EPIGENETIC HETEROGENEITY REVEALED THROUGH SINGLE-CELL DNA METHYLATION SEQUENCING

by

Zhao Kun (Tony) Hui

B.Sc. Honours with Distinction, The University of British Columbia, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2018

Abstract

Increasing evidence of functional and transcriptional heterogeneity in phenotypically similar single-cells has prompted interest in protocols for obtaining parallel methylome data. Despite appreciable advancements in experimental protocols for single-cell DNA methylation measurements, methods for analyzing the resulting data are still immature. To address the challenge of stochastic data loss associated with single cell measurements, current strategies average methylation in windows or region sets. However previous studies have demonstrated that single CpGs are functional and our analysis of single cell methylation measurements revealed a rapid decay in concordance neighbouring CpG states beyond 1kb. To leverage the information content of individual CpGs in the context of single cell methylation measurements we developed an analytical strategy for deriving single-cell DNA methylation states using individual CpGs, which we term PDclust. We validated PDclust on existing datasets and on data we generated from single index-sorted murine and human hematopoietic stem cells (HSCs) that are highly enriched in functionally defined stem cells. Using PDClust, we identified epigenetically distinct subpopulations within these HSC populations. Strikingly, human cord blood derived HSC populations were separable by donor specific methylation states whereas more differentiated hematopoietic cells separated solely by cell type. Interestingly, removal of methylation sites near genetic variants did not impact this separation, suggesting that these epigenetic states may be a consequence of environmental differences. Finally, through protocol optimization and deep sequencing we generated one of the most comprehensive sets of single cell methylome profiles (20% of CpGs on average) and from these were able to generate genomewide profiles from as little as 6 epigenetically related HSCs to derive subtype-specific regulatory states.

Lay Summary

Epigenetics is the study of heritable changes in gene expression that do not involve changes to the underlying DNA sequence. By measuring epigenetic states, one can infer biological function.

To date, most epigenetic measurements are generated from populations of cells. While useful, these measurements only reflect a population average. Recent discoveries revealed that most populations are comprised of unique cells with different cellular functions. These studies raise the question: do the epigenetic states of these cells also differ?

To answer this question, we developed a method to study epigenetics in single-cells. We applied our method to study blood stem cells responsible for the maintenance of blood cell hemostasis during the lifetime of an organism. Our study revealed rare blood stem cells with different epigenetic states, and we used this information to infer function. These insights may be useful to improve the process of generating blood stem cells in the future.

Preface

Dr. Martin Hirst initially identified the research program and I was the lead investigator that was responsible for designing all of the experiments, conceptualizing the analytical framework to analyze the data, and performed all of the data analysis. The initial invention of the PBAL method described here was done by me in the gap year preceding my Master's degree. Qi Cao helped design and perform the experiments outlined in chapter 2 and generated the sequencing data outlined in chapter 4. Joanna isolated cells for the data outlined in chapter 3. Colin and David isolated cells for the data outlined in chapter 4. Sam Aparicio, Aly Karsan, and Connie Eaves helped interpret the data.

Portions of chapters 2, 3, 4 and 5 has been submitted for publication:

Tony Hui, Qi Cao, Joanna Wegrzyn-Woltosz, Kieran O'Neil, Colin A. Hammond, David J.H.F. Knapp, Emma Laks, Michelle Moksa, Sam Aparicio, Connie J. Eaves, Aly Karsan, and Martin Hirst (2018) Epigenetic diversity of primitive hematopoietic subpopulations from single-cell DNA methylation data.

I generated the mouse single-cell libraries, analyzed and interpreted the data, and wrote the paper. QC helped optimize the single-cell methylation protocol and generated the human single-cell libraries. KO helped with the CpG adjacency analysis. JW, CH and DK isolated the cells analyzed and assisted in data interpretation. EL sorted test single-cells for technology development. MM assisted with technology development, interpreted the data, and wrote the paper. SA and AK interpreted data. MH designed the study, interpreted data and with CE prepared the final draft of the manuscript.

This study involved animal research and is approved by the University of British Columbia under ACC certificate number is 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).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DNA	Deoxyribonucleotide
HSC	Hematopoietic stem cell
HSPC	Hematopoietic stem and progenitor cell
WGBS	Whole-genome bisulfite sequencing
RRBS	Reduced representation bisulfite sequencing
PBAL	Post-bisulfite adapter ligation
PDclust	Pairwise dissimilarity clustering
qPCR	Quantitative polymerase chain reaction
DMR	Differentially methylated region
ESLAM	Murine HSCs sorted with the ESLAM markers (EPCR+CD45+CD48-CD150+)
LSK	Murine HSCs sorted with LSK markers (Lin- Sca1+c-Kit+)
CD49f	Human HSCs sorted with Lin-$CD34^{+}CD38^{-}CD90^{+}CD45RA^{-}CD49f^{+}$
MPP	Multipotent progenitors
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
MLP	Multi-lymphoid progenitor
GMP	Granulocyte macrophage progenitor
MK	Megakaryocytes
SNV	Single-nucleotide variants

Acknowledgements

I thank Dr. Sohrab Shah and Dr. Aly Karsan for being on my thesis advisory committee and their helpful advice, both in science and in life. I thank David Knapp, Colin Hammond, Emma Laks, and Joanna Wegrzyn-Woltosz for their help and expertise in sorting single-cells; without them, there would be no biology. I would also like to thank the other members of the Hirst lab for entertaining scientific discussions and laughing at my jokes.

I thank Dr. Martin Hirst for providing direction to the project in times of confusion, editing my sometimes terrible writing, supporting me through non-academic endeavors, and engaging with my philosophical speculations.

This work was supported by Terry Fox Research Institute Program Projects (Grant #1021, #1074 and #122869); Canadian Institutes of Health Research (CIHR), Genome Canada and Genome British Columbia (CIHR EP1-120589); Canadian Cancer Society grant generously supported by the Lotte & John Hecht Memorial Foundation (Grant #703489); a CIHR-National Science and Engineering Research Council of Canada grant (CHRP 413633); and a Terry Fox Research Institute New Investigator Award (Grant #1039). I was supported by a Canada Graduate Scholarship-Master's award (CGS-M). This research was enabled in part by support provided by WestGrid and Compute Canada (http://www.computecanada.ca/) and Canada Foundation of Innovation (#31343 & #31098). I also acknowledge Canada's Michael Smith Genome Sciences Centre, 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 http://www.bcgsc.ca/about/funding_support.

Dedication

I dedicate this thesis to all those around the world who fight tirelessly for peace; great scientific accomplishments (and funding for those endeavours) are much harder to come by in times of war and conflict.

Chapter 1: Introduction

1.1 The biological importance of single-cells

The genome of an organism provides the instructions to all possible phenotypes; individual genes serve as functional units, encoding for products that serve specific functions. However, the potential of genes can only be realized when a cell decides, through complex regulatory networks, to actively transcribe and translate genes. In multicellular organisms, individual cells utilize different parts of their genomes, thereby displaying different phenotypes and functional capacities. Therefore, the biology of multicellular organisms must be studied as the sum of its parts wherein the functional unit is the cell.

The description and classification of cell types is a primary focus of biologists, but defining a cell type is a difficult problem. First, cell types defined on the basis of anatomical location is insufficient as tissues and organs are comprised of many distinct operational units (e.g. the heart and its many sub-anatomical structures). An alternative approach is to isolate cell(s) with molecular markers (e.g. using fluorescent activated cell sorting) and assess function using a relevant assay. Through advances in technology, the definition of cell types using this approach has become increasingly specific with increasing number of markers. Despite these advances, using molecular markers to define cell types is not complete as functional heterogeneity and plasticity exist within marker defined cell types (e.g. CD4 T-cells and the plethora of T-cell subsets as reviewed in DuPage and Bluestone, 2016). Furthermore, markers used for isolating cell types were for the most part chosen by directed hypotheses, chance, or even convenience rather than by systematic understanding, leaving the potential for other important cellular markers to be discovered.

With the advent of single-cell omics approaches, molecular signatures of a cell can for the first time be comprehensively profiled. Moreover, profiling of entire organs (Macosko et al., 2015) or entire organisms (Cao et al., 2017) without needing to separate cells by classical cell type definitions is becoming a reality due to throughput of methods approaching thousands of cells per experiment. This in turn enables cell type definitions with unprecedented resolution and specificity without being constrained by historically defined markers. For example, using the entire transcriptional output of single-cells has led to discoveries of novel cell types in many organ systems such as the brain (Luo et al., 2017; Zeisel et al., 2015), the retina (Macosko et al., 2015), and blood (Nestorowa et al., 2016; Papalexi and Satija, 2017; Paul et al., 2015). Thus, comprehensive single-cell profiling studies show great promise in refining the definition of cell types.

1.2 Epigenetics and DNA methylation in health and disease

Epigenetics is the study of heritable changes in gene expression that do not involve changes in the DNA sequence. In mammals, single-cells orchestrate their gene expression programs by covalent modification of their chromatin and the study of these modifications and their consequences is the main goal of Epigenetics.

Covalent modifications occur on DNA itself and on the histones that package it. These modifications collectively control transcription factor binding and relax or constrict chromatin. Associations of epigenetic marks to chromatin states have been annotated; therefore, measuring Epigenetic marks in cells allow for annotation of chromatin state and function. For example, mono-methylation of histones at the 4^{th} lysine on the H3 subunit (H3K4me1) is correlated to enhancer activity, and is thought to permit binding of transcription factors.

The most studied modification in mammals is methylation of cytosine nucleotides in DNA (5mC) at CpG dinucleotides. DNA methylation is maintained through cell divisions by DNMT1 (Li et al., 1992), added to cytosines *de novo* by DNMT3a and DNMT3b (Okano et al., 1999), and removed from cytosines by the TET family of enzymes (Ito et al., 2010). The biological importance of DNA methylation is demonstrated by its importance in health and disease: DNA methylation is closely correlated with aging (Horvath, 2013; Sen et al., 2016); DNMT or TET knockout mice result in abnormal DNA methylation and embryonic lethality (Dawlaty et al., 2014; Okano et al., 1999); Mutations in DNMT and TET enzymes are commonly found in leukemia (The Cancer Genome Atlas Research Network, 2013); and DNA methylation levels as well as the expression of DNMT enzymes are abnormal in a variety of cancers (Subramaniam et al., 2014).

Classically, DNA methylation is thought to regulate expression by repression of gene promoters (Jones, 2012). Recent evidence suggests that DNA methylation dynamically regulates expression by modifying transcription factor affinity (Stadler et al., 2011; Yin et al., 2017). Overall, the proper maintenance of DNA methylation is crucial for physiological function.

1.3 Hematopoiesis

Hematopoiesis the process of blood formation wherein a stem cell can differentiate into every blood cell type. HSCs are defined by their ability to recapitulate the entire blood system when transplanted into a myelosuppressed permissive host (Doulatov et al., 2012; Eaves, 2015), and populations of cells enriched in HSC capacity may be isolated by cell surface markers. Recent clonal analyses of serially transplantable mouse HSCs have revealed that they are heterogeneously comprised of distinct subpopulations that have restricted abilities to produce

different types of mature blood cell types at differing frequencies (Benz et al., 2012; Dykstra et al., 2007; Kent et al., 2009; Sanjuan-Pla et al., 2013; Yamamoto et al., 2013).

Epigenetic modifications are critical for normal hematopoiesis as exemplified by the consequences of alterations incurred by disruption DNMT3 enzymes in primitive hematopoietic cells (Challen et al., 2012; Quivoron et al., 2011; Shlush et al., 2017). Moreover, in long-term HSC populations, lineage-specific regulatory regions appear to be epigenetically modified (Lara-Astiaso et al., 2014) and other regulatory regions show change in DNA methylation along differentiation from HSCs to their progeny (Bock et al., 2012; Cabezas-Wallscheid et al., 2014). Thus, individual HSCs may have distinct epigenetic profiles that correspond their heterogeneous functional outputs.

However, most of the epigenetic measurements underpinning these observations represent consensus values experimentally derived from thousands of cells enriched in HSCs or their progeny and thus unable to distinguish 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, evidence that epigenetic heterogeneity does exist amongst individual HSCs has come from genome wide DNA methylation measurements obtained from single-cell and 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 functional capacity remains poorly understood.

1.4 Profiling DNA Methylation at single-cell resolution

1.4.1 Existing molecular biology methods

Whole-genome bisulfite sequencing (WGBS) is currently the "gold standard" for DNA methylation analysis (Hirst and Marra, 2010). In this method, next-generation sequencing libraries are generated and subjected to bisulfite conversion (Cokus et al., 2008; Lister et al., 2008). This reaction converts unmethylated cytosine bases into uracil bases, which are subsequently converted to thymines for sequencing. Following sequencing and specialized alignment (e.g. with Novoalign or Bismark; Krueger and Andrews, 2011) of the resulting library, cytosine methylation can be quantitatively assessed genomewide at a single base-pair resolution. However, because the bisulfite conversion reaction introduces single-stranded nicks and renders 84-96% of the original library fragments unusable (Grunau et al., 2001), the protocol typically requires in excess of 1 microgram of genomic DNA.

We developed a protocol based off of the Post-bisulfite adapter tagging (PBAT, Miura et al., 2012) methodology that generates DNA methylation libraries from single-cells (Post Bisulfite Adapter Ligation (PBAL)). While PBAL was in development, other single-cell protocols based off the PBAT methodology were reported. **Table 1.1** qualitatively summarizes the key differences between published PBAT-based single-cell DNA methylation protocols. A major advance enabled by the PBAL methodology is increased library diversity, increased CpG recovery, and decreased cost compared to other single-cell DNA methylation protocols.

	Random priming	*Adapter attachment*	*Mappability*	*Diversity*
PBAL	2 rounds	Double-stranded ligation	30-40%	High
scBS-seq (Smallwood et al., 2014)	5 rounds	Tagged random primers	10-20%	High
snmC-seq (Luo et al., 2017)	1 round	Tagged random primers	50-60%	Moderate
scWGBS (Farlik et al., 2016)	None	Single-stranded ligation	50-60%	Low

Table 1.1 – Qualitative comparison between single-cell PBAT based methodologies

1.4.2 Existing computational methods

Regardless of the protocol used, analysis of single-cell methylome measurements are challenged by stochastic loss of data. This renders analytical strategies designed for bulk measurements unsuitable as they are not adapted for high amounts of missing data. To work around this, current analytical strategies for single-cell DNA methylation measurements average DNA methylation levels 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, multiple regulatory regions can be present within the same window and not be co-regulated by DNA methylation. Furthermore, the methylation state of a single CpG can impact 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) as well as the affinity of transcription factor binding to a given DNA sequence (Rishi et al., 2010; Yin et al., 2017). Imputation strategies may leverage sequence context and CpG methylation states across single cells to increase resolution (Angermueller et al., 2017); however, inference across cells partially relies on the assumption of homogeneity across cells, which may lead to a masking of rare subpopulations. To address these limitations, we developed an analytical framework that we term Pairwise Dissimilarity Clustering (PDClust) that leverages the

methylation state of individual CpGs in a pairwise fashion without inference to identify differentially methylated states across large numbers of single cells.

Chapter 2: Development and optimization of Post-Bisulfite Adapter Ligation (PBAL)

2.1 Development of PBAL

To obtain quantitative methylomes of single cells, we adapted and automated a post-bisulfite adapter tagging (PBAT) strategy (Miura et al., 2012). We sought to develop a single cell methylation strategy based off of PBAT that was optimized for diversity and CpG coverage under the hypothesis that single CpGs are informative.

PBAT and PBAT-derived single-cell strategies published since the start of this work all used random primers extended with Illumina sequences to enable direct amplification (Angermueller et al., 2016; Smallwood et al., 2014). However, when we compared this approach to untagged random priming we observed that extended randomers generated shorter double-stranded DNA fragments compared to untagged randomers suggesting inefficient priming **(Fig 2.1)**. To circumvent this problem, we used untagged random primers and ligated Illumina sequencing adapters to the double-stranded DNA fragments instead of using tagged adapters **(Fig 2.2)**.

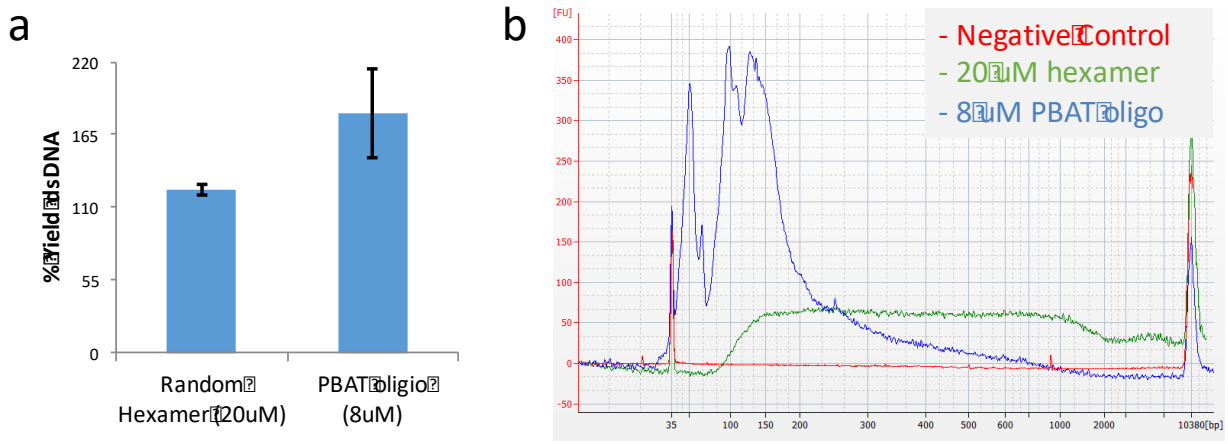

Figure 2.1 – Untagged primers generate more usable material than tagged primers

a) Yield of dsDNA generated after random priming, normalized by input dsDNA (100 ng). Libraries were made from bisulfite converted dsDNA extracted from HL60 cells.

b) Agilent bioanalyzer profiles of libraries generated in a).

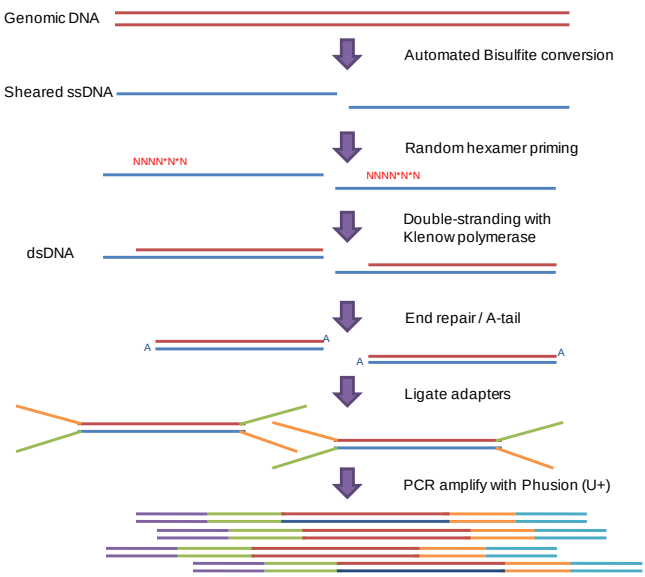

Figure 2.2 – Workflow of PBAL

We also sought to optimize the enzyme and primer concentrations during random priming. We found that the DNA polymerase used saturates at 1 U/uL **(Fig. 2.3a, b)** but did not

find a saturating effect on the primer concentration in the range of concentrations we tested **(Fig. 2.3c)**. However, we did notice that the average fragment size of the resulting dsDNA decreased as primer concentration was increased **(Fig. 2.3d)**. Therefore, we decided to use final concentrations of Klenow at 1 U/uL and random hexamers at 20 uM.

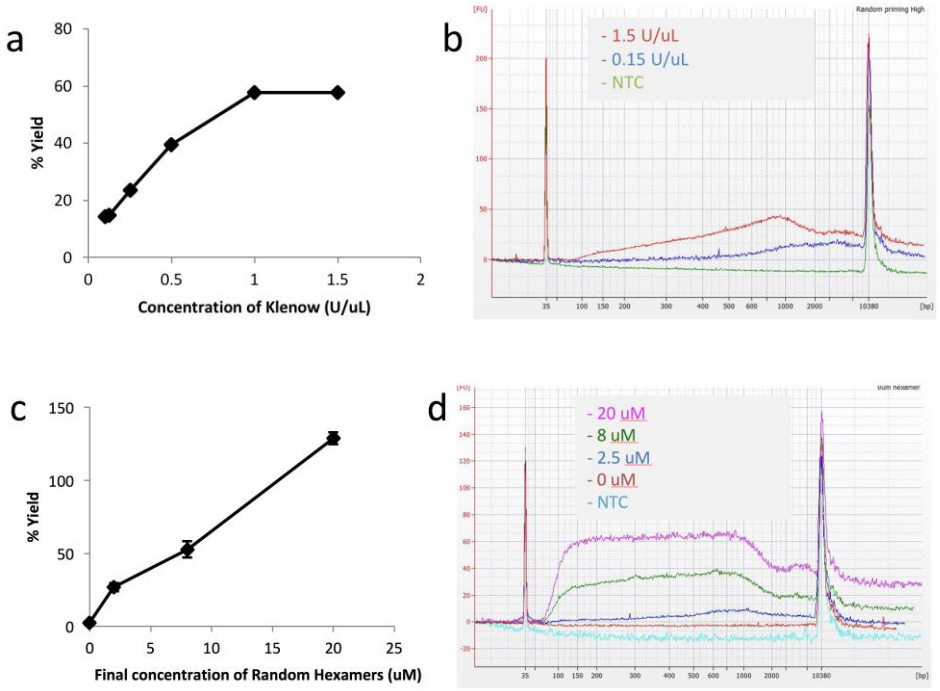

Figure 2.3 – Saturating quantities of enzyme and primer identified for the random priming reaction

When we generated libraries from quantities of DNA similar to single-cells (10 picograms), we noticed material in negative controls **(Fig 2.4a)**. When we sequenced these libraries, the alignment rate for the samples were equivalent to negative controls, suggesting that the majority of reads are contaminants. Upon close examination of the fragment size distributions, we noticed that negative controls were smaller than libraries from samples. We therefore adjusted our size selection step to select against smaller fragments. We tested this

adjustment on single-cells from a human breast cell and observed that single-cells generated more DNA fragments compared to negative controls **(Fig 2.4b)**.

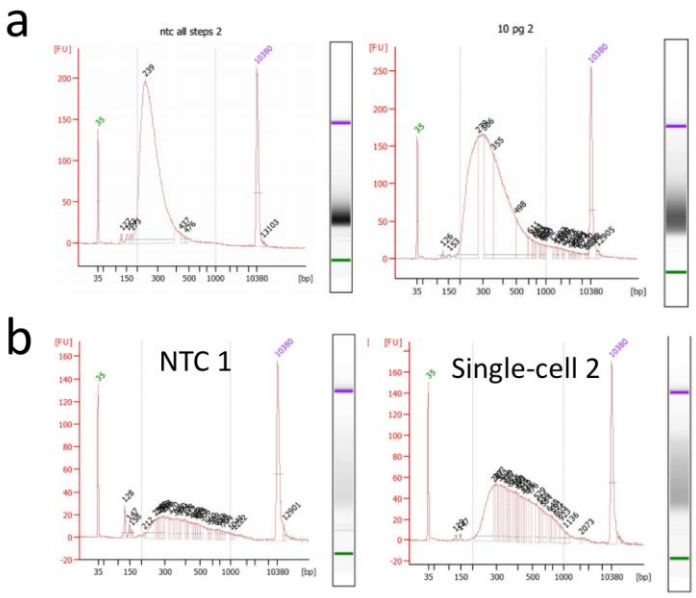

Figure 2.4 – Removal of contaminating DNA material present in single-cell libraries using a stringent bead cleanup

We then sequenced this batch of single-cells and aligned the resulting reads to the human genome (hg19) using Novoalign (http://www.novocraft.com), which we have previously demonstrated (data not shown) to be more sensitive than the conventionally used Bismark aligner (Krueger and Andrews, 2011). We observed that negative controls had 10% alignment rate while single-cells had 50% **(Fig 2.5)**. Although there is a clear separation between single-cells and negative controls, we could not confidently use the data in our single-cells as 20% of the data may be contaminating human DNA. Strikingly, all the reads in negative controls were unconverted, suggesting that the contaminating human DNA likely entered library construction after bisulfite conversion. To address this, we modified a parameter in Novoalign (-u) that

penalizes unconverted cytosines at CHG and CHH positions as these are less likely to be methylated than CGH sites. Instead of the default penalty of 8, we increased this penalty (to 50) until it was sufficient to decrease the alignment rate of negative controls by 10% to near 0. As expected, contaminates from single-cells were also removed, decreasing the alignment rate of single-cells by 10% as well **(Fig 2.5)**. We concluded that this strategy was sufficient to remove the contaminating human DNA from the libraries without removing true read fragments from single-cell libraries.

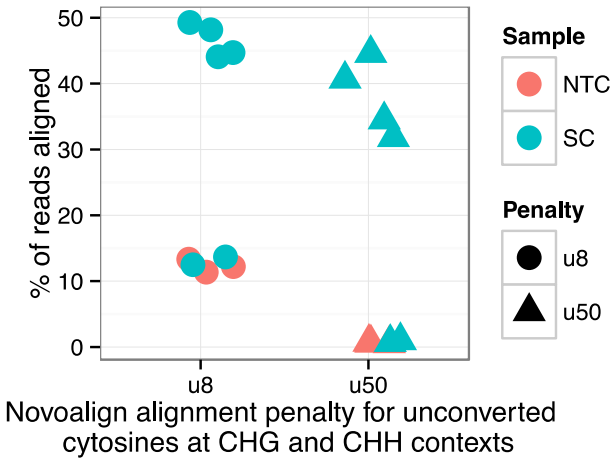

Figure 2.5 – Bioinformatic removal of contaminating human DNA alignment in single-cell samples

Although we were able to remove contaminating human DNA from single-cells, we were only able to align 40% of the reads. To understand what the other 60% of reads are, we ran a BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) search for 10,000 unaligned reads for 3 single-cell libraries and extracted the top result for each sequence (**Fig 2.6)**. We found that 50% of unaligned reads also have no hits by BLAST alignment to the NR database, suggesting that these reads may be bisulfite converted contaminants. We found that 30% of reads on average were

cloning vectors possibly introduced along with the enzymes. Sequences from human, bacteria, and viral genomes comprised the remaining alignments.

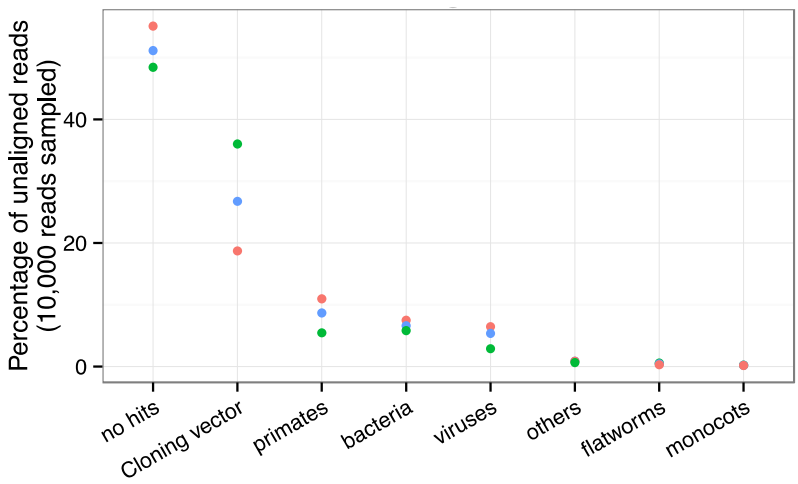

Figure 2.6 – Identities of unmapped reads are results of external sources of contamination

While PBAL was in development, another paper was published that contained a method for single-cell DNA methylation sequencing (scBS-seq; Smallwood et al., 2014). Upon close examination of the protocol, scBS-seq was an extension of the PBAT methodology with 5 rounds of random priming. We therefore tested to see if additional rounds of random priming could improve the efficiency of PBAL. We generated and sequenced batches of single-cell libraries with either 1 or 2 rounds of random priming and compared the results. We observed that libraries with 2 rounds of random priming had increased yield compared to libraries with one round (data not shown). Furthermore, the DNA fragment duplicate rate of libraries with 2 rounds of random priming was statistically significantly less than those of 1 round **(Fig 2.7a)** without impacting the efficiency of CpG recovery **(Fig 2.7b)**. We concluded that additional rounds of

random priming increased the diversity of single-cell libraries. However, the enzyme was prohibitively expensive and thus we compromised for 2 rounds of priming instead of 5.

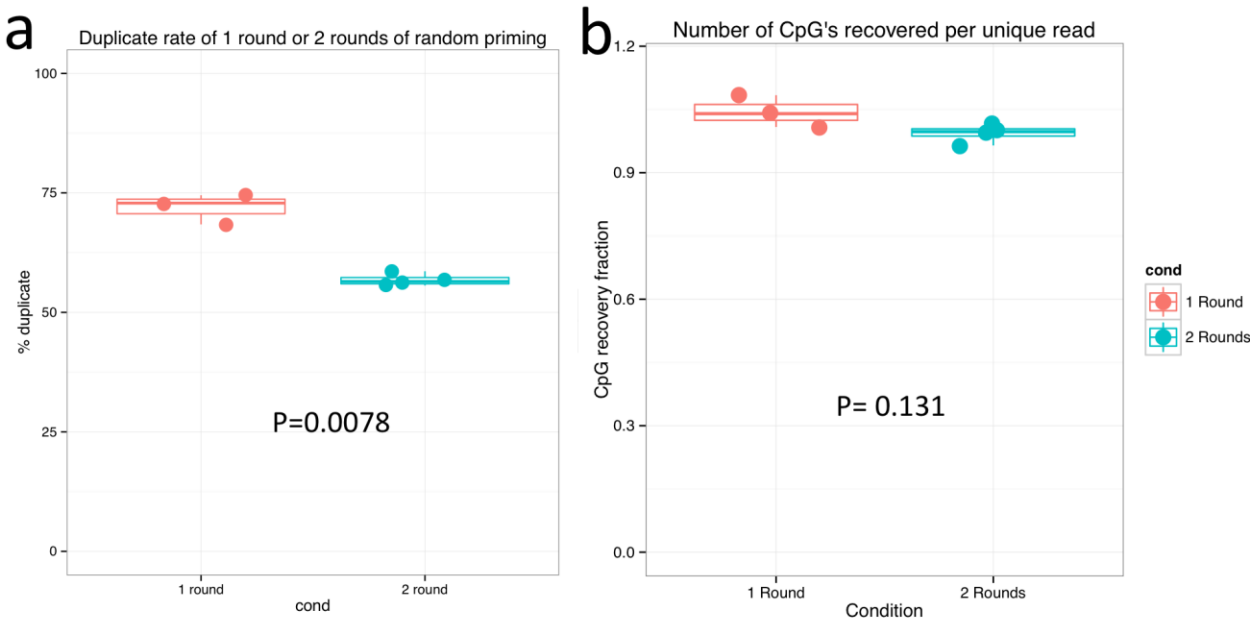

Figure 2.7 – 2 rounds of random priming significantly increase library diversity

2.2 Yield of single-cell libraries increased by optimization of library generation

During bulk PBAL library construction, we noticed that there was a yield discrepancy between the manual and automated versions of PBAL **(Fig 2.8)**. Upon further investigation, we ruled out confounding factors such as batch effect and input DNA quality (data not shown).

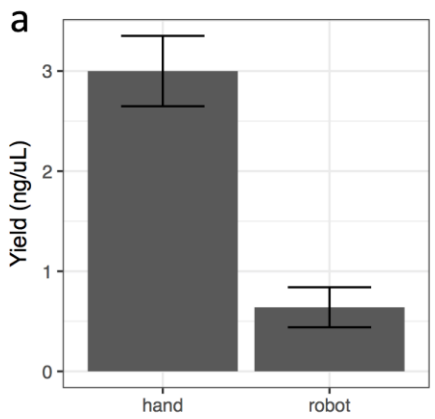

Figure 2.8 – Yield of bulk 100ng libraries is higher when generated by hand vs by robot.

To isolate the reason for the yield discrepancy, we systematically performed each step of PBAL by hand until a certain step before transferring the sample to the robot to perform the remainder of the protocol. We discovered that the loss of yield was primarily during the elution step which involved drying of the DNA-bead complex (data not shown). During this step, the beads are air-dried prior to elution so that residual ethanol does not also enter the sample and interfere with downstream reactions. However, drying the beads too much results in lower elution efficiency. We reduced the drying time for the beads, and was surprised to find that 1 minute of drying on the robot resulted in the same yield as doing the protocol manually which dries for 3 minutes **(Fig 2.9)**.

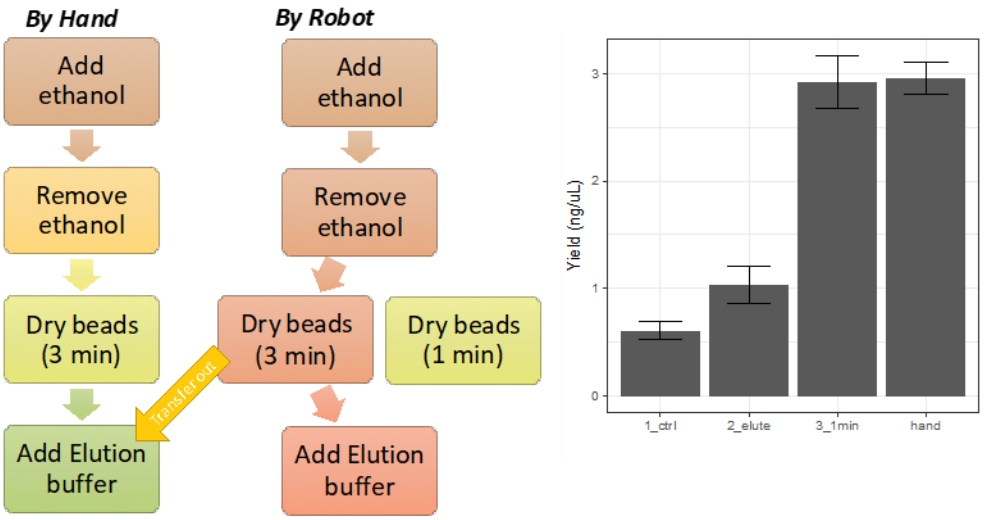

Figure 2.9 – Yield of 100ng libraries depending on length of air-drying step.

Yield (concentration) of libraries eluted in 35 uL generated from 100ng input. Error bars represent the standard deviation from three replicate experiments. Libraries were either eluted as normal (dried for 3 minutes, 2_elute), or dried for 1 minute and covered with tape (3_1min).

Because the robot performs the automated version of the protocol in a laminar flow hood, we reasoned that the amount of time required to dry the beads is less compared to the bench because of the lower humidity within the hood; therefore, 3 minutes of drying lead to the beads being over dried and prevented the efficient elution of DNA from the silica beads. After implementing this same correction for single-cell libraries, we gained significantly higher overall yield **(Fig 2.10)**.

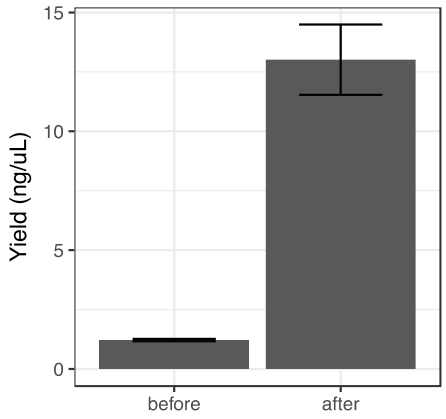

Figure 2.10 – Yield of single-cell libraries is significantly increased after optimizing for drying time.

Libraries were generated from single nuclei sorted, subject to 12 cycles of PCR and concentrated to 10 uL. Error bars represent standard deviation of two plates per condition.

2.3 A qPCR assay enables removal of failed libraries prior to pooling

The success rate of single-cell libraries generated by PBAL is typically around 80% (data not shown). Previously, there was no way to determine if a well contained a successful library until it was sequenced. However, sequencing failed wells takes away reads from successful wells and decreases the overall sequencing depth of libraries. We therefore sought to develop a method to identify failed wells prior to sequencing.

We hypothesized that a quantitative polymerase chain reaction (qPCR) assay would be sensitive enough to amplify organism-specific material which may be a good proxy for the amount of the target genome capture din the well. To find suitable genomic regions, we sought to first identify genomic regions within successful single-cell libraries with the highest coverage. Analysis of alignment rates from existing single cell libraries identified genomic locations with consistently high coverage in human **(Table 2.1)** and mouse genomes **(Table 2.2)**. We noticed that these regions were within highly repetitive elements. This means in a single-cell, there will

be many copies of these specific sequences which increases the probability of amplifying them in a qPCR reaction. Therefore, these regions are less susceptible to stochastic data dropout.

LIBRARY ID	ALIGNMENT RATE	CHROMOSOME	POSITION	COVERAGE
TCGAAG	30.50%	***chr1***	***121485205***	***29***
		chr1	121485240	28
		chr1	121485290	25
		chr1	121485291	24
GACGGA	9.40%	***chr1***	***121485205***	***46***
		chr19	27732041	37
		chr1	121485240	36
		chr19	27732032	34
CATGGC	17.50%	chr1	121485290	139
		chr1	***121485205***	***138***
		chr1	121485240	121
		chr19	27732041	88

Table 2.1 – Top 3 CpG sites with the highest coverage in three human libraries. The CpG site of interest is highlighted in yellow.

LIBRARY ID	ALIGNMENT RATE	CHROMOSOME	POSITION	COVERAGE
AACTTG	27.90%	chr2	98667168	1036
		chr2	98667149	1011
		chr2	***98667150***	***1008***
		chr2	98667208	1001
CACTCA	9.70%	chr2	98667196	378
		chr2	***98667150***	***360***
		chr2	98667168	355
		chr2	98667208	342
TAGCTT	16.00%	chr2	98667196	639
		chr2	98667168	611
		chr2	***98667150***	***591***
		chr2	98667208	589

Table 2.2 – Top 3 CpG sites with the highest coverage in three mouse libraries. The CpG site of interest is highlighted in yellow.

In silico bisulfite conversion was performed on the DNA sequences corresponding to selected genomic regions and Methprimer was run on the resulting converted sequences to generate 2 primer pairs per region.(Li and Dahiya, 2002). Using saved fractions of previously successful single-cell libraries, we performed PCR amplification to test the specificity of these primer sets in mouse and human. We identified primer sets for mouse and human negative controls that did not produce products while single-cells did. Overall, we observed a clear separation in melt curve of negative controls and hundred-cell pools. Some single-cell libraries appeared similar to negative controls while most single-cell libraries look like the positive controls **(Fig 2.11 left)**. The CT values of the hundred-cells were the lowest, followed by a bimodal distribution in the single-cells **(Fig 2.11 right)**. The negative controls failed to reach the threshold of detection after 45 cycles of qPCR. A similar result was obtained from murine single-cells using mouse-specific primers (data not shown). These results demonstrate the utility of the qPCR assay in distinguishing between failed and successful single-cells.

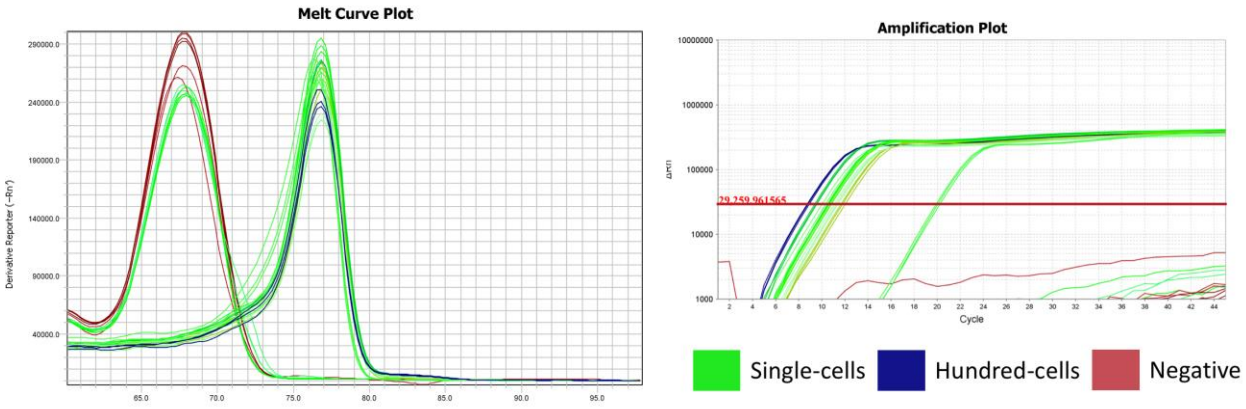

Figure 2.11 – Melt curve and amplification plot of human negative controls, single-cells and hundred-cell libraries allows for identification of failed single-cell wells.

In addition to the genome-specific primers, we also measured the amount of total library using Illumina sequencing adapter-specific (library-specific) qPCR primers. We used this information to standardize the organism-specific yield to the total yield. We calculated a qPCR score for each sample by subtracting the genome-specific CT value by a 1/100 dilution of the library-specific CT value, and compared these values to the sequencing results. We saw that qPCR scores of positive controls were low whereas qPCR scores for negative controls were high. These relationships allowed us to empirically determine a cut-off in qPCR score to remove failed single-cells **(Fig 2.12a)**.

We pooled together and sequenced negative controls, positive controls, and single-cells that passed the qPCR threshold and analyzed the resulting data. We observed that the qPCR score and the mappability of single-cell libraries were negatively correlated as expected **(Fig 2.12b)**. By removing failed wells, we were able to decrease wasted sequencing space and thus increase the sequencing depth of each successful library per lane of sequencing at the same sequencing cost.

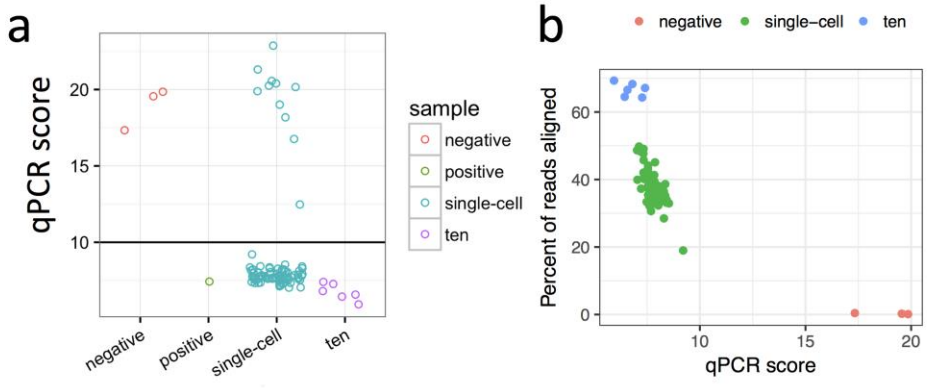

Figure 2.12 – qPCR scores of single-cells allow for identification of failed cells

a) The horizontal line represents an empirically determined cut-off for passed cells, where single-cells with a higher qPCR score are considered failed and removed prior to pooling.

b) There is a clear relationship between qPCR score and mappability of libraries.

2.4 Characteristics of single-cell DNA methylation libraries

2.4.1 Abnormal copy number profiles distinguish technical or biological outliers

Because methylomes are sampled in an unbiased manner, coverage of PBAL libraries generated from karyotypically normal single-cells should be uniform. After calling copy number variation using Control-FREEC (Boeva et al., 2012) in 5 MB bins, we noticed a small proportion of cells had highly uneven copy number patterns throughout their genome **(Fig 2.13)**. These cells may be cells technical failures during library construction, or cells that are actively cycling. In single-cell whole-genome sequencing methods, similar abnormal copy number profiles were also observed (Zahn et al., 2017). Regardless of the reason, a cut-off was determined empirically and the resulting cells are removed from analysis.

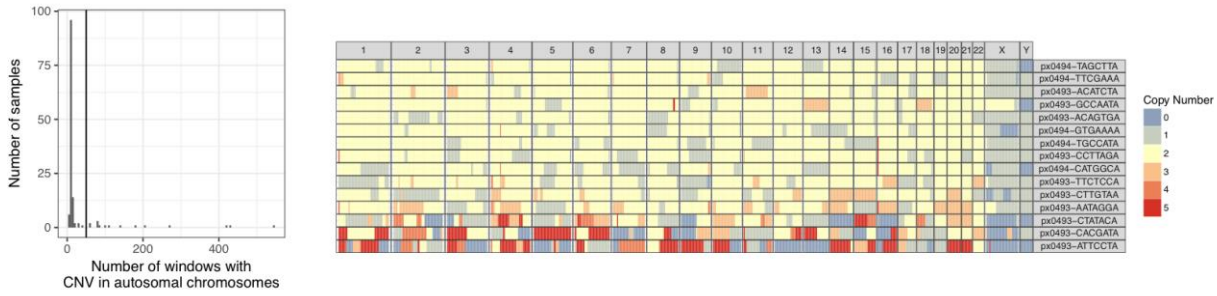

Figure 2.13 – Copy number profiles distinguish outlier cells with uneven coverage

2.4.2 PBAL libraries have improved CpG recovery over existing protocols

During the course of this work a number of single-cell DNA methylation protocols have been developed and published. To address whether the diversity of PBAL libraries were equivalent to existing protocols, we sequenced one set of single-cells to an average of 7 million mapped reads and down sampled the resulting reads. We observed that PBAL libraries can measure up to 5.7 million CpG sites in a single-cell and still not reach saturation **(Fig 2.14)**.

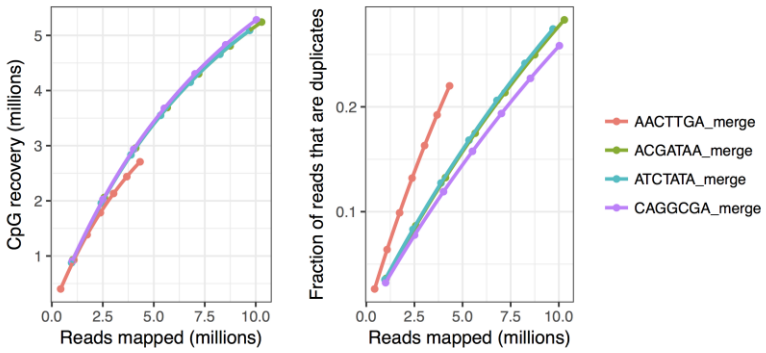

Figure 2.14 – Saturation curves from deeply sequenced set of PBAL libraries reveal that single-cells have not saturated.

We next sought to compare the performance of PBAL to existing single-cell methodologies (Angermueller et al., 2016; Luo et al., 2017; Smallwood et al., 2014). To do this, we downloaded raw sequence data from existing single-cell DNA methylation datasets and uniformly processed them with our pipeline. We observed that PBAL had the best predicted recovery of CpG sites **(Fig 2.15a)** amongst all the methodologies sampled and had lower variation of CpG recovery across samples. Furthermore, we noticed that PBAL libraries had similar duplicate rates comparing to existing methodologies **(Fig 2.15b)**. Altogether, these results support PBAL as a comparable single-cell DNA methylation protocol.

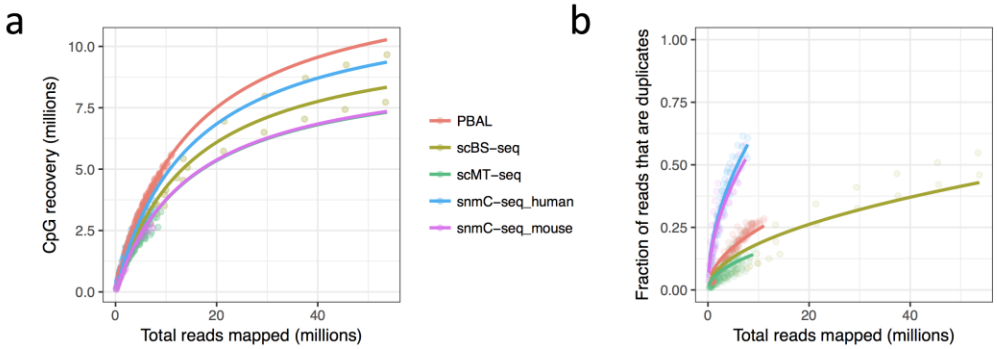

Figure 2.15 – PBAL has the highest CpG recover out of existing protocols.

Each dot represents one down sample of a library. The lines represent the y=1/(1+x) line of best fit.

2.4.3 Concordance of methylation of nearby CpGs in single-cells is restricted to 2kb

Genome-wide methylation levels derived from bulk cells are characterized by spatial correlations of 1-2kb (Eckhardt et al., 2006; Zhang et al., 2015) and this observation has been incorporated into algorithms designed to identify differentially methylated regions (DMRs) (e.g. Hansen et al., 2012). This observation has also provided a rationale for assigning methylation states from 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 nearby CpGs share the same methylation state in single-cells has yet to be studied. To address this, we calculated the probability that a CpG in single-cells was in the same methylation state with neighbouring CpGs as a function of their genomic separation **(Fig 2.16)**. A background concordance was also calculated as the probability of sampling 2 methylated or 2 unmethylated CpGs genomewide within each single-cell dataset.

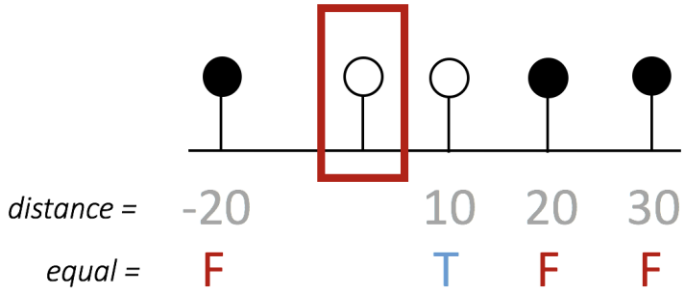

Figure 2.16 – Schematic for how CpG methylation concordance is calculated

This strategy revealed genome-wide concordance between CpGs up to 1 kb in *cis* in human CD49f single cells, which then decreased rapidly to near random chance at 2kb **(Fig 2.17)**. To be computationally efficient, we performed this calculation only for a subset of CpG sites in each single-cell. We demonstrate that sufficient subsampling (> 10,000 CpG sites) accurately captures the pattern of concordance genomewide **(Fig 2.17a)**. These measures of CpG methylation concordance also appear to be stable across cells of the same cell type **(Fig 2.17b**).

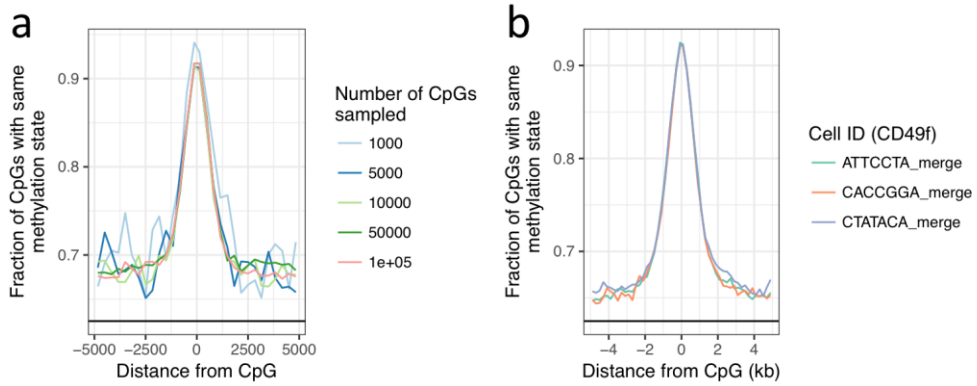

Figure 2.17 – Genomewide CpG concordance within single-cells decreases rapidly after 1kb

a) CpG concordance is accurately represented using a subsample of the data

b) CpG concordance does not vary between CD49f cells

We also computed CpG concordance using existing single-cell datasets to see if CpG concordance was a conserved feature across different cell types. In mouse embryonic stem cells (ESCs) grown in 2i or Serum media (Smallwood et al., 2014), concordance of methylation also rapidly decayed to near random chance at 2kb **(Fig 2.18a)**. ESCs grown in 2i media still had >85% concordance within 500 bp despite differing levels of background concordance (55% - 80%). In human leukemic cell lines (Farlik et al., 2015), we also observed high concordance within 500bp that was independent of background concordance **(Fig 2.18b)**. Moreover, K562

cells, despite having lower background concordance than HL60 cells, have the same minimum concordance as HL60 cells beyond 4kb. These results suggest that concordance of very close CpG neighbours (1kb) is highly conserved across species and cell types despite background concordance.

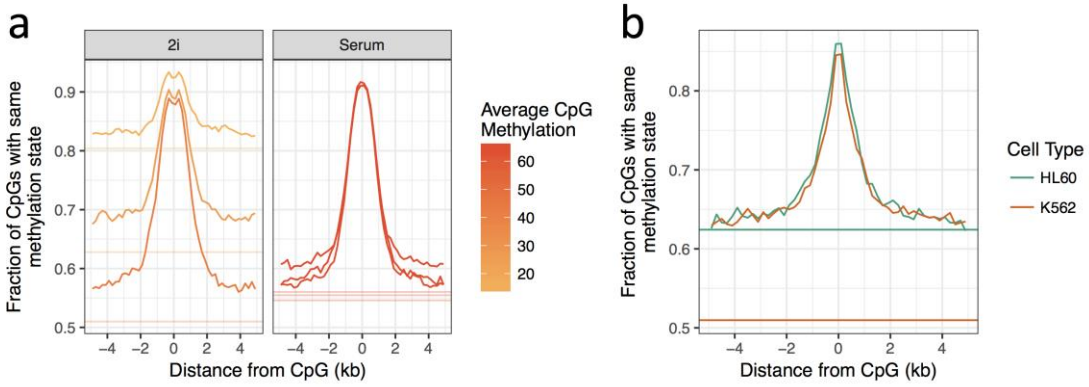

Figure 2.18 – Genomewide CpG concordance across cell types is stable within 500bp

a) CpG concordance of mouse embryonic stem cells grown in 2i or serum media

b) CpG concordance of two human leukemic stem cell lines

A direct comparison of concordance between bulk and single-cells is not possible because DNA methylation measurements in single-cells are binary while DNA methylation measurements in bulk samples are continuous. To estimate concordance in bulk samples we calculated the absolute difference in methylation between a given CpG and all nearby CpGs. We compared these values derived from bulk measurements of (10,000) LSK cells to concordance measures in single LSK cells and observed similar genomewide patterns **(Fig 2.19a).** However, genomewide CpG adjacency was consistently higher in bulk compared to single-cell measurements and did not approach random chance at 5kb, whereas single-cell concordance

reached background levels at 4kb. Within specific genomic contexts, concordance between CpGs for single-cell and bulk-cell measurements was nearly equivalent within CpG Islands and promoters, but was higher in bulk compared to single-cells in all other genomic contexts **(Fig 2.19b)**. In addition, we found that concordance for both single-cell and bulk samples decayed more rapidly within blood lineage enhancers (Lara-Astiaso et al., 2014) compared to other genomic regions. Taken together, these findings suggest that analytical strategies that infer cytosine methylation within fixed bins (up to 100 Kb, Luo et al., 2017) across the whole genome of single cells may lead to over smoothening of CpG methylation supporting a need for additional methodologies.

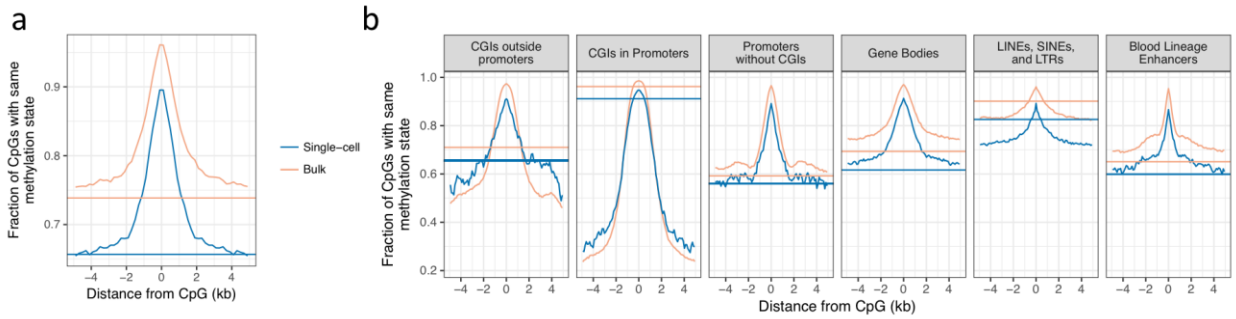

Figure 2.19 – CpG concordance is similar in bulk and single-cells

Chapter 3: Heterogeneity of murine hematopoietic stem and progenitor cells

3.1 Enhancers of single HSCs are hypo methylated and show increased variability compared to enhancers of lineage restricted cell types

From the bone marrow of mice, we used fluorescence activated cell sorting (FACS) to sort primary EPCR+CD45+CD48-CD150+ (ESLAM) cells, one of the most highly purified hematopoietic stem cell (HSC) populations found in mice (~40% repopulating capacity, Wilson et al., 2015). As a comparator progenitor population, we sorted linage negative Sca1+c-Kit+ (LSK) cells (~4% repopulating capacity, Warr et al., 2011).

To validate that the single-cells assayed were indeed HSCs, we sought to confirm that the epigenome of these single-cells corresponded with known HSC biology. In active regulatory regions such as enhancers, DNA methylation is typically lower than the genomewide average (Stadler et al., 2011). We hypothesized that hypomethylated enhancers in LSK and ESLAM cells should be associated with transcriptional states implicated in the maintenance of blood stem and progenitor populations, whereas enhancers that are active in terminally differentiated cell types should be closer to the genomewide average in methylation.

To address this hypothesis, we *in silico* merged single cells and calculated the average methylation in specific blood enhancer (defined by specific H3K4me1 signal) derived from bulk experiments (Lara-Astiaso et al., 2014). We then separated the enhancers into hypomethylated (<= 25% methylation) and hypermethylated (>= 75%), and plotted each enhancer's H3K4me1signal in each bulk blood cell type. We found that, on average, hypomethylated enhancers had increased H3K4me1 signal compared to hypermethylated enhancers, and each comparison was statistically significant for blood progenitor cell types **(Fig 3.1a and b).** These

results suggest that hypomethylated enhancer regions in our single cells correspond with increased enhancer usage in blood stem and progenitor cell types.

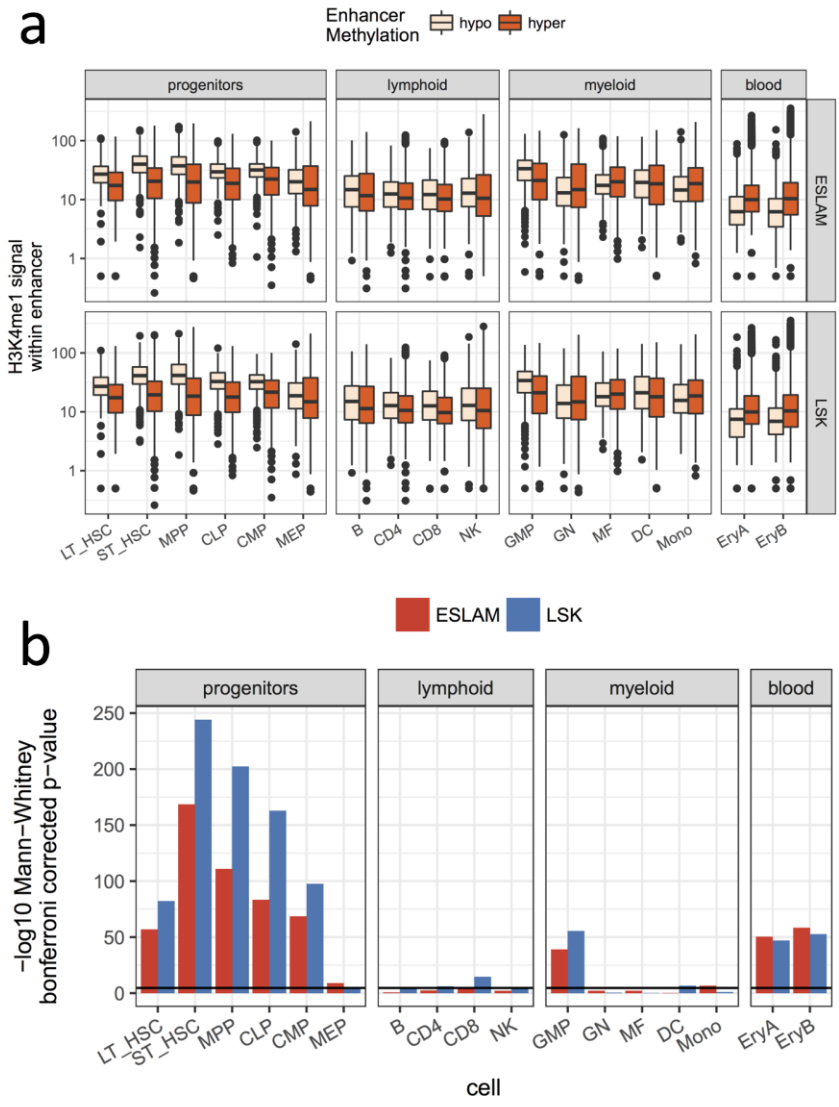

Figure 3.1 – LSK and ESLAM PBAL methylomes correlate with the enhancer landscape of HSCs

a) Hypomethylated enhancers in LSK and ESLAM cells have higher H3K4me1signal in progenitor populations but not terminally differentiated cell types

b) The statistical significance associated with the differences in H3K4me1 signal of hypomethylated and hypermethylated enhancers for ESLAM and LSK populations.

Next, we calculated the average enhancer CpG methylation in each single-cell. For each blood cell type, we classified enhancers as active (H3K4me1 signal >= 50) or inactive (signal <=25) as previously defined (Lara-Astiaso et al., 2014) and calculated the standard deviation of enhancer methylation of every single-cell. We observed that the methylation state of active enhancers in progenitor populations was more variable (35% vs. 20%) than inactive enhancers, while the variation of active and inactive enhancers in terminal populations were not statistically different **(Fig 3.2)**. These results suggest that progenitor enhancers are more heterogeneous across single ESLAM and LSK cells than enhancers active in terminally differentiated cells and further validate the biological relevance of our dataset.

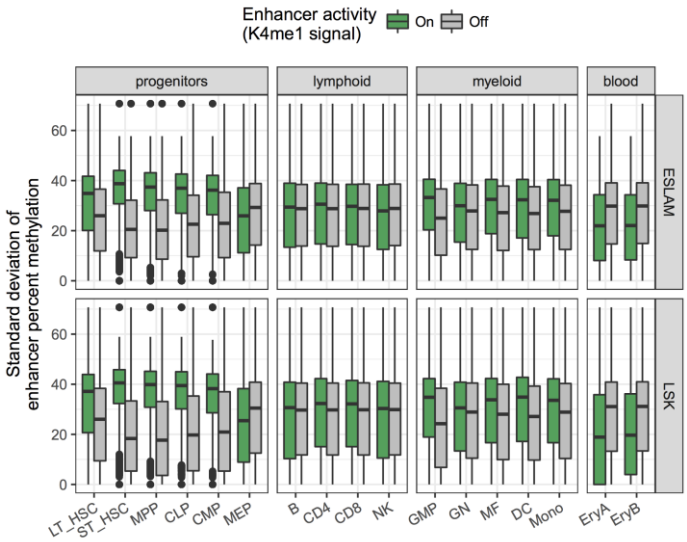

Figure 3.2 – Enhancers active in progenitor populations are heterogeneous than non-active enhancers

3.2 Pairwise dissimilarity clustering (PDclust) reveals subpopulations

Single-cell RNA-seq measurements of LSK (Paul et al., 2015) and ESLAM cells (Wilson et al., 2015) have identified subpopulations with different transcriptional profiles within each of these phenotypes. To determine whether analogous variability is also present in their methylomes, we first developed a measure of CpG methylation dissimilarity, as defined by the average of the absolute difference in methylation values at CpGs covered in each pairwise comparison **(Fig. 3.3a)**. We then used these PD values to cluster the data based on the extent of similarity in the CpG states measured. Application of this analysis to both the single ELSAM and LSK cell data showed that, on average, the ESLAM cells displayed less dissimilarity (p-value < 0.01), consistent with the knowledge that they are functionally less heterogeneous **(Fig. 3.3b)**.

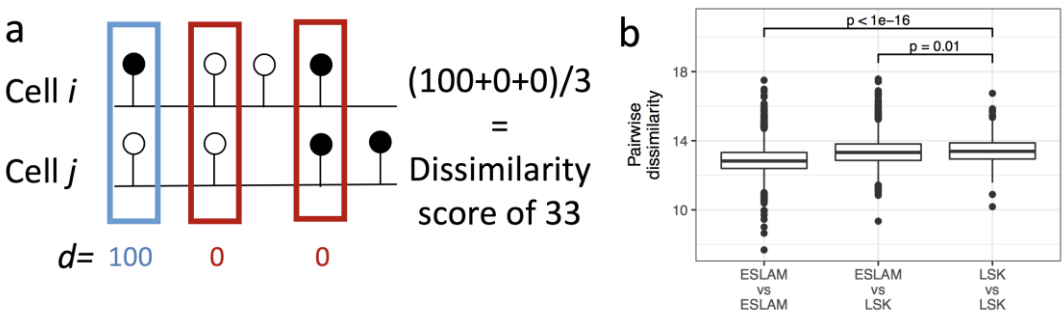

Figure 3.3 – ESLAM cells have lower PD values than LSK cells

a) A schematic showing the PD values determined between all paired comparisons of single ESLAM cells.

b) ESLAM cells have a lower overall PD compared to LSK cells and all cell types analyzed. Every pairwise comparison between cells denoted on the X-axis is summarized as a boxplot with the distribution of PD values shown on the Y-axis. P-values were calculated using a 2-sided t-test.

Unsupervised clustering on PDs (PDClust) derived from CpG sites within genomic regions previously implicated in the HSC to MPP transition (Cabezas-Wallscheid et al., 2014)

generated 2 distinct subsets that were differentially enriched in the ESLAM and LSK populations **(Fig. 3.4a).** Projection of PDs onto two-dimensional space with multidimensional scaling suggested 2 independent states defined by distinct DNA methylation signatures **(Fig. 3.4b)**; one that was most enriched in the ESLAM population and the other in the LSK cells, with a proportion of (13/64) LSK cells showing the "ESLAM profile".

We next calculated PDs without restricting the analysis to regulatory states. Instead, we considered all CpGs genome-wide to identify epigenetic subsets detected within the ESLAM and LSK populations. When all available CpG sites were used, PDClust revealed 2 subgroups with differing degrees of heterogeneity **(Fig. 3.4c)**. Multidimensional scaling (MDS) analysis confirmed these relationships by revealing a highly similar population (group1) and the rest of the cells appearing more dispersed(group2) **(Fig. 3.4d)**. Group1 included a higher proportion of ESLAM cells than LSK cells (26/84 versus 3/64 or 31% vs 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 **(Fig. 3.4d)**.

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 that showed no emerging clustering patterns **(Fig. 3.5).**

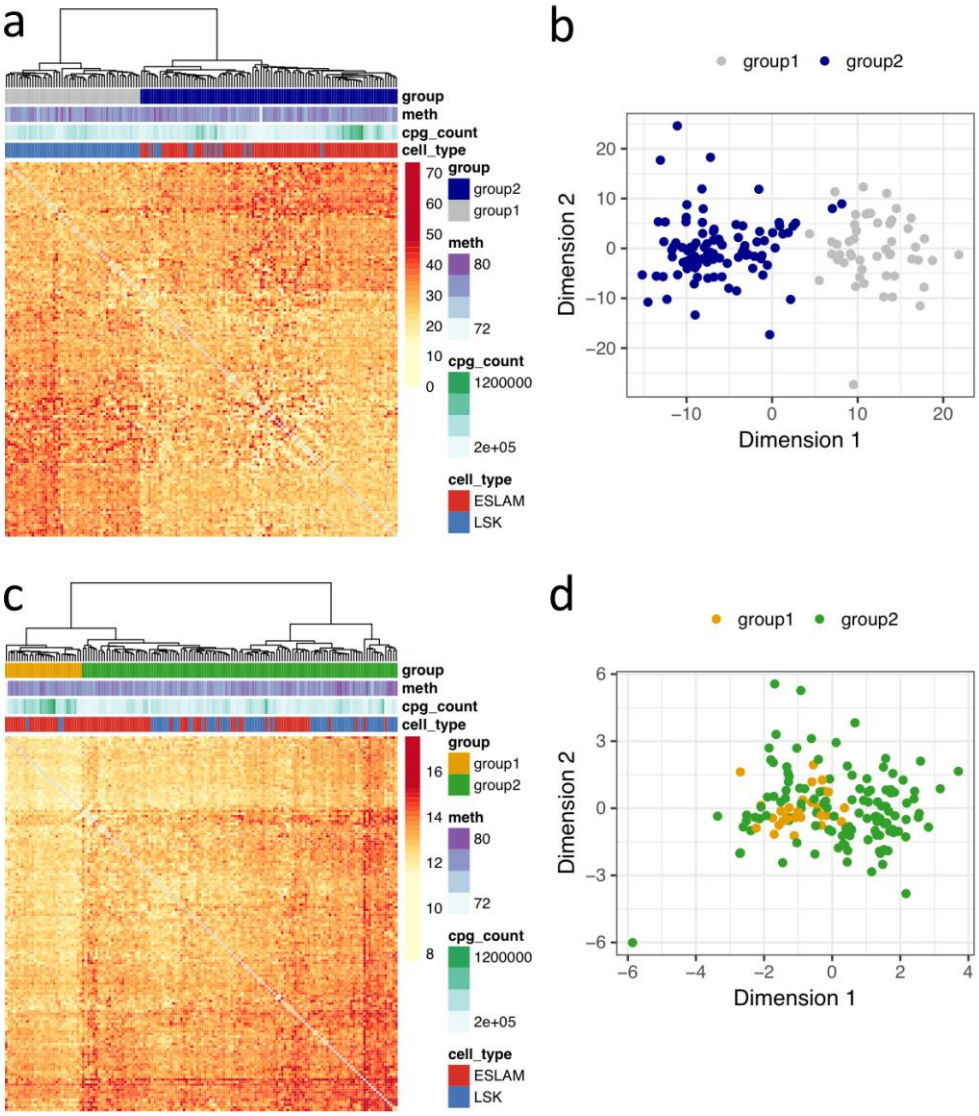

Figure 3.4 – Pairwise analysis of single cells reveals subsets within the murine ESLAM phenotype

a) PDClust of CpGs associated with genes important in HSC function separate ESLAM and LSK cells, with some LSK cells exhibiting an ESLAM epigenetic signature.

b) Multidimensional scaling using PD calculated from a) used directly as input.

c) Same as (**a**) but instead considering all CpGs regardless of their genomic position.

d) MDS analysis of **c**) reveals group 1 at the epicenter of single ESLAM cells with group 2 surrounding the central cluster.

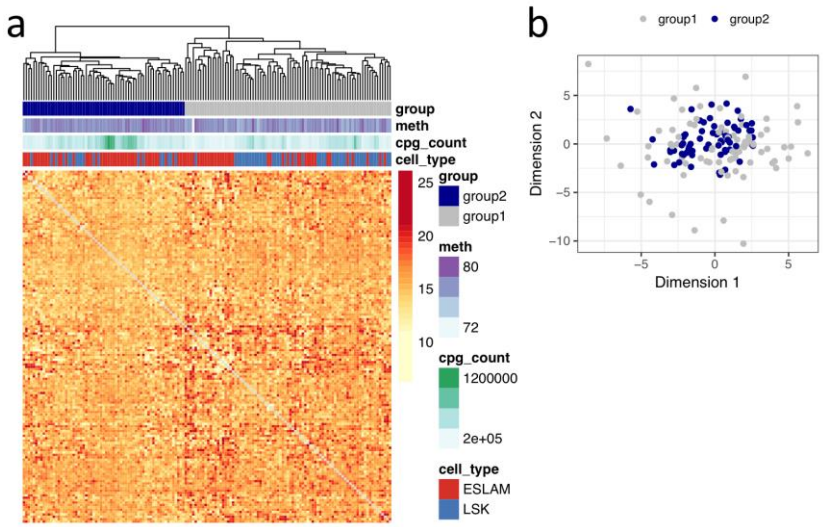

Figure 3.5 – PDClust of CpGs in an irrelevant region set reveals no clear clusters.

Clustering **(a)** and MDS scaling **(b)** of single-cells using CpGs that lie within an irrelevant genomic region set (cortex enhancers) show no pattern.

3.3 Epigenetic annotations of subpopulations reveal putative functional differences

To examine the functional significance of the epigenetically defined subsets of LSKs and ESLAM cells, we merged CpGs *in silico* separately across all cells belonging to group1 and group2. We then estimated the smoothed methylation values of all CpG sites in the genome for these two merged groups using BSmooth (Hansen et al., 2012). We computationally identified DMRs between the two groups by selecting regions that contain three or more CpGs that are statistically different at a 0.1 false discovery rate. (**Fig 3.6**). For each DMR, we then used HOMER (Heinz et al., 2010) to annotate its overlapping genomic feature(s) as well as the nearest transcript.

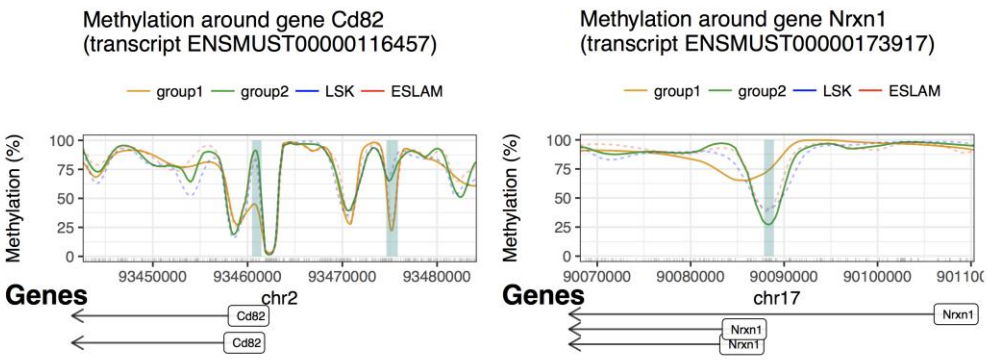

Figure 3.6 – Example DMRs between group1 and group2 single-cells

Ticks on the x axis represent individual CpG sites. Highlighted areas represent areas computationally defined as differentially methylated regions.

In the bulk cell data, CpG methylation is anti-correlated with expression in promoters and the first exon (Brenet et al., 2011) and active distal regulatory regions are typically hypomethylated (Stadler et al., 2011). We also found that DMRs were enriched in promoters (+/- 2Kb of TSS) and depleted in distal intergenic regions (data not shown), consistent with their function. We thus looked for subsets of DMRs that were either within the first exon, within the promoter, or were intergenic and within 20kb from the nearest TSS (Kundaje et al., 2015). We assigned each DMR to the subgroup that reported the lower methylation value and associated each DMR group to the nearest genes using HOMER (Heinz et al., 2010). As a control, we performed the same analysis for data merged *in silico* from the LSK and ESLAM single-cell data. To identify genes uniquely associated with only one population, we removed those associated with DMRs from both populations and subjected the results to gene set enrichment analysis (GSEA).

Genes associated with hypomethylated DMRs in group 1 were significantly enriched (FDR adjusted q-value < 0.1) in HSC proliferation terms as well as genes preferentially

expressed in long-term HSCs and erythrocytes **(Fig. 3.7)**. In contrast, genes associated with hypomethylated group 2 DMRs were enriched in genes specifically expressed in differentiated populations and genes that, when knocked out lead to increased HSC numbers. Together these results suggest that DMRs specifically hypomethylated in group 1 compared to group 2 are associated with genes implicated in HSC function.

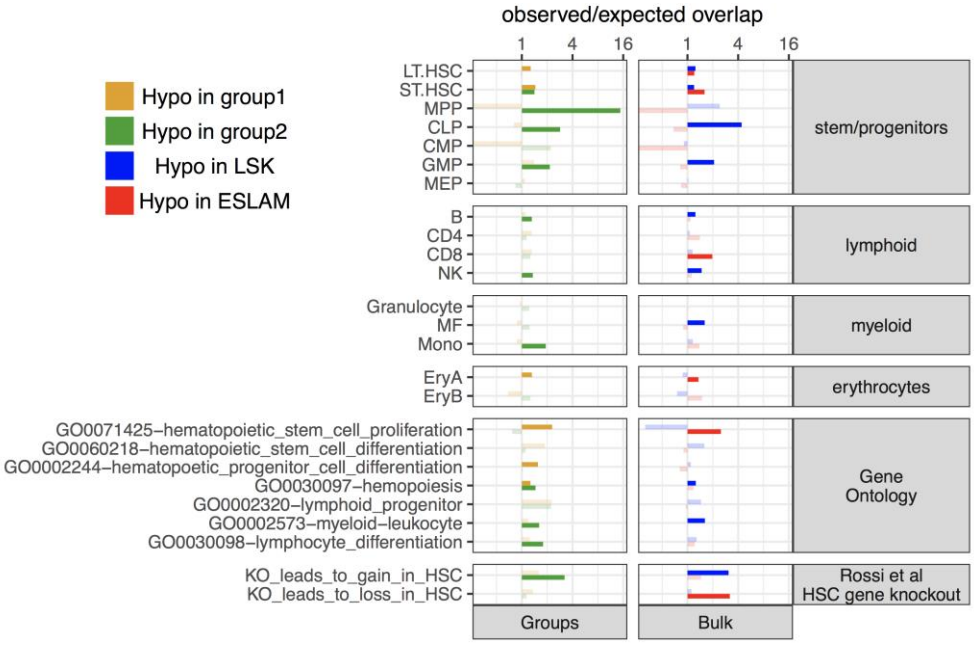

Figure 3.7 – Gene set enrichment analysis of DMR-associated genes reveal terms related to HSC function

3.4 Validation of single-cell epigenetic annotations by single-cell transcriptomics

After identifying genes associated with hypomethylated DMRs, we sought to examine their expression in single ESLAM cells. To do this, we made use of the expression data of $Lin^{-}c\text{-}kit^{+}Sca\text{-}1^{+}CD34^{-}Flt3^{-}CD48^{-}CD150^{+}$ HSCs available from 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 to all other genes ($p < 0.01$, Mann-Whitney test) **(Fig. 3.8)**. Further investigation of the levels of expression

of these DMR-associated genes within the ESLAM population (again using the data published by Wilson et al 2015) showed that a significant number showed considerable heterogeneous expression profiles (hypergeometric p-value <0.01) **(Fig. 3.8)**.

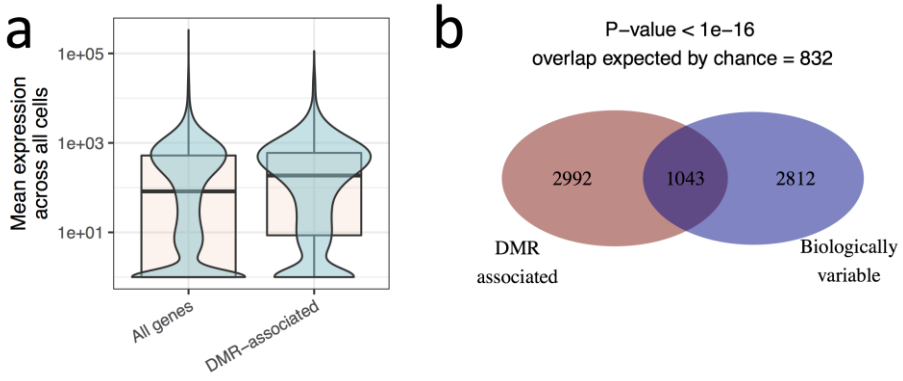

Figure 3.8 – Genes associated with DMRs in ESLAM cells are heterogeneously expressed in HSCs

(**a)** Genes associated with DMRs are expressed higher than the genome-wide average.

(**b)** There is a statistically significant overlap between DMR-associated genes and heterogeneously expressed genes in the ESLAM cells.

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* **(Fig. 3.9)**, a gene encoding a surface protein previously implicated in the maintenance of long-term HSCs *in vivo* (Hur et al., 2016). Interestingly, we do not see any obvious separation after clustering single-cells on the basis of RNA-seq data alone. This may be a reflection of the noise in single-cell RNA-seq, where stochastic dropout of low expressed genes skews the resulting analysis (Brennecke et al., 2013; Ding et al., 2015). Alternatively, there may be unique characteristics to DNA methylation signal that is not reflected in RNA-seq dataset, as previously annotated in bulk datasets (Kundaje et al., 2015).

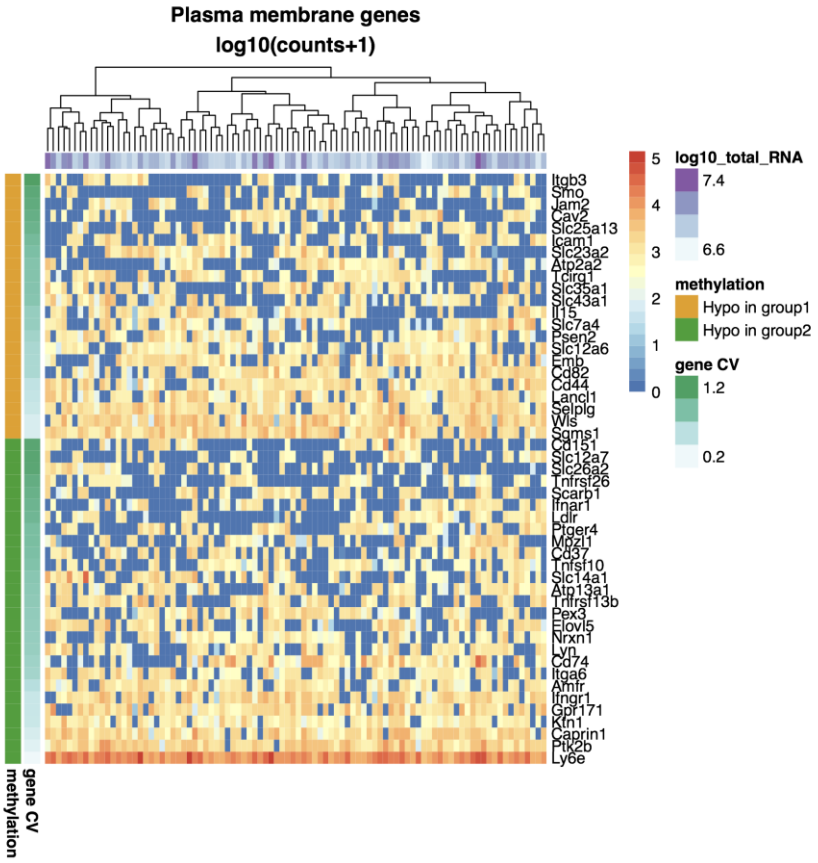

Figure 3.9 – No clear clustering pattern of single-cells using the expression of DMR-associated genes

Rows represent genes that encode plasma membrane proteins (as defined by GO), and the columns represent single cells.

Chapter 4: Heterogeneity of human hematopoietic stem and progenitor cells

4.1 Outliers can be identified and removed using PDclust

From human cord blood, we sorted primary Lin-CD34^{+}CD38^{-}CD90^{+}CD45RA^{-}CD49f^{+} HSCs (CD49f) (Knapp *et al*. submitted) that represent one of the most homogenous human HSC populations (~10% repopulation capacity in secondary mice, Notta et al., 2011). As a comparator, we downloaded previously published data for all the major CD34+ phenotypes in human cord blood (Farlik et al., 2016). Application of PDClust to the latter dataset showed that a majority of Megakaryocytes (MKs) and selected cells from other phenotypes had higher PD values and appeared to be outliers from the remaining phenotypes (groups 2 and 3) **(Fig 4.1a-c)**. These outlier cells also demonstrated significantly lower genome-wide average CpG methylation in comparison to all other cells, suggesting that they were either technical or biological outliers **(Fig 4.1d)**.

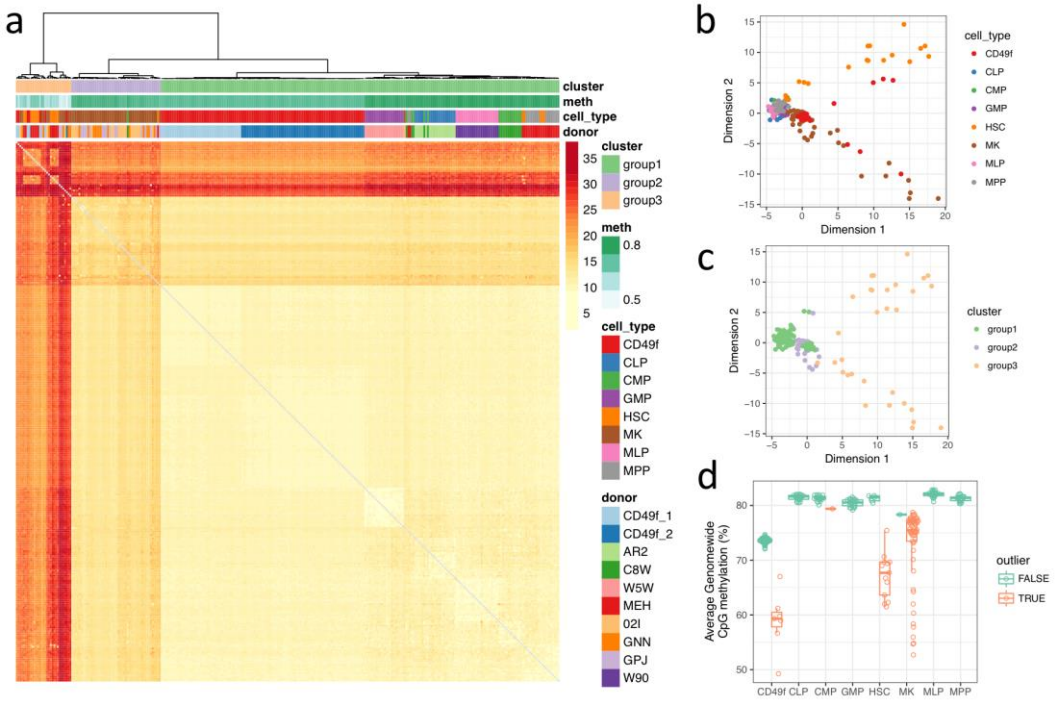

Figure 4.1 – PDclust can identify and remove outliers

4.2 Classical hematopoietic hierarchy reconstructed using single-cell DNA methylation

PDclust applied to the rest of the single cells separated the cells by phenotype with some overlap of the CD49f cells and MPPs **(Fig 4.2a)**. Multidimensional analysis confirmed separation of CD49f cells and MPPs from other phenotypes and revealed a clear separation of GMPs from all other phenotypes **(Fig 4.2b)**. This analysis also showed that cells within the same phenotype had a lower dissimilarity as compared to other phenotypes (p-value <0.01 for all comparisons) **(Fig 4.2c)**. Altogether, these results validate PDClust as an accurate way to segregate multiple primitive, phenotypically distinct, human hematopoietic cell populations from single CpG methylation measurements.

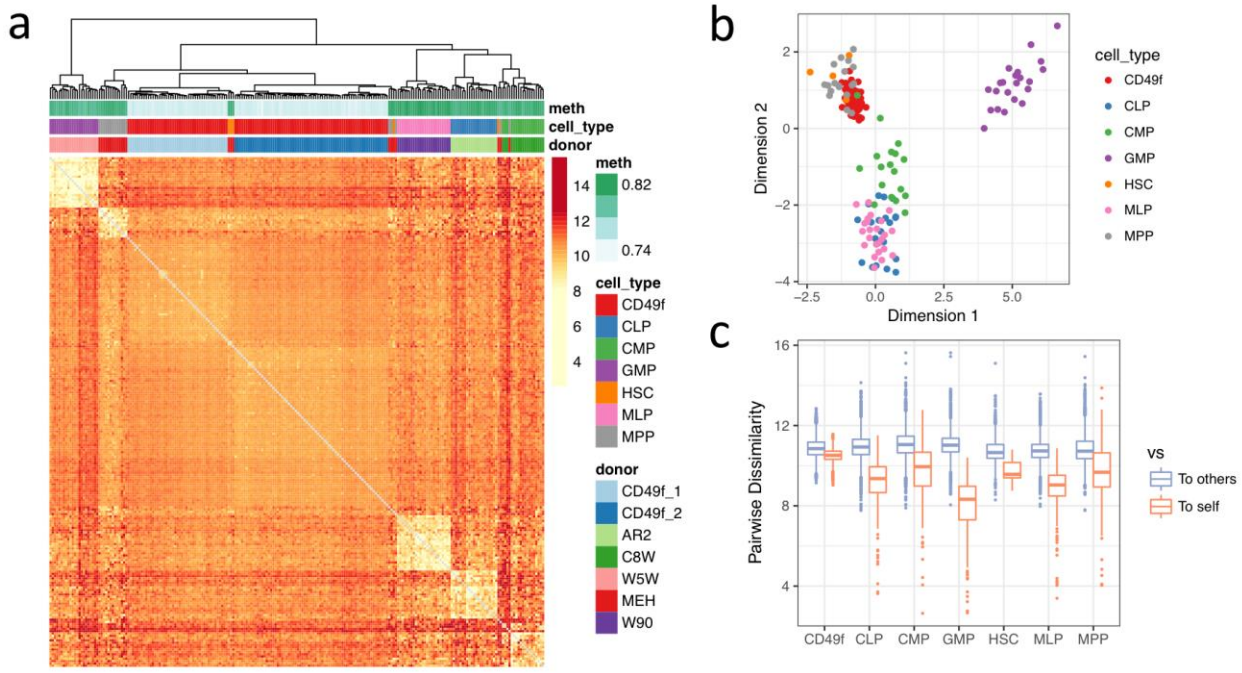

Figure 4.2 – Clustering of human hematopoietic progenitor and stem cells recapitulates known hierarchy

4.3 Epigenetic heterogeneity is dominated by donor-specific differences

Since the data for CD49f cells from our centre was derived from 2 individual donors, it was also possible to determine the donor-specific contribution to the methylomes obtained. Applying PDClust to only our CD49f cells showed that cells clustered separately by donor **(Fig 4.3)**.

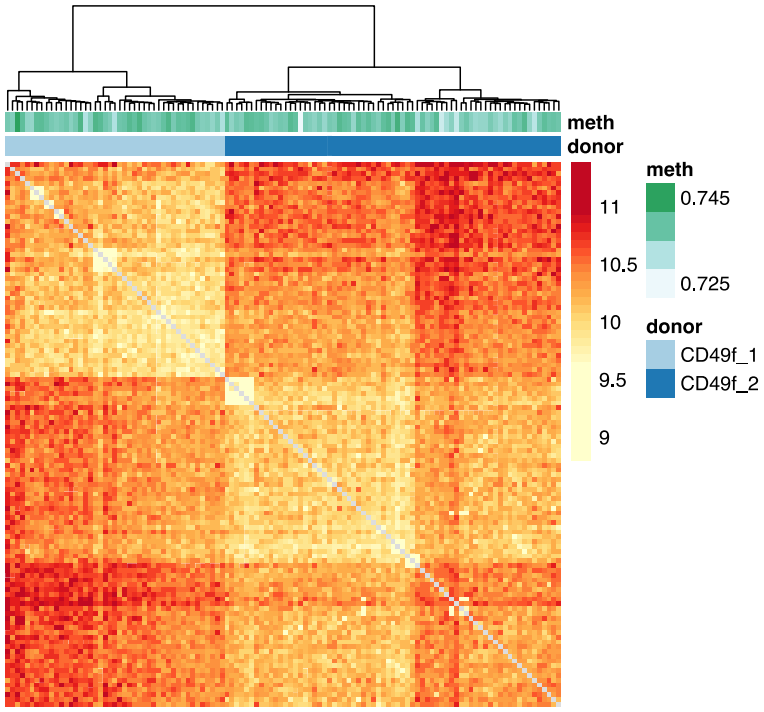

Figure 4.3 – Single CD49f cells separate by donor

To investigate if these donor-specific genetic differences contributed to this separation, we identified single-nucleotide variants (SNVs) by applying MethylExtract (Barturen et al., 2014) to the donor merged datasets. After combining non-identical homozygous variants for both donors (excluding SNVs involving a C or G), we looked for CpGs within 200 bp of the SNVs (polymorphic CpGs) and re-calculated PD values. We observed that donor-specific methylation states remained the major driver of variation within the CD49f population after removal of these

polymorphic CpGs **(Fig 4.4a)**. Single cells separated by donor **(Fig 4.4a)**, and PD values increased when **(Fig 4.4b)** when considering only polymorphic CpGs compared to all CpGs. Taken together, these results suggest that genetic variation accounts only partly for the donor-associated epigenetic variation observed and that donor-specific epigenetic variation is a dominant feature across single and highly purified human CD49f cells.

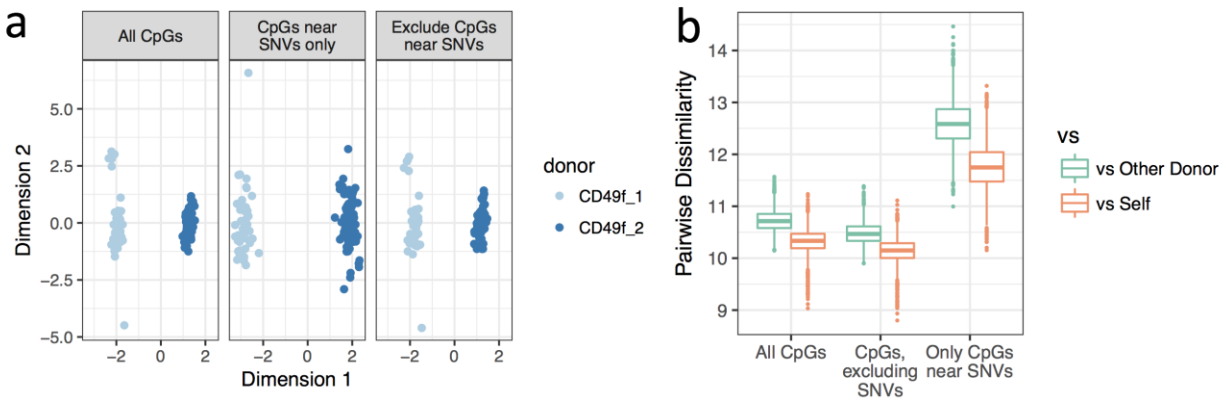

Figure 4.4 – Removal of CpGs near SNVs does not remove donor-driven epigenetic variation

a) MDS projection of pairwise dissimilarities onto two-dimensional space remains largely unchanged despite taking into account SNVs.

b) Pairwise dissimilarity values as a function of comparisons either versus self or the other donor.

4.4 Identification of human HSC subpopulations independently in two donors

Application of PDClust single CD49f cells from donors 1 and 2 separately enabled the identification in each of a consistent subpopulation (group 1, comprising 9% and 11%, respectively, **Fig 4.5a)** that did not correlate with expression levels of the surface markers used to isolate them (CD3, 11b, 19, 34, 38, 90, 45RA, 49f) **(Fig 4.5b)**.

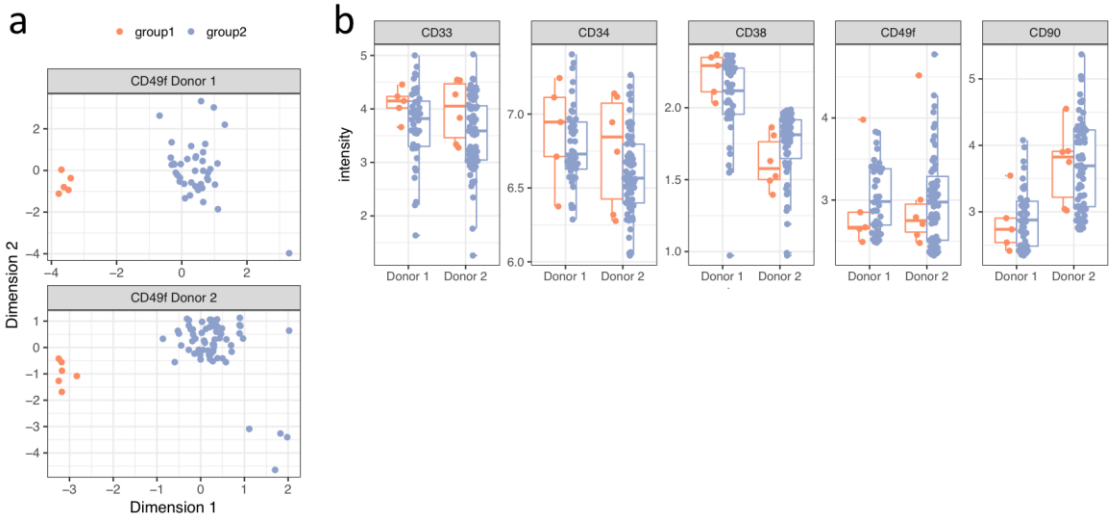

Figure 4.5 – Subpopulation of CD49f cells identified independently in two donors

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 multidimensional scaling.

b) Distributions of surface markers obtained during index sorting of single-cells belonging to cluster1 or cluster2, split by donor.

To control for donor-specific epigenetic variation we previously observed, we proceeded with only CD49f cells from donor 2. We then performed an *in silico* merging of the CD49f single-cell datasets previously identified as group 1 (n=6) and group 2 (n=63) cells. In silico merging of the data from group1 gave a nearly complete recapitulation of the human methylome (17.1 M CpGs with 27.1 M CpGs from group 2) that allowed for comprehensive epigenetic annotation of the 2 groups. As a comparator, we performed in silico merging of all available CD49f cell data as well as the published HPC data (CLPs, CMPs, GMPs, 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 showed that DMRs were enriched in promoters (i.e. sequences between 2 kb upstream and 500 bp downstream of coding gene TSSs)

and were depleted in intergenic regions, suggesting that these DMRs have functional relevance. GSEA of genes associated with DMRs that were hypomethylated in CD49f cells compared to HPCs showed these were enriched in pathways implicated in HSC differentiation **(Fig 4.6a)**. Interestingly, DMRs hypomethylated in group1 compared to group2 were enriched in genes that are upregulated in leukemia (Casorelli et al., 2006) and genes whose expression is upregulated in later types of hematopoietic progenitors (Ivanova et al., 2002) **(Fig 4.6b)**. 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.

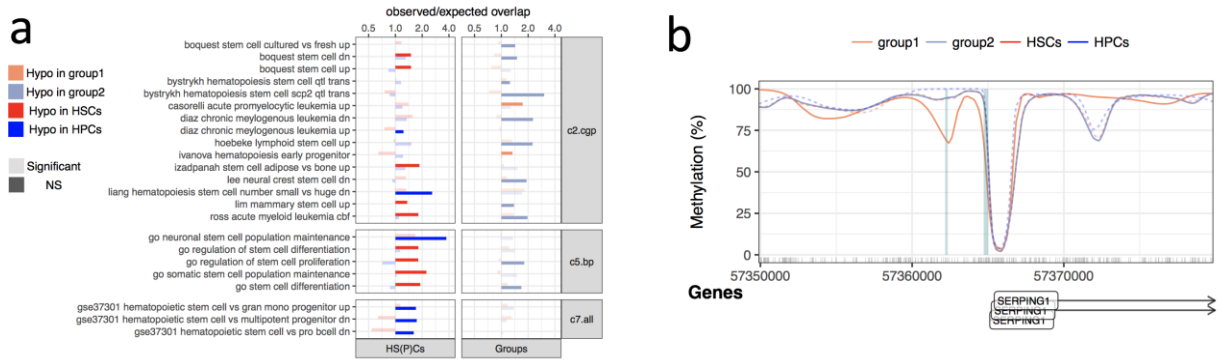

Figure 4.6 – Annotation of epigenetic differences in the CD49f subpopulation

a) 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.

b) Example of a DMR near the *SERPING1* gene.

Chapter 5: Conclusions and Future Directions

5.1 Advances in single-cell DNA methylation library generation

Current single-cell DNA methylation profiling technologies can be broadly categorized into two categories: those that include a pre-amplification step by random priming and those that do not. Within the non-amplification methods, the resulting CpG coverage is limited (up to ~1.5 million CpGs on average at saturation) in scRRBS due to enzyme bias (Guo et al., 2013; Hou et al., 2016; Hu et al., 2016) and in scWGBS due to low library diversity (Farlik et al., 2015, 2016). In contrast, methods that pre-amplify material by random priming, such as scBS-seq (Angermueller et al., 2016; Smallwood et al., 2014), snmC-seq (Luo et al., 2017) and PBAL, recovers an increased number of CpGs genomewide (up to ~10 million CpGs at saturation).

Libraries that were subjected to more rounds of random priming (5 in scBS-seq, 2 in PBAL, 1 in snmC-seq, and none in scWGBS) have higher library diversity (Luo et al., 2017). Interestingly, scBS-seq libraries with many rounds of random priming are highly diverse (~10% duplicates) despite almost twice the number of PCR cycles as PBAL (~20% duplicate at equivalent depth). These results suggest that pre-amplification is crucial to generating diversity in single-cell DNA methylation libraries.

The drawback of random priming methods is that of contamination. Libraries with more rounds of random priming suffer from lower mapping efficiency (average 25% in scBS-seq, 38% in PBAL, 53% in snmC-seq, and 55% in scWGBS). Because random priming is so sensitive, external sources of contamination is also amplified along with single-cell material which contributes to lower mapping efficiency of single-cells. Although not ideal, additional sequencing can mitigate this problem. In the case of scBS-seq, a single-cell was sequenced to 50

million reads which resulted in 9.4 million CpGs (Smallwood et al., 2014). However, this option is cost prohibitive for high-throughput experiments involving hundreds of cells.

Out of all existing protocols, we believe PBAL strikes a good balance between library diversity and mappability. PBAL improves library diversity by decreasing PCR cycles instead of increasing the rounds of random priming, and a complementary qPCR assay removes failed wells and results in higher capture of CpGs in successful cells for an equivalent sequencing cost. Compared to snmC-seq and scWGBS, these advances allow for increased library diversity conferred by random priming while partially offsetting the lower mappability.

In single-cell transcriptomics, major advances have been made in throughput by introducing genetic barcodes to cells and pooling them into a single reaction prior to library generation. One such example is drop-seq, which was used to profile over 40,000 single-cell transcriptomes in under a week (Macosko et al., 2015). Future experiments can apply barcoding principles to improve the throughput of PBAL. For example, it is possible to barcode the genomic DNA of single-cells prior to library construction similar to recently described single-cell chromatin accessibility assays (Buenrostro et al., 2015; Cusanovich et al., 2015). Indeed, this approach has already been successfully applied in bulk whole-genome bisulfite sequencing (Adey and Shendure, 2012).

5.2 Features of DNA methylation at single-cell resolution

The very first surveys of base-pair resolution DNA methylomes showed that nearby CpGs were concordant in their methylation states (Eckhardt et al., 2006) which gradually became a core assumption in most modern DNA methylation analytical strategies (e.g. Hansen et al., 2012). In single-cell data, we found that high CpG concordance of >85% within 1 kb is

highly conserved across many cell types across mouse and humans and is not dependent on background concordance. Furthermore, we found that single-cells have less concordance than bulk cells, but could rule out that the difference may be related to differences in background concordance. Finally, we noted that K562 cells uniquely have increased CpG concordance compared to background concordance even at distances of 4-5 kb. These observations further support the assumption that nearby CpGs are concordant, but suggests that current windowing approaches that use >3kb may be too large and lead to over smoothing.

PDClust consistently identified pairwise dissimilarities between cells functionally defined cell types to be 8-13%, but what fraction of this heterogeneity can be attributed to technical noise is uncertain. Furthermore, the ability of PDClust to distinguish different cell types from less than a few thousand pairwise common CpGs in single cells was unanticipated. As an explanation, we suggest that the information content in the epigenome of single cells may include extensive redundancy with only a few hundred or thousand non-redundant elements. For example, the unique components of a DNA methylation "age" signature could be observed in just 353 CpGs sites scattered across the genome. These CpGs likely represent a subset of a total "age" signature that may involve many more CpG sites not detected with array-based strategies (Horvath, 2013).

Future studies should aim to extract these cell-type specific signatures from whole-genome data. Subsampling bulk reference methylomes in directed ways may allow for identification of these signatures. Successful identification of these signatures would allow the development of a capture-based strategy that would assay only the necessary subset of CpGs genomewide and thus massively decrease the cost of "genomewide" DNA methylation profiling.

5.3 The contribution of genetics to epigenetic variation

Application of PDClust to datasets derived from single human hematopoietic cells characterized them according to classical phenotypes. However, single-cells separated by donor within the highly purified CD49f subset. Focusing on these cells, donor-specific differences was found to be a significant source of epigenetic heterogeneity even after masking CpG sites near SNV locations. This finding is consistent with previous reports that showed genetic diversity is related to but does not account for all DNA methylation differences (Gertz et al., 2011; Xie et al., 2012).

Aside from genetic variation, DNA methylation can be affected by the environment. As the CD49F cells were derived from cord blood, donor-specific DNA methylation differences may have occurred *in utero.* This finding is consistent with previous reports linking various prenatal exposures to changes in offsprings' DNA methylation profiles (Bommarito et al., 2017; Provençal and Binder, 2015). To address why donor-specific differences were only observed in HSCs and not in more differentiated progenitors, we hypothesize that degree of cell-to-cell heterogeneity is low enough in CD49f cells such that a "constant" amount of donor-specific differences dominates. This is in contrast to HPCs where cell-to-cell heterogeneity may be sufficiently prominent to mask donor-specific differences.

Future studies utilizing highly homogenous single-cells from other cell types should therefore aim to include samples from different donors to confirm if donor-specific epigenetic variation is a common feature. These studies can further aid in understanding the amount of homogeneity required in single-cells in order to observe donor-specific epigenetic differences. Furthermore, these observations give rise to unique opportunities to study environmentally

derived DNA methylation differences without being confounded by cellular composition and genetics variability.

5.4 Relationship between epigenetic heterogeneity and functional heterogeneity

Application of PDclust to murine ESLAM and LSK cells using CpGs within previously annotated regulatory regions (Cabezas-Wallscheid et al., 2014) identified a subset of single cells in 2 different compartments of adult mouse bone marrow (LSK and ESLAM cells) that closely resemble their known frequencies 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 in the literature: a majority of which have also been found to be transcriptionally active in mouse HSCs (Wilson et al., 2015). This aligns with previous evidence that CpG methylation status and expression are related 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). This might be a useful marker to use in conjunction with the ESLAM markers to obtain even higher purities of HSCs from adult mouse bone marrow.

Application of PDclust to human CD49f HSCs using CpGs genomewide revealed a consistent and rare (~10%) group of cells independently in two donors. Deep sequencing of one donor allowed for genomewide annotations of 17.1 M CpG sites for the 6 single-cells belonging to the rare subpopulation. Identification and association of subpopulation specific DMRs to

genes again revealed putative pathways in which these subpopulations differ in functional output; genes associated with hypomethylated DMRs in one group was found to be associated with pathways and gene sets related to leukemia.

In conclusion, this thesis presents compelling evidence of epigenetic heterogeneity at a single-cell level, and demonstrates examples of rare subpopulations defined by unique epigenetic states in blood stem cells from both mouse and human. These results add support to a growing body of scientific literature that suggests single-cells are unique in molecular characteristics and these differences likely contributes to the regulation of each cells' individual function. Similarly, epigenetic mechanisms also act on a single-cell level, and for each cell within a classically defined cell population these mechanisms differ slightly to give rise to phenotypic diversity. Altogether, the results from this thesis suggests that the definition of cell-types should continue to be reshaped and refined.

Bibliography

Adey, A., and Shendure, J. (2012). Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. Genome Res. *22*, 1139–1143.

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. F1000Research *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.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

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.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C.,

Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357,* 661–667.

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.

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.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science *348*, 910–914.

Dawlaty, M.M., Breiling, A., Le, T., Barrasa, M.I., Raddatz, G., Gao, Q., Powell, B.E., Cheng, A.W., Faull, K.F., Lyko, F., et al. (2014). Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. Dev. Cell *29*, 102–111.

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A., and Wang, W. (2015). Normalization and noise reduction for single cell RNA-seq experiments. Bioinformatics *31*, 2225–2227.

Domanico, M.J., Allawi, H., Lidgard, G.P., Aizenstein, B., Hunt, O., and Zutz, T.C. (2013). Modification of dna on magnetic beads (United States).

Doulatov, S., Notta, F., Laurenti, E., and Dick, J.E. (2012). Hematopoiesis: A Human Perspective. Cell Stem Cell *10*, 120–136.

DuPage, M., and Bluestone, J.A. (2016). Harnessing the plasticity of CD4+ T cells to treat immune-mediated disease. Nat. Rev. Immunol. *16*, 149–163.

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.

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.

Grunau, C., Clark, S.J., and Rosenthal, A. (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Res. *29*, E65-5.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. *23*, 2126–2135.

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.

Hirst, M., and Marra, M.A. (2010). Next generation sequencing based approaches to epigenomics. Brief. Funct. Genomics *9*, 455–465.

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.

Ito, S., D'Alessio, A.C., Taranova, O. V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

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.

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

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.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

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.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

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.

Li, L.-C., and Dahiya, R. (2002). MethPrimer: designing primers for methylation PCRs. Bioinformatics *18*, 1427–1431.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

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.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523–536.

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 (80-.). *357*, 600–604.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression

Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

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.

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

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.

Naik, S.H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R.J., and Schumacher, T.N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature *496*, 229–232.

Nestorowa, S., Hamey, F.K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., Wilson, N.K., Kent, D.G., and Gottgens, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood *128*, e20–e31.

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 (80-.). *333*, 218–221.

Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O.I., Wilson, G., Kaufmann, K.B., McLeod, J., Laurenti, E., Dunant, C.F., et al. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science *351*, aab2116.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell *99*, 247–257.

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.

Papalexi, E., and Satija, R. (2017). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell *163*, 1663–1677.

Perié, L., Duffy, K.R., Kok, L., de Boer, R.J., and Schumacher, T.N. (2015). The Branching Point in Erythro-Myeloid Differentiation. Cell *163*, 1655–1662.

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. *107*, 20311–20316.

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.

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.

Sen, P., Shah, P.P., Nativio, R., and Berger, S.L. (2016). Epigenetic Mechanisms of Longevity and Aging. Cell *166*, 822–839.

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.

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.

Subramaniam, D., Thombre, R., Dhar, A., and Anant, S. (2014). DNA methyltransferases: a novel target for prevention and therapy. Front. Oncol. *4*, 80.

The Cancer Genome Atlas Research Network (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N. Engl. J. Med. *368*, 2059–2074.

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.

Warr, M.R., Pietras, E.M., and Passegué, E. (2011). Mechanisms controlling hematopoietic stem cell functions during normal hematopoiesis and hematological malignancies. Wiley Interdiscip. Rev. Syst. Biol. Med. *3*, 681–701.

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 (80-.). *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.

Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S.P., Aparicio, S., and Hansen, C.L. (2017). Scalable whole-genome single-cell library preparation without preamplification. Nat. Methods *14*, 167–173.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

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.

Appendices

Appendix A : Materials and Methods

A.1 Single-cell isolation of murine HSPCs

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 samples were stained with LSK or ESLAM staining cocktails (Table S1). We verified that the repopulating potential of our sorted ESLAM and LSK cells were equivalent to previously published studies (40% and 4% respectively, data not shown) (Kent et al., 2009; Warr et al., 2011). Single cells were dry sorted into 384 well plates (ThermoFisherf, 4309849) using FACSAria (BD, Franklin Lakes), flash frozen using liquid nitrogen, and stored at -80°C until processing.

A.2 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 in GeneAmp 9700 PCR machine (ThermoFisher, 4413750) at 37˚C for 30 minutes, 70˚C for 10 minutes. After the treatment, we aliquoted the beads (2 uL per well) into a clean 96-well plate, mixed with 180 uL of MethylEdge Binding Buffer (Promega, N1301), and UV treated as per instructions below.

A.3 UV treatment

Prior to use, we UV treated using 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).

A.4 Cell Lysis and Bisulfite treatment

Plates with sorted cells were removed from -80°C storage and centrifuged at 3000 rpm, 4°C, for 2 minutes. Working on ice, 4 uL 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 single cell wells and three empty wells (negative controls), followed by centrifugation at 3000 rpm, 4°C, for 1min. 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. After the incubation, 1 uL of 60 fg/uL T7 Phage DNA (GeneON, #301-025) was added to each well except the negative control wells, and the mixture was subjected to bisulfite conversion using MethylEdge Bisulfite Conversion kit (Promega, N1301) following a bead-based protocol enabling automation.(Domanico et al., 2013) For this, the single-cell lysis/spike-in mixture (~5ul) was combined with 32.5 uL of MethylEdge Conversion reagent and incubated in GeneAmp 9700 PCR machine (98°C for 8 minutes, 54°C for 60 minutes). After the incubation, the plate was spun down at 3000 rpm for 1min. All of the subsequent steps were performed on Bravo Automated Liquid Handling Platform (Agilent Technologies, G5409A) with 96LT Disposable Tip head, 250uL sterile, filtered tips (Agilent Technologies, 19477-022) using custom programs created with VWorks Automation Control Software (Agilent Technologies, USA). Bisulfite converted DNA was mixed with 180 uL of MethylEdge Binding Buffer and 1.8 uL of 20 mg/mL of 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 a magnet, the beads were washed twice with 220 uL of 80% ethanol for 30 sec without resuspension. Next, 60 uL of MethylEdge desulfonation buffer was added to the beads and the mix was incubated at room temperature for 15 minutes. After the removal of desulfonation buffer, while still on a magnet, the beads were washed twice with 100 uL of 80% ethanol without resuspension and air-dried for 1 minute. To elute DNA, the beads were resuspended in 20 uL of 10 mM Tris-HCL, pH 8.5 (Qiagen, 19086) and incubated in a Thermomixer C (Eppendorf, 5382000015) at 56°C with 2,000 rpm for 15 minutes. The beads were then collected to the side on a magnet for 30 sec and the DNA containing supernatant was transferred to a new 96 well plate.

A.5 Double-stranding reaction

The bisulfite-converted scDNA was mixed with 1.25 uL of 10 mM dNTPs and 1 uL of 500 uM random hexamers (3' phosphothioate), incubated at 98°C for 1 minute, and then snap frozen on ice for 2 minutes. A mix of 0.5 uL of 50 U/uL Klenow exo- (NEB, M0212M) and 2.5 uL of 10X NEB Buffer 2 was added and reactions were incubated in a GeneAmp 9700 PCR machine at 4˚C for 10 minutes, + 4°C/s to 37C, 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 5uL of 2nd DNA synthesis mix (20 uM 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, 37°C for 30 minutes, and 70°C for 10 minutes. After the incubation, 20 uL of 10 mM Tris-HCL (pH 8.5) was added and reactions were purified at 1:1 ratio using house-made magnetic bead solution (1M NaCL, 20% PEG, Sera-Mag Speedbead (Fisher Scientific, 09981123)) following Beckman-Coulter Ampure XP PCR purification

protocol. The DNA was eluted in 35uL of 10 mM Tris-HCl (pH 8.5) and used for Illumina library generation.

A.6 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) using custom programs created with VWorks Automation Control Software (Agilent Technologies, USA). End Repair and 5' Phosphorylation reaction (35uL DNA sample, 5uL of 10X NEB 2 buffer, 2uL of 25mM ATP, 2uL of 10mM dNTP, 10U T4 Polynucleotide Kinase, 4.5U T4 DNA Polymerase, 1U Klenow Large Fragment DNA Polymerase, and ultrapure water to a total reaction volume of 50uL (NEB, E6000B-10)) was incubated at room temperature for 30 minutes. Single cell DNA samples were then purified at 1:1 ratio using house-made magnetic bead solution (1M NaCL, 20% PEG, Sera-Mag Speedbeads (Fisher Scientific, 09981123)) following Beckman-Coulter Ampure XP PCR purification protocol, and eluted in 25uL volume with 10mM Tris-CL (pH 8.5). To enable ligation to the adaptors, a single dA overhang was added to the 3' ends of DNA fragments (25uL DNA, 3.5uL of 10X NEB 2 buffer, 0.7uL of 10mM dATP, 3.5U Klenow Fragment ($3' \rightarrow 5'$ exo–), and ultrapure water to a total reaction volume of 35uL, (NEB, E6000B-10)). 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 (35uL DNA, 12uL of 5X Quick Ligation Buffer, 2000U Quick T4 DNA Ligase, 2uL of 0.5uM Illumina sequencing forked adaptor, and ultrapure water to a total volume of 60uL (NEB, E6000B-10)). Ligation reaction was performed at room temperature overnight. To remove adaptor dimers, the ligation product was purified twice using house-made

magnetic bead solution (1M NaCL, 20% PEG, Sera-Mag Speedbeads) - first at 0.8X (eluted in 50uL of 10mM Tris-Cl, pH 8.5) and then at 1:1 ratio (eluted in 25 uL of 10mM Tris-Cl, pH 8.5). Adaptor ligated libraries were PCR amplified and barcoded using custom indexing primers (25uL DNA, 10uL of 5X High fidelity buffer, 1uL of 10mM dNTPs, 1.5uL of DMSO, 1uL of 25uM forward primer, 2uL of 12.5uM custom reverse indexing primer (added separately to each well), 1U Phusion U Hot Start, and ultrapure water to a total volume of 50uL (ThermoFisher, F-555L)). PCR amplification was carried out using GeneAmp 9700 PCR machine with the following cycling conditions: 98^{0}C for 1 minute, 10 repeats of (98 ^{0}C 30 sec, 65 ^{0}C 15 sec, 72 ^{0}C 15 sec), 72 ^{0}C 5 minutes, and 4 ^{0}C hold. Barcoded libraries were size selected to remove primer dimers with 0.8X house-made magnetic bead solution (1M NaCL, 20% PEG, Sera-Mag Speedbeads) and eluted in 15uL of 10mM Tris-CL, pH 8.5. Based on Real-Time PCR analysis, 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. Samples were aligned and CpGs were called.

A.7 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 uL volume (1X KAPA SYBER FAST master mix, 1X Illumina Primer Premix (KAPA, KK4824), and 1 uL 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 uL volume (1X KAPA SYBER FAST master mix, 0.5 uM each of forward and reverse primers, and 1 uL of 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 sec, 60˚C for 30 sec, 72˚C for 45 sec); 72˚C for 5 minutes. Genome-specific CT values were normalized (by subtraction) to CT values of library, and a cut-off was determined empirically based on the distribution of single-cells and negative controls.

QC-passed single cell libraries were pooled together and concentrated by ethanol precipitation (0.1X 3M Sodium Acetate, 2.5 uL of 20mg/mL mussel glycogen (Sigma, 10901393001), 2.5X 100% Ethanol). After incubation at 20°C for 1 hour, centrifugation (4˚C, 20,000g for 40 minutes) and 75% Ethanol wash, the single cell library pool was air-dried for 5 minutes at room temperature, and then resuspended in 15 uL of 10mM 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.

A.8 Bulk PBAL construction

10,000 cells were FACS sorted into an Eppendorf tube, and DNA was extracted using AllPrep DNA/RNA Mini Kit (Qiagen, 80204) according to manufacturer's instructions. 4 uL of purified DNA (25 ng/uL) was mixed with 1 uL of T7 Phage DNA spikein (1 ng/uL), and the mixture was used directly as input to bisulfite conversion and desulphonation as described for single-cells, except neither DNase nor UV decontamination was performed on the reagents, and 5 uL of beads was used instead of 1.8 uL. After bisulfite conversion and on-bead desulphonation, the DNA was eluted in 40 uL of EB buffer and then used as input for one round of double-

stranding with 2.5 uL of 10 mM dNTPs, 2 uL of 500 uM random hexamers, 1 uL of 50 U/uL Klenow exo- (NEB, M0212M) and 5 uL of 10X NEB Buffer 2. Library construction was the same for single-cells except only 4 rounds of PCR was used. Libraries were sequenced directly without qPCR quality control.

A.9 Raw data processing

The first 6 bases of read1 and read2 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 version V3.02.10 (http://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 mice, we used–u8 (penalty for unconverted CHG or CHH cytosine) as recommended by Novoalign, while for humans we used –u50 due to 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 Phred quality score ≤15 were discarded. Methylation percentage of each CpG di-nucleotide (found by searching the genome with a custom java script) was calculated 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 only considered autosomal CpG sites and CpG sites with

methylation value of 0 or 100%. Copy number variation (CNV) in 5MB windows were called using Control_FREEC V7.0 (Boeva et al., 2012) with default parameters. Single cells with conversion rate < 96%, mappability < 5%, less than 130,000 CpGs, and containing more than 50 windows with CNVs were removed.

A.10 Obtaining genomic regions

For mouse, 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, we considered 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 HSC to MPP transition (HSC Regulatory Networks), we downloaded the list of DMRs from Cabezas-Wallscheid et al (Cabezas-Wallscheid et al., 2014); for unmethylated regions in HSCs, we downloaded the list of regions from Jeong et al (Jeong et al., 2014); and for blood enhancers, we downloaded the enhancer catalogue from Lara-Astiaso et al (Lara-Astiaso et al., 2014). For human, we considered gene bodies as every protein-coding V75 Gencode transcript with a support level of 1 or 2.

A.11 Methylation of adjacent CpGs in single-cells

For single cells, we randomly sampled up to 100,000 CpG sites either genomewide or within relevant genomic regions. For each randomly sampled CpG site (CpG1), we analyzed 100 CpGs sites (with coverage information) before and after CpG1 and calculated the distance to CpG1. 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 each CpG1. Then, after combining all these tables for every CpG1, distances were binned into 100bp bins and the mean concordance was calculated as the fraction of CpGs in each bin that were equal. For LSK single-cells, we did this for the 10 single-cells with the most CpG coverage. For bulk, since the data was continuous, we calculated the absolute difference in methylation between CpG1 and all nearby CpGs instead, using only CpGs with coverage >= 5 to avoid low coverage biasing potential CpG methylation values.

A.12 Epigenetic subpopulation discovery

For each pair of single-cells, we calculated the average difference in DNA methylation of all pairwise-common CpG sites as a measure of dissimilarity (pairwise dissimilarity, PD). 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 dissimilarity, we calculated Euclidean distances between each cell using the their PD values to other cells 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.

A.13 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 call differential methylated CpGs (dCpGs), we calculated the mean and standard deviation of the difference in CpG methylation between two groups. 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 two-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 two CpGs. Finally, the p-values were multiple test corrected using false discovery rate. To call DMRs, we grouped dCpGs together if they were within 500bp of each other and only considered regions with 3 or more CpGs.

To calculate the enrichment of DMRs in different genomic locations, first we calculated the expected overlap as the fraction of the genome each genomic feature (intergenic, promoter, etc) occupied. 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 a division of these two numbers.

A.14 Gene enrichment analysis

We first split DMRs into two 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 to (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 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 mouse, we downloaded gene sets from genes that control 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 human, 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.

A.15 Single-cell RNA-seq analysis

We downloaded processed read counts of 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).

Appendix B : Data Availability

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