10*, 273–283.

Bock, C., Beerman, I., Lien, W.-H., Smith, Z.D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D.J., and Meissner, A. (2012). DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Mol. Cell *47*, 633–647.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics *28*, 423–425.

Bommarito, P.A., Martin, E., and Fry, R.C. (2017). Effects of prenatal exposure to endocrine disruptors and toxic metals on the fetal epigenome. Epigenomics *9*, 333–350.

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A.J., Socci, N.D., and Scandura, J.M. (2011). DNA methylation of the first exon is tightly linked to transcriptional silencing. PLoS One *6*, e14524.

Cabezas-Wallscheid, N., Klimmeck, D., Hansson, J., Lipka, D.B., Reyes, A., Wang, Q., Weichenhan, D., Lier, A., von Paleske, L., Renders, S., et al. (2014). Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. Cell Stem Cell *15*, 507–522.

Casorelli, I., Tenedini, E., Tagliafico, E., Blasi, M.F., Giuliani, A., Crescenzi, M., Pelosi, E., Testa, U., Peschle, C., Mele, L., et al. (2006). Identification of a molecular signature for leukemic promyelocytes and their normal counterparts: focus on DNA repair genes. Leukemia *20*, 1978–1988.

Challen, G.A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J.S., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y., et al. (2012). Dnmt3a is essential for hematopoietic stem cell differentiation. Nat. Genet. *44*, 23–31.

Domanico, M.J., Allawi, H., Lidgard, G.P., Aizenstein, B., Hunt, O., and Zutz, T.C. (2013). Modification of DNA on magnetic beads. US patent US9315853, filed January 30, 2013, and published August 1, 2013.

Doulatov, S., Notta, F., Laurenti, E., and Dick, J.E. (2012). Hematopoiesis: a human perspective. Cell Stem Cell *10*, 120–136.

Dykstra, B., Kent, D., Bowie, M., McCaffrey, L., Hamilton, M., Lyons, K., Lee, S.-J., Brinkman, R., and Eaves, C. (2007). Long-term propagation of distinct hematopoietic differentiation programs in vivo. Cell Stem Cell *1*, 218–229.

Eaves, C.J. (2015). Hematopoietic stem cells: concepts, definitions, and the new reality. Blood *125*, 2605–2613.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet. *38*, 1378–1385.

Farlik, M., Sheffield, N.C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep. *10*, 1386–1397.

Farlik, M., Halbritter, F., Müller, F., Choudry, F.A.A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., et al. (2016). DNA methylation dynamics of human hematopoietic stem cell differentiation. Cell Stem Cell *19*, 808–822.

Fürst, R.W., Kliem, H., Meyer, H.H.D., and Ulbrich, S.E. (2012). A differentially methylated single CpG-site is correlated with estrogen receptor alpha transcription. J. Steroid Biochem. Mol. Biol. *130*, 96–104.

Gascard, P., Bilenky, M., Sigaroudinia, M., Zhao, J., Li, L., Carles, A., Delaney, A., Tam, A., Kamoh, B., Cho, S., et al. (2015). Epigenetic and transcriptional determinants of the human breast. Nat. Commun. *6*, 6351.

Gertz, J., Varley, K.E., Reddy, T.E., Bowling, K.M., Pauli, F., Parker, S.L., Kucera, K.S., Willard, H.F., and Myers, R.M. (2011). Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLoS Genet. *7*, e1002228.

Hansen, K.D., Langmead, B., and Irizarry, R.A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. *13*, R83.

Hashimoto, K., Otero, M., Imagawa, K., de Andrés, M.C., Coico, J.M., Roach, H.I., Oreffo, R.O.C., Marcu, K.B., and Goldring, M.B. (2013). Regulated transcription of human matrix metalloproteinase 13 (MMP13) and interleukin-1β (IL1B) genes in chondrocytes depends on methylation of specific proximal promoter CpG sites. J. Biol. Chem. *288*, 10061–10072.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D., and Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat. Genet. *45*, 1198–1206.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., et al. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. *26*, 304–319.

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.-Y., Xue, Z., and Fan, G. (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. Genome Biol. *17*, 88.

Hur, J., Choi, J.-I., Lee, H., Nham, P., Kim, T.-W., Chae, C.-W., Yun, J.-Y., Kang, J.-A., Kang, J., Lee, S.E., et al. (2016). CD82/KAI1 maintains the dormancy of long-term hematopoietic stem cells through interaction with DARC-expressing macrophages. Cell Stem Cell *18*, 508–521.

Ivanova, N.B., Dimos, J.T., Schaniel, C., Hackney, J.A., Moore, K.A., and Lemischka, I.R. (2002). A stem cell molecular signature. Science *298*, 601–604.

Jinno, Y., Ikeda, Y., Yun, K., Maw, M., Masuzaki, H., Fukuda, H., Inuzuka, K., Fujishita, A., Ohtani, Y., Okimoto, T., et al. (1995). Establishment of functional imprinting of the H19 gene in human developing placentae. Nat. Genet. *10*, 318–324.

Kent, D.G., Copley, M.R., Benz, C., Wöhrer, S., Dykstra, B.J., Ma, E., Cheyne, J., Zhao, Y., Bowie, M.B., Zhao, Y., et al. (2009). Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. Blood *113*, 6342–6350.

Knapp, D.J.H.F., Hammond, C.A., Miller, P.H., Rabu, G.M., Beer, P.A., Ricicova, M., Lecault, V., Da Costa, D., VanInsberghe, M., Cheung, A.M., et al. (2017). Dissociation of survival, proliferation, and state control in human hematopoietic stem cells. Stem Cell Rep. *8*, 152–162.

Knapp, D.J.H.F., Hammond, C.A., Hui, T., van Loenhout, M.T.J., Wang, F., Aghaeepour, N., Miller, P.H., Moksa, M., Rabu, G.M., Beer, P.A., et al. (2018). A novel subset of human CD33+ haematopoietic stem cells characterized at single-cell resolution. Nat. Cell Biol. https://doi.org/10.1038/s41556-018-0104-5.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Immunogenetics. Chromatin state dynamics during blood formation. Science *345*, 943–949.

Laverdière, I., Boileau, M., Herold, T., Rak, J., Berdel, W.E., Wörmann, B., Hiddemann, W., Spiekermann, K., Bohlander, S.K., and Eppert, K. (2016). Complement cascade gene expression defines novel prognostic subgroups of acute myeloid leukemia. Exp. Hematol. *44*, 1039–1043.e10.

Lorzadeh, A., Bilenky, M., Hammond, C., Knapp, D.J.H.F., Li, L., Miller, P.H., Carles, A., Heravi-Moussavi, A., Gakkhar, S., Moksa, M., et al. (2016). Nucleosome density ChIP-Seq identifies distinct chromatin modification signatures associated with MNase accessibility. Cell Rep. *17*, 2112–2124.

Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600–604.

Mamrut, S., Harony, H., Sood, R., Shahar-Gold, H., Gainer, H., Shi, Y.-J., Barki-Harrington, L., and Wagner, S. (2013). DNA methylation of specific CpG sites in the promoter region regulates the transcription of the mouse oxytocin receptor. PLoS One *8*, e56869.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res. *40*, e136.

Nile, C.J., Read, R.C., Akil, M., Duff, G.W., and Wilson, A.G. (2008). Methylation status of a single CpG site in the *IL6* promoter is related to *IL6* messenger RNA levels and rheumatoid arthritis. Arthritis Rheum. *58*, 2686–2693.

Notta, F., Doulatov, S., Laurenti, E., Poeppl, A., Jurisica, I., and Dick, J.E. (2011). Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. Science *333*, 218–221.

Osawa, M., Nakamura, K., Nishi, N., Takahasi, N., Tokuomoto, Y., Inoue, H., and Nakauchi, H. (1996). In vivo self-renewal of c-Kit+

Sca-1+ Lin(low/-) hemopoietic stem cells. J. Immunol. *156*, 3207–3214.

Provençal, N., and Binder, E.B. (2015). The effects of early life stress on the epigenome: From the womb to adulthood and even before. Exp. Neurol. *268*, 10–20.

Qu, W., Tsukahara, T., Nakamura, R., Yurino, H., Hashimoto, S., Tsuji, S., Takeda, H., Morishita, S., Guo, H., Smallwood, S.A., et al. (2016). Assessing cell-to-cell DNA methylation variability on individual long reads. Sci. Rep. *6*, 21317.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Quivoron, C., Couronné, L., Della Valle, V., Lopez, C.K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.-H., et al. (2011). TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell *20*, 25–38.

Rishi, V., Bhattacharya, P., Chatterjee, R., Rozenberg, J., Zhao, J., Glass, K., Fitzgerald, P., and Vinson, C. (2010). CpG methylation of half-CRE sequences creates C/EBP binding sites that activate some tissue-specific genes. Proc. Natl. Acad. Sci. USA *107*, 20311–20316.

Sanjuan-Pla, A., Macaulay, I.C., Jensen, C.T., Woll, P.S., Luis, T.C., Mead, A., Moore, S., Carella, C., Matsuoka, S., Jones, T.B., et al. (2013). Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. Nature *502*, 232–236.

Shlush, L.I., Mitchell, A., Heisler, L., Abelson, S., Ng, S.W.K., Trotman-Grant, A., Medeiros, J.J.F., Rao-Bhatia, A., Jaciw-Zurakowsky, I., Marke, R., et al. (2017). Tracing the origins of relapse in acute myeloid leukaemia to stem cells. Nature *547*, 104–108.

Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods *11*, 817–820.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., Tiwari, V.K., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Tsuboi, K., Nagatomo, T., Gohno, T., Higuchi, T., Sasaki, S., Fujiki, N., Kurosumi, M., Takei, H., Yamaguchi, Y., Niwa, T., et al. (2017). Single CpG site methylation controls estrogen receptor gene transcription and correlates with hormone therapy resistance. J. Steroid Biochem. Mol. Biol. *171*, 209–217.

Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sánchez Castillo, M., Oedekoven, C.A., Diamanti, E., Schulte, R., et al. (2015). Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. Cell Stem Cell *16*, 712–724.

Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L., and Ren, B. (2012). Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell *148*, 816–831.

Yamamoto, R., Morita, Y., Ooehara, J., Hamanaka, S., Onodera, M., Rudolph, K.L., Ema, H., and Nakauchi, H. (2013). Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell *154*, 1112–1126.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

Yu, V.W.C., Yusuf, R.Z., Oki, T., Wu, J., Saez, B., Wang, X., Cook, C., Baryawno, N., Ziller, M.J., Lee, E., et al. (2016). Epigenetic memory underlies cell-autonomous heterogeneous behavior of hematopoietic stem cells. Cell *167*, 1310–1322.e17.

Zhang, W., Spector, T.D., Deloukas, P., Bell, J.T., and Engelhardt, B.E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. *16*, 14.

Zhou, S., Shen, Y., Zheng, M., Wang, L., Che, R., Hu, W., Li, P., Zhou, S., Shen, Y., Zheng, M., et al. (2017). DNA methylation of METTL7A gene body regulates its transcriptional level in thyroid cancer. Oncotarget *8*, 34652–34660.

Stem Cell Reports, Volume *11*

Supplemental Information

High-Resolution Single-Cell DNA Methylation Measurements Reveal Epigenetically Distinct Hematopoietic Stem Cell Subpopulations

Tony Hui, Qi Cao, Joanna Wegrzyn-Woltosz, Kieran O'Neill, Colin A. Hammond, David J.H.F. Knapp, Emma Laks, Michelle Moksa, Samuel Aparicio, Connie J. Eaves, Aly Karsan, and Martin Hirst

Supplemental Figures

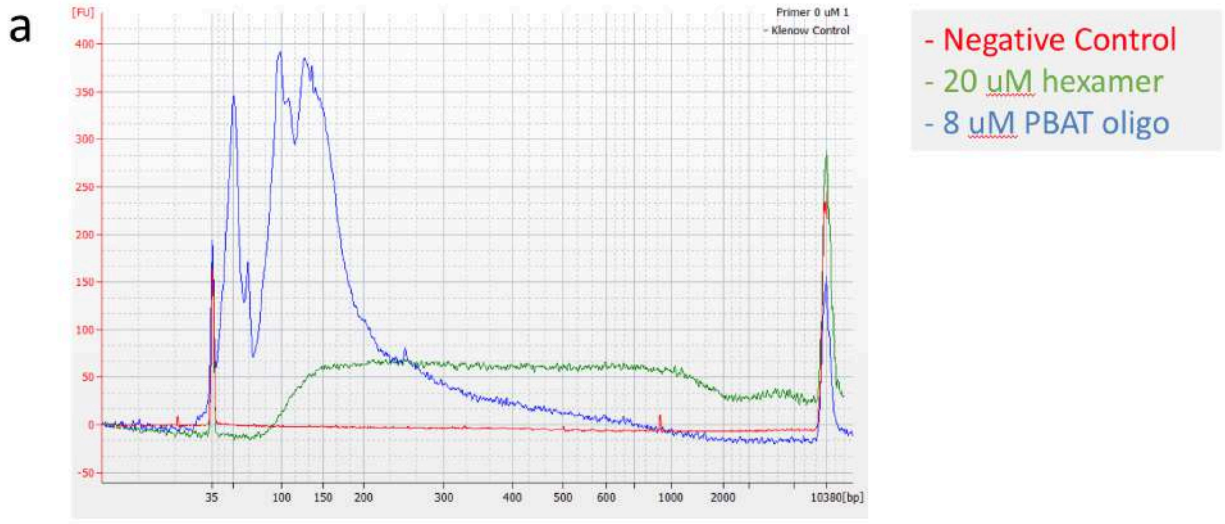

Figure S1. Increased length of library insert fragments using untagged random hexamers compared to end-tagged primers. Related to Figure 1. a) Illumina sequencing libraries were constructed from dsDNA generated using either an untagged random hexamer (**hexamer, green trace**), 5' tagged PBAT oligos (**PBAT oligo, blue**) or no primers (**Negative Control, red**) and the insert sizes and yields measured using an Agilent Bioanalyzer (Agilent, Santa Clara, CA).

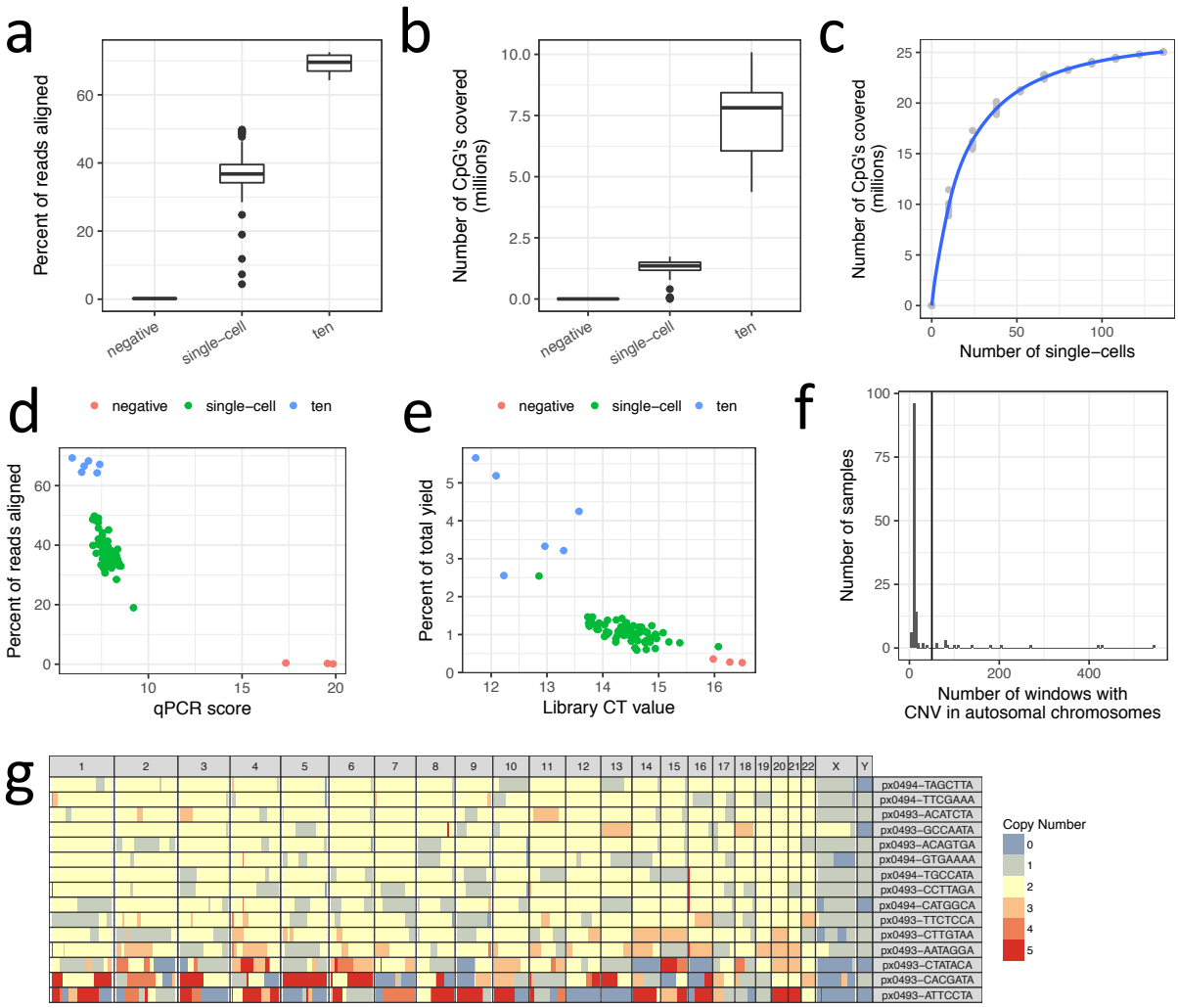

Figure S2. Quality control of human single-cell libraries. Related to Figure 1 a) Sequence reads generated from single-cells demonstrate a higher frequency of alignments compared to negative (no cell) controls. The number of reads that aligned to the reference genome (hg19) divided by the total number of sequenced reads is plotted. **b)** The number of unique CpG dinucleotides positions recovered from each library. **c)** In silico merging of CpG positions obtained from single HSCs recapitulates the bulk epigenome. Y-axis shows the number of unique CpG positions recovered by combining CpG positions from randomly selected HSCs *in silico* and the x-axis represents the number of cells merged. Each grey dot represents one iteration of combining *x* cells. **d)** Alignment rate correlates with the qPCR score. The qPCR score represents the Target (genome-specific) CT value minus the Library (Illumina adapter-specific) CT value. qPCR scores are plotted against alignment rate from panel **a** on the x-axis. Only CD49f cells from donor 2 are plotted. **e)** Single-cell and 10-cell samples result in CT values that are consistently lower than negative (no cell) controls. The percentage of total yield was calculated by the number of reads assigned to each sample divided by the total number of reads generated. Library CT value is the CT value of samples amplified using Illumina adapter-specific sequences **f)** Histogram of the number of autosomal CNV calls in 5 MB windows per sample. Samples with many CNV calls were considered technical and/or biological artifacts and removed. **g)** CNV profiles of the single cells with many CNVs from **f**. In each window, the color represents copy number, rows represent cells and columns representing chromosomes. Regions reported with a copy number >5 were rescaled to a maximum of 5.

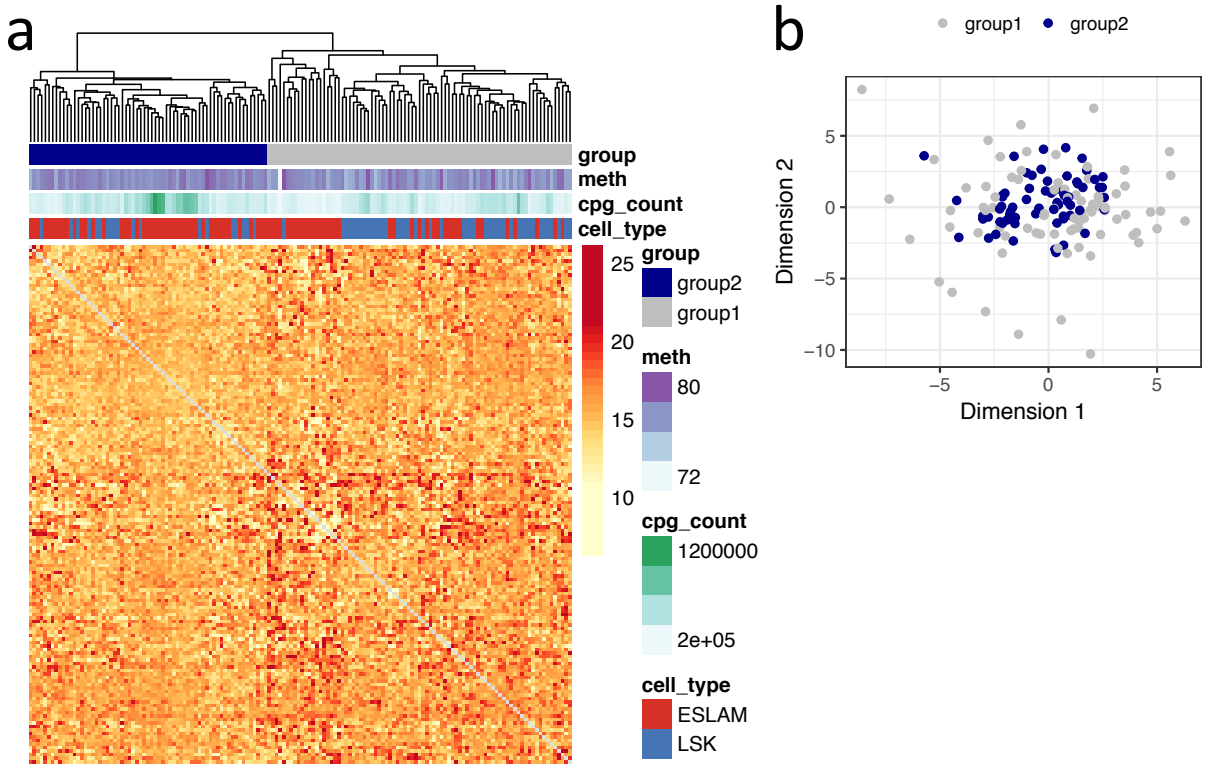

Figure S3. Application of PDClust to CpGs located in cortex specific enhancer states results in no clear structure. Related to Figure 3. Clustering (**a)** and MDS scaling (**b)** of single-cells using measurement for CpGs that lie within cortex enhancers show no pattern. **meth:** Average genomewide CpG methylation; **cpg_count**: Number of distinct CpG sites recovered; **cell_type:** cell from which CpG methylation was obtained.

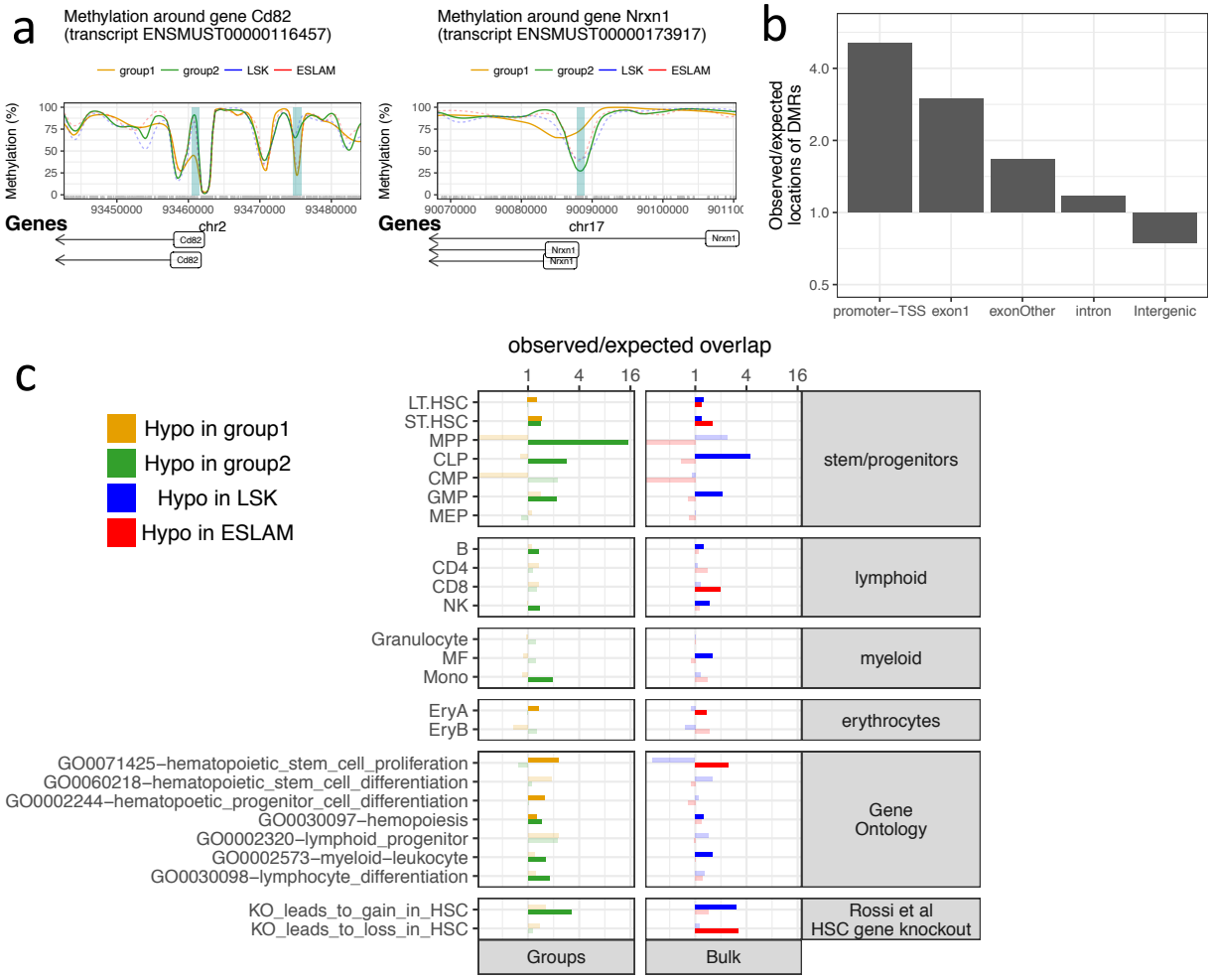

Figure S4. **DMR discovery and annotation reveals putative function of HSC subgroups. Related to Figure 3.** (**a**) Example DMRs near genes implicated in hematopoiesis. Methylation values were smoothed for each in silico merged population of cells with Bsmooth and plotted. Tick marks on the x-axis represent the location of CpG dinucleotides. DMRs identified using smoothed methylation calls are highlighted in blue. The genes track shows protein-coding transcripts obtained from Ensembl v85 with an evidence level of 1 or 2. (**b**) Observed over expected enrichment of DMRs calculated by dividing the fraction of all DMRs that overlap each region set by the fractional genomic occupancy of each region set. (**c**) Gene enrichment of DMRs specifically hypomethylated in each comparison. The comparisons between group 1 and group 2 were separate from the comparisons between ESLAM and LSK cells. Bars are faded if the FDR-corrected binomial q-value was >0.1.

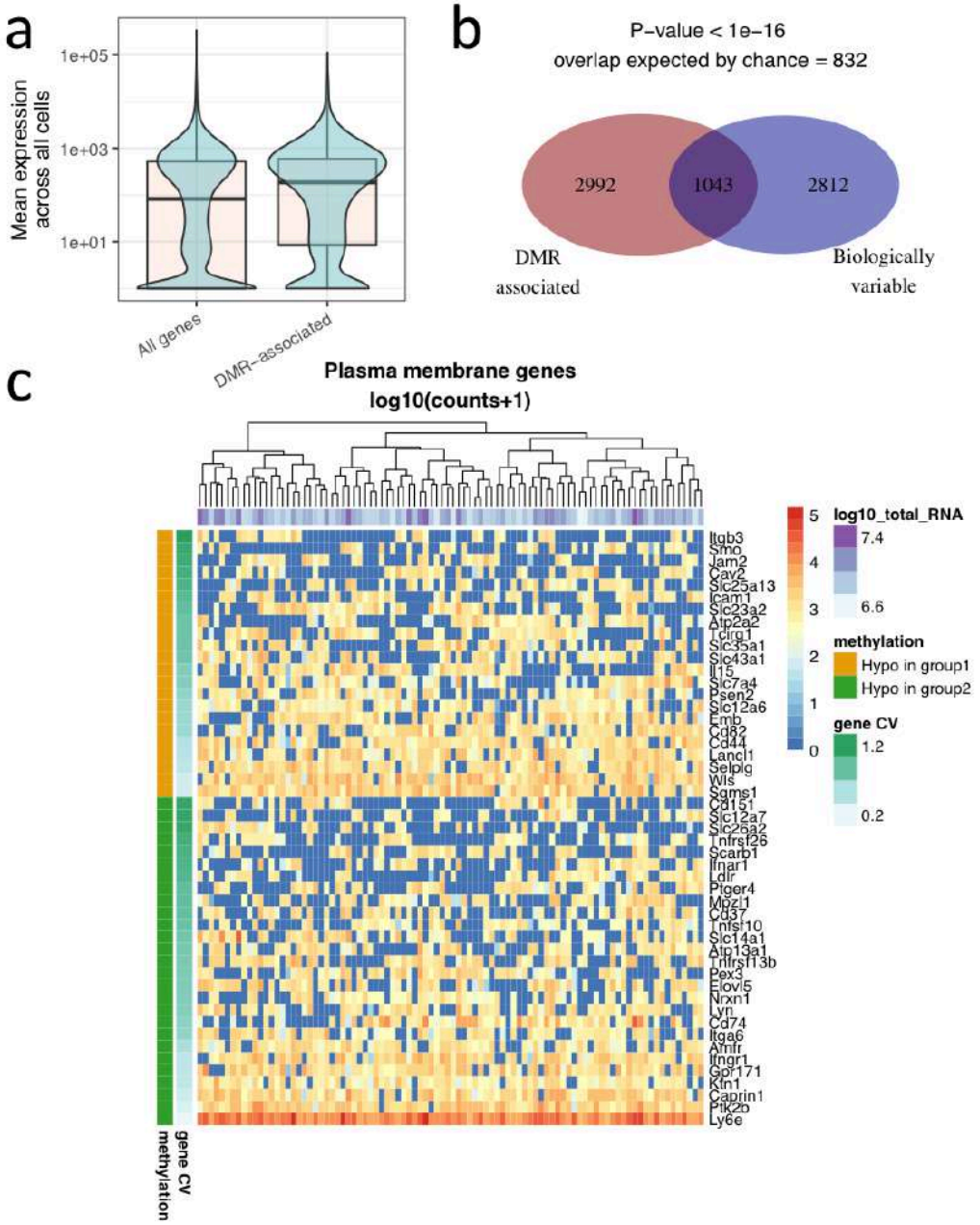

Figure S5. Single-cell expression of genes associated with DMRs in LSK and ESLAM cells. Related to Figure 3. (**a)** Genes associated with DMRs are expressed to a greater level than the genome-wide average. RNA-seq expression values were obtained from single-cell datasets exclusively and the read count for each single-cell was normalized and corrected (Wilson et al., 2015; see **supplemental methods**). The average expression across all single cells was calculated for each gene, plotted as a boxplot and associated with the mean expression on the y-axis. (**b)** A statistically significant overlap exists between DMR-associated genes and heterogeneously expressed genes in the ESLAM cells. Genes heterogeneously expressed at the single cell level were downloaded from (Wilson et al., 2015). (**c)** Heterogeneous expression of genes encoding cell surface markers associated with DMRs across single ESLAM cells. Rows represent genes that encode plasma membrane proteins (as defined by GO), and the columns represent single cells. Cell are coloured to represent the $\log_{10}$(count+1) expression of each gene. Gene CV is the coefficient of variation of the expression of each gene across all single cells.

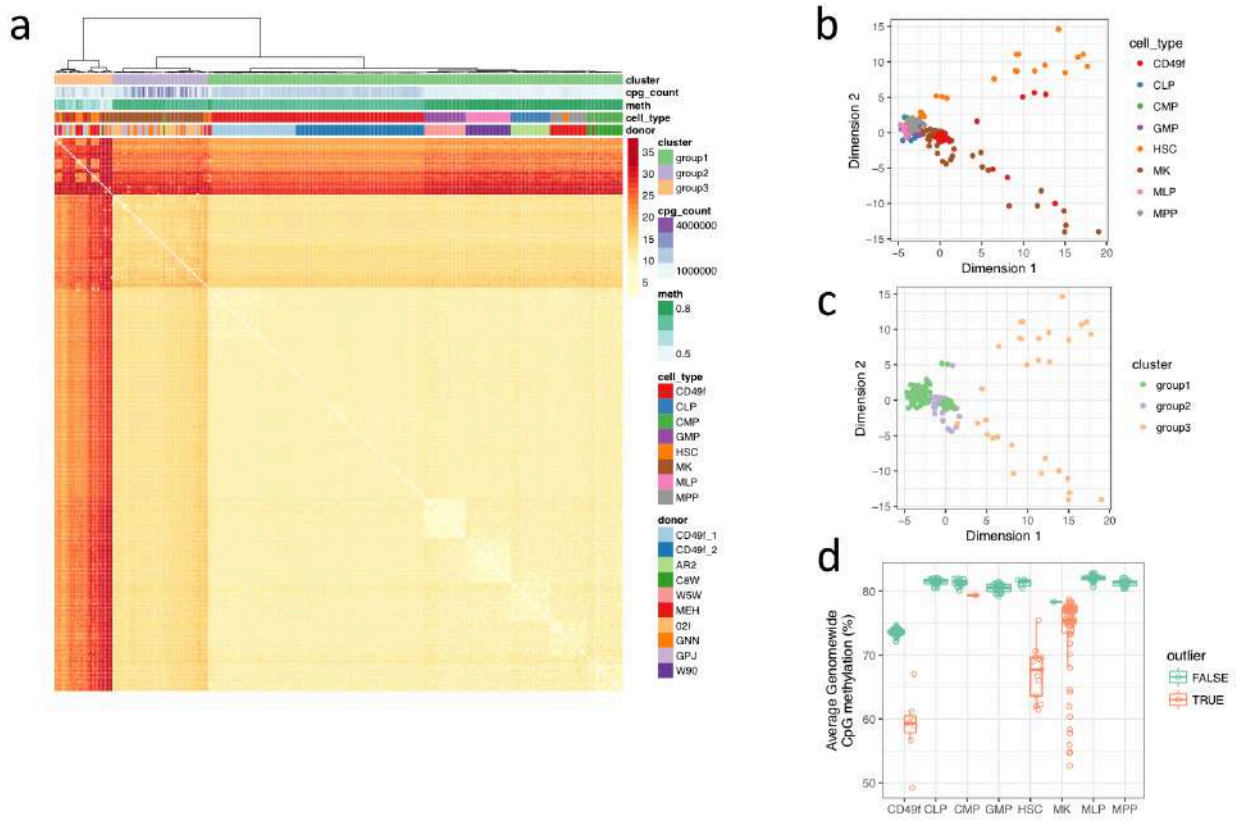

Figure S6. Clustering of different human hematopoietic phenotypes. Related to Figure 5. (**a)** Pairwise dissimilarities reveal group2 and group 3 consist largely of MKs and outlier hypomethylated cells. Cells not classified as group 1 display higher PD values overall. (**b,c)** Multidimensional scaling of PD values reveal group 2 and group 3 cells as outliers in 2D space. (**b)** Distribution of genome-wide methylation. Each dot represents one sample. Outliers are colored in orange remaining are colored in green.

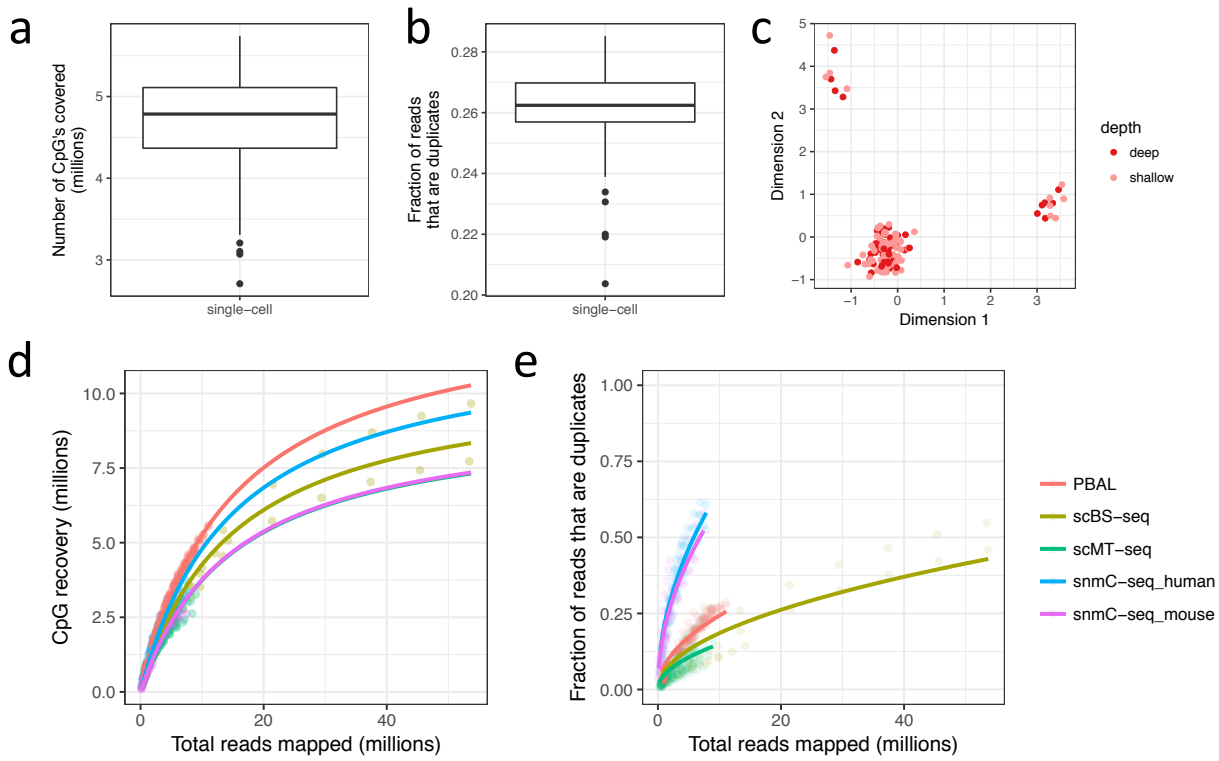

Figure S7. Library quality characteristics and CpG recovery of deeply sequenced single CD49f cells. Related to Figure 7. (a) Distribution of CpG recovery per single-cell library. **(b)** Distribution of duplicate rate per single-cell library. **(c)** PD values of CD49f cells from donor 2 were projected onto 2D space using MDS. Each dot represents a single CD49f cell. Sequencing depth of single-cells is independent of PD and indicated by color with dark red indicating increased sequencing. **(d)** A scatterplot of reads mapped versus the number of unique CpG sites recovered in single-cell libraries. For each method (PBAL **red**; scBS-seq, **gold**; scMT-seq, **green**; snmC-seq human, **blue**; snmC-seq mouse, **purple**), libraries were down-sampled and then CpG methylation called for each down-sampled library. Individual points represent a down-sample of a library within a method. The line represents the y = x/(1+ax) lines (where a is any constant) of a best fit for all subsampled data points for all libraries within each methodology. **(e)** The fraction of duplicate reads is plotted on the y-axis as a function of the number of reads mapped on the x-axis. The lines represent the y = sqrt(x) lines of best fit.

Supplemental Tables

Table S1. Related to Figure 1. Antibodies used for FACS

Cell population	Antibodies
LSK: Lin-$^{-}$Sca1+$^{+}$c-Kit+$^{+}$	Lineage: PerCpCy5.5-conjugated CD3, CD4, CD8a, B220, IL7R, Gr-1, Ter 119; Sca-1-PECy7, c-kit-APC
ESLAM: EPCR^{+}CD45^{+}CD48^{-}CD150^{+}	EPCR-PE; CD45-FITC; CD48-APC; CD150-PECy7

Table S2. Related to Figure 1. Sample metadata for mouse HSPCs

Table S3. Related to Figure 3. DMRs between *in silico merged* populations for murine HSCs

Table S4. Related to Figure 1. Sample metadata for human HSCs

Table S5. Related to Figure 7. DMRs between *in silico merged* populations for human HSCs

Supplemental Experimental Procedures

Isolation of cells

Bone marrow was harvested from the femur, tibia and pelvic bones of 8–10 week old, mixed sex C57BL/6 mice, pooled, and subjected to ammonium chloride-mediated red blood cell lysis. The cells were stained with LSK or ESLAM staining cocktails (Table S1). We verified that the 4-6 month repopulating potential of our sorted ESLAM and LSK cells were equivalent (40% and 4% respectively, data not shown) to previously published studies (Kent et al., 2009; Osawa et al., 1996). Single cells of each phenotype were sorted directly into 384 well plates (ThermoFisher, 4309849) using a FACSAria Fusion (BD, Franklin Lakes), flash frozen using liquid nitrogen, and stored at -80°C until processing.

Cord blood was obtained from 2 normal full-term deliveries with informed consent from the mothers and CD49f cells isolated within 48 hours by index sorting on a FACSAria as described previously and elsewhere (Knapp et al., 2017; Knapp et al., submitted)

DNase I treatment of silica beads

To decontaminate MagSi-DNA allround magnetic silica beads (MagnaMedics, MD02018), we aliquoted the required volume of beads into few wells of a 96-well plate (ThermoFisher, AB1400L), collected the beads on a magnet (Alpaqua, A000350), removed the supernatant, resuspended the beads in an equal volume of DNase I mix (0.08 U/uL DNAse I, 1X DNAse I reaction buffer (ThermoFisher, AM2222), and incubated the suspension in a GeneAmp 9700 PCR machine (ThermoFisher, 4413750) at 37˚C for 30 minutes, 70˚C for 10 minutes. We then aliquoted the beads (2 µL/well) into a clean 96-well plate, mixed with 180 µL of MethylEdge Binding Buffer (Promega, N1301), and UV treated each sample as described below.

UV treatment

Prior to use, we UV treated using a TL-2000 translinker (UVP, 95-0300-01, setting: UV crosslink: 60 minutes) the following reagents: DNase I treated silica beads (resuspended in MethylEdge Binding Buffer), MethylEdge Desulphonation buffer, 10mM Tris-CL pH 8.5 (Qiagen, 19086), 80% EtOH (home-made), and NEB 2 buffer (NEB, B7002S).

Cell Lysis and Bisulfite treatment

Plates containing frozen single cells were thawed and centrifuged at 3000 rpm at 4°C for 2 minutes. Four µL of lysis buffer containing 20 mM Tris-HCl, pH 8.0, 20 mM KCl, 0.3% Triton-X 100, 1 mg/mL Serine Protease (Qiagen, #19155) was added to each well with the plate on ice including 3 empty wells to serve as negative controls, followed by centrifugation at 3000 rpm at 4°C for 1 minute. Resulting lysates were transferred into a clean 96-well plate (ThermoFisher, AB1400L) and incubated in a GeneAmp 9700 PCR machine (ThermoFisher, 4413750) at 50°C for 30 minutes. One µL of unmethylated 60 fg/µL T7 Phage DNA (GeneON, #301-025) was then added to each murine single-cell well except for the negative control wells, whereas the human single-cells contained unmethylated lambda and fully methylated T7 Phage as conversion controls. The mixtures subjected to bisulfite conversion using the MethylEdge Bisulfite Conversion kit (Promega, N1301) using a bead-based protocol to enable its automation (Domanico et al., 2013). Accordingly, ~5 µl of each single-cell lysis/spike-in mixture was combined with 32.5 µL of MethylEdge Conversion reagent and incubated in the GeneAmp 9700 PCR machine (98°C for 8 minutes, 54°C for 60 minutes). Each plate was then spun down at 3,000 rpm for 1minute. All subsequent steps were performed on a Bravo Automated Liquid Handling Platform (Agilent Technologies, G5409A) with 96LT Disposable Tip head, 250uL sterile, filtered tips (Agilent Technologies, 19477-022) and custom programs created with VWorks Automation Control Software (Agilent Technologies, USA). Bisulfite-converted DNA was mixed with 180 µL of MethylEdge Binding Buffer and 1.8 µL of 20 mg/ml decontaminated MagSi-DNA allround silica beads (MagnaMedics, MD02018) and left at room temperature for 15 minutes. The DNA containing beads were collected to the side by placing the plate on a magnet (Alpaqua, A000350) for 3 minutes. While on the magnet, the beads were washed twice with 220 µL of 80% ethanol for 30 seconds without resuspension. Next, 60 µL of MethylEdge desulfonation buffer was added to the beads and the mix incubated at room temperature for 15 minutes. After the removal of the desulfonation buffer, the beads were washed twice with 100 µL of 80% ethanol while still on a magnet and hence without resuspension, and then air-dried for 1 minute. To elute the DNA, the beads were resuspended in 20 µl 10 mM Tris-HCL, pH 8.5 (Qiagen, 19086) and incubated in a Thermomixer C (Eppendorf, 5382000015) at 56°C while being centrifuged at 2,000 rpm for 15 minutes. The beads were then collected to the side using a magnet for 30 seconds and the DNA containing supernatant transferred to a new 96-well plate.

Double-stranding reaction

The bisulfite-converted scDNA was mixed with 1.25 µl 10 mM dNTPs and 1 µl 500 µM random hexamers (3' phosphothioate, NNNN*N*N), incubated at 98°C for 1 minute, and then snap frozen on ice for 2 minutes. A mix of 0.5 µl 50 U/µl Klenow exo- (NEB, M0212M) and 2.5 µl 10X NEB Buffer 2 was added and reactions incubated in a GeneAmp 9700 PCR machine at 4˚C for 10 minutes, + 4°C/second to 37ºC, and at 37°C for 30 minutes. Next, reactions were denatured again at 98˚C for 1 minute, snap frozen on ice for 2 minutes, and transferred to a new plate containing 5µl 2nd DNA synthesis mix (20 µM random hexamers, 0.5 mM dNTPs, 1X NEB 2 buffer, 25 U Klenow fragment exo-). Samples were then incubated in a GeneAmp 9700 at 4˚C for 10 minutes, +4°C/s to 37°C, at 37°C for 30 minutes, and at 70°C for 10 minutes. 20 µl 10 mM Tris-HCL (pH 8.5) was then added and reactions purified at a 1:1 ratio using an in-house prepared magnetic bead solution (1M NaCL, 20% PEG, Sera-Mag Speedbead (Fisher Scientific, 09981123)) following a Beckman-Coulter Ampure XP PCR purification protocol. The DNA was then eluted in 35 µl 10 mM Tris-HCl (pH 8.5) and used for Illumina library generation.

Library construction

All liquid handling steps were carried out on the Bravo Automated Liquid Handling Platform (Agilent Technologies, G5409A) with 96LT Disposable Tip head and 250uL sterile, filtered tips (Agilent Technologies, 19477-022) and custom programs created with VWorks Automation Control Software (Agilent Technologies, USA). An End Repair and 5' Phosphorylation reaction (35 µl DNA sample, 5 µl 10X NEB 2 buffer, 2 µl 25mM ATP, 2 µl 10mM dNTP, 10 U T4 Polynucleotide Kinase, 4.5 U T4 DNA Polymerase, 1 U Klenow Large Fragment DNA Polymerase, and ultrapure water to a total reaction volume of 50 µl (NEB, E6000B-10)) was carried out at room temperature for 30 minutes. Single cell DNA samples were then purified at a 1:1 ratio using an in-house prepared magnetic bead solution (1M NaCL, 20% PEG, Sera-Mag Speedbeads (Fisher Scientific, 09981123)) following Beckman-Coulter Ampure XP PCR purification protocol, and eluted in 25 µl volume with 10 mM Tris-CL (pH 8.5). To enable ligation to the adaptors, a single dA overhang was added to the 3' ends of DNA fragments (25 µl DNA, 3.5 µl 10X NEB 2 buffer, 0.7 µl 10 mM dATP, 3.5 U Klenow Fragment (3´→5´ exo–), and ultrapure water to a total reaction volume of 35 µl, (NEB, E6000B-10)). The dA-addition reaction was incubated in a GeneAmp 9700 PCR machine at 37 ^{0}C for 30 minutes. Next, short adaptors containing sequences required downstream in the sequencing workflow were ligated to the dA-tailed DNA fragments (35 µl DNA, 12 µl 5X Quick Ligation Buffer, 2,000 U Quick T4 DNA Ligase, 2 µl 0.5 µM Illumina sequencing forked adaptor, and ultrapure water to a total volume of 60 µl (NEB, E6000B-10)). The ligation reaction was performed at room temperature overnight. To remove adaptor dimers, the ligation product was purified twice using in-house prepared magnetic bead solution (1M NaCL, 20% PEG, Sera-Mag Speedbeads) - first at 0.8X (eluted in 50 µl 10mM Tris-Cl, pH 8.5) and then at a 1:1 ratio (eluted in 25 µl 10mM Tris-Cl, pH 8.5). Adaptor ligated libraries were PCR-amplified and barcoded using custom indexing primers (25 µl DNA, 10 µl 5X High fidelity buffer, 1 µl 10 mM dNTPs, 1.5 µl DMSO, 1 µl 25uM forward primer, 2 µl 12.5 µM custom reverse indexing primer (added separately to each well), 1 U Phusion U Hot Start, and ultrapure water to a total volume of 50 µl (ThermoFisher, F-555L)). 8 cycles of PCR amplification was carried out using GeneAmp 9700 PCR machine with the following cycling conditions: 98^{0}C for 1 minute, 8 repeats of (98^{0}C for 30 seconds, 65^{0}C for 15 seconds, 72^{0}C for 15 seconds), 72^{0}C for 5 minutes, and 4^{0}C hold. Barcoded libraries were size-selected to remove primer dimers with 0.8X in-house prepared magnetic bead solution (1M NaCL, 20% PEG, Sera-Mag Speedbeads) and eluted in 15 µl 10 mM Tris-CL, pH 8.5. Based on Real-Time PCR analysis (below), successful single-cell libraries were selected for sequencing and pooled together. Pooled libraries were volume-reduced by ethanol precipitation, visualized using High-Sensitivity DNA chip on the Agilent Bioanalyzer, and sequenced 125-bp paired-end on a HiSeq2500 using V3 chemistry.

qPCR quality control and pooling of constructed libraries

Libraries were checked by qPCR before pooling. Real-Time PCR using primers against Illumina library sequences was performed in a 384 well MicroAmp plates (ThermoFisher, 4309849) in a 10 µl volume (1X KAPA SYBER FAST master mix, 1X Illumina Primer Premix (KAPA, KK4824), and 1 µl of 1/100 diluted PCR-amplified single-cell library). Real-Time PCR using primers specific to bisulfite converted DNA (Mouse forward: AAAATTGAAAATTATGGAAAATGAG, reverse: CCAAATCCTTCAATATACATTTCTC; Human forward: TGTTTGTAAAGTTTATAAGTGGATA, reverse: CAAAAAAATTACTAAAAATTCTTCT) was performed in a 10 µl volume (1X KAPA SYBER FAST master mix, 0.5 µM each of forward and reverse primers, and 1 µl PCR-amplified single-cell library). Reactions were cycled using ViiA 7 Real-Time PCR system (ThermoFisher, 4453536) as follows: 98˚C for 1 minute, 40 repeats of (98˚C for 30 seconds, 60˚C for 30 seconds, 72˚C for 45 seconds); and

72˚C for 5 minutes. Genome-specific CT values were normalized (by subtraction) to CT values of the library, and a cutoff determined empirically based on the distribution of single-cells and negative controls.

Single-cell libraries that passed QC were pooled together and concentrated by ethanol precipitation (0.1X 3M Sodium Acetate, 2.5 µl 20 mg/ml mussel glycogen (Sigma, 10901393001), 2.5X 100% ethanol). After incubation at 20°C for 1 hour, centrifugation (4˚C, 20,000 g for 40 minutes) and a 75% ethanol wash, the single cell library pool was air-dried for 5 minutes at room temperature, and then resuspended in 15 µl 10 mM Tris-Cl, pH 8.5. The pool was assessed for quality and size using High-Sensitivity DNA Chip on the Agilent Bioanalyzer and sequenced 125-bp paired-end on Illumina HiSeq2500 sequencing platforms (v3 chemistry) following the manufacturer's protocols.

Bulk PBAL construction

10,000 cells were deposited directly by the flow cytometer into an eppendorf tube and DNA extracted using an AllPrep DNA/RNA Mini Kit (Qiagen, 80204) according to manufacturer's instructions. 4 µl purified DNA (25 ng/µl) was mixed with 1 µl T7 Phage DNA spikein (1 ng/µl), and the mixture used directly as input to bisulfite conversion and desulphonation as described for the single-cells, except that neither DNase nor UV decontamination was performed on the reagents, and 5 µl of beads was used instead of 1.8 µl. After bisulfite conversion and on-bead desulphonation, the DNA was eluted in 40 µl EB buffer and then used as input for one round of double-stranding with 2.5 µl 10 mM dNTPs, 2 µl 500 µM random hexamers, 1 µl 50 U/µl Klenow exo- (NEB, M0212M) and 5 µl 10X NEB Buffer 2. Library construction was the same as for the single cells except that only 4 rounds of PCR were used. Libraries were sequenced directly without qPCR QC.

Raw data processing

The first 6 bases of read 1 and read 2 were trimmed using Trimgalore v0.4.0 and Cutadapt v1.2.1 (Martin, 2011) using the parameters --clip_R1 6 --clip_R2 6 –paired. Trimmed fastq files were aligned using Novoalign V3.02.10 (www.novocraft.com) to the mouse assembly GRCm38 (mm10) or human assembly GRCh37 (hg19). Alignments were done in paired-end mode using the options -b4 (non-directional), -t 20,3 (optimized for SNP concordance), –a (adapter trimming), --hlimit 8 (homopolymer filter), and -H 20 (removes trailing bases with quality ≤20). For the mouse cells, we used –u8 (penalty for unconverted CHG or CHH cytosine) as recommended by Novoalign, whereas for human cells we used –u50 due to the presence of unconverted human contaminants. Aligned reads were sorted using SAMtools V0.1.17 and deduplicated with Picard V1.31 (http://picard.sourceforge.net). Methylation of each cytosine was called using SAMtools mpileup (-B -C 0 -q 30 -d 500) followed by Novomethyl V1.01 (-o Consensus -%) (www.novocraft.com). Cytosine calls with a Phred quality score ≤15 were discarded. Fractional methylation was merged for each CpG di-nucleotide (found by searching the genome with a custom java script) by taking a weighted average of each cytosine using the map command from the bedtools suite (Quinlan and Hall, 2010). In most cases, only one base within a CpG dinucleotide had coverage. In these cases, the methylation information of the covered base was extrapolated to the other base. Processed CpG calls were imported into R V3.3.2 for downstream analysis. We considered only autosomal CpG sites and CpG sites with a methylation value of 0% or 100%. CNVs in 5 MB windows were called using Control_FREEC V7.0 (Boeva et al., 2012) with default parameters. Single cells with conversion rates <96%, mapability <5%, <130,000 CpGs, and containing more than 50 windows with CNVs were removed.

Obtaining genomic regions

For the mouse data analysis, we downloaded BED file annotations for each of our genomic region sets as follows: for CpG Islands, we downloaded the CpG Island track from the UCSC genome browser; for CpG Island shores, we took the flank of each CpG island by extending each island by 2kb; for LINEs, SINEs, and LTRs, we downloaded the Repeatmasker track from the UCSC genome browser and filtered by category; for gene bodies, considering every protein-coding V85 Ensembl transcript with evidence level of 1 or 2; for DMRs in blood differentiation (Blood Lineage DMRs), we downloaded the list of DMRs from Bock et al (Bock et al., 2012); for DMRs in the ESLAM to MPP transition (HSC Regulatory Networks), we downloaded the list of DMRs from Cabezas-Wallscheid et al (Cabezas-Wallscheid et al., 2014); for blood enhancers, we downloaded the enhancer catalogue from Lara-Astiaso et al (Lara-Astiaso et al., 2014); for human cells, we considered gene bodies as every protein-coding V75 Gencode transcript with a support level of 1 or 2.

Differentially methylated regions analysis

To group cells that belonged to the same cluster, we treated coverage of every CpG site as the number of cells with coverage at that site, and treated methylation fraction as the fraction of cells that had a methylated CpG at that site.

We used BSmooth (Hansen et al., 2012) to obtain estimated CpG methylation at all CpG sites in the genome. To identify differentially methylated CpGs (dCpGs), we calculated the mean and standard deviation of the difference in CpG methylation between the 2 groups of cells being compared. Then, for each CpG comparison between the two groups, we calculated the z-score (how far away from the mean the difference in methylation was, in units of standard deviation) and calculated a 2-tailed p-value using the z-score assuming a normal distribution (pnorm function in R) with the null hypothesis that the methylation is not different between the 2 CpGs. Finally, the p-values were multiple test-corrected using a false discovery rate estimate. To call DMRs, we grouped dCpGs together if they were within 500 bp of each other and only considered regions with ≥3 CpGs.

To calculate the enrichment of DMRs in different genomic locations, we first calculated the expected overlap as the fraction of the genome occupied by each genomic feature (intergenic, promoter, homologous, etc.). To calculate the observed overlap, we calculated the total genomic occupancy of DMRs overlapping each feature divided by the total genomic occupancy of all DMRs. The observed/expected ratio was then obtained as the ratio of these 2 numbers.

GSEA

We first split DMRs into 2 groups depending on which population had the lower methylation in each pair of comparisons. To associate DMRs to genes, we found the nearest protein coding transcript for each DMR and calculated the distance to the TSS of that transcript. However, for DMRs that lie within exons or introns, we used the TSS of the transcript that the DMR was found in instead of the TSS of the closest transcript. We further filtered DMRs based on our criterion of distance to TSS, coverage, and genomic context. For the remaining DMRs, we represented them based on the gene they were associated with (DMR gene list). We removed genes associated with DMRs from both groups and used the remaining DMRs for further analysis. For each group, we first calculated the expected proportion of overlap by dividing the number of genes in each gene set by the total number of autosomal genes. We then calculated a binomial p-value as the probability that an equal or higher number of DMRs that overlap with each gene set by chance given the number of tries as the number of DMRs for each group. P-values were multiple-test corrected using the false discovery rate method.

For the mouse cell data, we downloaded gene sets from genes published as controlling HSC numbers (Rossi et al., 2012), the relevant gene sets from the Gene Ontology database (Carbon et al., 2009), and built a list of preferentially expressed genes for each cell type if their expression was more than 20% compared to any other cell type (Lara-Astiaso et al., 2014). For the human cell data, we downloaded MsigDB (Liberzon et al., 2015) and considered all terms that included “HEMATO”, “STEM_CELL” or “LEUKEMI” in the term name. For each gene set, we only considered autosomal genes for analysis.

Single-cell RNA-seq analysis

We downloaded processed read counts of $Lin^{-}c\text{-}kit^{+}Sca\text{-}1^{+}CD34^{-}Flt3^{-}CD48^{-}CD150^{+}$ HSCs from GSE61533, and downloaded the list of biologically variable genes from the supplemental materials of Wilson *et al* (Wilson et al., 2015). We removed failed cells according to the criteria the authors established, and normalized reads to transcripts per million reads sequenced. Mouse plasma membrane genes were identified by their membership to GO:0031226 (intrinsic component of plasma membrane).

Supplemental References

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., and Lewis, S. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288–289.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. *1*, 417–425.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rossi, L., Lin, K.K., Boles, N.C., Yang, L., King, K.Y., Jeong, M., Mayle, A., and Goodell, M.A. (2012). Less is more: unveiling the functional core of hematopoietic stem cells through knockout mice. Cell Stem Cell *11*, 302–317.